MEMOIRE

présenté à
L'UNIVERSITE Pierre et Marie Curie (Paris 6)

pour obtenir le diplôme
d'**HABILITATION A DIRIGER DES RECHERCHES**

Spécialité "Sciences de l'Ingénieur"

APPRENTISSAGE STATISTIQUE ET FOUILLE DE DONNEES POUR L'ANALYSE TEMPS-REEL D'IMAGES ET LES SYSTEMES DE TRANSPORT INTELLIGENTS

par
Fabien MOUTARDE

Soutenue le 29 août 2013 devant le jury constitué de :

M. Gérard DREYFUS	Rapporteur	Professeur à ESPCI ParisTech
M. Didier AUBERT	Rapporteur	Directeur de Recherches à IFSTTAR
M. Pierre VANDERGHEYNST	Rapporteur	Professeur Associé à EPFL
Mme Catherine ACHARD	Examinateur	Maître de Conférences (HdR) à l'UPMC
M. Thierry CHATEAU	Examinateur	Professeur à Univ. Clermont-Ferrand
M. Arnaud de La FORTELLE	Examinateur	Professeur à Mines ParisTech
M. Nikolas GEROLIMINIS	Examinateur	Professeur assistant à EPFL

Remerciements

Je tiens avant tout à remercier chaleureusement Claude LAURGEAU, fondateur et ancien directeur du centre de Robotique (CAOR) de Mines ParisTech de m'avoir accueilli fin 2006 dans son laboratoire, où j'ai pu rapidement redémarrer une activité soutenue de Recherche, après mes 10 premières années aux Mines de Paris consacrées principalement à l'enseignement informatique. Merci aussi à Gladys HUBERMAN de m'avoir recruté en 1996 comme Maître-assistant au Centre de Calcul de MinesParis, me permettant ainsi de revenir à une carrière académique après un détour de 5 ans dans la R&D privée à Alcatel Alsthom Recherche. Un merci particulier aussi à Fawzi NASHASHIBI, maintenant à INRIA-IMARA après de nombreuses années au CAOR, avec qui j'ai co-encadré deux de mes doctorants, dont le premier, et collaboré pour la plupart des recherches pour Valeo.

Ensuite, ce manuscrit n'existerait tout simplement pas sans le travail des stagiaires et surtout des 4 doctorants que j'ai encadrés, Alexandre BARGETON, Omar HAMDOUN, Ayet SHAIEK et Anne-Sophie PUTHON : merci à eux de m'avoir fait confiance pour guider leurs premiers pas dans le monde de la Recherche, et d'avoir ainsi apporté une part essentielle à la plupart des contributions scientifiques résumées dans ce manuscrit. Un grand merci aussi à Yufei HAN dont les travaux dans le cadre du post-doctorat effectué sous ma supervision ont constitué un apport décisif sur le plus récent de mes axes de recherche.

Par ailleurs, je tiens à remercier aussi le professeur Alexandre BAYEN de l'Université de Californie à Berkeley qui m'y a accueilli en court séjour sabbatique début 2012, et a permis d'initier une collaboration scientifique fructueuse. Plus généralement, je remercie tous les partenaires académiques et industriels (en particulier à Valeo, PSA, et Safran-MORPHO) avec lesquels j'ai été amené à collaborer durant mes travaux de Recherche.

Merci enfin à tous les collègues actuels et anciens du centre de Robotique et du CCSI, et plus généralement à toutes les personnes de l'Ecole des Mines que j'ai été amené à côtoyer durant les 17 dernières années.

INTRODUCTION

Ce manuscrit présente mes travaux de recherche effectués ou encadrés de 2006 à 2013, notamment dans le cadre de ***4 thèses*** que j'ai supervisées : ***celle d'Alexandre BARGETON*** (la première que j'ai co-encadrée) soutenue en décembre 2009, **celle d'Omar HAMDOUN** soutenue en décembre 2010, **celle d'Ayet SHAIEK** soutenue en mars 2013, ***et celle d'Anne-Sophie PUTHON*** (que j'ai co-encadrée) soutenue en avril 2013.

Par soucis d'unité, sont volontairement omises dans ce document les recherches que j'ai effectuées antérieurement, à savoir : d'une part ma thèse en astrophysique à l'Observatoire de Meudon, portant sur des simulations numériques et analyses statistiques en cosmologie (formation des grandes structures de répartition de matière dans l'Univers) ; d'autre part mes travaux comme ingénieur de recherche à Alcatel Alsthom Recherche, concernant des applications des techniques neuronales en modélisation de procédé et en classification de signal, ainsi que la compression « intelligente » d'images fixes.

L'ensemble des recherches présentées ici couvre diverses applications des techniques d'apprentissage statistique, en analyse temps-réel d'images et en fouille de données :

- détection et reconnaissance de panneaux routiers, pour des applications d'aide avancée à la conduite, menées en lien avec l'équipementier automobile Valeo ;
- détection et catégorisation visuelle d'objets tels que voitures et piétons, aussi pour les applications embarquées d'aide à la conduite, pour Valeo et PSA ;
- ré-identification de personnes entre caméras à champs disjoints, et identification d'objets sur des images 3D de profondeur ;
- analyse, fouille de données et prédiction de trafic routier.

Dans les 3 premiers cas, il s'agit de problèmes de reconnaissance de formes, avec une variabilité intra-classe soit très faible (panneaux), soit au contraire très grande (catégorisation, notamment dans le cas des piétons), et la difficulté spécifique pour les applications d'identification, de présenter une variabilité inter-classes souvent basse (ce qui différencie 2 personnes est assez subtil par rapport à la ressemblance globale de leur aspect). Le dernier domaine, outre que contrairement aux autres il n'intègre pas de problématique de traitement d'images, relève plutôt de l'apprentissage non-supervisé (clustering) pour identifier des motifs typiques spatiaux et temporels, ainsi que des techniques de prédiction de série temporelle multi-variée.

1. DETECTION ET RECONNAISSANCE DE PANNEAUX ROUTIERS (TSR)

Les travaux présentés dans ce chapitre ont été réalisés de 2006 à 2012 dans le cadre d'un stage ingénieur, puis d'un stage de Master2 Recherche, et enfin surtout de ***2 thèses*** : ***celle d'Alexandre BARGETON*** (la première que j'ai encadrée) soutenue en décembre 2009, ***et celle d'Anne-Sophie PUTHON*** (que j'ai co-encadrée), soutenue en avril 2013.

Le contexte de ces travaux est une **collaboration avec l'équipementier automobile *Valeo***, d'abord dans une étude bilatérale financée directement par celui-ci, puis au sein du **projet *SPEEDCAM* subventionné par *l'ANR*** et en lien avec le constructeur automobile allemand ***Daimler***. Dans les 2 cas, le contexte général est celui des Système Avancés d'Aide à la Conduite (« Advanced Driving Assistance Systems », ADAS, en anglais). L'objectif applicatif précis pour l'industriel est le prototypage d'un système de détection et reconnaissance des panneaux de limites de vitesse, afin de proposer un système capable d'informer en permanence le conducteur (voire un futur régulateur intelligent de vitesse) de la vitesse limite courante, et ce même dans les zones de travaux ou à limite variable pour lesquelles la cartographie GPS, statique par nature, ne peut évidemment pas fournir d'information pertinente.

1.1 INTRODUCTION SUR LA RECONNAISSANCE DE PANNEAUX

La détection et reconnaissance de panneaux routiers (« Traffic Sign Recognition », TSR, en anglais) est historiquement un des premiers domaines d'application pour les systèmes d'aide à la conduite (ADAS) de l'analyse intelligente temps-réel de vidéo. Cela s'explique en partie par le fait que les objets qu'il s'agit de détecter et reconnaître visuellement sont relativement bien normalisés, avec des formes, des couleurs, et un contenu, variant peu entre instances physiques distinctes d'un même panneau. Ceci rend leur détection et reconnaissance potentiellement assez aisée, en théorie faisable par simple seuillage de couleur suivi d'une comparaison aux prototypes normalisés de panneaux. Toutefois le grand défi lié à cette application (comme pour toutes les aides à la conduite exploitant le flux d'une caméra) est l'extrême variabilité des conditions d'éclairage et de contraste ; à ceci vient s'ajouter la taille relativement faible des objets dans les vidéos « large champ » utilisées, et le potentiel vieillissement et dégradation des panneaux (tags, par exemple !), ainsi que leur occlusion partielle possible (par exemple par des branchages voisins). En résumé, obtenir une détection et reconnaissance temps-réel robuste et fiable en toute circonstance s'avère un problème difficile qui requiert l'usage de techniques avancées de reconnaissance de formes.

La quasi-totalité des travaux actuels de recherche découpent le processus de TSR en trois parties :

1. Détection
2. Reconnaissance
3. Intégration temporelle

La *détection* des panneaux routiers dans l'image consiste à y trouver et segmenter des régions d'intérêt (RoI) susceptibles de correspondre à un panneau. Les approches pour cette étape se scindent en deux grandes catégories : segmentation par couleur, ou segmentation par contours de forme. Comme rappelé par (Bahlmann, et al. 2005), la segmentation par couleur est la méthode la plus couramment utilisée, compte tenu des couleurs « franches » et normalisées (anneau ou triangle rouge en bordure, notamment) des panneaux ; divers travaux, par exemple (de la Escalera, Moreno, et al. 1997) se contentent d'ailleurs d'un seuillage sur la composante rouge du modèle RGB, plus rarement sur la chrominance rouge de YUV/YCbCr. Cependant, ces modèles de couleurs sont sensibles aux changements de conditions lumineuses, même en passant dans l'espace de couleur HSV comme indiqué par (Loy et Barnes 2004), sans oublier que la couleur des panneaux s'estompe avec le temps. Ces méthodes de segmentation par couleur peuvent donc poser problème pour la robustesse d'un système TSR, ce qui a conduit plusieurs équipes de recherche à effectuer la détection plutôt à l'aide de la forme des contours. Pour les nombreux panneaux européens de forme *circulaire*, la méthode la plus connue et la plus robuste employée est l'algorithme de Hough qui est

utilisé par exemple dans (Garcia, Sotelo et Martin-Gorostiza 2006). Cependant cette technique peut s'avérer trop lente, en particulier sur des images haute-résolution. C'est pourquoi de nombreux travaux essayent de développer d'autres méthodes. A partir de la *Distance Transform* (DT) définie dans (Borgefors 1986) et qui calcule la distance pour chaque pixel de l'image au contour le plus proche, il est possible d'effectuer une corrélation « rapide » (Gavrila 1999) pour identifier la position et la taille de panneaux potentiels dans une image. Pour les panneaux d'autres formes (notamment triangulaires, et aussi rectangulaires), une variante de Hough a été proposée dans (Loy et Barnes 2004). Certains travaux essayent de combiner les deux techniques (détection par couleur et par contours de forme) de façon soit parallèle (Fang, Chen et Fuh 2003), soit successive (Broggi, et al. 2007). Enfin, il est possible d'utiliser des méthodes de type apprentissage statistique pour apprendre à un algorithme à trouver les régions d'une image contenant des panneaux. Ainsi, (Bahlmann, et al. 2005) utilise la technique du boosting, dont l'idée est de combiner beaucoup de petits classificateurs « faibles », pour localiser les panneaux potentiels.

La phase de *reconnaissance* des panneaux consiste à identifier, pour chaque zone d'intérêt (RoI) de l'image trouvée précédemment, quel panneau elle contient (si c'est bien un panneau), ou à supprimer les faux positifs (les ROI où aucun panneau n'est en fait présent). De nombreux travaux, par exemple (Barnes et Zelinsky 2004) (Miura, Kanda et Shirai 2000) (Escalera et Radeva 2004), utilisent le processus de corrélation croisée, l'algorithme le plus simple à mettre en œuvre. Il sera de plus ici assez rapide car les zones à couvrir sont petites, puisque limitées aux RoIs issues de l'étape de détection. Cependant, le fait que la dimension des panneaux en pixels dans l'image peut être très variable oblige à utiliser plusieurs masques à différentes échelles pour chaque type de panneau. Le travail est toutefois facilité si la phase de détection renvoie des RoIs approximativement cadrées sur les panneaux, car alors la taille du masque à appliquer peut s'en déduire directement. Comme souligné dans (Piccioli, et al. 1996), l'étape de classification est en fait beaucoup moins critique que celle de la détection, et le choix de l'algorithme d'apprentissage pourra être guidé par d'autres considérations (facilité pour gérer le multi-classes, disponibilité d'outils, familiarité avec telle ou telle technique, …). Dans le domaine de la reconnaissance de panneaux, les divers types de réseaux neuronaux sont les plus couramment utilisés. Un réseau de neurones de type ART1 est proposé par (de la Escalera, Armingol et Mata 2003) et un autre de type ART2 dans (Fang, Fuh, et al. 2003), mais la topologie la plus fréquemment utilisée est le perceptron multi-couches (PMC, ou multi-layer perceptron - MLP) comme par exemple dans (Torresen, Bakke et Sekanina 2004), (Garcia, Sotelo et Martin-Gorostiza 2006) ou (Medici, et al. 2008). L'entrée du PMC est une imagette normalisée (ses dimensions en pixels sont fixes, et son histogramme de niveaux de gris est normalisé). Les RBF (Radial Basis Functions neural network) sont une autre topologie de réseaux de neurones rencontrée dans le domaine, comme par exemple dans (Gavrila 1999). Etonnamment les « Séparateurs à Vaste Marge » (SVM, « Support Vector Machines » en anglais), bien qu'ayant partiellement supplanté les réseaux neuronaux dans beaucoup de domaines, sont encore assez peu utilisés pour les panneaux ; peut-être en raison de la nécessité de réaliser autant d'apprentissages que de classes (les SVM étant binaires par nature, il faut a priori combiner N classificateurs SVM « un contre tous »). L'un des points critiques de toutes ces techniques d'apprentissage statistique est la nécessité d'avoir une base de données d'exemples (i.e. la base d'apprentissage) assez conséquente afin d'obtenir de bons résultats. (Soetedjo et Yamada 2005), pour diminuer cette contrainte, ont développé une technique utilisant la corrélation d'histogrammes, ce qui leur permet de n'avoir besoin que de quelques exemples dans leur base d'apprentissage ; ils partitionnent leurs images en sous-régions circulaires concentriques et effectuent le processus de reconnaissance séparément sur les histogrammes de chacune ; la faiblesse principale réside dans la grande variabilité pratique desdits histogrammes selon luminosité et contraste. Pour être robuste aux déformations, rotations (etc…), il existe trois grandes techniques comme indiqué dans (Cecotti, Choisy et Belaid 2004). Une première approche consiste à ajouter des images dans la base d'apprentissage en transformant les exemples déjà présents, par déformation, ajout de bruit, changement de contraste et de luminosité (etc..), comme dans (de la Escalera, Moreno, et al. 1997). Une autre possibilité est de « redresser » les images comme dans

(Escalera et Radeva 2004) où ils détectent des ellipses et non des cercles dans l'image, ce qui leur permet de rendre à un panneau déformé dans l'image une apparence circulaire normale. Enfin, une dernière technique envisageable est de faire la reconnaissance plutôt sur la transformée de Fourier des RoIs, qui est invariante au bruit, à la rotation, à la translation et à l'échelle des panneaux détectés, comme le font (Kang, Griswold et Kehtarnavaz 1994). Une approche similaire consiste à utiliser la transformée de Fourrier-Melin, qui selon (Cecotti, Choisy et Belaid 2004) est plus performante, pour cette application, que la simple transformée de Fourrier. Enfin, pour améliorer les performances de ces algorithmes, comme signalé par (Soetedjo et Yamada 2005), il est possible de combiner différents algorithmes d'apprentissage / reconnaissance (technique du boosting), ce qui permet de mieux identifier et rejeter des exemples ambigus, et / ou aussi de générer un plus grand nombre d'exemples comme déjà indiqué plus haut.

1.2. RECHERCHES MENEES EN TSR

Les recherches que j'ai menées et/ou encadrées dans ce domaine ont porté d'une part sur des méthodes de détection par la forme de panneaux circulaires et rectangulaires, et d'autre part sur une approche originale de reconnaissance dédiée aux panneaux de limites de vitesse exploitant la reconnaissance des chiffres contenus dans le panneau, ainsi que sur une architecture pour la reconnaissance des panonceaux. Ces derniers sont de petits panneaux rectangulaires parfois situés sous les panneaux, et qui en précisent ou limitent la portée, par exemple à certaines voies ou catégories de véhicules (cf. Figure 1.19).

1.2.1. AMELIORATION DE LA DETECTION PAR CONTOURS DE PANNEAUX CIRCULAIRES

Un des problèmes soulevés par l'utilisation de l'algorithme de Hough pour la détection des cercles est sa relative lenteur d'exécution sur des images en pleine résolution. Ceci est fréquemment contourné, pour permettre le fonctionnement en temps réel, en appliquant l'algorithme sur une version sous-échantillonnée du flux vidéo. Cependant, réduire la résolution de l'image engendre évidemment une perte d'information rendant l'algorithme peu exploitable dans certains cas et notamment si le contraste est faible, et quand les panneaux à détecter sont éloignés ou petits, comme illustré sur la Figure 1.1.

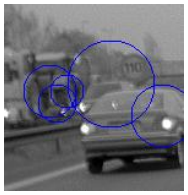


Figure 1.1 - Détection de cercles sur une image sous-échantillonnée

Pour pouvoir appliquer ensuite une reconnaissance « globale » du panneau, il est nécessaire d'avoir une détection plus robuste, et en particulier mieux cadrée sur le panneau. L'approche que nous avons développée consiste à effectuer, après la transformée de Hough appliquée à la totalité de l'image mais significativement sous-échantillonnée, une étape de raffinement de la recherche par Hough, en réappliquant l'algorithme à pleine résolution mais uniquement sur les zones correspondant aux premières détections (imprécises) issue de la première étape.

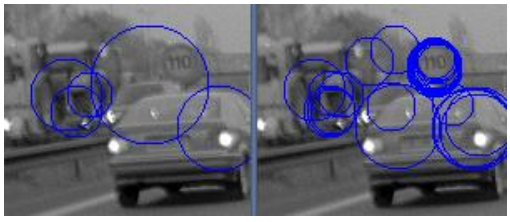


Figure 1.2 - A gauche détection de cercles avec la transformée de Hough circulaire appliquée sur toute l'image, mais à faible résolution, à droite résultat en réappliquant Hough à pleine résolution mais uniquement sur les zones détectées à gauche.

C'est ainsi que la nouvelle recherche de cercles en renvoie de nombreux, et affine correctement les premiers résultats dans la plupart des cas (cf. Figure 1.2, le panneau qui est maintenant correctement détecté). En revanche, la nouvelle technique fournit de nombreux « candidats cercles » qu'il faut donc filtrer en ne gardant que les « meilleurs » cercles. Se contenter de conserver celui ayant la plus grande valeur d'accumulateur n'étant pas satisfaisant, une approche originale a été proposée : celle-ci est inspirée de ce qui est fait dans le domaine des contours actifs (souvent nommés *snakes* dans la littérature, cf. (Kass, Witkin et Terzopoulos 1987)) où l'idée est, à partir d'une courbe de base, de minimiser une certaine énergie afin de déterminer la version déformée de cette courbe épousant au mieux le contour recherché. Ainsi la courbe se déplace et épouse lentement les contours des objets en fonction de divers paramètres comme l'élasticité, la tolérance au bruit, etc. Dans notre cas, il ne s'agit pas de faire évoluer une courbe, mais de rechercher parmi les cercles du groupe de cercles pré-détectés celui dont « l'énergie » est la plus faible. ***L'énergie est définie afin de minimiser la variance du cercle par rapport au gradient de l'image***. L'idée sous-jacente est que pour le cercle le mieux dimensionné et centré sur le panneau, le gradient de luminance entre l'intérieur et l'extérieur devrait être assez uniforme tout le long du bord du cercle. Pour que les calculs restent rapides, le calcul effectué en pratique se limite à la variance sur les 8 « points cardinaux » autour du cercle. Cette nouvelle et dernière approche donne de bons résultats (détection assez précise) comme en témoignent les Figure 1.3 et 1.4. De plus, le calcul ne se faisant plus que sur 8 points et non plus sur l'ensemble de l'accumulateur, le temps de calcul s'en trouve considérablement réduit, ce qui rend cette nouvelle approche de l'algorithme de Hough exploitable pour des applications temps réel. Cette amélioration originale de l'algorithme de Hough a fait l'objet d'un ***brevet déposé par la société Valeo, et dont je suis co-inventeur*** : "*Method of circle detection in images for round traffic sign identification and vehicle driving assistance device*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi and Benazouz Bradaï, Brevet PCT/EP2010/007569 déposé par VALEO le 11/12/2010.

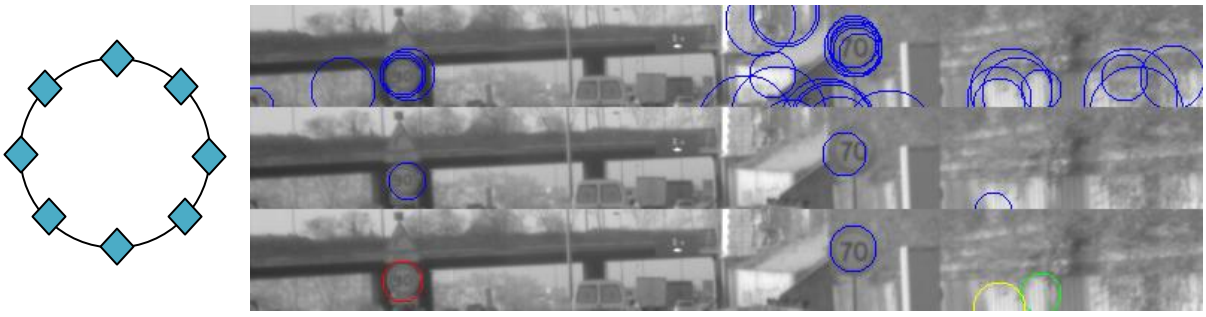


Figure 1.3 – A gauche, les 8 « points cardinaux » où est calculé le gradient de luminance entre intérieur et extérieur du cercle ; c'est le cercle pour lequel la variance de ces 8 gradients est minimale qui est conservé. A droite, résultat de la minimisation de variance de gradient : par groupe d'images et de haut en bas, toutes les détections (groupes de cercles), meilleur cercle au sens de l'accumulateur, meilleur cercle au sens de la variance

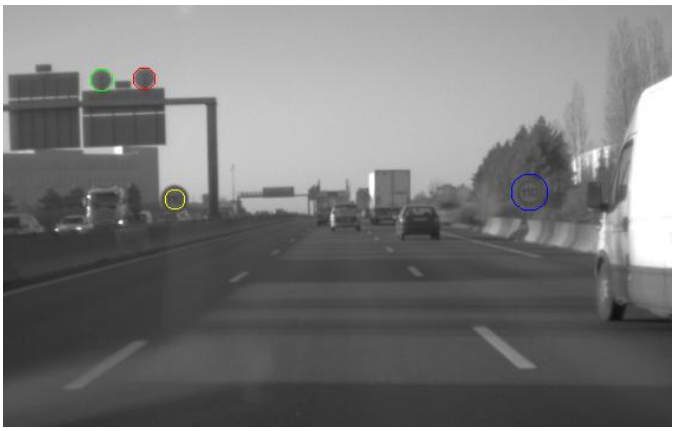


Figure 1.4 - Exemple de résultats de la nouvelle approche de l'algorithme de Hough

1.2.2. *Detection basee contours (en niveaux de gris) de panneaux et panonceaux rectangulaires*

La détection des panneaux rectangulaires est plus difficile à effectuer que celle des panneaux circulaires. A noter que ce point est critique pour la détection des panneaux de limite de vitesse aux Etats-Unis, où ces panneaux (comme beaucoup d'autres d'ailleurs) sont de cette forme (cf. Figure 1.10). Ce problème se pose aussi avec acuité pour la détection et reconnaissance des panonceaux parfois situés sous les panneaux, et qui en précise ou limite la portée, par exemple à certaines voies ou catégories de véhicules (cf. Figure 1.19). Nous avons donc mené des recherches sur ce point dur de la détection de panneaux.

Une première approche a été développée en 2006 par Lowik CHANUSSOT dans le cadre de son stage de fin d'études d'ingénieur (Centrale Nantes) que j'ai encadré. La recherche des rectangles s'effectue à partir de la carte de contours obtenus à l'aide du filtre de Canny et fait appel à différentes heuristiques. Les étapes de l'algorithme sont résumées sur la Figure 1.5, et détaillées ci-dessous :

1. ***Application du filtre de Canny*** pour extraire les pixels de contour
2. ***Filtrage des contours suivant leur orientation***.
   En supposant que les panonceaux sont majoritairement horizontaux, il est possible de filtrer la carte obtenue pour ne conserver que les contours horizontaux et verticaux. Les intervalles correspondant sont $[-\varepsilon;+\varepsilon][90-\varepsilon\ ;\ 90+\varepsilon\ ]$, où ε correspond à la tolérance sur les orientations et permet de prendre en compte de petites rotations.
3. ***Regroupement des contours en segments***.
   Cette étape sert d'intermédiaire entre bas (pixel) et haut niveau (rectangle). Il s'agit de déterminer les segments dans l'image à partir d'un ensemble d'heuristiques et de paramètres. Ces derniers doivent être soigneusement choisis pour garder un compromis robustesse/précision acceptable.
4. ***Association des segments en paires parallèles***
   Pour réduire la complexité de la tâche de recherche de rectangles, les paires de segments parallèles sont d'abord extraits. Certains a priori sur la taille des objets recherchés permettent encore de réduire le nombre de candidats.
5. **Formation des rectangles**

Figure 1.5 – Algorithme de recherche de rectangles basée contours.

Le principal avantage de cette méthode est sa rapidité par rapport à d'autres approches utilisant les contours. Cette technique a fait l'objet d'un autre ***brevet déposé par Valeo, et dont je suis co-inventeur*** : "*Procédé de détection d'un objet cible*", Herbin Anne, Chanussot Lowik et Moutarde Fabien, Brevet FR2917869 déposé par VALEO, éd. INPI. - France, 25 06 2007.

1.2.3. DETECTION DE PANONCEAUX RECTANGULAIRES BASEE CROISSANCE DE REGIONS A PARTIR DE GRAINES CONTRASTEES

La technique précédente est d'une utilisation assez complexe du fait des nombreux paramètres. Elle souffre aussi des inconvénients liés aux méthodes utilisant les bords, à savoir sensibilité aux bruits, occultations ainsi qu'aux rotations. ***C'est pourquoi nous avons développé, dans le cadre du projet SPEEDCAM et de la thèse d'Anne-Sophie Puthon, une nouvelle technique*** qui s'appuie sur la caractérisation des panneaux ou panonceaux par le fait qu'ils contiennent des pictogrammes ou du texte fortement contrastés (typiquement, noir sur blanc-crème). D'où l'idée d'utiliser une croissance de région (Chang et Li 1994) à partir de ces zones de fort contraste, pour englober la zone relativement claire et homogène constituant le fond du panneau ou panonceau. La méthode proposée consiste à partir de « graines » auxquelles sont ensuite agglomérés les pixels voisins tant que leur inclusion respecte les critères d'homogénéité de la zone obtenue. Le principe est le suivant :

1. Sélection des "graines" initiales
   Il s'agit de choisir un ensemble initial de pixels depuis lesquels sera effectuée la croissance de région ; la difficulté est de faire en sorte que ceux-ci soient bien à l'intérieur du (ou des) panonceau(x), s'il y en a. Pour cela nous exploitons le fait que tous les panonceaux contiennent des informations (pictogramme ou texte) de couleur noire qui contrastent fortement avec le fond du panonceau qui est blanc-crème. En pratique, nous détectons ces extrema locaux de contraste en appliquant l'opérateur morphologique de reconstruction à partir de marqueurs (Vincent 2003), puis en effectuant une soustraction et un filtrage, puis l'extraction des composantes connexes, et enfin la détermination des bords extérieurs de ces composantes. Cette succession d'opérations est illustrée sur la Figure 1.6.

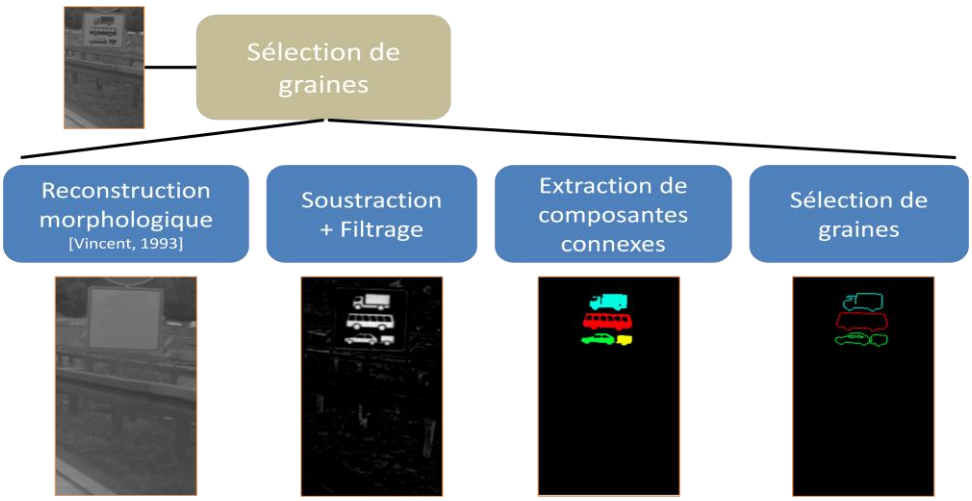


Figure 1.6 – Illustration de l'algorithme de sélections de « graines contrastées ».

2. Itération d'un opérateur d'agrégation
   A partir de la région courante R (initialisée avec les « graines »), parmi les pixels p adjacents à R (au sens 4-connexité ou 8-connexité), sont présélectionnés ceux qui vérifient le critère d'homogénéité suivant :

$$\exists q \in N_v(p) \cap R \ tq \left| \frac{I(p)}{I(q)} - 1 \right| \leq \kappa_L \qquad \text{(Eq. 1.1)}$$

   où $N_v(p) \cap R$ est l'ensemble des pixels de R adjacents à p, et I(q) est la valeur du pixel q.

Ensuite, parmi ces pixels adjacents vérifiant le critère d'homogénéité locale, seuls sont effectivement ajoutés à la région R ceux qui vérifient *en plus* le critère d'homogénéité global suivant :

$$\left| \frac{I(p)}{\mu_0} - 1 \right| \leq \kappa_G \qquad \text{(Eq. 1.2)}$$

où μ_0 est la moyenne des pixels de la région initiale R_0.

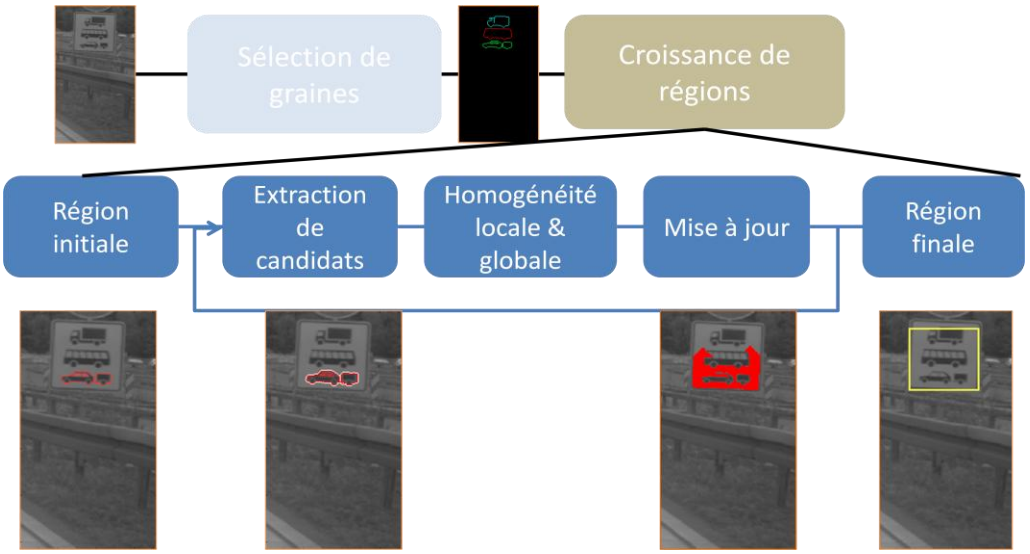


Figure 1.7 – Illustration de la croissance de région à partir des « graines contrastées ».

Cette approche a été évaluée, et comparée à des techniques alternatives, sur une large base de panonceaux. Elle s'avère fournir un taux de bonne détection légèrement inférieur à une approche de segmentation basée Graphe, mais un taux de fausses alarmes (détections erronées) sensiblement plus bas, comme illustré à droite sur la Figure 1.8.

Plus de détails sur cette méthode peuvent être trouvés dans un article présenté à la conférence ITSC'2012 : (Puthon, Moutarde et Nashashibi, Subsign detection with region-growing from contrasted seeds 2012)

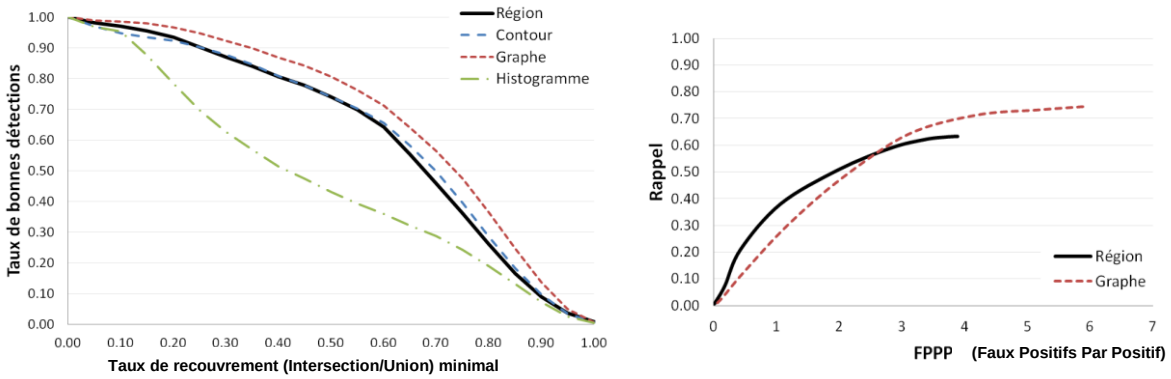


Figure 1.8 – Evaluation comparative de la détection de panonceaux par croissance de région à partir des « graines contrastées ».

1.2.4. *Reconnaissance de panneaux de limites de vitesse*

Du fait de l'application en vue (« Assistant de Limite de Vitesse ») nos recherches en TSR ont été axées sur les panneaux de *limites de vitesse*. Nous avons conçus dans cet objectif une approche originale et spécifique pour la reconnaissance des panneaux de limite de vitesse, passant par

l'extraction puis la reconnaissance des chiffres de la limite. Curieusement, dans la littérature, seuls (Torresen, Bakke et Sekanina 2004) utilisent une technique similaire, tandis que toutes les autres recherches publiées effectuent une reconnaissance *globale* des panneaux.

Figure 1.9 - Illustration de la variabilité de l'aspect des panneaux de limitation de vitesse : en gris, des panneaux allemands (à gauche de chaque paire) comparés à leur équivalent français ; en couleur, à droite, 2 variantes d'un même panneau français.

Un avantage de notre technique est qu'elle permet de réaliser une reconnaissance robuste de la totalité des variantes d'un même panneau d'un pays à l'autre (cf. Figure 1.9), et s'adapte aussi facilement à la reconnaissance des panneaux américains, qui sont d'un type totalement différent, rectangulaires et avec de nombreuses variantes (cf. Figure 1.10).

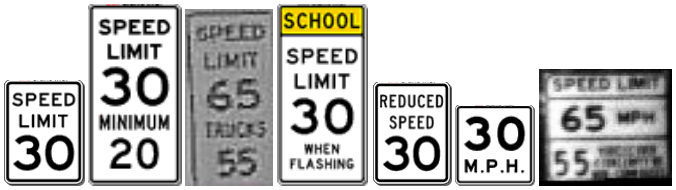


Figure 1.10 - Illustration de quelques uns des panneaux de limitation de vitesse américains qui prouvent la complexité d'utiliser un processus de reconnaissance globale des panneaux dans ce pays.

Pour isoler les chiffres, une extraction des composantes connexes sera utilisée. Le choix de l'algorithme de reconnaissance d'image / de classification (SVM, RN,…) est beaucoup moins critique, ces algorithmes donnant tous des résultats satisfaisants. De plus, notre approche passant par la segmentation et reconnaissance des chiffres pourrait permettre, dans une version ultérieure, d'utiliser un algorithme d'OCR (reconnaissance de caractère) bien éprouvé et robuste, issu par exemple de la reconnaissance des documents scannés.

Le fonctionnement de notre approche de détection et reconnaissance des panneaux de limite de vitesse, utilisant une caméra monochromatique embarquée dans le véhicule, est schématisé sur la Figure 1.11, et se compose des 6 étapes suivantes :

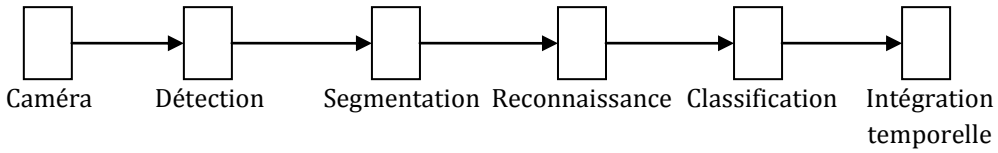


Figure 1.11 - Fonctionnement général du système

1. Un flux d'images en niveaux de gris est acquis par une caméra embarquée monochrome (Figure 1.12a).
2. De ce flux d'images sont extraits les panneaux potentiels grâce à l'algorithme de Hough (i.e. sont détectés tous les cercles dans chacune des images) pour les panneaux circulaires de type européen (Figure 1.12b).
3. Chaque imagette issue d'un cercle subit une normalisation d'histogramme suivie d'un seuillage binaire local adaptatif (Figure 1.12c).
4. Une segmentation des caractères, via une extraction de composantes connexes (blobs) est effectuée sur chacune de ces imagettes en noir et blanc (Figure 1.12d).
5. Les composantes respectant certains critères géométriques (ex : ratio hauteur/largeur) sont analysées par un réseau de neurones (de type perceptron multi-couches) afin d'en reconnaître le chiffre (Figure 1.12e). La topologie du réseau de neurones a été déterminée de

façon empirique (Tableau 1.1) : il apparaît qu'avec 20 neurones sur la couche cachée, l'apprentissage est optimal (i.e. aussi bon qu'avec plus de neurones, et meilleur qu'avec moins). La taille de l'entrée est de 64 neurones, pour reconnaître des imagettes de 8x8 pixels contenant un chiffre. Le réseau de neurones a été entrainé sur une base d'exemples contenant des imagettes de chiffres extraits de nos enregistrements (qui constituent donc de vrais exemples et non des exemples artificiels) dont 2772 images de chiffres (provenant de panneaux français, allemands et italiens) et 2790 exemples négatifs (des « non-chiffres »).

6. Le flux de chiffres reconnus est classifié (i.e. assemblé et validé 1 + 1 + 0 → 110 ou invalidé 1 + 9 + 0 → 190, limitation qui n'existe pas) puis intégré temporellement afin d'extraire un niveau de confiance géométrique et temporel de la détection pour valider ou invalider la détection d'un panneau (Figure 1.12f).

Figure 1.12a - Image initiale

Figure 1.12b - Détection de cercles

Figure 1.12c - Image binaire

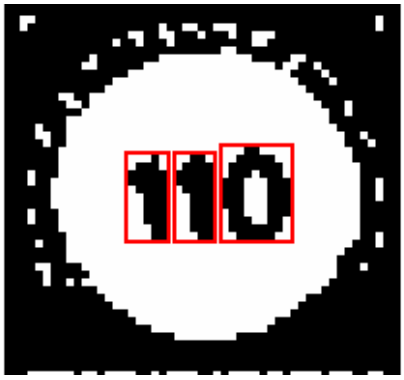
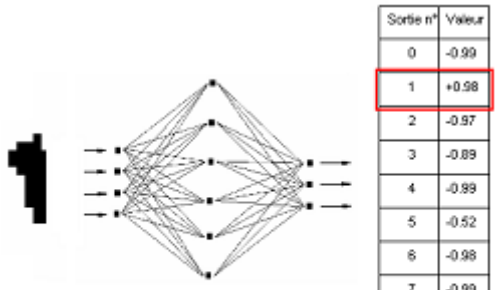


Figure 1.12d - Extraction des blobs



Sortie n°	Valeur
0	-0.99
1	+0.98
2	-0.97
3	-0.89
4	-0.89
5	-0.52
6	-0.98
7	-0.99
8	-0.97
9	-0.88

Figure 1.12e - Reconnaissance des blobs

Figure 1.12f – Résultat

Tableau 1.1 - Taux de bonne classification du réseau de neurones sur les bases de chiffres

Taille de la couche cachée	Taux de bonne classification sur la base d'apprentissage	**Taux de bonne classification sur la base de test**
5	74%	**72%**
10	95%	**93%**
15	96%	**95%**
20	97%	**96%**
25	97%	**96%**
30	97%	**96%**
35	97%	**96%**

Quelques commentaires s'imposent sur cette première version, présentée au symposium IV'2007 (Moutarde, Bargeton, et al. 2007). Le système fonctionne globalement bien et est très prometteur. La segmentation des caractères par extraction de composantes connexes est assez robuste aux transformations géométriques, par exemple quand les panneaux sont inclinés et non droits (Figure 1.13).

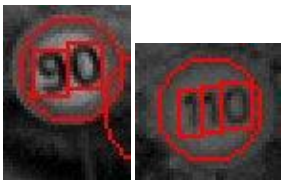


Figure 1.13 - Illustration de la robustesse de la segmentation par extraction de composantes connexes.

Segmentation globale des chiffres

La segmentation est en revanche inefficace quand deux chiffres sont trop proches (ce qui arrive quand le panneau à reconnaître est loin du véhicule). Par conséquent, dans les cas où le panneau est toujours trop petit dans l'image, ce qui arrive quand il est physiquement trop petit et/ou quand il est toujours loin du véhicule (par exemple sur une route à 3 voies, quand le panneau est à l'opposé du véhicule), le système ne peut le détecter. Nous avons donc proposé, dans le cadre de la thèse d'Alexandre Bargeton que j'ai encadrée, un algorithme permettant d'éviter ce problème en commençant par effectuer la segmentation globale de la boîte englobant l'ensemble des chiffres. Cette méthode utilise une technique de propagation orientée de pixel en pixel pour déterminer d'abord les limites haute et basse, puis un balayage des colonnes de pixels pour estimer les bords droit et gauche de l'ensemble des 2 ou 3 chiffres situés au centre du panneau. Le principe de propagation pour trouver les limites verticales est illustré sur la Figure 1.14a, et celui du balayage pour estimer les bords sur la Figure 1.14b.

Fig. 1.14a – Recherche des limites verticales : à gauche, propagation (en rouge) dans chaque quart ; à droite, le résultat est la borne haute et la borne basse du nombre.

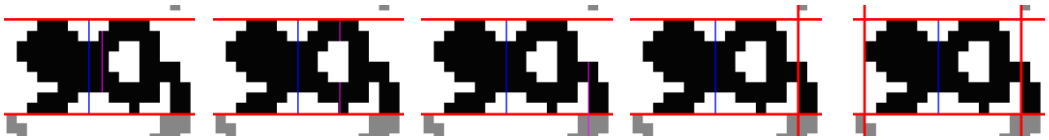


Figure 1.14b - Recherche des limites horizontales : exploration colonne par colonne (cf le segment pourpre) jusqu'à arrêt sur colonne blanche (à gauche) ou dépassement du noir (à droite).

Cet algorithme a été testé et validé sur des enregistrements effectués dans diverses conditions : jour et nuit (voir Figures 1.15 et 1.16), ainsi qu'en cas de très fort éclairage du panneau, ou au contraire de contre-jour sévère induisant une image binarisée de mauvaise qualité (cf Figure 1.15).

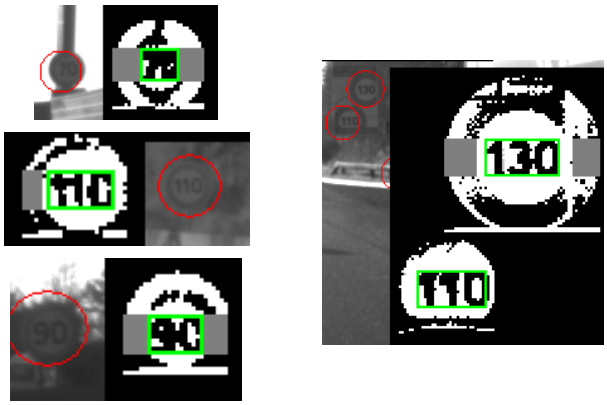


Figure 1.15 - Segmentation globale du nombre.

Une fois déterminée la boîte englobante du groupe de chiffres, la binarisation adaptative est à nouveau effectuée, mais en la restreignant à cette zone (au lieu de la totalité du panneau), ce qui permet d'obtenir une bien meilleure segmentation des chiffres (voir Figure 1.16).

Figure 1.16 - A gauche, encadrement du nombre ; à droite, binarisation restreinte à la zone du nombre.

L'amélioration apportée par notre approche a été effectuée tout d'abord uniquement sur les cas "difficiles" qui n'étaient pas correctement reconnus avec notre technique initiale de simple segmentation et reconnaissance des chiffres. Cette évaluation a montré que notre « segmentation globale » permet de reconnaître correctement 61% de ces cas difficiles tel que celui illustré sur la Figure 1.17.

Figure 1.17 – Exemple d'un panneau tagué mais correctement reconnu grâce à notre approche avec segmentation globale.

Expérimentations et résultats

L'évaluation d'un système de détection et reconnaissance de panneaux peut se faire de diverses manières. Elle est en particulier souvent faite uniquement sous forme de taux de bonne reconnaissance sur des bases d'images, ou encore en mesurant ce même taux image-par-image. Mais ce qui compte réellement dans les applications est de détecter, reconnaître correctement et

« valider » au moins une fois le panneau durant sa visibilité sur la vidéo. C'est donc cette dernière mesure que nous utilisons. L'évaluation a été effectuée sur un ensemble d'enregistrements effectués en France et en Allemagne, dans les divers environnements (autoroute, ville, routes de campagne) et une grande variété de conditions d'illuminations. L'ensemble contient environ 140 passages de panneaux, couvrant 11 types de limites (30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130). Comme indiqué dans le Tableau 1.2, le taux de validation correcte obtenu est de 94% avec notre système amélioré incluant la segmentation globale du groupe de chiffres, contre 85% pour notre algorithme initial effectuant une simple segmentation chiffre par chiffre. De plus, le taux de *mauvaise* classification est inférieur à 1%, et la quasi-totalité des 6% non correctement reconnus sont donc simplement des panneaux non détectés (ceci la plupart du temps à cause d'un contraste trop faible qui empêche de détecter correctement le cercle de bordure du panneau). Encore plus important, il est à noter qu'aucune fausse alarme validée (panneau « vu » là où il n'y en a en fait aucun) n'a été constatée sur plusieurs heures de vidéo testées.

Tableau 1.2 - Evaluation globale de la détection/reconnaissance avec notre approche des panneaux de limites de vitesse européen sur les routes allemandes et française.

Technique pour la reconnaissance de panneau	Panneaux détectés, reconnus, et validés avec le type correct	Panneaux mal classés
Segmentation directe des chiffres	85 %	0,7%
Segmentation préalable globale du nombre	94 %	0,7%

Plus de détails sur cette approche utilisant une segmentation globale préalable du nombre peuvent être trouvés dans l'article qui a été présenté au symposium IV'2008 (Bargeton, et al. 2008).

Notre approche fondée sur l'extraction et reconnaissance des chiffres, au lieu d'une reconnaissance globale, facilite clairement la gestion de la variabilité des panneaux au sein des pays européens. En utilisant un module universel de reconnaissance de chiffres, il serait a priori possible d'obtenir une reconnaissance robuste et totalement pan-Européenne, sans pour autant avoir besoin de collecter des exemples des panneaux dans tous les pays.

Figure 1.18 – Reconnaissance réussie pour un panneau de type LED, obtenue avec notre système sans refaire d'apprentissage, en faisant une simple inversion de luminance des pixels à l'intérieur des cercles détectés.

Une autre illustration de l'avantage de notre approche est la facilité avec laquelle nous avons pu l'adapter pour reconnaître aussi bien les panneaux type LED (voir Figure 1.18). C'est d'autant plus important que ces derniers sont de plus en plus répandus, en particulier sur les autoroutes à vitesse

limite variable, pour lesquelles il est évidemment indispensable d'avoir une reconnaissance de panneau puisque la cartographie GPS statique ne peut pas contenir la prescription réellement en vigueur.

1.2.5. *RECONNAISSANCE DES PANONCEAUX*

La signification précise des panneaux routiers, en particulier ceux de limite de vitesse, est souvent modifiée par un (ou plusieurs) panonceau(x) placés dessous (voir exemples sur la Figure 1. 19). Ces panonceaux sont absolument essentiels pour une interprétation correcte du panneau principal, car ils indiquent son "domaine" d'applicabilité, par exemple type de véhicules concernés (voiture avec caravane, camion, etc…), extension en distance de la limitation, voie particulière (notamment les sorties d'autoroute) à laquelle la vitesse limite s'applique, dates ou heures spécifiques, condition météorologique particulière (pluie, neige, etc…).

Figure 1.19 - Exemples de panneaux qui ne peuvent pas être interprétés correctement sans prendre en compte le panonceau situé dessous.

La tâche de reconnaissance des panonceaux n'est pas très différente de celle des panneaux principaux, mais une particularité est le nombre potentiellement très grand de types de panonceaux, puisque certains contiennent du « texte libre ». Tout ceci nous a conduit, tout comme (Liu, et al. 2011), à nous orienter vers une classification hiérarchique. En revanche, afin de ne pas avoir besoin de ré-entraîner tous les classifieurs à chaque ajout d'un nouveau type de panonceau, nous proposons plutôt une définition « manuelle » du premier niveau hiérarchique, en 4 « méta-catégories » de panonceaux : « Flèche », « Pictogramme », « Texte » et « Mixte » (voir Figure 1.20).

Figure 1.20 - Illustration des 4 "méta-classes" utilisées pour les panonceaux.

Pour mettre au point un classifieur, la première étape consiste à choisir les bons "descripteurs" à calculer sur l'imagette et fournir en entrée du classifieur. Une première approche possible, qui a été utilisée dans plusieurs travaux antérieurs, d'abord nous-mêmes (Hamdoun, et al., 2008), puis (Nienhüser, et al. 2010), consiste à simplement redimensionner l'imagette vers une taille fixe telle que 16x16 ou 12x12 pixels. Cependant, cette représentation présente le gros inconvénient d'être

extrêmement sensible à tout décalage des bords estimés ou erreur sur la boîte englobante : ainsi une légère imperfection de l'étape de détection-localisation du panonceau peut conduire à un vecteur d'entrée du classifieur totalement différent, et donc un résultat de classification aberrant. Il nous a donc paru préférable de nous tourner vers un descripteur qui « intègre » des mesures sur des portions assez larges de l'imagette, tout en conservant suffisamment d'information discriminante. Nous avons donc adopté le « pyramid-HoG » (p-HoG) proposé par (Bosch, Zisserman et Munoz 2007), qui calcule des histogrammes d'orientation du gradient sur hiérarchie de sous-images de type quad-tree (voir Figure 1.21).

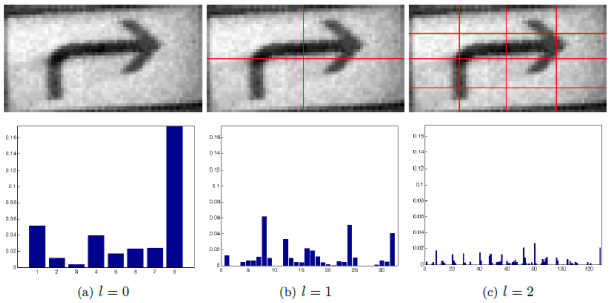


Figure 1.21 - Illustration du descripteur pyramid-HoG (pHoG) utilisé : à gauche, histogramme global ; au milieu niveau l=1 de subdivision (histogramme sur chaque quart) ; à droite, niveau l=2 de subdivision.

Après avoir entraîné un premier classifieur de type « SVM à noyau Gaussien » sur le descripteur p-HoG, nous avons constaté que les résultats étaient décevant, et ce probablement du fait que le p-HoG n'intègre que des mesure de gradient, et pas de mesures colorimétriques. Plutôt que d'ajouter un histogramme usuel de 256 niveaux de gris intégré sur toute l'image, qui perd toute information de distribution spatiale, nous avons préféré utiliser une information colorimétrique extrêmement simple (la proportion de pixels sombres), mais calculée sur la même hiérarchie de sous-images : comme illustré sur la Figure 1.22, nous effectuons une binarisation fondée sur la reconstruction morphologique (la même que celle exploitée pour la phase de détection).

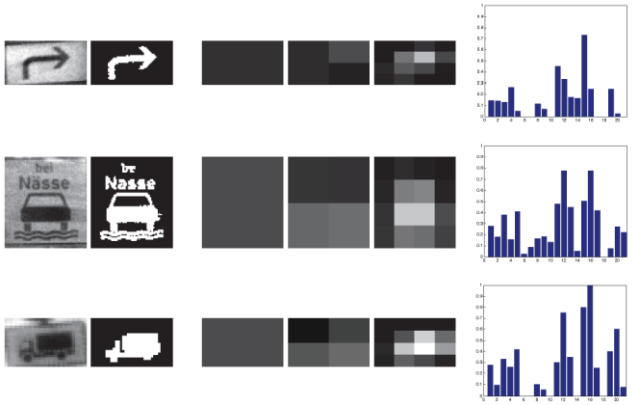


Figure 1.22 - Illustration de la "proportion de pixels sombres" calculée sur la même pyramide que le p-HoG : à gauche image originale et son inverse binarisée, au milieu visualisation de la "proportion de sombre" dans chaque sous-région des 3 niveaux de la pyramide ; à droite, vecteur résultant.

Ce nouveau descripteur, que nous nommons p-HoG_Dp, concatène donc le p-HoG standard puis les « proportions de pixels sombres » calculées sur les mêmes sous-images. Comme le montre la Figure 1.23, ceci nous permet d'obtenir une amélioration significative des performances en terme de précision-rappel.

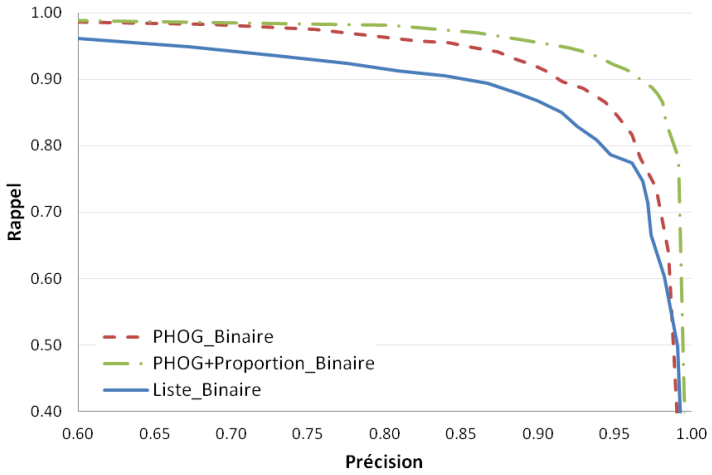


Figure 1.23 - Comparaison des performances de classification obtenues avec 3 descripteurs : pixels de l'imagette redimensionnées (courbe du bas, en bleu), p-HoG (courbe du milieu, en pointillés rouges), et p-HoG_Dp (courbe du haut en traits mixtes verts).

Expérimentations et résultats

Afin d'évaluer le taux de bonne reconnaissance atteignable avec ce descripteur, nous avons effectué des tests sur une large base d'images de panonceaux. Comme mentionné plus haut, nous catégorisons, dans une première étape, les panonceaux en 4 « méta-classes » (« Texte », « Fleche », « Pictogramme » et « Mixte »).

Tableau 1.3 – Caractéristiques des bases d'exemples de panonceaux (apprentissage et test).

Catégorie	Exemples	Apprentissage	Test
Texte		*1693*	*697*
Flèche		*267*	*99*
Pictogramme		*2030*	*473*
Mixte		*1293*	*521*
Négatif		*13856*	*4944*

Après apprentissage de classifieurs type SVM à noyau Gaussien, nous obtenons sur la base de test des taux satisfaisants de bonne classification, comme l'illustre la Figure 1.24. Cette dernière montre les courbes précision-rappel évaluées séparément pour chaque méta-classe (et sur les imagettes encadrées manuellement, ce qui émule une étape de détection parfaite).

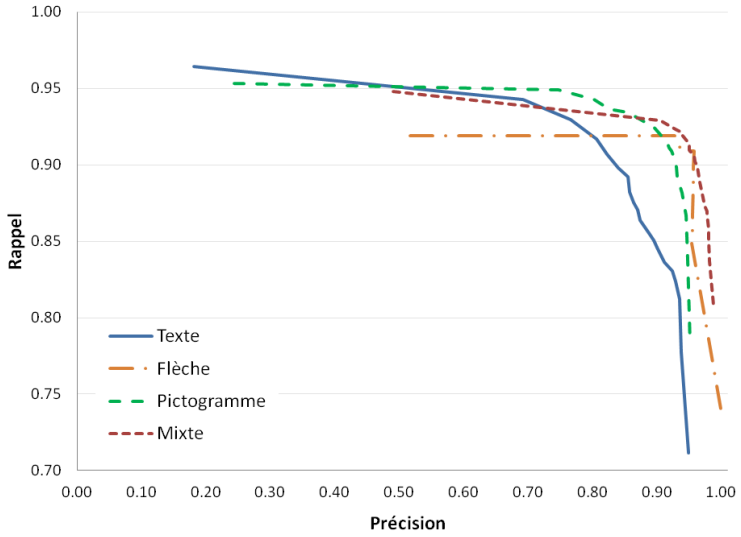


Figure 1.24 – Courbes précision-rappel pour chacune des méta-classes, calculées sur la base de test (en utilisant les boîtes englobantes de la vérité-terrain, créée manuellement).

Plus de détails sur les résultats de catégorisation sont donnés dans la matrice de confusion présentée dans le Tableau 1.4. La catégorie « Texte » est celle qui présente la moins bonne précision, ce qui s'explique aisément par des confusions difficilement évitables avec notamment des glissières de sécurité qui ont le même type de régularité. Pour plus de précisions sur ce travail, le lecteur peut se reporter à l'article accepté à un workshop de IV'2013 (Puthon, Moutarde et Nashashibi 2013).

Tableau 1.4 – Matrice de confusion pour la reconnaissance des types de panonceaux.

	Sortie du classifieur					
Vérité-terrain	*Nég.*	*Texte*	*Flèche*	*Pict.*	*Mixte*	**Rappel**
Négatif	***4766***	*107*	*0*	*49*	*22*	**96%**
Texte	*48*	***632***	*2*	*13*	*2*	**91%**
Flèche	*2*	*4*	***90***	*2*	*1*	**91%**
Pictogramme	*23*	*8*	*0*	***440***	*2*	**93%**
Mixte	*24*	*14*	*0*	*9*	***474***	**91%**
Précision	**98%**	**83%**	**98%**	**86%**	**95%**	**95%**

1.3. BILAN ET PERSPECTIVES SUR MES RECHERCHES EN DETECTION ET RECONNAISSANCE DE PANNEAUX ROUTIERS

Les recherches que j'ai effectuées et/ou encadrées (notamment via 2 thèses sur le sujet) sur la détection et reconnaissance de panneaux routiers ont donc apporté les contributions suivantes :

- amélioration de la détection de panneaux circulaires (avec un brevet déposé par Valeo) ;
- deux techniques originales et très différentes pour la détection de panneaux ou panonceaux rectangulaires (là encore, un brevet déposé par Valeo) ;
- approche originale et robuste pour la reconnaissance des panneaux de limite de vitesse ;
- cadre général pour la reconnaissance des très nombreux types de panonceaux.

Ces travaux ont donné lieu à ***2 brevets déposés par Valeo et dont je suis « co-inventeur »*** :

- "*Method of circle detection in images for round traffic sign identification and vehicle driving assistance device*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi and Benazouz Bradaï, Brevet PCT/EP2010/007569 déposé par VALEO le 11/12/2010.
- "*Procédé de détection d'un objet cible*", Herbin Anne, Chanussot Lowik et Moutarde Fabien, Brevet FR2917869 déposé par VALEO, éd. INPI. - France, 25 06 2007.

ainsi que ***6 publications dans des conférences internationales à comité de lecture*** :

- "*Recognition of supplementary signs for correct interpretation of traffic signs*", Anne-Sophie Puthon, Fabien Moutarde and Fawzi Nashashibi, proc. Workshop on Environment Perception and Navigation for Intelligent Vehicles, Intelligent Vehicle symposium (IV'2013), Gold Coast (Australia), Juin 2013.
- "*Subsign detection with region-growing from contrasted seeds*", Anne-Sophie Puthon, Fabien Moutarde and Fawzi Nashashibi, proc. 15th IEEE Intelligent Transportation Systems Conference (ITSC'2012), Anchorage (USA), 16-19 septembre 2012.
- "*Joint interpretation of on-board vision and static GPS cartography for determination of correct speed limit*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi and Anne-Sophie Puthon, proc. 17th ITS world congress (ITSwc'2010), Busan (Korea), 25-29 octobre 2010.
- "*Detection and Recognition of End-of-speed-limit and Supplementary Signs For Improved European Speed Limit Support*", Omar Hamdoun, Alexandre Bargeton, Fabien Moutarde, Benazouz Bradai and Lowik Chanussot, proc. of 15th World Congress on Intelligent Transport Systems (ITSwc'2008), New York City (USA), 16-20 novembre 2008.
- "*Improving pan-European speed-limit signs recognition with a new 'global number segmentation' before digit recognition*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi and Benazouz Bradai, proc. of IEEE Intelligent Vehicles Symposium (IV'2008), Eindhoven (Netherlands), 4-6 juin 2008.
- "*Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system*", Fabien Moutarde, Alexandre Bargeton, Anne Herbin and Lowik Chanussot, proc. of IEEE Intelligent Vehicles Symposium (IV'2007), Istanbul (Turkey), 13-15 juin 2007.

Toutes ces recherches m'ont permis à la fois d'acquérir une certaine expertise sur les techniques d'analyse temps-réel de vidéos embarquées (ainsi que sur le cadre plus général des systèmes d'aide à la conduite), et de développer un bon savoir-faire sur l'application pratique des systèmes de reconnaissance de formes les plus standards (réseaux neuronaux, SVM, etc…).

En ce qui concerne la suite de ces travaux, elle sera pour une bonne part assurée directement par l'industriel Valeo car le niveau de maturité de la technologie fait qu'elle relève désormais plus du développement et de l'industrialisation que de la Recherche à proprement parler. En revanche, ***un nouvel axe de recherche intéressant qui se fait jour (et que je commence à explorer dans le cadre d'une nouvelle thèse démarrée en octobre 2012 que je supervise) consiste à utiliser les détections de panneaux comme des indices qui, en combinaison avec d'autres, vont permettre d'inférer (de manière probabiliste et statistique) une information plus abstraite de contexte de conduite***, tel que « approche d'intersection », « zone de travaux », « chaussée séparée », etc… Ce contexte peut d'une part être une information indispensable pour la mise en œuvre des futures aides à la conduite telles que « Aide au franchissement d'Intersection » (« Intersection Assist ») et « Aide au Dépassement » (« Overtake Assist »), et d'autre part contribuer en retour à rendre plus robuste la reconnaissance ou l'interprétation des panneaux (cf. par exemple l'intérêt potentiel, pour une bonne fusion vision+GPS des informations de vitesse limite, de savoir si le véhicule se trouve ou non dans une zone de travaux).

2. DETECTION ET CATEGORISATION VISUELLE D'OBJETS (VOITURES, PIETONS, ETC…)

Un autre défi important dans la perception pour les systèmes avancés d'aide à la conduite (ADAS) est la détection, en temps-réel dans les vidéos, de catégories d'objets particuliers tels que véhicules et piétons. ***Ces travaux ont été menés notamment de 2006 à 2012 dans le cadre de 2 stages de Master2 recherche, un stage de fin d'étude et un stage d'initiation à la Recherche que j'ai encadrés.*** Le cadre de ces travaux a été d'une part ***deux collaborations directes avec des industriels*** (une avec ***PSA sur la détection de véhicule,*** et une autre avec l'équipementier automobile ***Valeo pour la détection de piétons),*** et d'autre part la contribution du laboratoire au ***projet collaboratif français « LOVe » sur la détection des vulnérables.***

2.1. INTRODUCTION A LA DETECTION VISUELLE DE CATEGORIES D'OBJETS

La détection visuelle, dans les images, d'objets d'une certaine catégorie se fait généralement en combinant trois « ingrédients » :

- une technique de « balayage multi-échelle » de l'image (voir Figure 2.1) ;
- un classifieur permettant d'estimer pour chacune des sous-fenêtres testées si oui ou non elle contient principalement (et entièrement) un objet de la catégorie cherchée ;
- un post-traitement des détections pour fusionner en une seule celles qui correspondent au même objet.

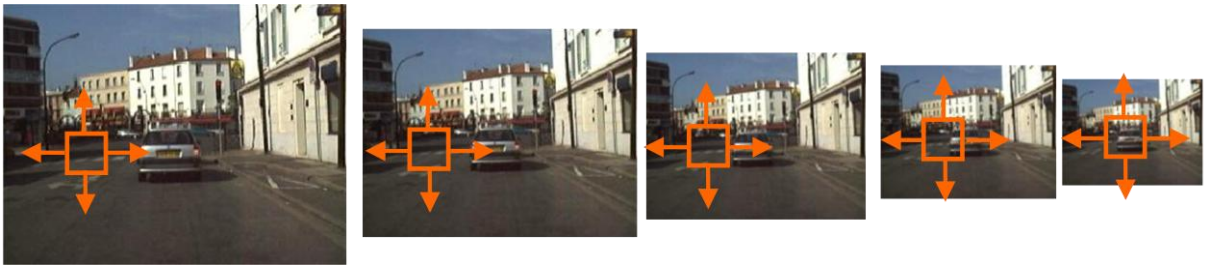


Figure 2.1 - Principe du balayage multi-échelles d'une image pour chercher des objets d'une catégorie.

L'élément le plus critique est naturellement le classifieur, et ses performances en terme de taux de reconnaissance et de fausses alarmes dépendent de divers éléments :

- les exemples d'apprentissage utilisés ;
- les primitives visuelles calculées sur la sous-fenêtre et fournies en entrée du classifieur ;
- l'algorithme d'apprentissage proprement dit (SVM, boosting, réseau neuronal, …).

Les primitives visuelles sont particulièrement critiques, car elles déterminent l'espace dans lequel va se faire la classification, et donc la plus ou moins grande difficulté de la tâche d'apprentissage. De plus, elles ont une influence déterminante sur le temps de calcul de la chaîne globale. Ce dernier point est essentiel dans les applications qui nous intéressent, puisque nous voulons pouvoir analyser en temps-réel des flux vidéos (notamment celui d'une caméra embarquée dans un véhicule).

Historiquement, une des premières approches publiées capable de détecter en temps-réel dans des vidéos des objets d'une certaine catégorie est celle de (Viola et Jones 2001), qui utilise le boosting avec sélection de primitives visuelles, pour localiser avec succès les visages dans les images. Dans leur cas, les primitives visuelles sont choisies dans une famille de filtres type Haar, qui calculent la différence entre somme de pixels dans des rectangles adjacents (voir Figure 2.2). Leur temps de calcul est très faible grâce à l'astuce consistant à pré-calculer une « image intégrale » qui contient en chaque « pixel » la somme des valeurs de pixels situés en haut et à gauche.

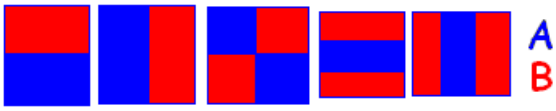


Figure 2.2 – Primitives visuelles de type filtres de Haar, utilisées par Viola & Jones.

Pour chaque type, position et taille de filtre, un « classifieur faible » est obtenu en appliquant la règle suivante :

if $\left|Area(A) - Area(B)\right| > Threshold$ **then True** (Eq 2.1)

else False

Ces primitives fournissent des indices qui peuvent être combinés sous forme de vote pondéré par l'algorithme adaBoost, de manière à obtenir un classifieur capable de très bonnes performances de reconnaissance sur une grande variété d'objets de la même catégorie. La détection de visages dans des images, en particulier, est un problème très bien traité par cette approche. De même d'excellents résultats peuvent être obtenus pour la détection et reconnaissance de voitures, par exemple. Quand la catégorie recherchée présente une variabilité d'aspects beaucoup plus grande, comme les piétons, la méthode n'est en revanche que partiellement satisfaisante. Ceci est probablement dû au fait que les primitives visuelles type Haar n'extraient pas une information suffisamment générale, puisque par nature elles repèrent surtout des contrastes verticaux ou horizontaux.

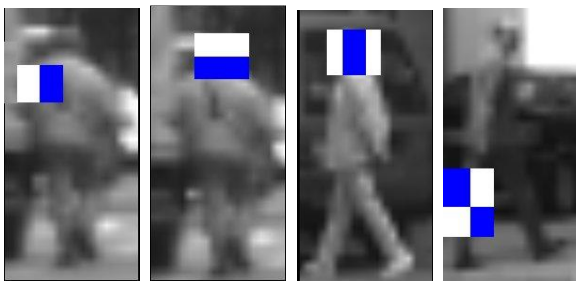


Figure 2.3 – Primitives type Haar sélectionnées par boosting pour la détection de piétons.

Des progrès considérables de performance ont été apportés par l'introduction dans (Dalal et Triggs 2005) des histogrammes d'orientation de gradient (HoG). Ces derniers consistent à calculer les angles des gradients dans l'image, puis à les agréger sur diverses sous-fenêtres sous forme d'histogrammes, qui contiennent une information très riche sur la statistique géométrique des contrastes dans l'image. Le principal inconvénient des HoG a été leur temps de calcul, mais divers travaux plus récents (Pettersson, Petersson et Andersson 2008) (Dollar, et al. 2009) ont proposé des manières de les calculer plus efficacement.

Figure 2.4 – Illustration des Histogrammes de Gradient orientés (HoG) : l'image est subdivisée en sous-régions, puis dans chacune les proportions des diverses orientations sont évaluées.

En parallèle de nombreuses recherches sur la catégorisation visuelle d'objet ont visé à identifier la

catégorie par détection conjointe de sous-parties spécifiques (par exemple les roues etc pour une voiture, les jambes la tête etc pour un piéton). Une part importante des meilleurs résultats sur les benchmarks publics tels que le « Pascal Visual Object Class challenge » ont ainsi été obtenus (voir par exemple (Csurka, et al. 2004), (Winn, Criminisi et Minka 2005) et (Perronnin, et al. 2006)) en exploitant la statistique des « points d'intérêt » (voir Figure 2.5). Ces « points d'intérêt » (PI en abrégé, ou keypoints en anglais), sont le plus souvent soit les "Scale Invariant Feature Transform" (SIFT) introduits dans (D. Lowe 2004), soit leur variante plus rapide à calculer Speeded-Up Robust Features (SURF) présentés dans (Bay, Tuytelaars et van Gool 2006), pour caractériser les catégories d'objets. Ces dernières utilisent généralement une représentation de type « sac de mots » (voir Figure 2.6) : les descripteurs de points d'intérêt sont partitionnés (par un apprentissage non-supervisé de type « clustering ») en groupes correspondant chacun à un « mot visuel », et chaque catégorie est caractérisée par la statistique des nombres d'occurrences de ces différents « mots visuels ».

Figure 2.5 – Illustration de points d'intérêt détectés (à gauche), et de leur relative stabilité par changement d'échelle et d'angle de vue qui permet de mettre en correspondance un même objet dans 2 images.

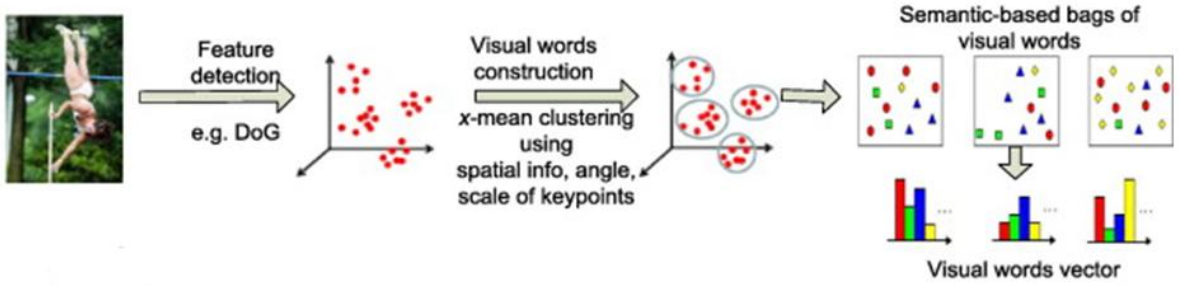


Figure 2.6 – Illustration du principe du « sac de mots visuels » pour décrire une image afin de la catégoriser (extrait de (Kesorn, et al. 2011)**).**

2.2. TRAVAUX REALISES EN DETECTION ET RECONNAISSANCE VISUELLE DE CATEGORIES D'OBJETS

2.2.1. DEVELOPPEMENT DE NOUVELLES « PRIMITIVES VISUELLES »

Avant que je commence mes travaux, (Abramson, Steux et Ghorayeb, YEF (Yet Even Faster) Real-Time Object Detection 2005) avaient proposé au sein du laboratoire une méthode inspirée de la désormais classique technique de Viola&Jones, mais avec une famille de primitives visuelles beaucoup plus rapides à calculer que les filtres type Haar. Ces primitives, baptisées « control points » effectuent uniquement quelques comparaisons de valeurs de pixels. Un classifieur faible de ce type est défini par 2 listes P^+ et P^- de positions de pixels testés, et il classe comme positive toute imagette telle que :

$$\left(\min\left(\left\{P_i^+\right\}_{i=1,2,..m+}\right) - \max\left(\left\{P_j^-\right\}_{j=1,2,...,m-}\right)\right) > S \text{ OU } \left(\min\left(\left\{P_j^-\right\}_{j=1,2,...,m-}\right) - \max\left(\left\{P_i^+\right\}_{i=1,2,...m+}\right)\right) > S \quad \text{(Eq. 2.2)}$$

où les P^+_i sont les points "positifs" de contrôle de la liste P^+ , les P^-_j les points de contrôle "négatifs" de la liste P-, et S est un seuil positif séparant toutes les valeurs de pixels de P- de toutes celles de P+ (voir Figure 2.7a).

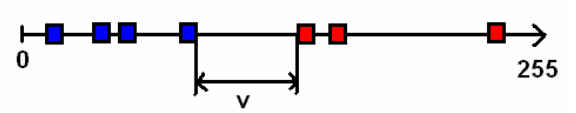


Figure 2.7a – Exemple de distribution de valeurs des pixels de P- et P+ classée positif (pour tout seuil S< V).

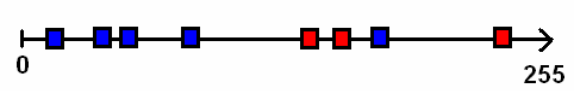


Figure 2.7b - Exemple de distribution de valeurs des pixels de P- et P+ classée négatif (quelle que soit la valeur du seuil S).

Cette famille de primitives visuelles, utilisée comme classifieurs faibles dans le cadre du boosting avec sélection de primitives, permet de faire des tests quelconques sur l'image à catégoriser (voir illustration sur la Figure 2.8), contrairement aux filtres type Haar qui se focalisent sur une seule région rectangulaire. Les performances obtenues sur la détection/reconnaissance de voitures se sont avérées très intéressantes, avec le 2° meilleur résultat au « Pascal VOC challenge » de 2006 (voir Figure 2.9).

Figure 2.8 – "Control points" sélectionnées par adaBoost pour la détection de piétons (à gauche) ou de voitures (à droite).

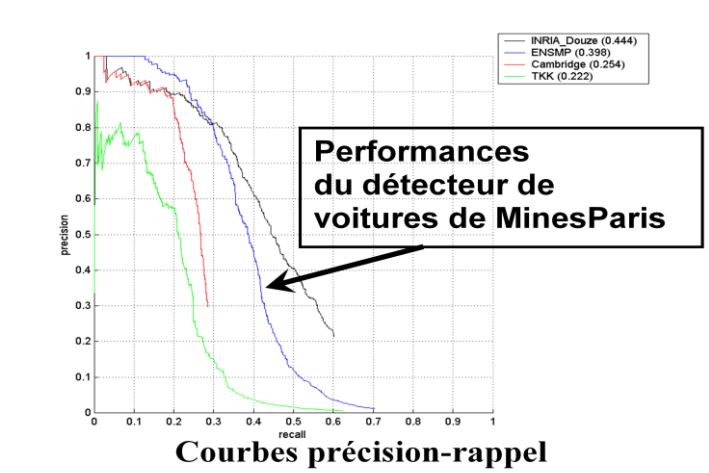


Figure 2.9 - Comparaison de notre détecteur de voitures (adaBoost + primitives type Control-Points) avec l'état de l'art, lors de l'édition 2006 du « Pascal Visual Object Class » challenge.

En revanche, un des inconvénients des « control points » est la cardinalité colossale de l'ensemble de toutes les primitives possibles (environ 10^{36} pour une fenêtre de détection de 32x32) qui oblige à recourir à une heuristique évolutionniste pour chercher une « bonne » primitive à chaque itération du boosting. J'ai d'ailleurs commencé mes recherches sur le sujet en me focalisant sur l'optimisation de cette recherche heuristique, comme rapporté dans ma publication (Abramson, Moutarde, et al. 2006).

Ensuite, dans les premiers travaux que j'ai encadrés sur ce sujet, nous avons proposé de restreindre cette famille en imposant une contrainte de contigüité des pixels testés, ce qui permet de réduire significativement l'espace de recherche (~10^{19} pour une fenêtre de détection de 32x32) et donc d'obtenir des meilleurs classifieurs faibles pour un temps de recherche identique durant l'apprentissage. En prime, les primitives obtenues, baptisées "connected-control-points", sont plus faciles à interpréter tout en restant de forme plus générale que les filtres type Haar (voir Figure 2.10).

Figure 2.10 – Primitives types "connected control-points" pour la détection de piétons (à gauche) ou de voitures (à droite).

Encore plus intéressant, il s'avère qu'au final les classifieurs forts obtenus par boosting avec cette nouvelle famille de primitives visuelles obtiennent des performances de reconnaissance significativement meilleures, tant sur la détection de voitures (cf. Figure 2.11) que sur le plus difficile problème de détection de piétons (voir Figures 2.12a et 2.12b).

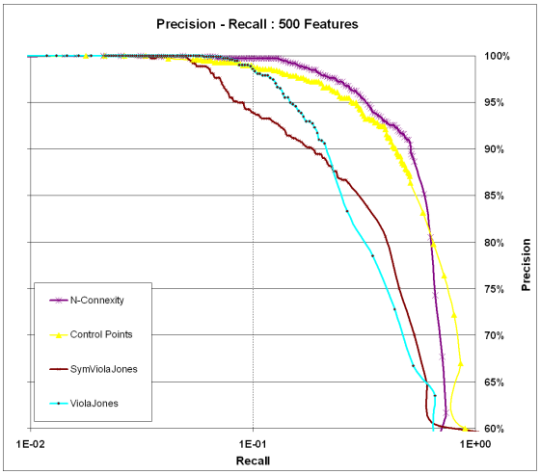


Figure 2.11 – Courbes précision-rappel pour la reconnaissance de voiture avec boosting de diverses primitives : « connected_control-points » (au-dessus), « control-points » simples (2° courbe), etc…

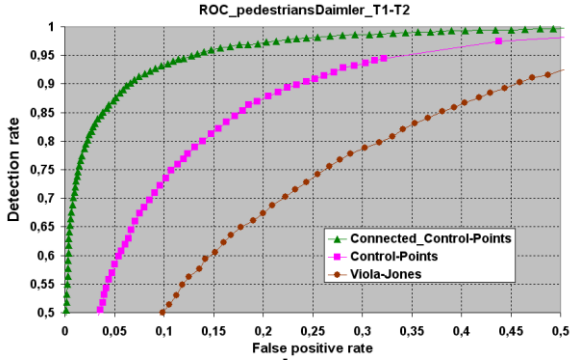


Figure 2.12a – Courbes ROC pour reconnaissance de piétons par boosting : les "connected control-points" sont nettement meilleurs que les simples « control points », et encore plus que les « Haar » de Viola-Jones.

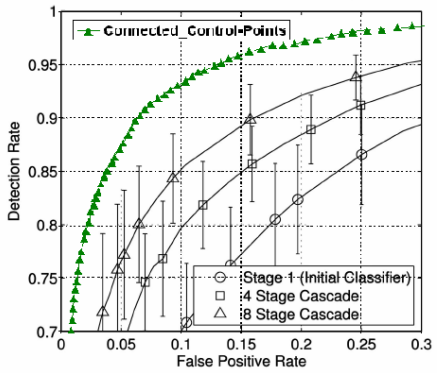


Figure 2.12b – Courbe ROC obtenue pour la détection de piétons avec boosting de nos "connected control-points", comparée à celles publiées sur la même base dans (Munder et Gavrila 2006) **pour une cascade de Haar.**

De plus, en comparant aux résultats des meilleures techniques (quadratic SVM et RBF SVM, ainsi que NN-LRF) rapportés dans (Munder et Gavrila 2006), il s'avère que le boosting avec nos « connected control-points » a des performances supérieures (Figure 2.13), ceci avec un temps de calcul de l'ordre de ~0.4ms par image-test sur un ordinateur portable Intel Core2 à 2 GHz, à comparer aux 250ms sur un PC Pentium IV.2 Ghz PC indiqués pour les meilleures méthodes de l'article de référence.

Plus de détails sur ces travaux peuvent être trouvés dans nos 2 publications suivantes : (Moutarde, Stanciulescu et Breheret 2008) et (Stanciulescu, Breheret et Moutarde 2007).

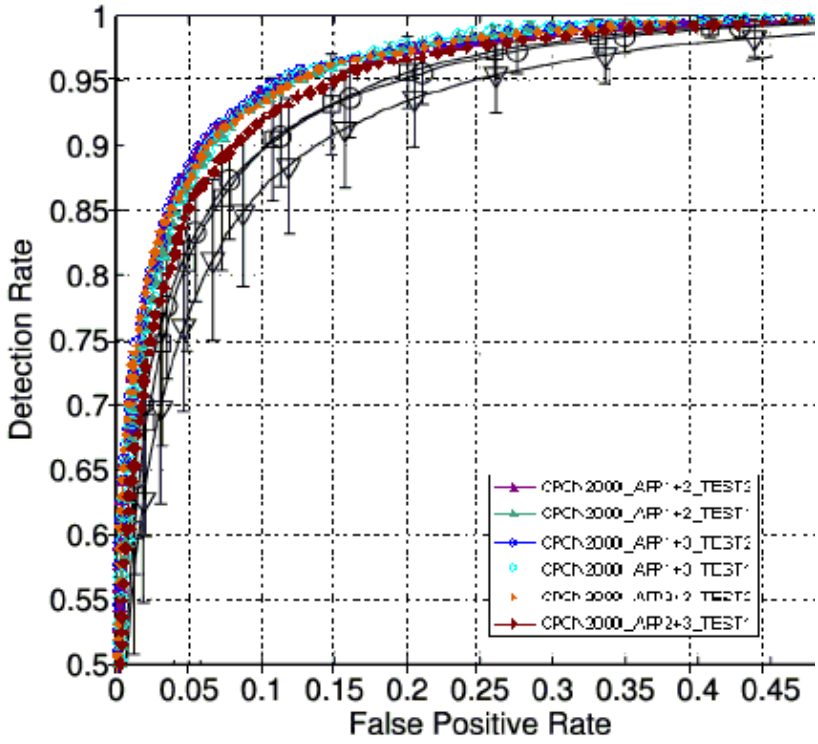


Figure 2.13 - Courbe ROC pour la détection de piétons avec nos «connected control-points» (courbes colorées, au-dessus), comparée aux performances des meilleures techniques (3 courbes du bas) testées sur la même base dans (Munder et Gavrila 2006)**.**

2.2.2. *Detection et reconnaissance a l'aide de points d'interet*

Par la suite, nos recherches en catégorisation d'objets se sont orientées vers l'utilisation des « points d'intérêt ». En effet, tant nos propres travaux (voir section précédente) que d'autres approches publiés récemment (Zhu, et al. 2006) (Pettersson, Petersson et Andersson 2008), ont montré que le résultat du boosting avec sélection de primitives dépend énormément de la famille de primitives utilisée. Par ailleurs, de nombreuses publications (voir par exemple (Csurka, et al. 2004), (Perronnin, et al. 2006), et (Leibe, Leonardis et Schiele 2008)) ont montré la puissance des techniques exploitant les points d'intérêt SIFT ou SURF pour caractériser les catégories d'objets.

Nous avons donc cherché à combiner les avantages du boosting par sélection de primitives et ceux de la caractérisation par les descripteurs de points d'intérêt. Notre idée a consisté à définir une famille de classifieurs faibles de type « présence de points d'intérêt », en espérant que chaque catégorie puisse être caractérisée par la présence simultanée de plusieurs types de points d'intérêt. Formellement, nous définissons chaque primitive par un vecteur D de descripteur SURF (dans $\Re^{64}$) associé à un seuil scalaire d, et le classifieur faible correspondant répond positivement pour une image I si et seulement si I contient au moins un point d'intérêt dont le descripteur D' vérifie $|D-D'|<d$, où |.| correspond à la norme L1 (somme des différences absolues) dans $\Re^{64}$. Ce principe est illustré schématiquement sur la Figure 2.14.

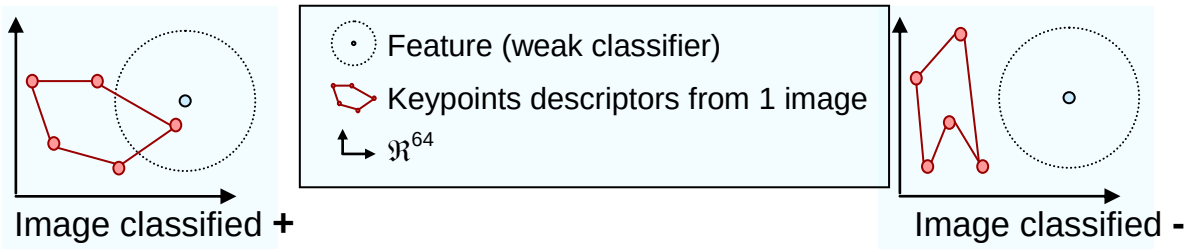


Figure 2.14 – Illustration schématique du nouveau type de classifieur faible proposé : « présence de point d'intérêt ayant un descripteur donné » (i.e., contenu dans une petite boule L1 de $\Re^{64}$).

La technique d'apprentissage utilisée est le boosting avec sélection de primitive, avec l'apprenant faible suivant : à chaque itération de boosting, le classifieur faible est construit en choisissant D parmi tous les descripteurs trouvés dans toutes les images d'exemples positifs, et en déterminant le seuil d à l'aide d'une matrice pré-calculée M contenant les distances M_{ij} entre chaque descripteur de point K_i trouvé dans un des exemples positifs et chaque image d'apprentissage I_j (la « distance » entre descripteur et image étant définie comme la plus petite distance entre K_i et les divers descripteurs de l'image I_j. La matrice M est de taille Q×N où Q est le nombre de points d'intérêt trouvés dans *l'ensemble de tous les exemples positifs*, et N est le nombre total d'images d'apprentissage.

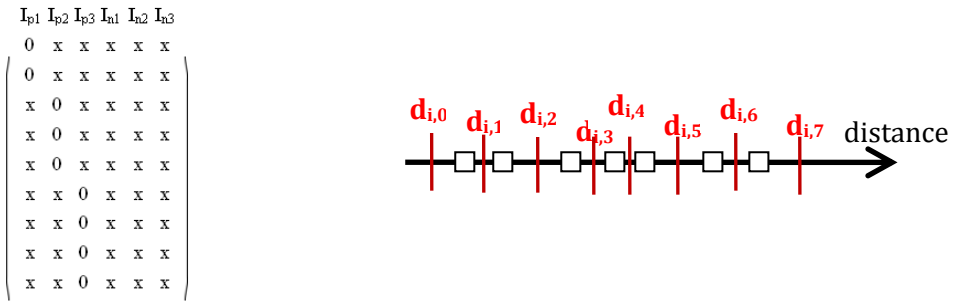
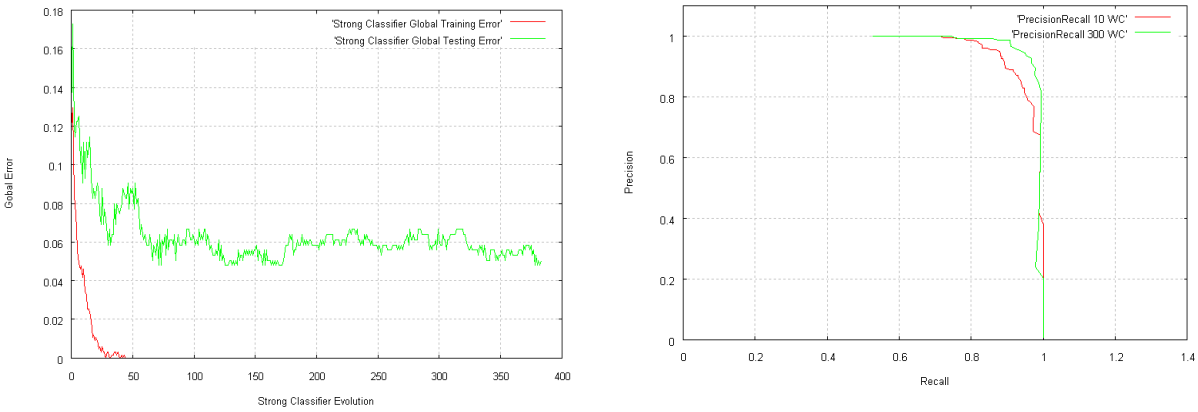


Figure 2.15 – A gauche, illustration de la matrice des « distances » entre points d'intérêts trouvés dans les exemples positifs (un par ligne) et toutes les images d'apprentissage (positives et négatives, une par colonne). A droite, détermination des seuils de distance à essayer pour le point d'intérêt i : une fois triées les distances (figurées par les carrés) contenues dans la ligne i, on choisit les $d_{i,k}$ comme les milieux entre les valeurs successives de distance.

Comme illustré sur la Figure 2.15, la matrice M contient au minimum un zéro sur chaque ligne (sur la colonne correspondant à l'image dont provient le point d'intérêt K de la ligne). Chaque ligne de cette matrice M est triée par ordre croissant de distance, ce qui permet ensuite de construire pour chaque ligne i l'ensemble $\{d_{ik},\ k=1,\ldots,N\}$ des valeurs candidates pour le seuil à associer au point d'intérêt K_i. Enfin, à chaque itération de boosting, une recherche exhaustive est faite parmi tous les couples (K_i,d_{ik}) pour déterminer le classifieur faible qui donne la plus faible erreur *pondérée* sur la base d'apprentissage :

$$(i^*,k^*) = \text{argmin}_{ik} \left(\sum_{j=1}^{N} w_j \, |h(K_i,d_{ik},I_j) - l_j| \right) \qquad \text{(Eq. 2.3)}$$

Un premier essai effectué sur la base UIUC de voitures vues latéralement (collectée par (Agarwal, Aatitf Awan et Roth 2004), et disponible à http://l2r.cs.uiuc.edu/~cogcomp/Data/Car/) a permis de montrer que notre approche permet effectivement de construire un classifieur fort, avec de très bons résultats de reconnaissance : ~95% de rappel (pourcentage bien reconnu de voitures latérales) pour ~95% de précision (pourcentage correct parmi toutes les images que le classifieur considère du type « voiture latérale »).



(a) *(b)*

Figure 2.16 – (a) Evolution typique de l'erreur de classification sur les "voitures latérales" en fonction des itérations de boosting ; (b) Courbe précision-rappel, calculée sur la base de test des voitures vues latéralement, du classifieur fort obtenu par boosting et assemblant 10 et 300 classifieurs faibles de « présence de point d'intérêt avec un certain descripteur ».

Encore mieux, il s'avère que la plupart des descripteurs de points d'intérêt sélectionnés dans le classifieur fort ont une correspondance avec des points situés sur une région spécifique de la catégorie, telle que roue, capot, dessous de caisse, etc… Ceci montre que chaque descripteur incorporé dans le classifieur a une signification sémantique liée à la nature de la catégorie.

Weak Classifier	Image 1	Spécificité
40		Points d'intérêt concentrés sur le capot de la voiture
230		Points d'intérêt concentrés sur le "bas de caisse" entre les roues
300		Points d'intérêt concentrés sur les roues

Figure 2.17 - Positions de divers points d'intérêt sélectionnés par adaBoost et répondant positivement, accumulés sur tous les exemples positifs d'apprentissage : les points correspondant à un classifieur faible donné semblent correspondre à une partie spécifique de voiture.

Les essais effectués sur d'autres benchmarks publics montrent aussi des résultats comparables à l'état de l'art : nous obtenons notamment ~87% de rappel pour ~87% de précision sur la base de piétons de (Munder et Gavrila 2006).

Localisation des objets directement à partir des points d'intérêt

Une autre de nos motivations pour la catégorisation par points d'intérêt est la possibilité potentielle de détecter et localiser les instances de la catégorie cherchée directement à partir des positions des points d'intérêt, donc sans recourir à un balayage multi-échelles de l'image qui est toujours une étape coûteuse en calcul dans la détection visuelle d'objets.

Le principe que nous avons proposé est inspiré et adapté des travaux de (Leibe, Leonardis et Schiele 2008) : durant l'apprentissage nous collectons la statistique des positions (relativement au centre de l'objet) des points d'intérêt sélectionnés comme « discriminants » ; durant la phase de détection, les descripteurs similaires « voteront » alors pour des positions possibles/probables de centre. Un lissage de ces votes suivi d'un seuillage permet ensuite d'estimer des boîtes englobantes pour les diverses instances d'objets de la catégorie cherchée. Ce principe est illustré sur la Figure 2.18.

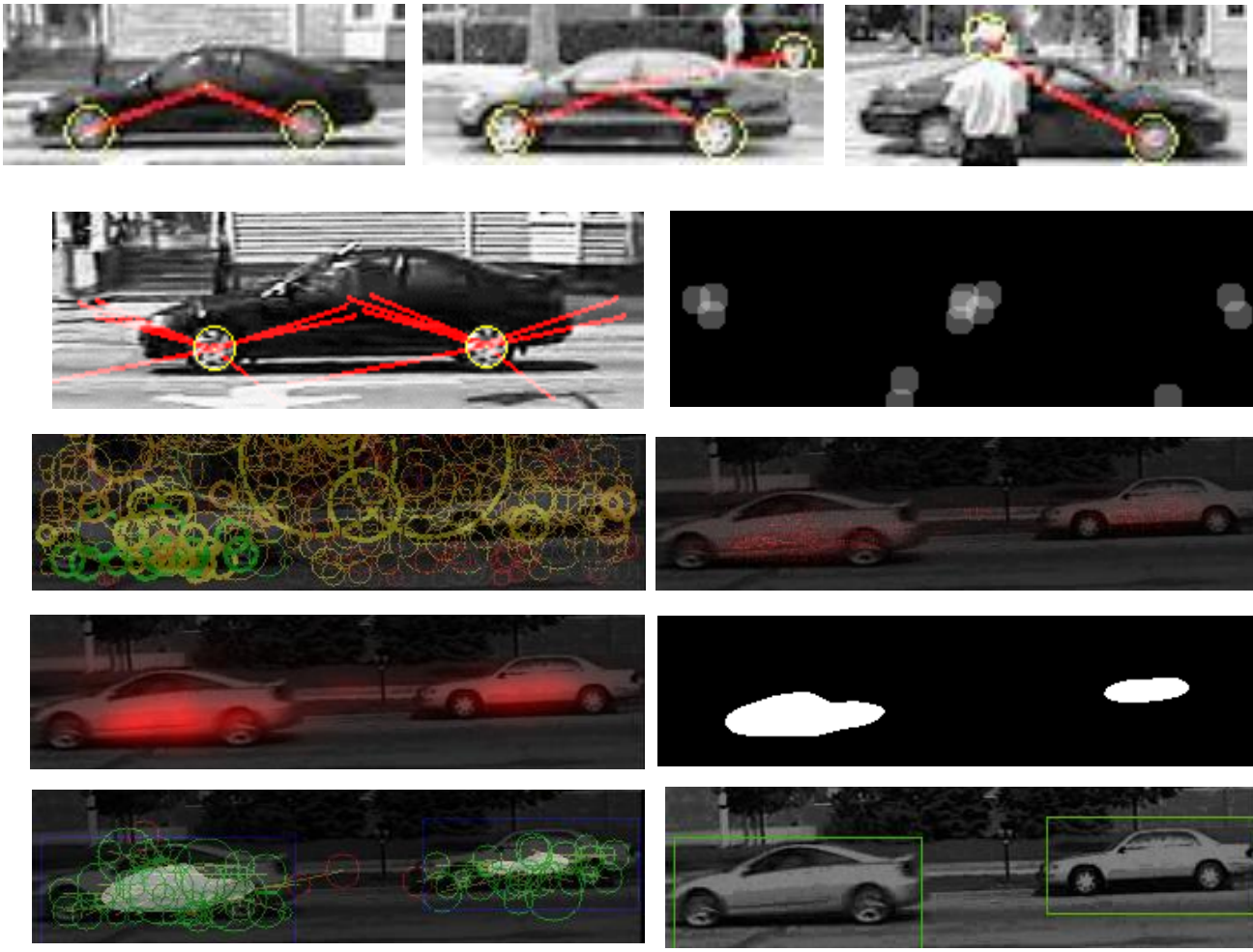


Figure 2.18 – Illustration de notre principe de localisation des objets à partir des points d'intérêt (PI) : (a) en haut, collecte durant l'apprentissage des vecteurs reliant les PIs au centre de l'objet ; (b) deuxième ligne : votes par chaque PI pour les positions possibles du centre, et accumulation sur une carte type transformation de Hough généralisée ; (c) troisième ligne, PIs détectés, et résultat brut des votes pour la/les position(s) de centre(s) ; (d) quatrième ligne, lissage Gaussien des votes, suivi d'un seuillage pour segmenter position approximative du/des centre(s) ; (e) dernière ligne, PIs correspondant aux objets, et boîtes englobantes obtenues.

Si ce principe de détection/localisation fonctionne bien sur des images simples, il s'avère moins robuste sur des scènes plus complexes, pour lesquelles les points d'intérêt sélectionnés par le boosting ne sont pas toujours assez spécifiques. Néanmoins, en ajoutant une étape de filtrage préliminaire des points d'intérêt pour supprimer ceux de l'arrière-plan, nous parvenons à contourner cette difficulté. Le principe final de notre détection est donc illustré sur la Figure 2.19.

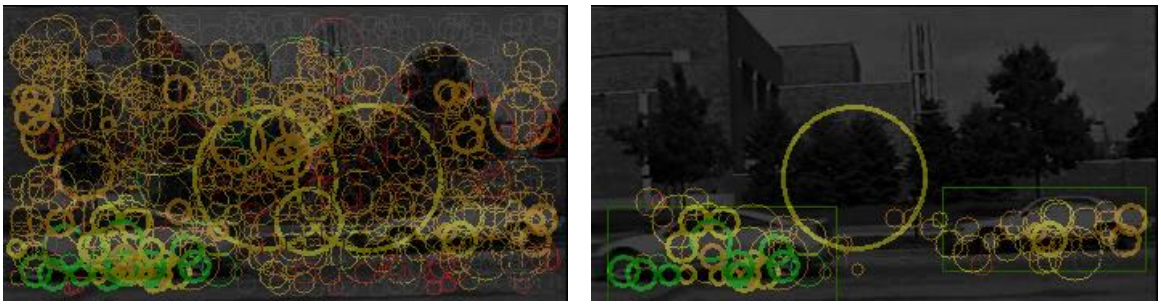


Figure 2.19a – Illustration du filtrage préliminaire appliqué : à gauche, les points d'intérêt (PI) répondant positivement sont disséminés sur l'image (même si les réponses de ceux situés sur les voitures vues latéralement sont plus fortes) ; à droite, le résultat après filtrage par un classifieur de PIs entraîné à discriminer entre les PIs d'arrière-plan et ceux de voitures vues de côté.

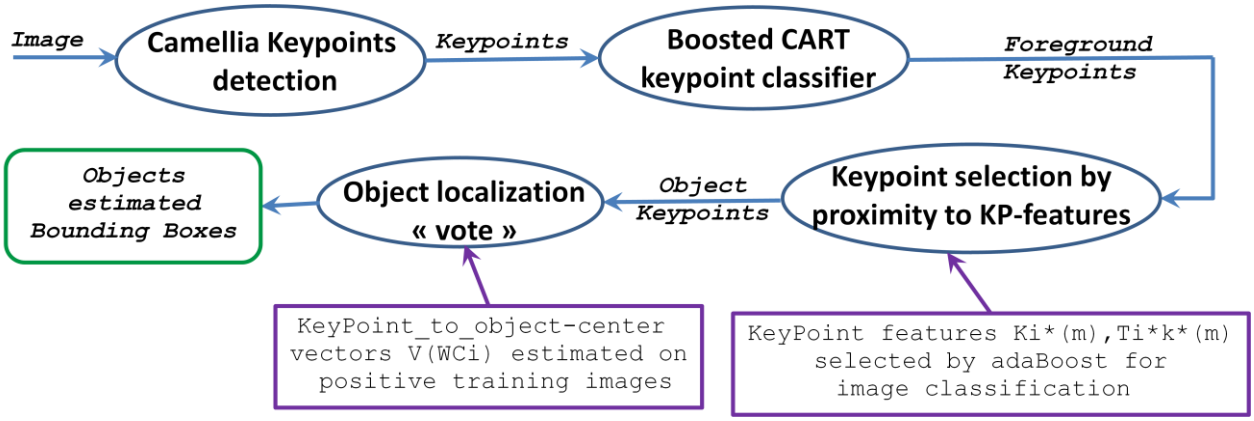


Figure 2.19b – Chaîne de traitement de notre technique de localisation d'objet à partir des points d'intérêt.

Les performances, avec et sans l'étape de filtrage préalable, ont été calculées de la base de « voitures vues latéralement » de UIUC, où une précision moyenne de 83% est obtenue, la courbe complète étant présentée sur la Figure 2.20. Les résultats finaux, sous forme de boîtes englobantes sont illustrés sur la Figure 2.21.

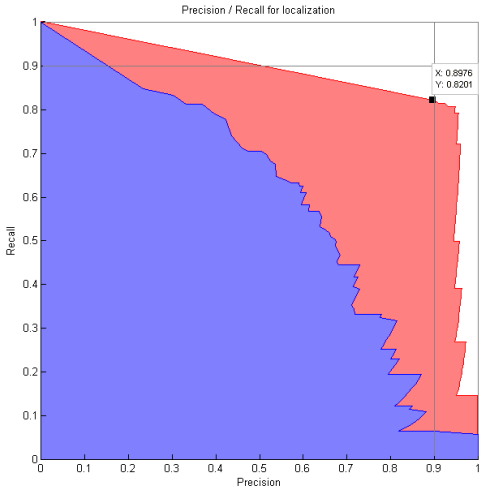


Figure 2.20 – Amélioration apportée à la localisation de voitures vues latéralement par le filtrage préliminaire : la courbe la plus basse est la précision-rappel initiale, et la plus haute (avec une précision moyenne de 83% au lieu de 59%) est celle obtenue en n'utilisant pour la localisation que les points d'intérêt conservés après filtrage de ceux de l'arrière-plan.

Figure 2.21 – Exemples typiques de détections correctes (à gauche), et de positifs manqués et fausses alarmes (à droite).

Sur un PC à 3.2GHz, notre algorithme de détection/localisation traite une image de 400x250 pixels en environ 300ms. Ce n'est pas du véritable temps-réel, mais cela suffit à faire des détections au rythme d'environ 3 images/seconde. Le résultat d'un test préliminaire sur une vidéo est illustré sur la Figure 2.22.

Ces travaux ont fait l'objet des publications suivantes aux conférences internationales IV'2009 et ITSwc'2009 : (Bdiri, Moutarde et Steux, Visual object categorization with new keypoint-based adaBoost features 2009), (Bdiri, Moutarde et Bourdis, et al. 2009).

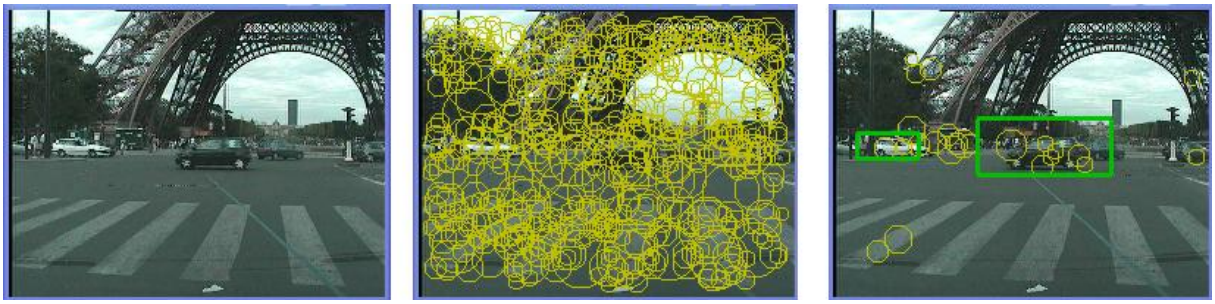


Figure 2.22 – Premiers tests de détection sur une vidéo : à gauche, l'image courante ; au milieu, tous les points d'intérêt détectés ; à droite, uniquement les points classifiés "voiture vue latéralement", et les boîtes englobantes résultant de notre technique de localisation.

2.3. BILAN ET PERSPECTIVES SUR MES RECHERCHES EN DETECTION ET CATEGORISATION VISUELLE D'OBJETS

Les recherches présentées dans ce deuxième chapitre sur la détection et catégorisation visuelle d'objets ont donc apporté les contributions suivantes :

- une nouvelle famille de primitives visuelles, les « connected control-points », qui permettent, quand ils sont utilisés comme « classifieurs faibles » pour adaBoost, d'égaler ou dépasser l'état de l'art en taux de reconnaissance, au moins pour les catégories voitures et piétons ;
- une approche originale s'appuyant uniquement sur les descripteurs des points d'intérêt détectés, et qui permet d'effectuer *simultanément* la catégorisation et la localisation visuelle d'objets.

Ces travaux ont par ailleurs fait l'objet de ***5 publications dans des conférences internationales à comité de lecture*** :

- "*adaBoost with 'keypoint presence features' for real-time vehicle visual detection*", Taoufik Bdiri, Fabien Moutarde, Nicolas Bourdis and Bruno Steux, proc. of 16th World Congress on Intelligent Transport Systems (ITSwc'2009), Stockholm (Sweden), septembre 2009.
- "*Visual object categorization with new keypoint-based adaBoost features*", Taoufik Bdiri, Fabien Moutarde, and Bruno Steux, proc. of IEEE Symposium on Intelligent Vehicles (IV'2009), held in XiAn (China), juin 2009.
- "*Real-time visual detection of vehicles and pedestrians with new efficient adaBoost features*", Fabien Moutarde, Bogdan Stanciulescu, and Amaury Breheret, proc. of 'Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)' of "2008 International Conference on Intelligent RObots and Systems (IROS'2008)", Nice (France), 26 septembre 2008.
- "*Introducing New AdaBoost Features for Real-Time Vehicle Detection*", Bogdan Stanciulescu, Amaury Breheret and Fabien Moutarde, proc. of COGIS'07 conference on COGnitive systems with Interactive Sensors, held in Stanford University, California (USA), November 26-27, 2007.
- "*Combining adaBoost with a Hill-Climbing evolutionary feature search for efficient training of performant visual object detectors*", Yotam Abramson, Fabien Moutarde, Bruno Steux and Bogdan Stanciulescu, proc. of FLINS2006 on Applied Computational Intelligence, Gênes (Italie), 29-31 août 2006 (pp. 737-744).

Ces recherches, tout en ayant en ligne de mire la même préoccupation applicative (l'amélioration des systèmes d'aide à la conduite, notamment dans le cadre de collaborations directes avec PSA et Valeo), sont de portée potentiellement plus générale que mes travaux sur la reconnaissance de panneaux, et très « en pointe » du point de vue de la Recherche « académique ». Elles m'ont ainsi permis de compléter mon expertise en reconnaissance de formes sur la problématique de la catégorisation, plus difficile du fait de la grande variabilité intra-classe. Cela m'a aussi amené à élargir le spectre de mes connaissances en terme de descripteurs visuels (par exemple avec les HoG, désormais incontournables en détection de piétons), et à confronter nos travaux à l'état de l'art des algorithmes de catégorisation, via des benchmarks et compétitions publics tels que ceux du « Pascal VOC challenge ». J'ai enfin pu acquérir de bonnes compétences sur les techniques de type détecteurs et descripteurs de points d'intérêt (SIFT ou SURF notamment).

En ce qui concerne les perspectives, une ***collaboration a été initiée avec une équipe de l'Université de Yeungnam en Corée, sur l'utilisation, pour la catégorisation visuelle, de primitives visuelles type Haar mais fondées sur des statistiques d'ordre supérieur*** (variance, skewness, kurtosis, etc..). Et, avec l'ajout probable à court terme d'un item « Evitement de collision piétons » dans les critères d'évaluation euroNcap de sécurité des voitures, il est plus que probable que nous serons amenés dans les années qui viennent à relancer et approfondir nos recherches en particulier pour la détection et reconnaissance de piétons.

3. IDENTIFICATION DE PERSONNES OU D'OBJETS EN 2D OU 3D

Dans les 2 chapitres précédents ont été présentés des travaux centrés sur la détection et reconnaissance d'objets d'apparence figée (les panneaux routiers), puis sur la détection et *catégorisation* visuelle d'objets présentant au contraire une grande variabilité intra-classe (tels que voitures, ou piétons). Un troisième grand type d'applications des algorithmes de type reconnaissance de formes consiste à *identifier* une personne ou un objet déjà vu(e) précédemment. Les techniques utilisées peuvent être similaires, mais du fait qu'il s'agit de reconnaître comme identique un même objet pouvant se présenter sous des aspects fluctuants, tout en distinguant des instances différentes d'une même catégorie (par exemple deux silhouettes de piétons) ou des objets de formes potentiellement similaires, l'accent doit être mis sur la recherche des spécificités différenciantes.

J'ai mené des recherches sur des problématiques de type identification dans 2 domaines très différents :

- la ré-identification de personnes entre caméras à champs disjoints ;
- l'identification, à partir d'une seule vue 3D, d'objets connus.

3.1. IDENTIFICATION DE PERSONNES ENTRE CAMERAS

De 2007 à 2010, j'ai effectué, notamment via l'encadrement (à 100%) de la ***thèse d'Omar HAMDOUN*** (financée par les fonds Carnot de ressourcement, et soutenue en décembre 2010), des recherches dans un autre domaine d'application : la vidéo-protection. L'axe exploré s'est concentré sur l'exploitation pour l'identification (et non plus la catégorisation) des techniques de type points d'intérêt SURF.

3.1.1. INTRODUCTION A LA RE-IDENTIFICATION DE PERSONNES

Pendant longtemps la recherche pour les applications de vidéosurveillance s'est concentrée sur l'analyse automatisée limitée au champ d'une unique caméra : détection d'intrusion, repérage de l'abandon de bagage, reconnaissance de scène de bagarre ou de vol, etc… Avec le déploiement de plus en plus courant de *réseaux* de caméras couvrant toute une installation (aéroport, gare, métro, centre commercial, etc…), de nouvelles problématiques ont émergé : en particulier, il est souvent utile et souhaitable de pouvoir dire si une personne ou un véhicule présent dans le champ d'une caméra a été vu(e) ailleurs précédemment, et si oui, où et quand. Ce problème de suivi à long terme entre caméras à champs de vue disjoints est connu sous le nom de « ***ré-identification*** », et illustré sur la Figure 3.1. Une bonne présentation générale du domaine, en particulier pour le cas de suivi inter-caméras de personnes, se trouve dans le §7 de (Tu, et al. 2007). La difficulté de la ré-identification réside dans les grandes variations qui surviennent souvent, tant de l'angle de vue, de l'échelle, des conditions lumineuses, que de l'attitude des personnes.

Toute une catégorie des méthodes existantes tentent de s'affranchir de ces problèmes en se fondant sur des techniques biométriques (reconnaissance du visage, de la démarche, etc…). Cependant, même si les algorithmes d'identification de visages atteignent désormais d'excellent taux de bonne reconnaissance sur des images haute résolution en environnement d'éclairage contrôlé, et avec une orientation donnée de visage (voir (Belhumeur, Hespanha et Kriegman 1997) et (Draper, et al. 2003)), leurs performances sont beaucoup moins satisfaisantes dans un contexte ouvert et « non-coopératif » comme le suivi dans un grand réseau de caméras.

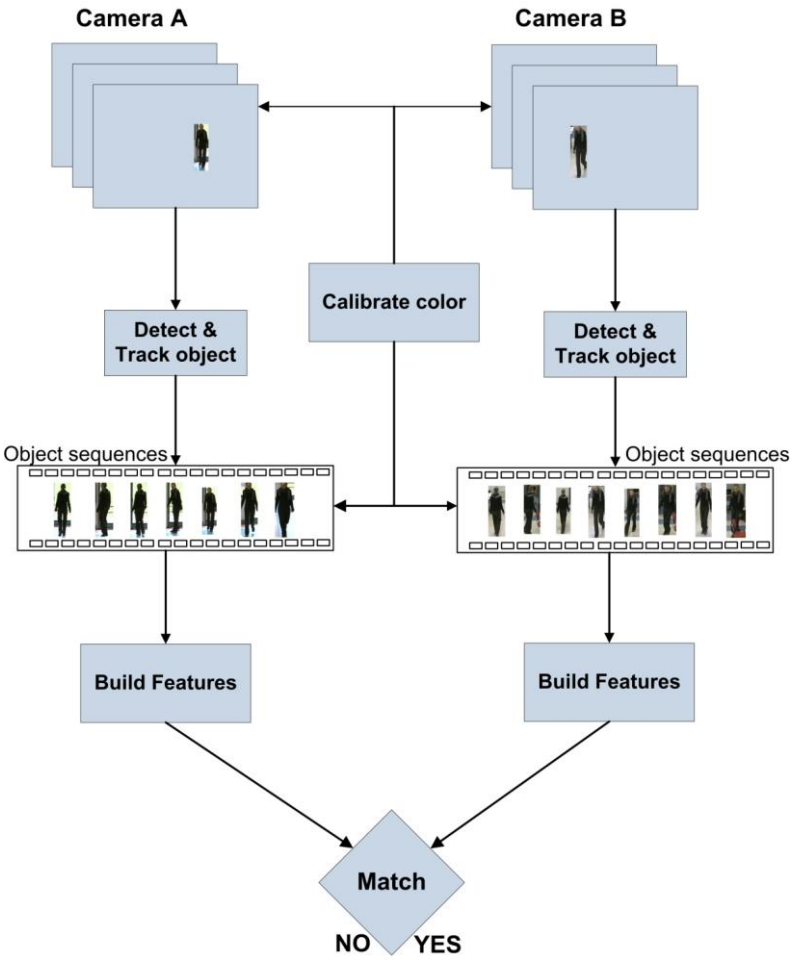


Figure 3.1 – Diagramme illustrant la structure générale de la problématique de ré-identification : des personnes sont détectées et suivies dans chaque caméra. Après éventuelle calibration des couleurs, les caractéristiques des personnes vues sont mémorisées. Ceci permet ensuite, pour toute apparition ultérieure d'une personne, de faire la mise en correspondance, afin de déterminer s'il s'agit ou non d'une personne déjà vue précédemment dans une autre caméra du réseau de surveillance.

Un autre groupe de méthodes effectue la ré-identification uniquement sur la base de l'apparence globale, par exemple la silhouette ou l'histogramme des couleurs, sans recours à la biométrie. Une hypothèse généralement faite dans ce cas est que les personnes ne changent pas de vêtements entre les diverses caméras, ce qui est assez raisonnable quand il s'agit d'un suivi en continu, et non pas une recherche sur plusieurs jours dans une base de vidéos enregistrées. Les travaux de ce type peuvent se subdiviser en 3 grandes sous-catégories, selon la technique employée :

- comparaison à prototypes (template-matching) ;
- histogrammes de couleur ;
- descripteurs locaux de points ou régions d'intérêt.

Les premières approches consistent à créer pour chaque personne un prototype intégrant l'information spatiale et l'apparence (Lipton, Fujiyoshi et Patil 1998). Ceci étant peu robuste aux changements d'orientation, plusieurs auteurs ont proposé d'exploiter un a priori sur la forme pour construire « hors-ligne » un modèle de chaque personne ou objet (Ning, et al. 2004), (Moeslund et Granum 2001), (Stauffer et Grimson 2001). Une autre faiblesse de l'approche "prototypes" (template-matching) est sa sensibilité aux conditions d'éclairage, ce que les travaux de (Huttenlocher, Klanderman et Rucklidge 1993) tentent de compenser en utilisant des prototypes des *contours,* avec les distances de Hausdorff ou de Chamfer, pour comparer robustement les silhouettes même en cas d'alignement imparfait.

La seconde catégorie de méthodes, de loin la plus couramment employée, consiste à utiliser l'histogramme des couleurs. Un avantage de cette signature est son invariance aux translations, rotations et changements d'échelle. (Jaffré et Joly 2004) et (Pham, Worring et Smeulders 2007) exploitent l'histogramme RGB d'une zone *sous le visage* (trouvé par un détecteur de visages). Le principal inconvénient de l'histogramme est son manque de spécificité du fait que toute l'information spatiale est perdue. Ceci peut être amélioré en subdivisant l'objet en plusieurs régions dont les histogrammes sont calculés et comparés séparément, comme font (Park, et al. 2006) qui découpent la personne détectée en 3 parties de haut en bas (à 1/5 et 3/5 de sa hauteur), et n'utilisent que le milieu et le bas. Une autre approche intéressante pour intégrer de l'information géométrique dans l'histogramme de couleur a été proposée dans (Huang, et al. 1999) : le corrélogramme de couleurs qui mesure la corrélation de la couleur avec la distance entre pixels. Enfin, (Madden, Cheng et Piccardi 2007) décrivent chaque personne par ses couleurs principales, déterminées automatiquement avec l'algorithme K-means, le résultat étant un histogramme de couleurs principales (Major Color Spectrum Histogram, MCSHR). A noter que cet histogramme est calculé sur une petite fenêtre temporelle contenant plusieurs images successives. Une des grandes difficultés pratiques pour l'utilisation des couleurs entre caméras est le fait que celles-ci varient fortement selon l'éclairage (et même le modèle de capteur caméra). Divers travaux ont donc proposé des techniques pour calibrer les couleurs entre caméras : (Porikli 2003), (Finlayson, et al. 2005), (Javed et Shah 2005) et (Prosser, Gong et Xiang 2008). Dans beaucoup de travaux, des algorithmes d'apprentissage statistique sont utilisés pour déterminer les primitives discriminantes : (Nakajima, et al. 2003) emploient des SVMs multi-classes calculés sur des primitives globales de couleurs pour faire l'identification de personnes. (Bak, et al. 2010) proposent quant à eux de recourir à adaBoost pour trouver les primitives discriminantes, parmi les « Haar features » et le Dominant Color Descriptor (DCD) qui décrit les couleurs principales du haut et du bas de la personne. (Schwartz et Davis 2009) divisent plutôt le piéton en un ensemble de blocs se chevauchant, sur lesquels sont calculés les co-occurrences de l'histogramme d'orientation du gradient (HoG) et des couleurs dans l'espace HSV. Enfin, (Cong, et al. 2009) caractérisent la silhouette avec des histogrammes normalisés de couleurs et des « spatiogrammes ».

Enfin, un nombre croissant de travaux utilisent des caractéristiques locales. Par exemple (Lantagne, Parizeau et Bergevin 2003) divisent la silhouette en 3 parties correspondant respectivement à la tête, au tronc et bras, et aux jambes. Chaque partie est segmentée en régions, sur lesquelles sont calculées des descripteurs de couleur utilisant l'espace HSV. Ceci est combiné avec un descripteur de texture qui somme les réponses de 4 détecteurs de bords d'orientations différentes. (Gheissari, Sebastian et Hartley 2006) segmentent aussi la silhouette, mais avec un algorithme d'inondation (watershed). Pour éviter la sur-segmentation due aux plis des vêtements, ils construisent un graphe spatio-temporel où chaque noeud représente une région, et les arêtes correspondent aux relations spatiales et temporelles entre régions adjacentes. Pour finir ce graphe est partitionné afin de sélectionner les "régions saillantes", dont les histogrammes sont calculés. Une autre manière de sélectionner des sous-régions « stables » est proposée par (Farenzena, et al. 2010) qui recherchent les régions maximalement stables (Maximally Stable Color Regions, MSCR). Le descripteur de chaque région combine un histogramme de couleur HSV et une mesure de texture basée sur l'entropie. Récemment, le succès en reconnaissance/identification d'objets des approches à base de points d'intérêt (type SIFT/SURF/etc…) a suscité plusieurs travaux les appliquant en ré-identification de personnes. (Gheissari, Sebastian et Hartley 2006) utilisent un détecteur "Hessien affine" qui produit un grand nombre de points. Chacun des points est caractérisé par le calcul sur une zone locale de taille fixe d'un histogramme d'orientation du gradient (HOG) et d'un histogramme de couleur HSV. (Arth, Leistner et Bischof 2007), quant à eux, exploitent le "PCA SIFT", combiné avec un « arbre de vocabulaire » des descripteurs. Ceci leur permet de coder le descripteur de chaque point simplement par la position dans l'arbre du plus proche voisin. Ce panorama des méthodes de ré-identification est résumé dans le Tableau 3.1, où elles sont catégorisées selon le type de descripteur (forme, couleur, ou primitives locales).

Tableau 3.1 - Panorama des principales recherches en suivi et ré-identification entre multiples caméras.

Système	Objectif	Technique de ré-identification utilisée
VIP (Vision tool for comparing Images of People) (Lantagne, Parizeau et Bergevin 2003)	Reconnaître en temps-réel une personne vue sous divers angles	Descripteur fondé sur les couleurs et la texture
ViSE: Visual Search Engine Using Multiple Networked Cameras (Park, et al. 2006)	Aider à la surveillance, et à la recherche de personnes	La silhouette de la personne est divisée en 3 parties, de haut en bas. Le modèle concatène les 3 histogrammes HSV correspondants.
MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras (Albu, et al. 2006)	Aider à la surveillance, et à la recherche de personnes	Modèle combinant histogramme de couleurs et caractéristiques du visage
KNIGHT: A real time surveillance system for multiple overlapping and non-overlapping camera (Javed, et al. 2003)	Détecter, suivre, et catégoriser les objets mobiles dans le réseau de cameras	Histogramme de couleurs mais combine à une estimation de trajectoire pour restreindre les correspondances potentielles.
A Multi-Camera Visual Surveillance System for Tracking of Reoccurrences of People (Pham, et al. 2007)	Détection des ré-occurrences de personnes	Histogramme de couleurs
Person Re-identification Using Spatiotemporal Appearance (Gheissari, Sebastian et Hartley 2006)	Ré-identification de piétons	Points d'intérêt (opérateur Hessien) et un modèle graphique de décomposition triangulaire.
Object reacquisition and tracking in large-scale smart camera networks (Arth, Leistner et Bischof 2007)	Ré-identification de véhicules (le long d'une route)	Points d'intérêt, descripteurs PCA-SIFT, et sac de mots visuels sous forme d'un « arbre de vocabulaire »

3.1.2. *Travaux de recherche sur la re-identification inter-caméras de personnes*

Les travaux de recherche que j'ai encadrés se sont concentrés sur la mise au point et l'évaluation d'une technique originale de ré-identification utilisant l'accumulation de points d'intérêt durant la/les période(s) d'apparition des personnes, aussi bien dans la base permettant l'apprentissage que sur chaque séquence de requête dans laquelle on souhaite ré-identifier une personne vue antérieurement.

Construction des modèles

Pour chaque piéton suivi dans une séquence, nous construisons un modèle en accumulant les descripteurs de points d'intérêt collectés sur lui. Pour éviter trop de redondance, nous effectuons tout d'abord une sélection temporelle pour ne conserver que les images de la personne apportant suffisamment d'informations nouvelles : pour cela, les descripteurs des points d'intérêt extraits de chaque nouvelle image de la séquence sont comparés avec ceux de la précédente « image-clef », et l'image courante n'est conservée comme nouvelle image-clef que si le pourcentage de ses points d'intérêt ayant une correspondance avec l'image-clef précédente est *inférieur* à un certain seuil. Si ce n'est pas le cas, l'image courante est considérée comme n'apportant pas assez d'information nouvelle, et est donc ignorée, à part une mise à jour de la fréquence des points d'intérêts pour ceux qui ont engendré des correspondances. Ce principe de sélection est illustré sur la Figure 3.2, et des exemples de suites d'images-clef sélectionnées dans des séquences sont présentés dans la Figure 3.3.

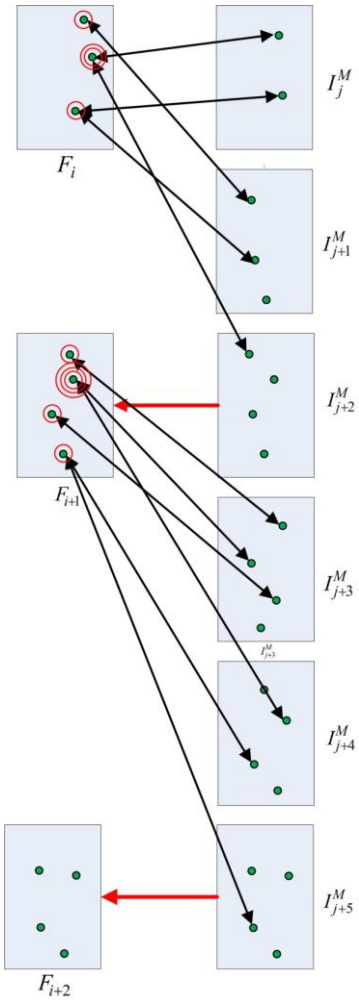


Figure 3.2 – Sélection des images "clefs" dans une séquence : les points d'intérêt de l'image courante I_j^m sont comparés à ceux de la dernière « image-clef » F_i. Si le pourcentage de points mis en correspondance est inférieur à un certain seuil, alors l'image courante devient l'image-clef suivante F_{i+1}. Sinon, nous mettons juste à jour le nombre de correspondances des points (illustré par des cercles sur la figure).

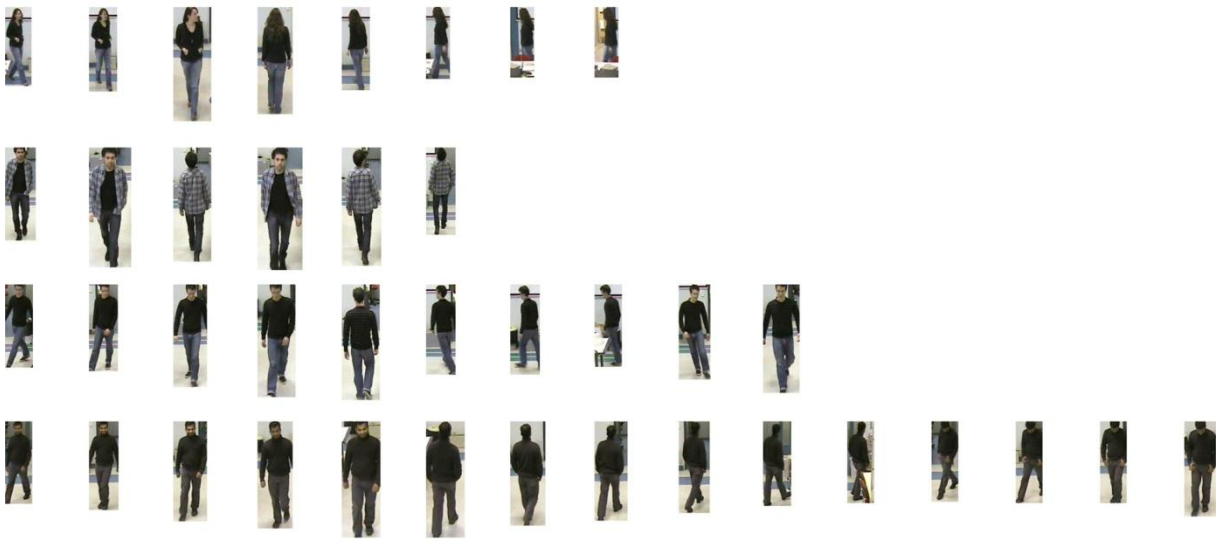


Figure 3.3 – Exemples de résultats de notre technique de sélection d'images-clef dans une séquence (dans chaque cas, la séquence complète initiale contient 200 images).

Une fois sélectionnées les « images-clefs » de la séquence, nous collectons tous leurs points d'intérêt et les accumulons pour construire le « modèle » de la personne. Au lieu de juste ajouter au modèle tous les points d'intérêt trouvés, nous cherchons à estimer les *fréquences* de points « similaires ». Aussi, pour chaque nouveau point d'intérêt, nous recherchons s'il existe déjà dans le modèle courant un autre point d'intérêt dont le descripteur soit assez similaire, et dont la position et l'échelle (relative à la taille de la personne) soient aussi proches (afin d'éviter des correspondances erronées entre des points sans lien « physique »). S'il n'y a aucune correspondance, nous ajoutons le point au modèle. Par contre s'il y a une correspondance, nous effectuons uniquement une mise à jour des « paramètres agrégés » du point, et de sa fréquence, comme explicité dans l'algorithme ci-dessous :

Algorithme de construction du modèle

1. $L \leftarrow 0\ (nombre\ de\ points\ d'intérêt\ du\ modèle), M \leftarrow \phi\ (ensemble\ des\ keypoints\ du\ modèle)$
2. $Pour\ t = 0$ à $N\ (le\ nombre\ de\ frames)$
   I. Extraire les points d'intérêt de l'image I_t.
   II. ***Pour chaque point p_i trouvé, FAIRE :***
      a) Chercher K_m dans $M = \{K_j | 1 \leq j \leq L\}$ qui corresponde à p_i selon 3 conditions suivantes :
         - $(x_{p_i} - x_{K_m})^2 + (y_{p_i} - y_{K_m})^2 \leq \varepsilon_1$
         - $\left(scale_{p_i} - scale_{K_m}\right)^2 \leq \varepsilon_2$
         - Les descripteurs (D_{p_i}, D_{K_m}) se correspondent *mutuellement* .
      b) Si L=0 ou si aucune correspondance trouvée, alors L ←L+1 et ajouter dans M nouveau point :
         - $D = D_{P_i}$
         - $f = 1, x_{M_L} = x_{p_i},\ y_{M_L} = y_{p_i},\ scale_{M_L} = scale_{p_i}$
      c) Sinon, mettre à jour le K_m trouvé comme suit :
         - $D_{K_m} = \frac{D_{K_m} * f + D_{p_i}}{f+1}$
         - $f = f + 1$
         - $x_{K_m} = \frac{x_{K_m} * f + x_{p_i}}{f+1}$, $y_{K_m} = \frac{y_{K_m} * f + y_{p_i}}{f+1}$, $scale_{K_m} = \frac{scale_{K_m} * f + scale_{p_i}}{f+1}$

      FIN FAIRE

FIN POUR

De plus, au final nous associons à chaque point d'intérêt un poids de type « tf-idf » (« term frequency – inverse document frequency »), comme utilisé couramment en recherche de documents, et comme (Sivic et Zisserman 2008) ont aussi proposé pour la recherche visuelle d'objets dans des vidéos :

$$w_i = \frac{n_{im}}{n_m} log \frac{N}{N_i} \qquad \text{(Eq. 3.1)}$$

où n_{im} est le nombre d'occurences du point d'intérêt i trouvées dans toutes les images du modèle m, et n_m est le nombre total de points d'intérêt du même modèle ; par ailleurs, N_i est le nombre de modèles (parmi toutes les personnes) contenant le point d'intérêt i, et N est le nombre total de modèles de la base. Le premier facteur $\frac{n_{im}}{n_m}$ correspond donc à la fréquence du point d'intérêt dans toutes les images de la personne m, tandis que le deuxième facteur mesure la spécificité du point i. La logique sous-jacente est de donner à i un poids d'autant plus grand qu'il apparaît souvent pour la personne m, et d'autant plus faible qu'il se retrouve en fait sur beaucoup d'autres personnes.

Principe de notre technique de ré-identification

Le principe que nous utilisons pour la ré-identification est un vote pondéré : pour chaque point d'intérêt de la requête, nous cherchons les correspondances avec toutes les images de tous les modèles. Si une correspondance est trouvée entre un point d'intérêt R_i de l'image t de la requête et un point d'intérêt K_m du modèle k, alors le vote pour le modèle k est incrémenté comme suit :

$$V_t(k) += p_{kt}(R_i, K_m) \qquad \text{(Eq. 3.2)}$$

Pour le poids p_{kt}, nous utilisons une formule similaire à celle proposé par (Schmid 1999), en y intégrant en plus la cohérence spatiale :

$$P_{kt}(R_i, K_m) = P_c(R_i, K_m) \times P_s(t,k) \times P_t(t,k) \qquad \text{(Eq. 3.3)}$$

Les 3 facteurs P_c, P_s et P_t correspondent respectivement à :

- La qualité de la correspondance P_c
- La cohérence spatiale P_s
- La cohérence temporelle P_t

$P_c(R_i,K_m)$ estime la ***qualité de la correspondance entre les 2 points d'intérêt*** R_i et K_j, selon la similarité de leurs descripteurs, et le poids (déduit des fréquences par l'Equation 3.1) W_{Km} dans le modèle k du keypoint K_m en correspondance, avec la formule suivante :

$$P_c(R_i, K_m) = \frac{1}{dist\left(D_{R_i}, D_{K_m}\right)} \times W_{K_m} \qquad \text{(Eq. 3.4)}$$

$P_s(k,t)$ caractérise la ***cohérence spatiale***, au sens d'évaluer si les divers points d'intérêt de la requête t, quelles que soient leur position, s'appairent de manière cohérente ou non avec le même modèle k. Nous l'évaluons simplement par la ***proportion représentée par k parmi les appariement***s :

$$P_s(t,k) = \frac{N_t(k)}{N_t} \qquad \text{(Eq. 3.5)}$$

où $N_t(k)$ est le nombre de points de la requête t qui s'apparient avec un point du modèle k, et N_t est le nombre *total* de point d'intérêt de t pour lesquels un appariement a été trouvé.

Enfin, $P_t(t,k)$ estime la ***cohérence temporelle***, i.e. la stabilité du vote pour le modèle k au fil des images successives de la requête :

$$P_t(t,k) = \alpha V_{t-1}(k) + (1-\alpha) P_t(t-1,k) \qquad \text{(Eq. 3.6)}$$

où α spécifie avec quelle rapidité les nouvelles correspondances supplantent les anciennes.

A noter que toutes les correspondances entre points d'intérêt sont déterminées avec la distance L1 entre descripteurs, et que nous utilisons un « KD-tree » pour stocker tous les points des modèles, afin d'obtenir une recherche très rapide. Un KD-tree est une structure de données permettant de stocker des vecteurs de Rd, initialement proposée dans (Friedman, Bentley et Finkel 1977) pour une recherche efficace de type K-plus-proches voisins.

Evaluation de notre technique de ré-identification

L'évaluation des performances de ré-identification a été faite sur la base publique ETHZ de vidéos dédiée à l'évaluation de ré-identification entre caméras. Cette évaluation a montré que notre approche permet d'atteindre un taux de reconnaissance d'environ 96%, au lieu de 87% pour les meilleurs résultats publiés sur le même benchmark, soit un gain de presque +10% (voir Figure 3.4). La comparaison à d'autres descripteurs (tels que SIFT, HOG, etc…) montre aussi que le descripteur SURF que nous avons utilisé est bien plus efficace pour cette problématique de ré-identification de personnes (voir Figure 3.5).

Pour plus de détails sur ces travaux, le lecteur peut se reporter à nos publications suivantes : (Hamdoun, Moutarde et Stanciulescu, et al. 2008), (Hamdoun, Moutarde et B., et al. 2008) et (Hamdoun et Moutarde 2009).

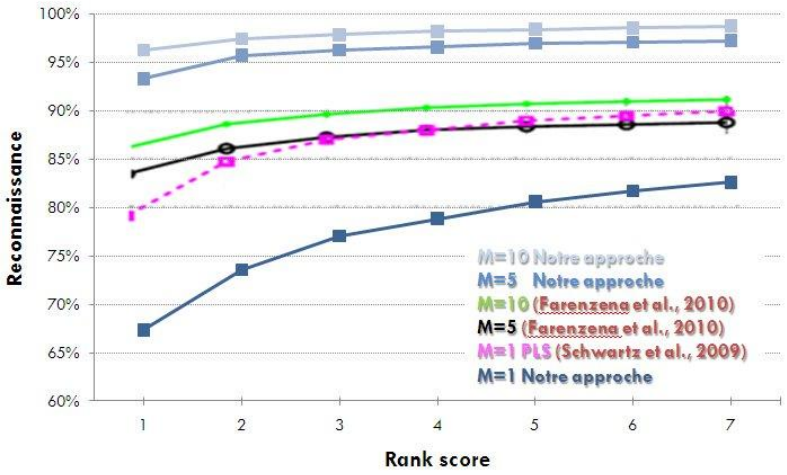


Figure 3.4 – Evaluation sur la base publique ETHZ des performances de notre technique de ré-identification (2 courbes du haut) comparées aux meilleurs résultats publiés antérieurement sur la même base.

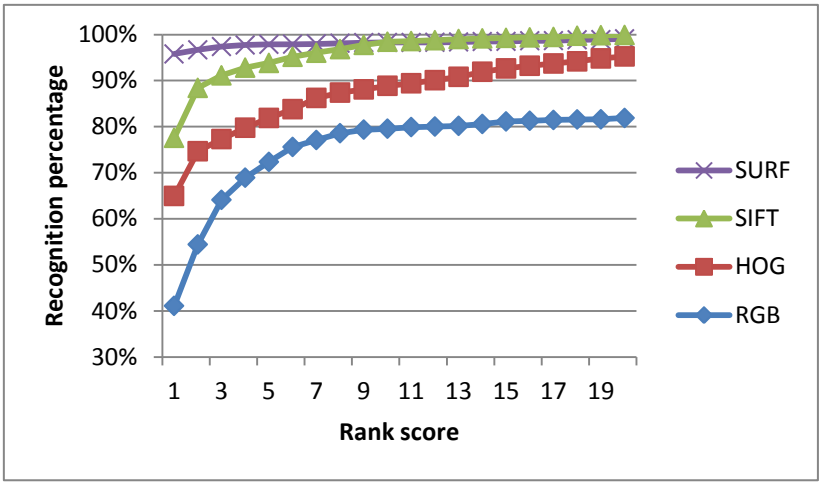


Figure 3.5 – Comparaison sur la base publique ETHZ des performances de notre approche, avec ce qui peut être obtenu en utilisant des descripteurs autres que les points d'intérêt SURF.

3.2. RECONNAISSANCE/IDENTIFICATION 3D D'OBJETS

Plus récemment, compte tenu du développement croissant des capteurs 3D (e.g. Kinect™) et de l'utilisation importante de nuages de points 3D dans d'autres équipes du laboratoire, j'ai diversifié mes recherches vers la reconnaissance *tridimensionnelle* d'objets. Ces travaux ont été menés de 2008 à 2012 dans le cadre d'un stage de Master2 Recherche, puis de la **thèse d'Ayet SHAIEK** soutenue en mars 2013, que j'ai entièrement encadrée.

3.2.1. INTRODUCTION A LA RECONNAISSANCE EN 3D

Parmi les nombreuses techniques de reconnaissance/identification d'objet à partir de données 3D, on peut distinguer celles qui utilisent directement l'information tridimensionnelle (par exemple via des courbures ou normales), et celles qui s'appuient plutôt sur une/des projection(s) ou image(s) de profondeur. Ces dernières présentent l'avantage de permettre l'application quasi-directe des algorithmes de reconnaissance de forme 2D, mais les méthodes réellement 3D exploitent mieux la totalité de l'information disponible.

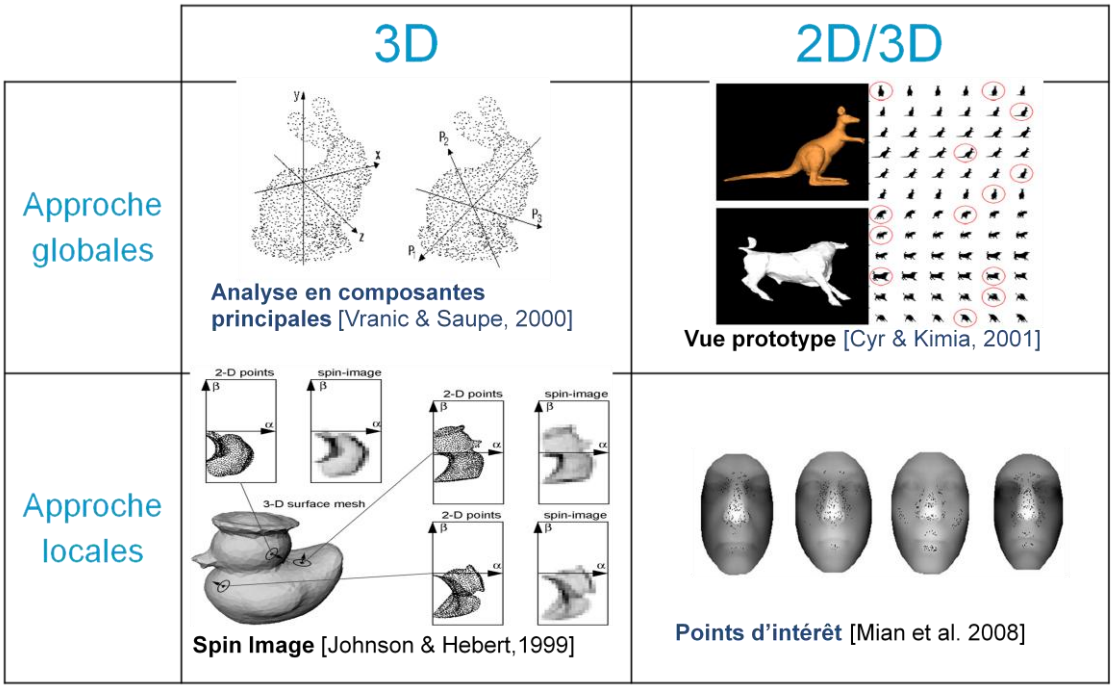


Figure 3.6 - Typologie des algorithmes de reconnaissance 3D d'objets

Indépendamment, on peut aussi séparer les approches globales qui évaluent certaines quantités sur la totalité d'une vue 3D (voire même de l'enveloppe complète), et celles qui exploitent plutôt des descripteurs locaux calculés au voisinage de « points saillants ». Le gros avantage de ces dernières est leur potentielle robustesse aux occultations partielles ; en revanche puisque seules des informations partielles locales sont extraites, il est évidemment critique que les zones choisies et leurs descripteurs capturent néanmoins l'essentiel de l'information sur la vue 3D de l'objet à reconnaître.

Tableau 3.2 - Panorama des principales recherches en reconnaissance/identification d'objets 3D.

Approches 3D			**Approches 2D/3D**	
Globales		**Locales**	**Globales**	**Locales**
Surfacique	**volumique**			
EGI (Horn 1984) Distribution de forme (distance, angle, taille) (Osada, et al. 2001) SF3D (Zaharia et Prêteux 2002) 3D DFT (Vranic et Saupe 2001) DH3DO (Zaharia et Prêteux 2002)	Géons (Medioni et François 2000) SAI (Hebert, Ikeuchi et Delingette 1995)	Harmoniques sphériques (Kazhdan et Funkhouser 2002) RANSAC + Histogramme des normales et des tailles des patches (Swadzba et Wachsmuth 2008) Spin Images (Johnson et Hebert 1999) 3D Shape Context (Frome, et al. 2004) Pairwise Geometric Histogram (PGH) (Ashbrook, et al. 1998) MRG (Hilaga, et al. 2001) Différence des normales sur points saillants (Li et Guskov 2005) Point saillant et représentation en tenseur (Mian, Bennamoun et Owens 2008) Détection des lignes crêtes et ravines (Song, Liu et Yang 2005)	Depth Buffer-based Descriptor (Vranic et Saupe 2000) Lignes de profondeur (Chaouch et Verroust-Blondet 2007) Courbes de niveau des profondeurs (Samir, Daoudi et Srivastava 2006) Eigenfaces (Tsalakanidou, Tzovaras et Strintzis 2003) (Chang, Bowyer et Flynn 2003) Ondelettes de Gabor (Diego, et al. 2010)	Filtre rectangulaire de Viola et Jones (Jones et Viola 2003) Représentation en tenseur (Mian, Bennamoun et Owens 2006) Histogramme de profondeurs, normales, et courbures (Hetzel, et al. 2001)

3.2.2. *TRAVAUX REALISES EN IDENTIFICATION D'OBJETS EN 3D*

Les travaux de recherche que j'ai encadrés dans ce domaine ont principalement consisté à développer et tester une nouvelle famille de « points d'intérêt 3D », baptisée SKIN. Cette famille consiste :

- d'une part en plusieurs variantes de détecteurs utilisant comme critère de saillance le type de surface 3D locale estimé conjointement par les indices de forme S et C, et les courbures H et K.
- d'autre part, chaque point d'intérêt SKIN est ensuite caractérisé par un des 2 nouveaux descripteurs suivants :
  - ***indThrift***, qui est un histogramme 2D (inspiré du LSP) comptant les occurrences jointes d'angles entre normales (cf Thrift) et d'indice de forme SI
  - ***indSHOT***, qui est un histogramme 1D (inspiré du C-SHOT) concaténant l'histogramme des angles entre normales du SHOT, et l'histogramme de l'indice de forme.

Détecteurs SKIN : SC_HK_xxx

En ce qui concerne les détecteurs, nous nous sommes concentrés sur ceux qui découlent de la classification de surface 3D locale. Il existait 2 classifications de ce type : SC et HK. La première s'appuie principalement sur le calcul de l'indice de forme S qui se déduit des courbures principales κ_1 et κ_2 selon la formule ci-dessous :

$$S = \frac{2}{\pi} \cdot \arctan\left(\frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2}\right) \quad (\kappa_1 > \kappa_2) \qquad \text{(Eq. 3.7)}$$

Quand l'intensité de courbure $C = \sqrt{\kappa_1^2 + \kappa_2^2}$ est assez grande ($C>C_0$), la classification dépend juste de la valeur de S, comme illustré sur la Figure 3.7.

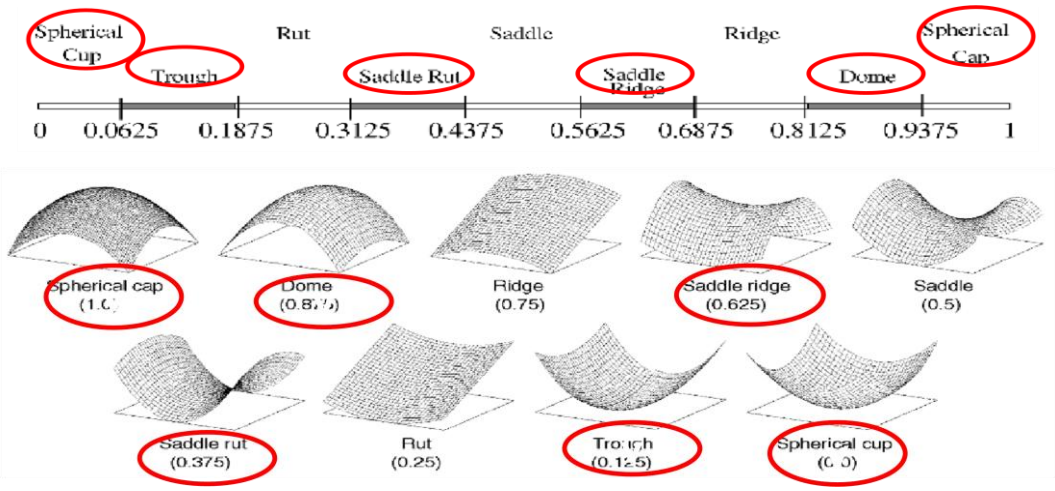


Figure 3.7 - Types de surfaces 3D selon valeur de S, pour $C>C_0$ (illustration extraite de (Dorai et Jain 1997)**)**

En ce qui concerne la classification HK, elle s'appuie sur le couple de valeur (H,K), H étant la courbure moyenne et K la courbure Gaussienne, qui se déduisent toutes deux des courbures principales κ1 et κ2 selon les formules ci-dessous :

$$H = \frac{\kappa 1 + \kappa 2}{2} \qquad K = \kappa 1 \times \kappa 2 \qquad \text{(Eq. 3.8)}$$

Le type de forme locale dépend ensuite des 2 valeurs, comme explicité sur la Figure 3.8.

Mean curvature H	Gaussian curvature K		
	$K>0$	$K=0$	$K<0$
$H<0$	Peak $T_p=1$	Ridge $T_p=2$	Saddle ridge $T_p=3$
$H=0$	None $T_p=4$	Flat $T_p=5$	Minimal $T_p=6$
$H>0$	Pit $T_p=7$	Valley $T_p=8$	Saddle valley $T_p=9$

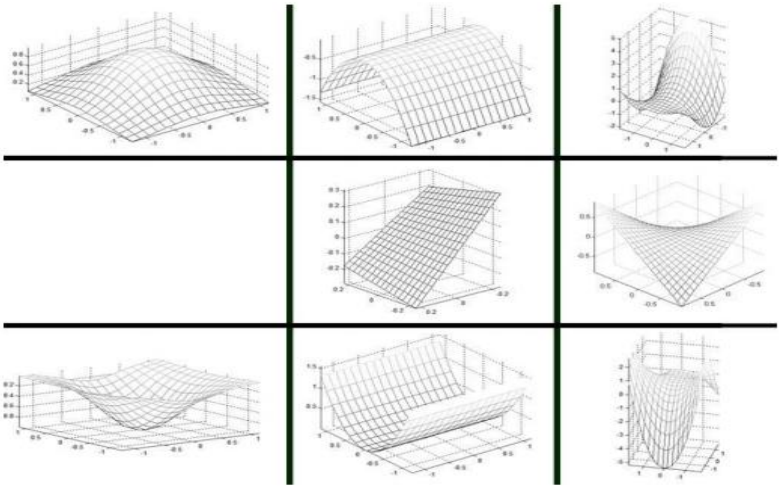


Figure 3.8 - Types de surfaces 3D selon valeurs de H et K (illustration extraite de (Akagündüz 2011)**)**

Partant du constat que les classifications HK et SC devraient en théorie conduire au même type de surface, nous avons proposé le principe de conserver comme points saillants uniquement ceux qui sont estimés du MEME type dôme, bassin ou selle, par les DEUX approches SC et HK, comme illustré sur la Figure 3.9.

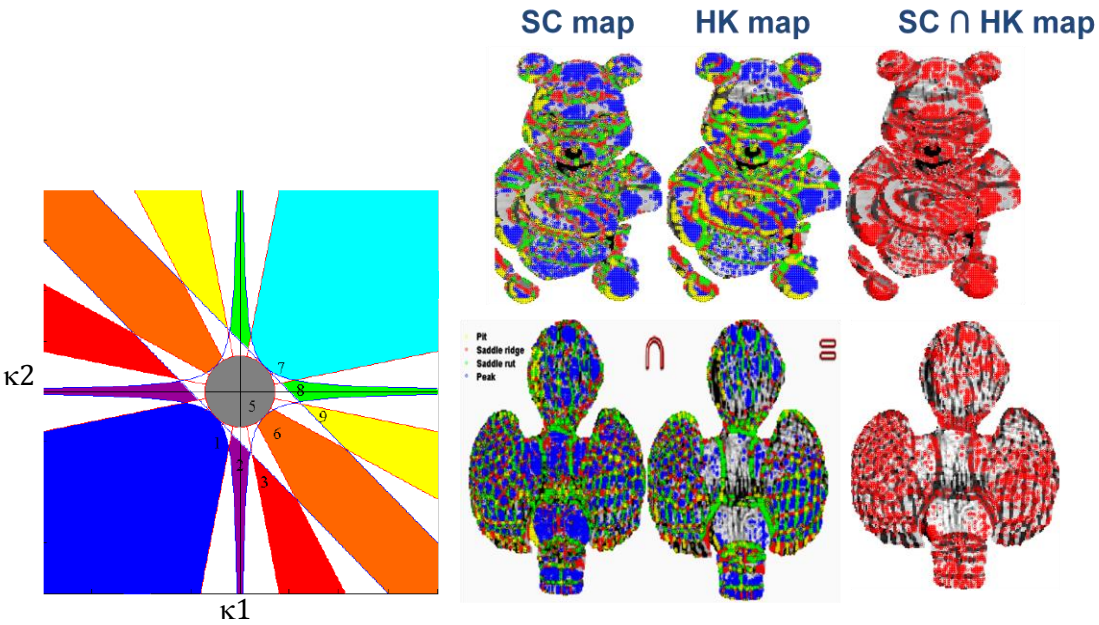


Figure 3.9 - A gauche, dans le plan des courbures principales (κ1,κ2) les zones en couleurs correspondent au cas où le type de surface est identique selon SC et HK (illustration extraite de (Akagündüz 2011)**) ; à droite, notre principe de critère « joint » SC_HK pour présélectionner les points d'intérêt potentiels**

Ensuite, parmi tous les points fournis par le critère SC_HK, ne sont conservés qu'une certaine proportion présentant les plus grandes valeurs soit de Courbure, soit de FQ (Facteur Qualité), soit de « confiance ». Le « facteur qualité » proposé dans (Mian, Bennamoun et Owens 2009) est donné par la formule ci-dessous, calculée sur l'ensemble du voisinage V entourant le point d'intérêt :

$$Q_k = \frac{1000}{n^2} \sum_V |K| + \max_V(100K) + |\min_V(100K)| + \max_V(10\,\kappa_1) + |\min_V(10\kappa_2)| \qquad \text{(Eq. 3.9)}$$

La confiance, quant à elle, est une estimation, proposée dans (Ho et Gibbins 2009), du niveau de fiabilité des points, fondée sur une mesure de la déviation de leur intensité de courbure par rapport à la moyenne et la variance de celle-ci dans le voisinage :

$$\gamma(\mathbf{p}, r_k) = \frac{|c_{\mathbf{p}} - \mu_{N_{\mathbf{p}}}|}{\sigma_{N_{\mathbf{p}}}} \qquad \text{(Eq. 3.10)}$$

Au final, selon le critère de tri final, nous obtenons diverses variantes de détecteurs 3D, comme récapitulé dans le Tableau 3.3.

Tableau 3.3 – Synthèse de nos variantes de détecteurs de points d'intérêt 3D

	Critère de sélection finale	**Méthode de regroupement**
SC_HK_C	Plus grande courbure	Tri et clustering
SC_HK_FQ	Plus grands Facteur Qualité	Tri et clustering
SC_HK_Conf	Plus grands index de confiance	Tri et clustering
SC_HK_Conn		Composantes connexes

Notre procédure complète de détection de points d'intérêt 3D est résumée ci-dessous :

Etapes générales des algorithmes de détection :

1. Mailler le nuage de points (profiter de la structure régulière+ Filtrage bilatéral si bruit)
2. Calculer les courbures principales $\kappa 1$ et $\kappa 2$ en chaque point
3. Pour chaque point , extraire le voisinage N_p (zone sphérique proportionnelle à la diagonale de boite englobante + un seuil sur la distance)
4. Calculer les mesures de saillances sur chaque support (classifications SC et HK selon les courbures)
5. Calculer une évaluation (C ou FQ ou Conf) des PIs
6. Regrouper et trier/filtrer les points détectés :

- **HK ∩ SC & Conf**
- **HK ∩ SC & FQ**
- **HK ∩ SC & C**

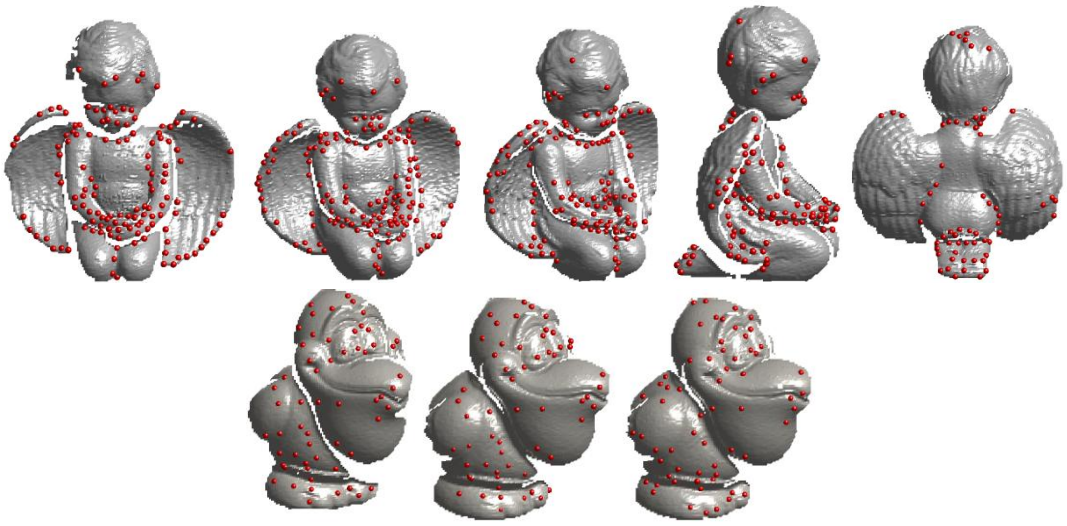


Figure 3.10 – Exemples de points d'intérêt SKIN obtenus avec notre détecteur SC_HK_FQ.

Une des caractéristiques souhaitables pour un détecteur de points d'intérêt est de sélectionner des zones répétables même en cas de changement d'angle de vue ou d'échelle. Une évaluation de cette répétabilité a été menée sur un benchmark public de nuages 3D issus d'un scanner, et a montré une supériorité assez nette de nos détecteurs en ce qui concerne les variations d'angle (cf. Figure 3. 11).

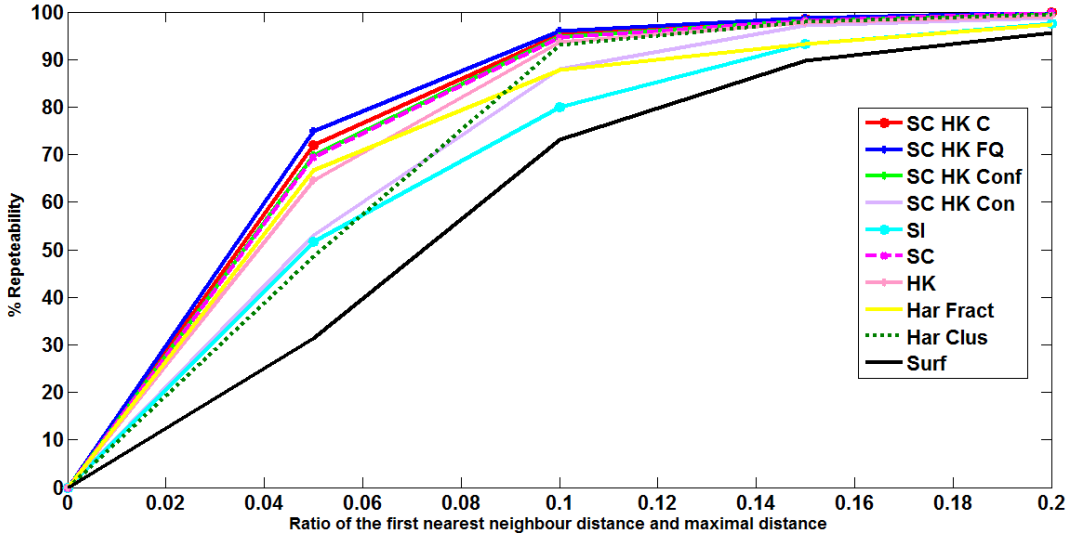


Figure 3.11 - Evaluation et comparaison (sur 9 objets de la base Minolta) de la répétabilité de position des points d'intérêt en cas de changement d'angle de vue.

Descripteurs SKIN : IndSHOT ou IndThrift

Après étude de l'état de l'art des descripteurs pour keypoints 3D, nous avons proposé deux nouveaux descripteurs :

- ***indSHOT***, qui concatène l'histogramme des cosinus de normales du SHOT de (Salti, Tombari et Di Stefano 2010), et un histogramme des indices de forme (SI) ; en pratique, nous avons repris la structure du C-SHOT de (Tombari, Salti et Di Stefano 2011), et remplacé la seconde partie (histogramme de couleurs) par notre histogramme de SI, comme illustré sur la Figure 3.12 ;

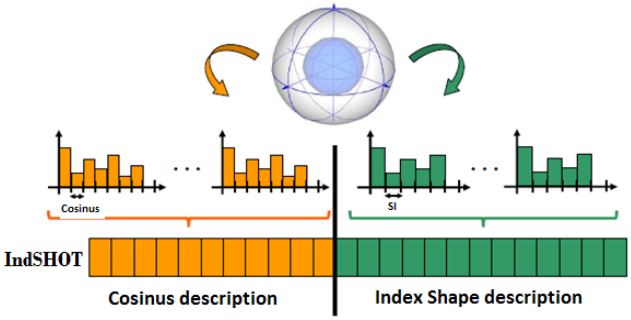


Figure 3.12 – Structure et contenu de notre descripteur IndSHOT

- ***indThrift***, qui mixe la formulation du descripteur « Local Surface Patch » (LSP) de (Chen et Bhanu 2007) avec celle du Thrift (Flint, Dick et van den Hengel 2007) : nous reprenons la structure du LSP, mais en remplaçant la mesure du cosinus entre la normale référence et les normales du voisinage par celle proposée dans le Thrift (voir Figure 3.13).

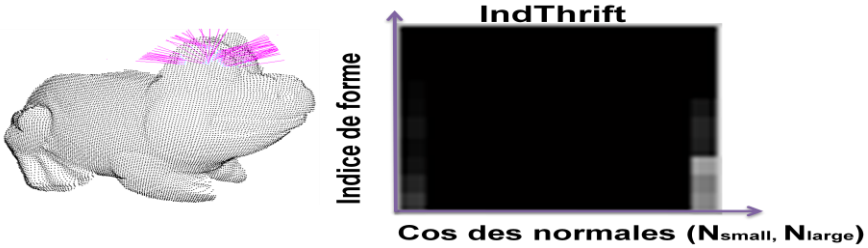


Figure 3.13 - Illustration de la structure de notre descripteur IndThrift

En ce qui concerne les descripteurs, il est souhaitable qu'ils fournissent une signature locale non seulement discriminante, mais aussi stable vis-à-vis des rotations et des changements d'échelle. L'évaluation, toujours sur un benchmark public de vues 3D réelles, nous a permis de mettre en évidence la supériorité, de nos descripteurs en terme de stabilité, en comparaison des principaux descripteurs de points d'intérêt 3D de l'état de l'art.

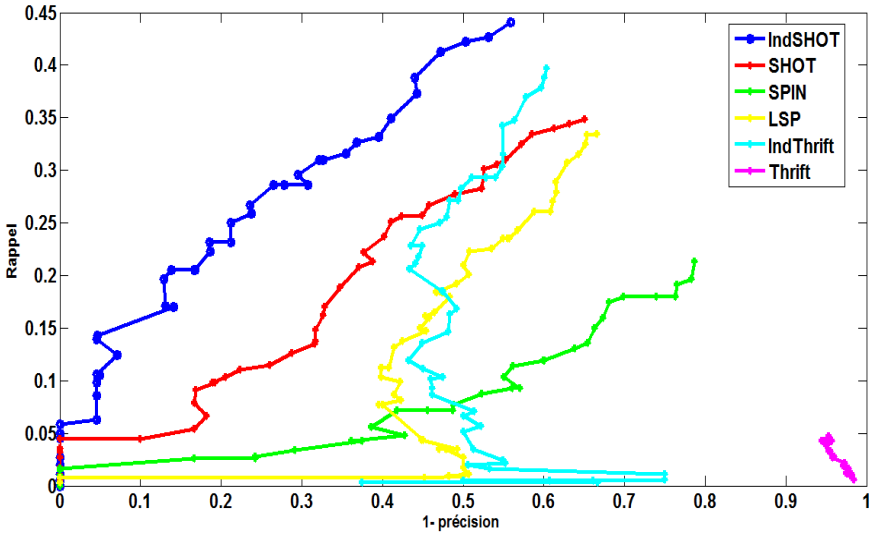


Figure 3.14 - Evaluation et comparaison de la stabilité, vis-à-vis du changement d'angle de vue, de divers descripteurs

Une autre manière d'évaluer la stabilité de nos descripteurs consiste à vérifier si, calculé sur un même point physique d'un objet vu sous 2 angles différents, nous obtenons bien des descripteurs suffisamment similaires pour que la mise en correspondance se fasse correctement. Il s'avère que c'est bien le cas, comme illustré sur la Figure 3.15.

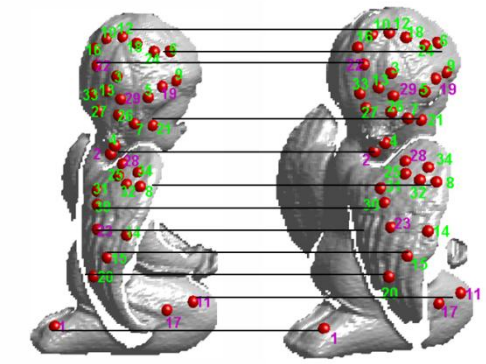


Figure 3.15 - Mise en correspondance correcte des descripteurs indSHOT après rotation des points entre la vue initiale à 100° et la vue à 120°.

Evaluation de nos points d'intérêt 3D « SKIN » pour l'identification d'objets 3D

Au final, ce qui compte réellement est bien sûr la performance de reconnaissance/identification d'objets 3D atteignable en utilisant une combinaison SKIN d'une de nos variantes de détecteurs, et d'un de nos 2 descripteurs. Cette évaluation a été menée sur plusieurs bases 3D contenant quelques dizaines d'objets (voir Figure 3.16), et montre que notre nouvelle famille de points d'intérêt 3D permet de dépasser significativement les performances de l'état de l'art.

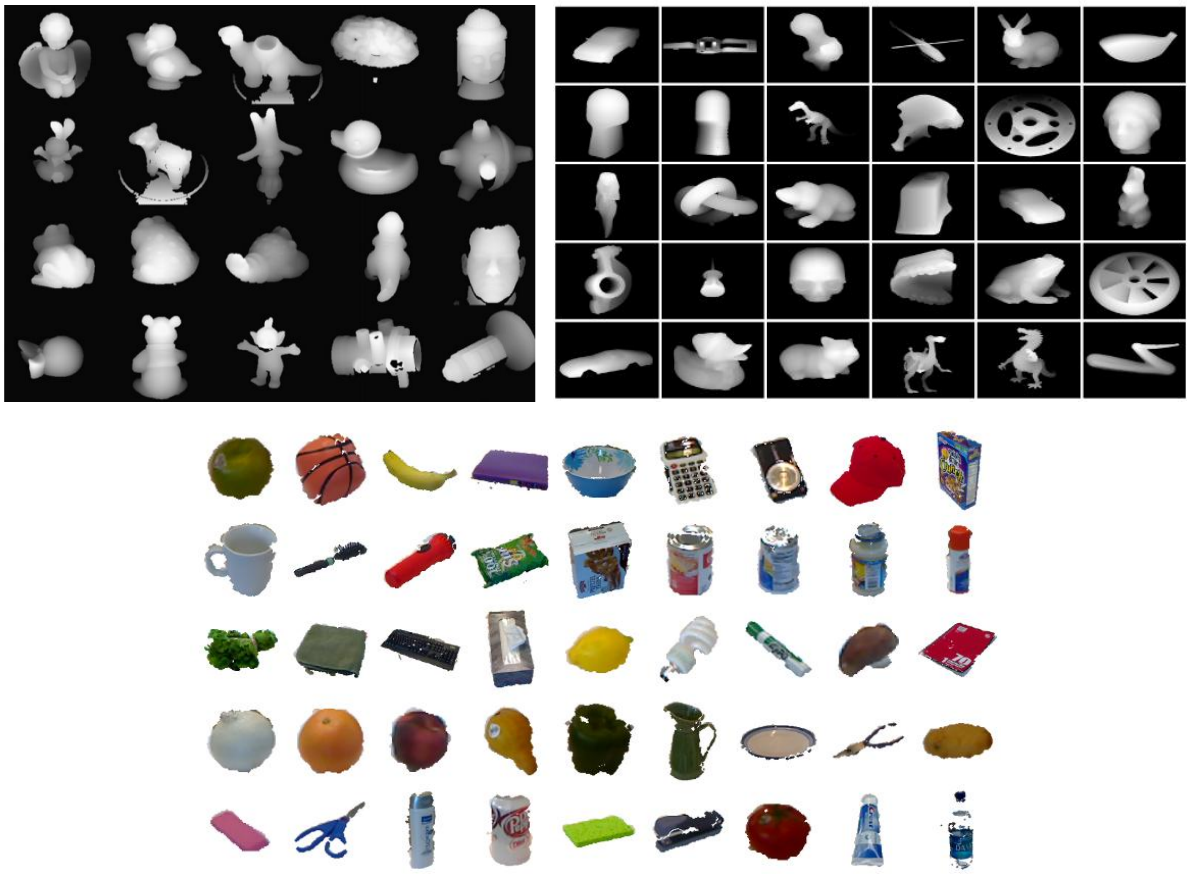


Figure 3.16 – Les 3 benchmarks publics utilisés : en haut à gauche la base Minolta, contenant des scans réels par laser de 20 objets vus sous divers angles ; en haut à droite, la base Stuttgart constituée d'images de profondeurs synthétiques générées à partir des modèle 3D complets de 42 objets ; en bas, la base RGB-D de 46 objets « communs » enregistré avec un capteur Kinect™.

A noter en particulier une performance de 95% d'identification correcte sur les 42 objets de la base Stuttgart avec notre SC_HK_C+IndSHOT (voir les détails dans le Tableau 3.4), soit un gain de 4% par rapport aux meilleurs résultats publiés antérieurement sur cette même base.

Tableau 3.4 – Performances de reconnaissance avec nos points d'intérêt SKIN, comparées aux résultats publiés sur la même base Stuttgart

Nombre d'objets	5		25	**42**
Méthode	E1	E2		
HK (Eskizara, 2009)	**99.5%**		93.0%	91,0%
SI (Hozatlı, 2009)		99,0%		91.0%
Notre méthode SC_HK_C- IndSHOT	95.4%	**99.6%**	**94.9%**	**94.9%**

De même, sur la base Minolta, en comparant nous-mêmes de nombreuses combinaisons détecteur+descripteur, nous constatons que les meilleures performances d'identification correcte (~89%) en variation d'échelle pour la base Minolta sont obtenues avec SC_HK_C+IndSHOT ou SC_HK_FQ+IndThrift (voir les détails dans le Tableau 3.5).

Tableau 3.5 – Performances de reconnaissance avec nos points d'intérêt SKIN, comparées à d'autres combinaisons détecteur+descripteur, sur la base Minolta

	IndSHOT	IndThrift	SURF
SC_HK_Conf	86.3%	**86.7%**	
SC_HK_C	**88.6%**	86.7%	
SC_HK_FQ	86.2%	**88.6%**	
SI	79.0%	**81.0%**	
Harris_cluster	67.6%	**68.6%**	
Harris_fract	79.4%	**82.8%**	
SURF			**71.9%**

Enfin, sur les données réelles plus bruitées et basse résolution (capteur Kinect™) de la base RGB-D, nos points d'intérêt SKIN (variante SC_HK_FQ+IndThrift) atteignent environ 78% d'identification correcte parmi les 37 objets (voir aussi diagramme de Hinton sur la Figure 3.17).

Pour plus de précisions, le lecteur pourra se reporter à nos publications suivantes : (Shaiek et Moutarde 2011), (Shaiek et Moutarde, 3D Keypoints Detection for Objects Recognition 2012), et (Shaiek et Moutarde 2012).

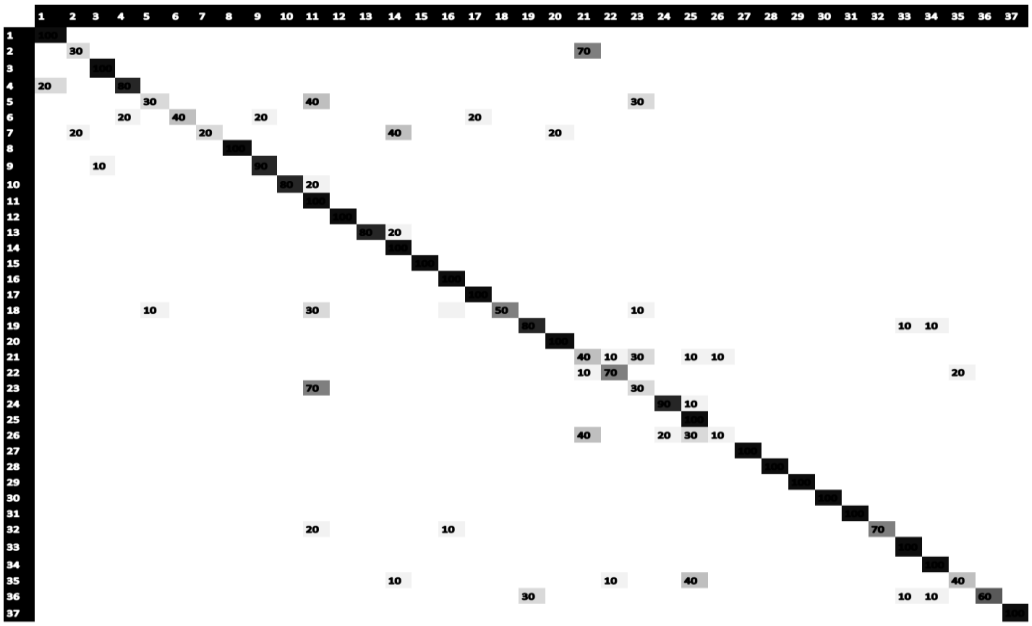


Figure 3.17 – Matrice de confusions sous forme de diagramme de Hinton pour l'identification des 37 objets de la base RGB-D, avec notre variante SC_HK_FQ+IndThrift de points d'intérêt SKIN, qui donne les meilleurs résultats sur cette base.

3.3. BILAN ET PERSPECTIVES SUR MES RECHERCHES EN IDENTIFICATION DE PERSONNE OU D'OBJET

Les divers travaux de recherche que j'ai réalisés et encadrés (principalement via 2 thèses sur le sujet) dans le domaine de l'identification de personnes en 2D, ou d'objets en 3D, ont donc apporté les contributions suivantes :

- une approche originale pour la ré-identification de piétons entre caméras à champs disjoints, fondée sur l'accumulation des points d'intérêt collectés sur les personnes, et dont les performances dépassent l'état de l'art ;
- la famille « SKIN » de détecteurs et descripteurs de points d'intérêt 3D.

Ces recherches ont fait l'objet d'***un article soumis à un journal (Pattern Recognition Letters) et en cours de relecture***, et de ***7 articles dans des conférences internationales*** à comité de lecture :

- "*Fast 3D keypoints detector and descriptor for view-based 3D objects recognition*", Ayet Shaiek and Fabien Moutarde, proc. International Workshop on Depth Image Analysis (WDIA'2012) of 21st International Conference on Pattern Recognition (ICPR'2012), Tsukuba (Japan), 11 nov. 2012.
- "*3D Keypoints Detection for Objects Recognition*", Ayet Shaiek and Fabien Moutarde, proc. 16th Int. Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'2012), Las Vegas (USA), 16-19 juillet 2012.
- "*3D keypoint detectors and descriptors for 3D objects recognition with TOF camera*", Ayet Shaiek and Fabien Moutarde, proc. of IS&T/SPIE Electronic Imaging conference on 3D Image Processing (3DIP) and applications, San Francisco (USA), 26-27 janvier 2011.
- "*3D Object Recognition and Facial Identification Using Time-averaged Single-views from Time-of-flight 3D Depth-Camera*", Hui Ding, Fabien Moutarde, and Ayet Shaiek, proc. 3rd Eurographics Workshop on "3D Object Retrieval" (3DOR2010), Norrköping (Sweden), mai 2010.
- "*Keypoints-based background model and foreground pedestrians extraction for future smart cameras*", Omar Hamdoun and Fabien Moutarde, proc. of 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2009), Como (Italy), 30 août - 3 septembre 2009.
- "*Interest points harvesting in video sequences for efficient person identification*", Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu and Bruno Steux, proc. of '8th international workshop on Visual Surveillance (VS2008)' of "10th European Conference on Computer Vision (ECCV'2008)", Marseille (France), 17 octobre 2008.
- "*Person Re-identification in Multi-camera System by Signature based on Interest Point Descriptors Collected on Short Video Sequences*" Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu and Bruno Steux, proceedings of ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC-08), Stanford University, California (USA), 7-11 septembre 2008.

Ces travaux, et l'encadrement scientifique total des 2 thèses correspondantes, m'ont permis à la fois d'approfondir mon expertise sur les techniques de type « points d'intérêt », d'acquérir une certaine expérience sur la spécificité des données 3D et descripteurs liés aux courbures, et d'élargir mon spectre de compétences vers d'autres champs applicatifs : vidéoprotection d'une part, et robotique d'autre part.

En ce qui concerne les perspectives de cet axe de recherches, une suite assez naturelle a déjà débuté via le ***co-encadrement scientifique d'une thèse CIFRE chez Aldebaran Robotics, pour laquelle un des objectifs est la localisation approximative d'un robot en environnement intérieur, par ré-identification visuelle de l'arrière-plan***. Par ailleurs, un autre prolongement de ces travaux sont les ***recherches que j'ai récemment démarrées sur l'identification de gestes, dans le cadre de la Chaire PSA « Robotique et réalité virtuelle ».***

4. FOUILLE DE DONNEES ET PREDICTION DE TRAFIC ROUTIER

Les travaux présentés dans ce chapitre sont de nature assez différente du reste de nos recherches, puisqu'ils ne concernent plus l'analyse d'images, ni la reconnaissance de formes. En revanche, il s'agit toujours d'analyser des données structurées (en graphe, au lieu de matrice), et les techniques mises en œuvre restent liées à l'apprentissage statistique, même s'il s'agit plus d'apprentissage non-supervisé (« clustering ») et de régression (prédiction) plutôt que de classification.

Les recherches présentées ci-dessous ont été réalisés de 2009 à 2012 dans le cadre du ***projet « TRAVESTI » financé par l'ANR, ainsi que du projet « PROBEX » subventionné par le Ministère de l'Environnement et du Développement Durable.*** En pratique une part importante de ces travaux ont été faits en collaboration avec le **post-doctorant Yufei HAN**, dont j'ai supervisé les travaux au CAOR durant 18 mois.

4.1. INTRODUCTION SUR LA FOUILLE DE DONNEES DE TRAFIC ROUTIER

La plupart des travaux publiés dans le domaine de l'analyse de données de trafic routier se focalisent sur la modélisation de la dynamique temporelle d'une seule section de route, que ce soit une portion d'autoroute où des modèles physiques approchés sont souvent utilisés, ou en zone urbaine où la régulation par les feux et le franchissement des intersections ont une importance critique. Des exemples d'approches fondées sur des modèles sont par exemple de type cinématique comme dans (Herty, Klar et Pareschi 2005), ou celui (classique) de suivi de véhicule précédent comme dans (Rakha 2009), à base d'automates cellulaires comme dans (Nagel et Schreckenberg 1992), ou encore issu de la physique comme le modèle de transition de phase de (Blandin, et al. 2011), ou enfin déduits d'une analyse des files d'attentes aux feux (Hofleitner, Herring et Bayen 2012). Par ailleurs, des approches dirigées par les données se sont aussi développées, comme par exemple : filtre de Kalman étendu dans (Wang et Papageorgiou 2005), analyse multivariée de série temporelle dans (Ghosh, Basu et O'Mahony 2009) ou (Min, et al. 2009), algorithmes neuronaux ou neuro-flous comme dans (Quek, Pasquier et Lim 2006), ou enfin propagation de croyances dans (Furtlehner, Lasgouttes et de La Fortelle, A belief propagation approach to traffic prediction using probe vehicles 2007).

A l'exception de la dernière référence, ces approches dirigées par les données ont été fortement influencées par la nature des données disponibles, qui jusque récemment étaient plutôt issues de boucles magnétiques situées sous les chaussées de quelques axes principaux (principalement autoroutiers), et assez espacées. Le développement récent de systèmes permettant d'obtenir des données issues de capteurs GPS de nombreux véhicules faisant office de « sondes » (Herring, et al. 2010) (Hofleitner, Herring et Bayen 2011) permet non seulement d'accéder directement aux informations de type « temps de parcours » (contrairement aux boucles magnétiques qui mesurent essentiellement des flux), mais aussi de disposer d'informations trafic sur des zones géographiques beaucoup plus étendues.

Une conséquence intéressante du développement de mesures disséminées par les véhicules traceurs, outre l'abondance croissante des données de trafic, est la possibilité, nouvelle, d'obtenir à tout instant une image du trafic sur la *totalité* d'une agglomération, y compris le réseau artériel. En collectant ces données sur une période de temps assez longue (plusieurs semaines au minimum), il devient possible d'analyser la dynamique du trafic considéré d'une manière globale : identifier les divers « motifs » spatiaux globaux de congestion (à l'échelle de toute la ville ou agglomération), ainsi que les divers types d'évolution temporelle (sur un horizon de plusieurs heures, voire d'une journée) entre ces motifs. Assez peu de travaux ont été publiés jusque maintenant concernant l'analyse et modélisation de la dynamique du trafic sur une large échelle géographique, à part ceux de Geroliminis (Geroliminis et Daganzo, Macroscopic Modeling of Traffic in Cities 2007) qui utilisent la notion de diagramme fondamental macroscopique (Geroliminis et Daganzo 2008) pour segmenter une agglomération en zones ayant chacune un diagramme flux-densité moyen spécifique (Geroliminis et Sun 2011).

4.2. NMF POUR L'ANALYSE DES ETATS DE TRAFIC ROUTIER

La problématique nouvelle de l'analyse de données trafic à grande échelle nous a conduit à développer une approche originale pour effectuer cette fouille de données sans se heurter au « fléau de la dimensionnalité ». Dans le principe, identifier les configurations typiques de congestion est un simple problème de partitionnement ou regroupement (« clustering » en anglais) : étant donnés l'ensemble des vecteurs x_t de l'état de trafic mesuré à divers moments sur une période assez longue, l'analyse de la distribution de ces x_t dans $\Re^n$ (où n est le nombre d'arcs routiers du réseau) doit permettre de trouver les états se reproduisant fréquemment ainsi que les zones de $\Re^n$ balayées par x_t selon les heures, les jours, et les incidents. Cependant, avec plusieurs milliers (voire dizaines de milliers) d'arcs routiers, pour chacun desquels est fournie une estimation de congestion mise à jour environ toutes les 10 minutes (donc ~150 mesures par jour), il serait illusoire d'appliquer aveuglément des algorithmes de type « clustering » directement sur ces données de très grande dimension.

Nous proposons donc pour analyser ces données d'en réduire la dimensionnalité tout en conservant leurs propriétés topologiques, grâce à une décomposition matricielle nommée « Non-negative Matrix Factorization » (NMF, voir notamment (Lee et Seung 2000) et (Lin 2007)). Celle-ci consiste à approximer l'énorme matrice X de dimension $n \times m$ des états de congestion (avec n le nombre d'arcs et m le nombre de mesures temporelles) comme un produit $M.V$ de 2 matrices non-négatives, de dimensions respectives $n \times s$ et $s \times m$, où s est la dimension du sous-espace sur lequel sont « projetées » les données ; le couple (M, V) est donc la solution du problème de minimisation sous contrainte suivant :

$$(M,V) = \underset{M \geq 0, V \geq 0}{\arg\min} \|X - MV\|_F \qquad \text{(Eq. 4.1)}$$

La formule donnant la reconstruction approximative de l'état de trafic est :

$$X_j = \sum_{i=1}^{s} M_i V_{i,j} \qquad \text{(Eq. 4.2)}$$

Une des spécificités de la NMF est la contrainte de non-négativité sur M et V. Du fait de celle-ci, la formule ci-dessus est une superposition *strictement additive* de s composantes, chacune d'elle (une colonne M_i de la matrice M) correspondant à une configuration spatiale particulière de congestion. La décomposition NMF met donc en lumière à la fois un ensemble de s « états prototypes » de trafic (chacun des M_i), et une décomposition de chaque état observé comme un barycentre particulier entre ces prototypes : par exemple, si l'état global de trafic X_j est très similaire à M_i, alors la pondération $V_{i,j}$ sera la plus grande.

De plus, en raison de la contrainte de non-négativité, M et V tendent à être « creuses » (« sparse » en anglais), ce qui fait de chacune des « composantes » de la décomposition une sorte de sous-ensemble du graphe routier (voir Figure 4. 5) dont la dynamique temporelle est relativement homogène. La décomposition $X \approx M.V$ revient ainsi à approximer l'évolution de la congestion sur les milliers d'arcs comme une superposition d'évolutions indépendantes de s « composantes » du réseau routier.

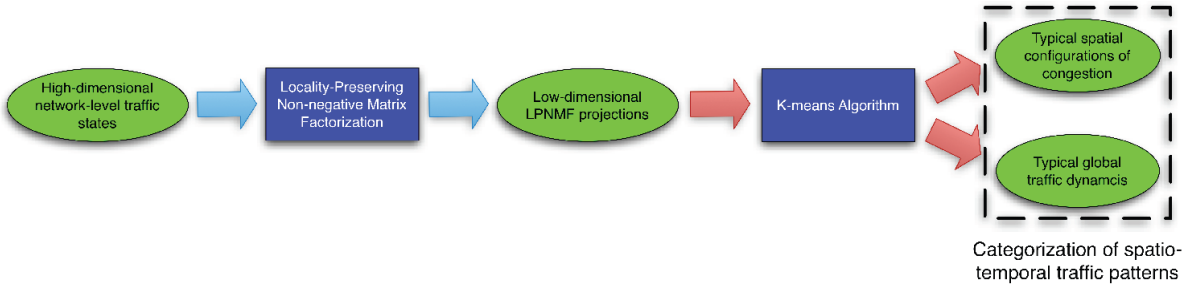


Figure 4.1 - Schéma de la méthodologie proposée de clustering.

En pratique, nous utilisons une variante de NMF qui préserve le voisinage local (« Locality-Preserving » NMF) :

$$O = \|X - MV\|_F^2 + \lambda Tr(VLV^T) \qquad \text{(Eq. 4.3)}$$

où L est le « Graph Laplacian », défini comme L=D−W avec W matrice des similarités entre états (W_{ij} = similarité entre *i-ème* et *j-ème* observation du trafic) et D est une matrice diagonale contenant les sommes par colonne de W ($D_{ii} = \Sigma_j (W_{ij})$). En pratique, le couple (M, V) minimisant la fonction de coût est obtenu par une méthode itérative. L'algorithme NMF, ou ici sa variante LP-NMF, approxime **X** comme le produit d'une matrice non-négative M de dimension $n \times s$ par une autre matrice non-négative V de dimension $s \times m$. Chaque colonne de V est donc assimilable à la projection d'un des états de trafic sur un espace de dimension s défini par les colonnes de M. En général, s est choisi beaucoup plus petit que la dimension n des états de trafic, aussi V fournit une représentation de l'état de trafic dans un espace de petite dimension. L'espace de projection formé par ces s composantes est ensuite très commode pour effectuer par exemple un clustering des diverses configurations géographiques de congestion, ou encore une typologie des évolutions temporelles globales.

Expérimentations sur des données simulées

Nous avons tout d'abord testé l'intérêt et la validité de notre approche sur des données simulées. Ces dernières ont été obtenues par un logiciel nommé Metropolis (Marchal 2001) (De Palma et Marchal 2002), qui a été développé pour la modélisation et simulation du trafic en vue de la planification des infrastructures. Les simulations qu'il permet de réaliser sont d'échelle "mésoscopique", c'est-à-dire permettant simulant les parcours individuels de chaque véhicule (contrairement aux simulateurs macroscopiques qui ne calculent que des quantités agrégées type débits et flux), mais en estimant uniquement le temps de parcours sur chaque arc (et non le détail du parcours et des interactions avec les autres véhicules, comme les font les simulateurs « microscopiques » type VISSIM). Outre la topologie du réseau et la vitesse « libre » moyenne sur chaque arc, les simulations se fondent sur des matrices « Origine-Destination » qui précisent les nombres de trajets par zones de départ et d'arrivée des véhicules, ainsi que l'heure désirée d'arrivée. Le temps de parcours effectif de chaque véhicule sur chaque arc est ensuite estimé selon la densité instantanée de véhicules par une simulation de type file d'attente. Les simulations que nous avons utilisées couvrent les routes et rues principales de la totalité de la région Ile-de-France, soit 13627 arcs en tout (voir topologie du réseau sur la Figure 4.2).

Nous avons utilisé une collection de 146 simulations couvrant chacune 8h d'évolution de trafic (uniquement la matinée) avec une période d'échantillonnage de 10mn, soit 48 pas de temps par simulation. Dans nos travaux, nous utilisons une mesure unique simplifiée et agrégée par arc et par période de 10mn du niveau de fluidité du trafic :

$$x_{pr} = \frac{\Delta t_p^0}{\Delta t_{pr}} \in]0,1] \qquad \text{(Eq. 4.4)}$$

Le numérateur est le temps de parcours « nominal », i.e. quand la fluidité est totale, pour l'arc p. Le dénominateur est le temps de parcours moyen sur ce même arc p mais au temps r (±7,5mn). Notre indice de trafic x_{pr} est donc une mesure de fluidité qui vaut 1 quand le trafic est fluide, et tend vers 0 en cas de congestion totale.

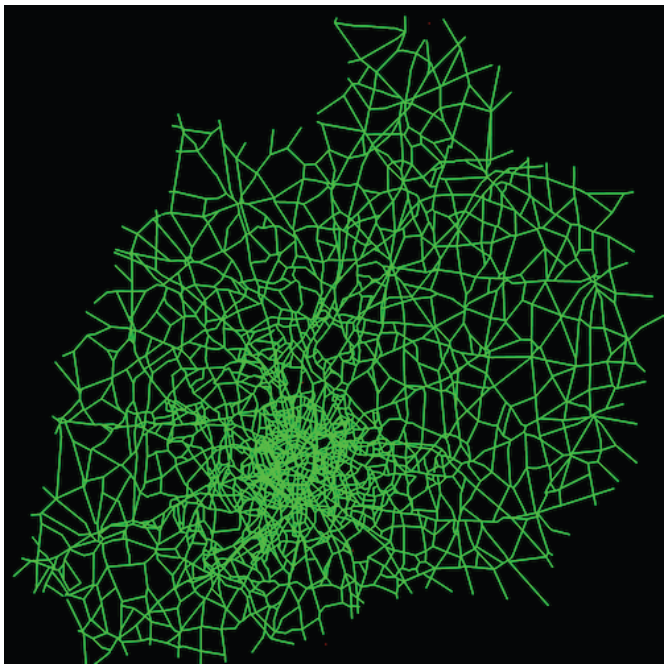


Figure 4.2 - Topologie du réseau francilien IAURIF simulé avec le logiciel Metropolis : ce modèle simplifié contient 13627 arcs, et couvre la totalité de l'Ile de France, depuis Paris intra-muros au centre, jusqu'à la lointaine banlieue en périphérie.

La taille totale de notre matrice X d'états de trafic est 13627×7008. Les 146 demi-journées simulées sont en fait subdivisées en 3 sous-ensembles avec des variantes de matrice Origine-Destination :

- des simulations avec principalement une demande « isotrope » de déplacement depuis la banlieue vers le centre de Paris (ITD);
- des simulations « anisotropes » avec plus de déplacements depuis la banlieue Nord vers le centre (ATD) ;
- des simulations de congestion extrême (ETD), obtenues en augmentant les nombres de déplacements à l'intérieur de la région centrale (Paris + petite couronne).

Les trajectoires typiques de l'état global de trafic des 3 types de simulation sont visualisées, dans l'espace 3D obtenu par Analyse en Composante Principales (ACP), sur la Figure 4.3. La position des trajectoires dans cet espace ACP est très différente selon le type de demande simulée sur le réseau. Les formes de trajectoires sont par contre assez similaires, à part quand la demande sature le réseau et que donc le système ne revient pas à l'état fluide initial en fin de demi-journée (cas ETD).

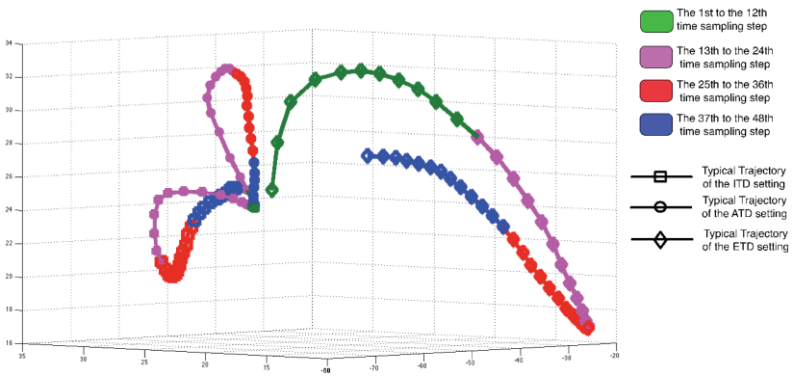


Figure 4.3 - Evolutions temporelles typiques du trafic (visualisées dans l'espace ACP 3D) pour chacun des 3 types de simulations.

Après avoir effectué la décomposition de la grande matrice X par LP-NMF avec 7 composantes, nous avons effectué un clustering par K-means sur le vecteur de projection dans $\Re^7$ (au lieu de travailler dans $\Re^{13627}$!). La partition obtenue pour K=3 est illustrée sur la Figure 4.4. Cette partition retrouve bien les 2 types différents de congestion correspondant aux cas isotropes ou anisotropes de congestion, ce qui s'avère impossible en effectuant le clustering dans l'espace obtenu par une simple Analyse en Composante Principale (ACP), qui ne fait que subdiviser en niveaux croissants de congestion (voir Figure 4.5).

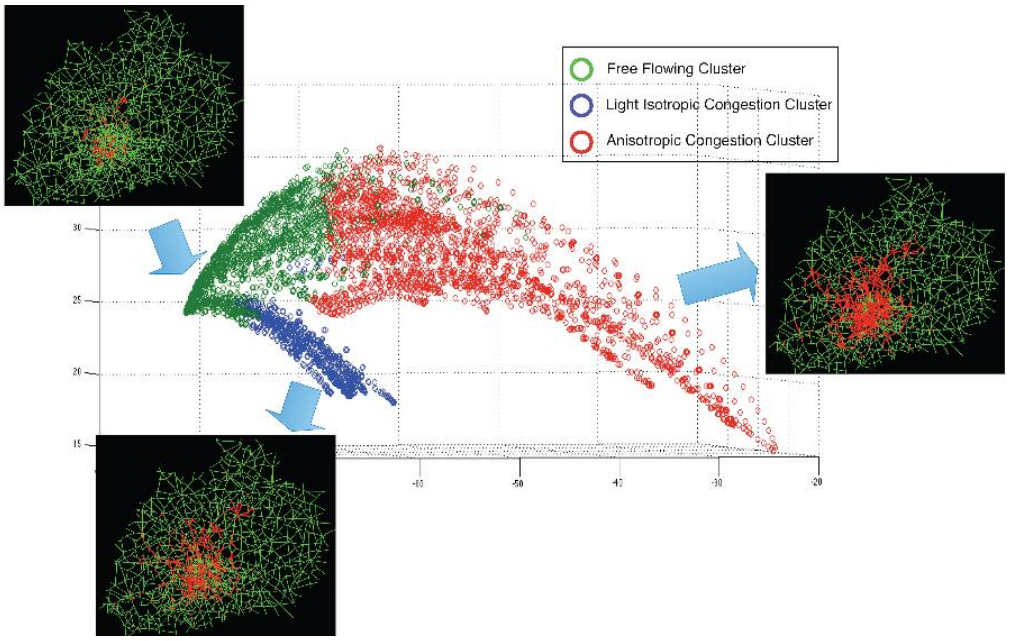
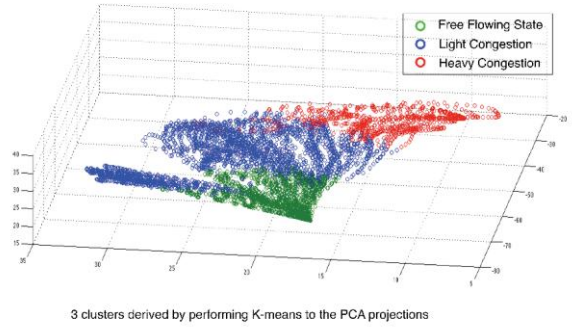
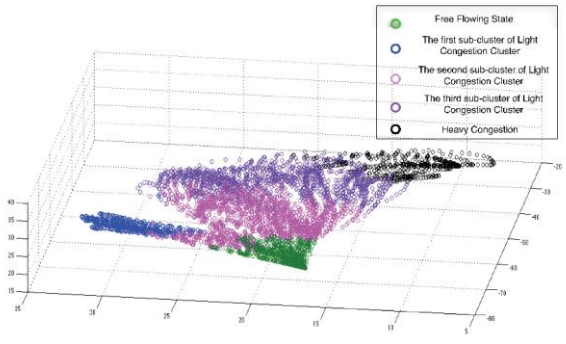


Figure 4.4 - Partition en 3 groupes des états de trafic, obtenue par K-means appliqué dans l'espace de projection LP-NMF de dimension 7 (visualisées dans l'espace ACP 3D) : nous obtenons un cluster correspondant à l'état fluide, et 2 clusters correspondant à 2 types différents de congestion.



3 clusters derived by performing K-means to the PCA projections



5 clusters derived by performing K-means to the PCA projections

Figure 4.5 – Résultat du clustering par K-means appliqué dans l'espace ACP 3D : ceci produit une subdivision selon le niveau moyen de congestion, mais pas selon la configuration spatiale de congestion.

Outre cette mise en lumière des diverses configurations spatiales typiques de congestion, la décomposition NMF nous permet aussi, via les « composantes NMF », d'identifier des portions du réseau routier qui ont des évolutions temporelles de congestion similaires. Pour cela, nous visualisons les composantes M_i en coloriant en rouge les arcs correspondant aux 20% de valeurs les plus grandes. Comme illustré sur la Figure 4.6, ceci met en évidence des zones qui ont bien une interprétation physique : par exemple la petite couronne d'une part, et la grande couronne d'autre part ressortent comme des régions cohérentes du point de vue évolution temporelle du trafic.

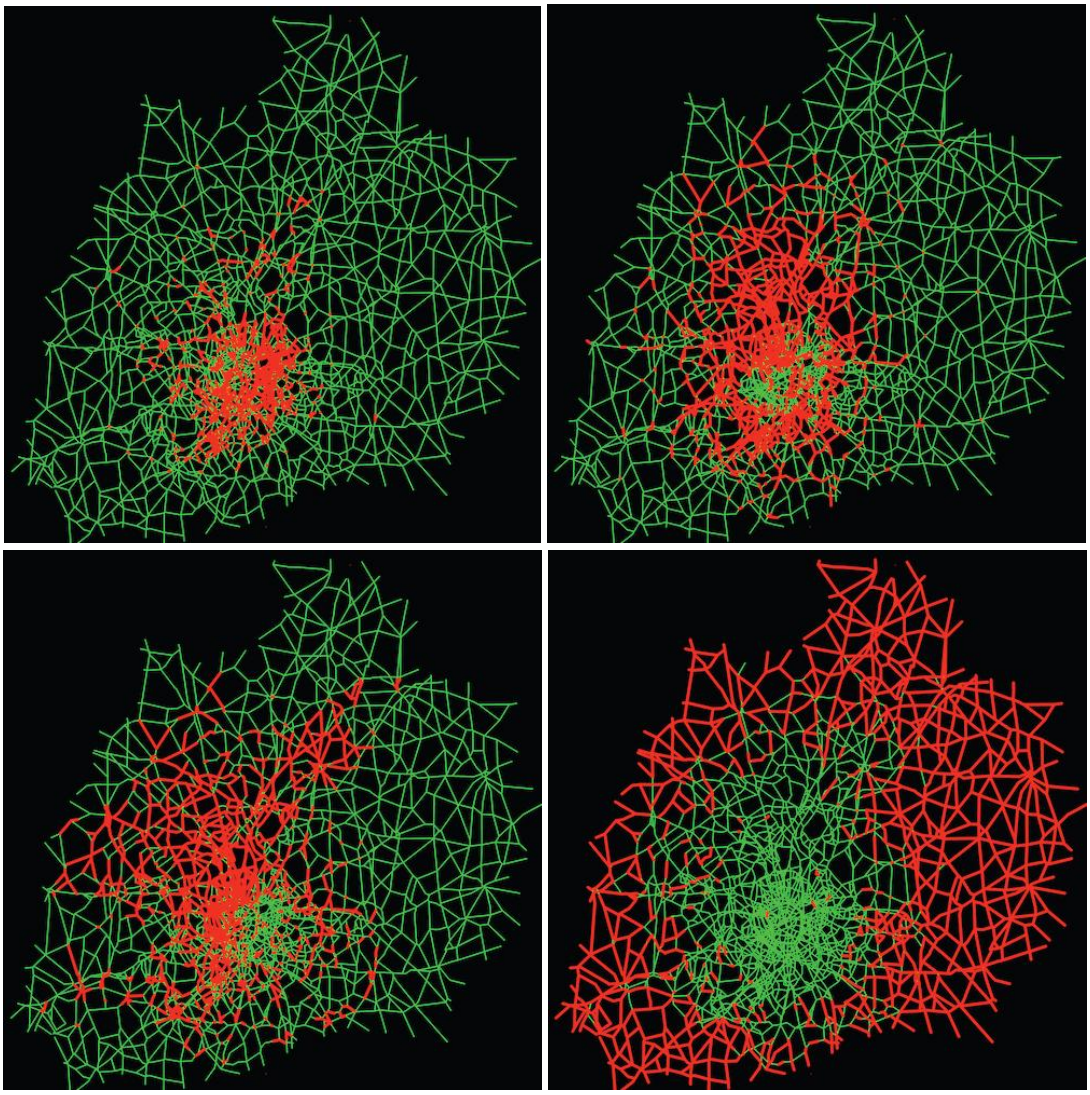


Figure 4.6 - Visualisation de 4 des composantes NMF obtenues ; grâce à la "sparsité" due à la non-négativité de la décomposition, chaque composante correspond bien principalement à une sous-partie du réseau (région centrale en haut à gauche, banlieue lointaine en bas à droite, etc…)

Expérimentations sur des données réelles (en collaboration avec UC Berkeley)

Nous avons aussi évalué notre approche sur des données réelles de trafic, dans le cadre d'une collaboration avec l'équipe du Pr Bayen à l'Université de Californie à Berkeley. Cette équipe a mis en place depuis 2008 une plate-forme d'acquisition des données de trafic issues de positions GPS de véhicules traceurs. A partir de ces données brutes, ils estiment, par tranche de 5mn, les temps de parcours sur chaque arc des principales artères de San Francisco et sa région. Dans le cadre de notre collaboration, UC Berkeley nous a fourni ces estimations pour un sous-réseau de 2626 arcs (ville de San Francisco) sur une durée de 6 mois (184 jours, du 1er mai 2010 au 31 octobre 2010). En

appliquant notre décomposition LP-NMF sur la matrice 2626×52292 correspondante, nous avons pu tout d'abord identifier 5 groupements (clusters) des états de trafic. Comme indiqué dans le Tableau 4.1 et visualisé sur la Figure 4.7, ces états prototypiques correspondent grosso-modo à des moments différents d'une journée. Ce qui est intéressant, c'est que cette partition de l'espace des états de congestion parvient à distinguer, malgré des fluidités moyennes similaires, les configurations spatiales différentes d'une part de la congestion croissante du matin (MIC) et du retour progressif à fluidité le soir (ADC), et d'autre part de la quasi-fluidité du soir (EFF) différente de celle de la nuit et aube (NFF).

Tableau 4.1 – Les 5 clusters d'états de trafic identifiés par clustering dans espace NMF sur les données trafic réelle de San Francisco

Marker symbol	Average fluidity	Cluster_name
Green_stars	0.7757	'Night + early morning Free-Flow' (NFF)
Blue_circles	0.7185	'Morning Increasing Congestion' (MIC)
Red_stars	0.6393	'Mid-Day Congestion' (MDC)
Yellow_diamonds	0.6730	'Afternoon Decreasing Congestion' (ADC)
Cyan_squares	0.7420	'Evening Free-Flow' (EFF)

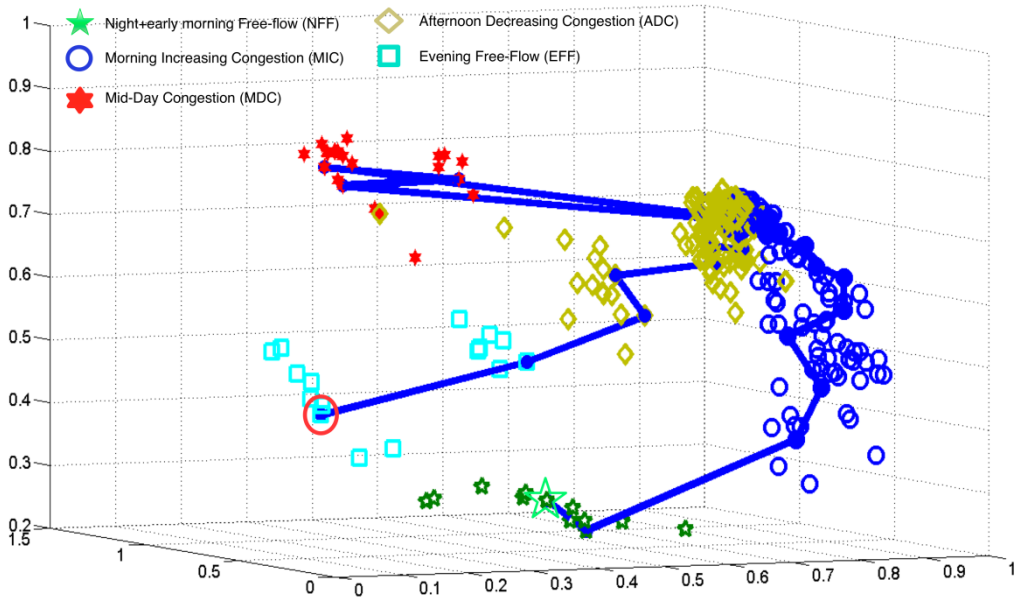
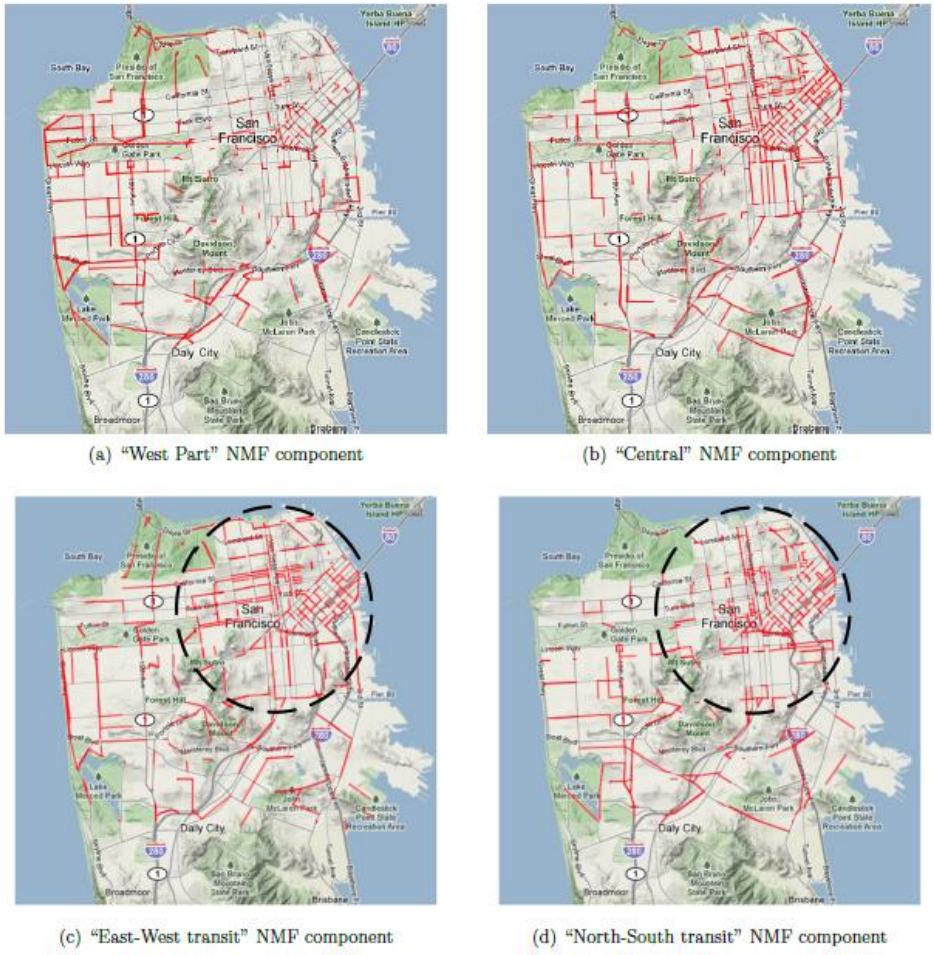


Figure 4.7 - Visualisation des 5 clusters d'états de trafic obtenus par K-means dans espace NMF sur les données trafic réelles Mobile Millenium de San Francisco.

De plus, l'analyse de la décomposition NMF obtenue nous a permis de constater que les composantes NMF (voir Figure 4.8) correspondent bien à des zones distinctes du point de vue de la réalité de l'évolution journalière usuelle du trafic : zone « Ouest » v.s. zone centrale, et axes Est-Ouest v.s. axes Nord-Sud.

Pour plus de détails sur ces travaux, le lecteur peut se référer à nos publications suivantes : (Furtlehner, Han, et al. 2010), (Han et Moutarde, Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization 2011), (A. Hofleitner, A. Herring, et al. 2012), et (Han et Moutarde 2013).



(a) "West Part" NMF component (b) "Central" NMF component

(c) "East-West transit" NMF component (d) "North-South transit" NMF component

Figure 4.8 - Visualisation de 4 composantes NMF obtenues sur les données trafic réelles Mobile Millenium fournies par UC Berkeley dans le cadre d'une collaboration.

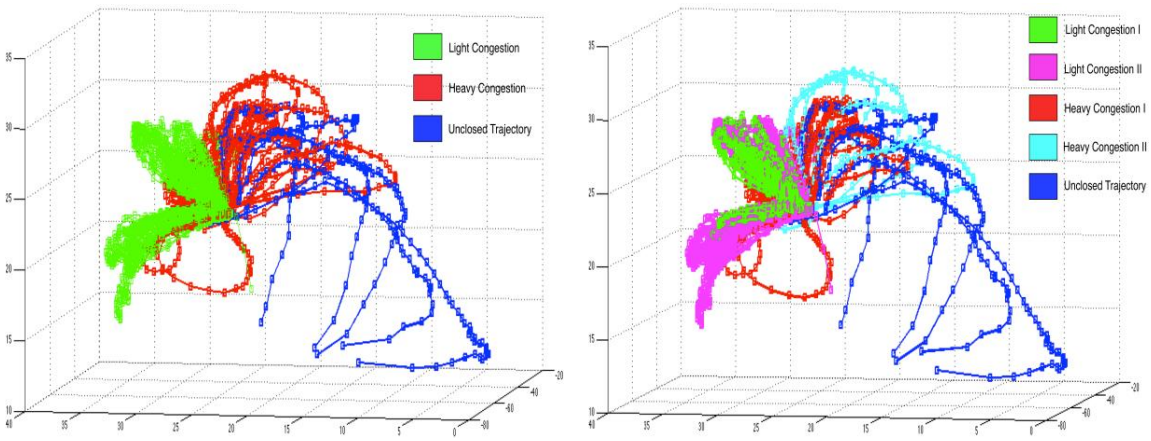
L'analyse d'une part des configurations de congestion, et d'autre part des sous-ensembles du graphe routier qui ont des évolutions temporelles similaires (composantes NMF), renseignent sur la distribution spatiale des propriétés du trafic. Ces résultats peuvent être exploités dans un but de modélisation simplifiée de la dynamique du trafic : par exemple, sous forme d'un modèle de Markov avec des transitions entre des « états- prototypes » de congestion ; et/ou sous forme d'évolutions parallèles relativement indépendantes pour chacune des « composantes ». Cependant, le plus intéressant au final est de parvenir à fournir des prédictions *quantitatives* de l'évolution du trafic.

4.3. NMF POUR LA TYPOLOGIE DES *EVOLUTIONS* DE TRAFIC, ET LA *PREDICTION* A MOYEN TERME

Pour analyser non plus la configuration spatiale de la congestion, mais sa dynamique temporelle, nous sommes tournés tout d'abord vers une typologie des évolutions temporelles globales. Pour ce faire, nous utilisons encore la projection LP-NMF, en représentant chaque demi-journée d'évolution comme une séquence de 48 vecteurs de $\Re^7$ (l'espace des composantes NMF). Pour effectuer un clustering sur ces trajectoires, nous utilisons comme mesure de similarité entre 2 séquences $p=\{p_i, i=1:48\}$ et $q=\{q_i, i=1:48\}$ la « distance cosinus » définie comme suit :

$$D(p,q) = 1 - \sum_{i=1}^{48} \frac{p_i \cdot q_i}{\|p_i\| \|q_i\|} \qquad \text{(Eq. 4.5)}$$

Cette “distance” est bornée dans l’intervalle [0 ;1], et vaut 0 pour deux séquences identiques, et tend vers 1 quand les 2 séquences sont très différentes. En appliquant K-means avec cette distance, nous obtenons facilement une segmentation de l’ensemble des trajectoires en plusieurs types d’évolution, comme cela se voit sur la Figure 4.9.



(a) (b)

Figure 4.9 - Résultats de clustering, dans l’espace LP-NMF, des évolutions globales du trafic sur les simulations IAURIF : (a) Découpage en 3 clusters ; (b) Découpage en 5 clusters.

Chacun des « groupes de trajectoire » correspond à la fois à des configurations spatiales différentes de la congestion au moment de l’heure de pointe, et à des évolutions temporelles moyennes différentes, tant en terme d’amplitude que de l’heure précise du pic de congestion, comme illustré sur la Figure 4.10.

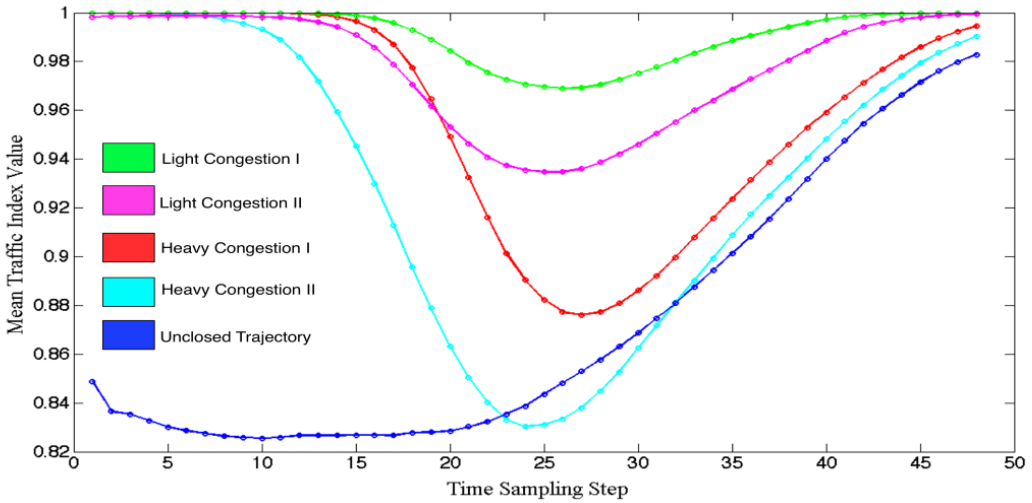


Figure 4.10 – Evolutions temporelles de l’indice de trafic (moyenné sur tous les arcs) pour les centroïdes de chacun des groupes de trajectoires obtenus sur la Figure 4.9b.

Dans le cas des données trafic réelles à San Francisco, le clustering permet de retrouver l’influence déterminante du jour de semaine sur le niveau de trafic, comme cela se voit sur la Figure 4.11.

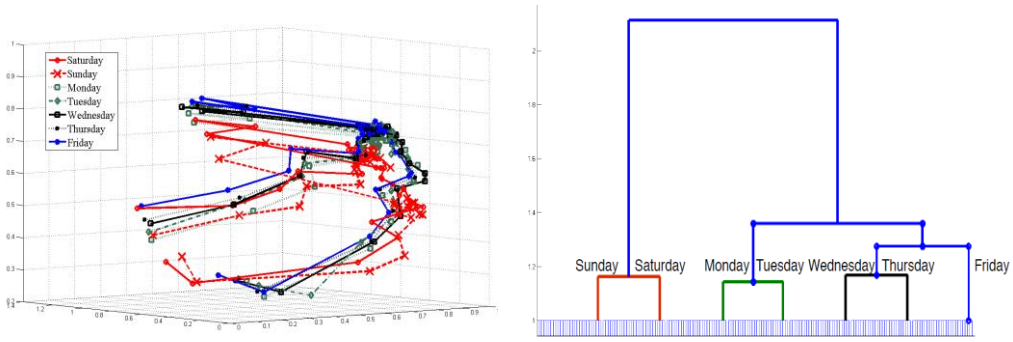


Figure 4.11 - Résultats du clustering des évolutions journalières du trafic réel sur San Francisco (données Mobile Millenium) ; à droite le dendrogramme qui retrouve bien les jours de week-end comme un groupe à part, et le vendredi comme un jour particulier.

En regardant les trajectoires sur la Figure 4.9, il semble possible, si par exemple le premier quart de la matinée a déjà été observé, de prédire la suite « naturelle » (i.e. si aucun incident spécifique ne vient perturber le système) de l'évolution de trafic. Plus formellement, nous avons testé l'approche suivante : étant donné un début d'évolution, nous cherchons dans l'historique les trajectoires dont les débuts sont les plus similaires (au sens de la distance cosinus). Puis nous utilisons ces « plus proches trajectoires » comme des indications de futurs possible/probable pour le trafic dans heures à venir.

Nous avons expérimenté ceci sur nos simulations IAURIF, de la façon suivante : la séquence temporelle de 5x15mn correspondant à la montée en charge de 7h à 8h15 de la congestion du matin est considérée comme observée, et nous cherchons à prédire les 26x15mn suivantes de la matinée. L'historique nous servant à effectuer notre prédiction est constitué de 106 simulations choisies aléatoirement, et les 40 restantes servent à évaluer l'erreur de prédiction. Notre procédure de prédiction est la suivante : étant donné le début de séquence observé, nous recherchons dans l'historique les K débuts de séquence les plus proches (au sens distance cosinus) dans l'espace de projection NMF. Ensuite nous utilisons ces K plus proches voisins pour estimer au mieux (comme un barycentre optimal) les coordonnées de projection NMF du début observé de séquence. Une fois cette estimation de projection NMF effectuée, nous les utilisons pour prédire la suite de l'évolution à l'aide de la formule 4.2 de reconstruction des états de trafic. Comme le visualise la Figure 4.12, cette technique de prédiction permet d'obtenir une erreur d'estimation du futur non seulement beaucoup plus faible qu'en utilisant simplement la moyenne de l'historique pour les mêmes jour/heure, mais aussi deux fois moins grande que si la moyenne est restreinte aux K trajectoires les plus similaires.

Pour plus de détails sur ces travaux, le lecteur peut se référer à nos publications suivantes : (Moutarde et Han 2011), (A. Hofleitner, A. Herring, et al. 2012), et (Han et Moutarde 2012).

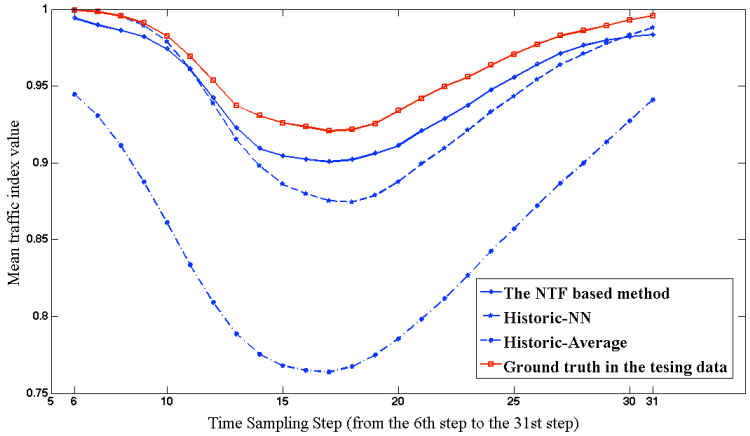


Figure 4.12 - Comparaison de la prédiction d'évolution du trafic par notre approche, avec les résultats de méthodes plus simples.

4.4. BILAN ET PERSPECTIVES SUR MES RECHERCHES EN ANALYSE ET PREDICTION DE TRAFIC ROUTIER

Les divers travaux de recherche que j'ai réalisés et supervisés (principalement via ***1 post-doc de 18 mois encadré sur le sujet***) dans le domaine de l'analyse et prédiction de trafic routier ont donc apporté les contributions suivantes :

- une méthode originale, utilisant la décomposition NMF des données, pour à la fois identifier les états prototypiques de congestion, ainsi que des zones ayant des comportements temporels cohérents, et catégoriser les évolutions journalières de trafic ;
- une technique pour faire de la prédiction d'évolution moyen-terme (quelques heures à l'avance) du trafic en fonction du début observé de journée et de l'historique disponible des évolutions temporelles.

Ces recherches ont fait l'objet d'***un article accepté dans le journal IET Intelligent Transport Systems*** :

- "*Statistical Traffic State Analysis in Large-scale Transportation Networks Using Locality-Preserving Non-negative Matrix Factorization*", Yufei Han and Fabien Moutarde, accepted for publication in IET Intelligent Transport Systems journal (2013).

et de ***6 articles dans des conférences internationales*** à comité de lecture :

- "*Analysis of Large-scale Traffic Dynamics using Non-negative Tensor Factorization*", Yufei Han and Fabien Moutarde, proc. 19th World Congress on Intelligent Transport Systems (ITSwc'2012), Vienna (Austria), 22-26 octobre 2012.
- "*Large scale estimation of arterial traffic and structural analysis of traffic patterns using probe vehicles*", Aude Hofleitner, Ryan Herring, Alexandre Bayen, Yufei Han, Fabien Moutarde and Arnaud de la Fortelle, proc. of Transportation Research Board 91st Annual Meeting (TRB'2012), Washington DC (USA), 22-26 janvier 2012.
- "*A new traffic-mining approach for unveiling typical global evolutions of large-scale road networks*", Fabien Moutarde and Yufei Han, proc. 18th World Congress on Intelligent Transport Systems (ITSwc'2011), Orlando (USA), 16-20 octobre 2011.
- "*Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization*", Yufei Han and Fabien Moutarde, proc. 14th IEEE Intelligent Transport Systems Conference (ITSC'2011), Washington DC (USA), 5-7 octobre 2011.
- "*Clustering and Modeling of Network level Traffic States based on Locality Preservative Non-negative Matrix Factorization*", Yufei Han and Fabien Moutarde, proc. 8th Intelligent Transport Systems (ITS) European congress, Lyon (France), 6-9 juin 2011.
- "*Spatial and Temporal Analysis of Traffic States on Large Scale Networks*", Cyril Furtlehner, Yufei Han, Jean-Marc Lasgouttes, Victorin Martin, Fabrice Marchal and Fabien Moutarde, proc. 13th Int. IEEE Conf. on Intelligent Transportation Systems (ITSC'2010), Madeira Island (Portugal), 19-22 septembre 2010.

Ces travaux m'ont permis à la fois de diversifier mon expertise scientifique vers les techniques de fouille de données, et d'élargir mon spectre de compétences vers le domaine de l'analyse, modélisation et prédiction de trafic routier.

En ce qui concerne les perspectives de cet axe de recherche, ***une évolution logique, en partie commencée durant mon court séjour sabbatique à l'UC Berkeley, serait d'exploiter la prédiction moyen-terme de trafic afin d'optimiser le routage des véhicules dans le réseau pour diminuer la congestion***. Un des points durs, mais très intéressant à la fois par son application pratique et le challenge algorithmique qu'il représente, est de concevoir une méthode permettant des re-routages « collaboratifs » pour éviter de créer des congestions secondaires importantes, comme cela risque fort de se produire si chaque véhicule pris dans un même bouchon recalcule indépendamment le *même* itinéraire alternatif.

CONCLUSIONS ET PERSPECTIVES

Les recherches présentées dans ce manuscrit ont pour point commun et fil conducteur le fait de mettre en œuvre des techniques d'apprentissage statistique. De plus, la majorité de ces travaux se situe dans le domaine de la vision par ordinateur. Enfin un dénominateur commun à une part importante de ces recherches est de concerner des applications de Système de Transport Intelligents. Elles m'ont permis d'apporter des contributions significatives :

- en détection et reconnaissance de panneaux routiers ;
- en détection et catégorisation visuelle d'objets (notamment voitures et piétons) ;
- en ré-identification visuelle de personnes entre caméras à champs disjoints ;
- en identification d'objets sur des images 3D issues de capteurs de profondeur (scanner laser ou Kinect) ;
- en analyse et prédiction de trafic routier.

Par ailleurs, ces travaux m'ont permis d'acquérir une bonne expertise couvrant un champ assez large :

- les techniques d'analyse temps-réel de vidéos embarquées (ainsi que sur le cadre plus général des systèmes d'aide à la conduite) ;
- l'application pratique des systèmes de reconnaissance de formes les plus standards (réseaux neuronaux, SVM, etc…) ;
- les descripteurs visuels pertinents pour effectuer reconnaissance, catégorisation, ou identification ;
- les techniques de type détecteurs et descripteurs de points d'intérêt (SIFT ou SURF notamment) ;
- les spécificités du traitement des « images 3D » issues de capteurs de profondeur ;
- les algorithmes de réduction de dimension type NMF ;
- les spécificités de l'analyse et prédiction de données de type trafic routier sur un large graphe ;
- les techniques d'apprentissage non supervisés (clustering).

Projet de recherche

Mon projet de recherche pour les années à venir s'inscrit dans la continuité des travaux présentés dans ce manuscrit, à savoir concevoir et mettre au point des applications innovantes utilisant les techniques d'apprentissage statistique et de fouille de données, principalement dans les domaines de la vision temps-réel et des Systèmes de Transport Intelligents. Plus précisément mes recherches actuelles s'organisent selon 3 axes principaux :

- Détection et catégorisation d'objets en temps-réel dans un flux vidéo, en embarqué dans un véhicule, pour des systèmes avancés d'aide à la conduite (ADAS, Advanced Driving Assistance Systems)
- Identification de personnes ou d'objets, en 2D ou 3D (images de profondeur)
- Fouille de données et prédiction de trafic routier à grande échelle

En ce qui concerne le premier axe (correspondant aux deux premiers chapitres de ce manuscrit), la suite naturelle est de continuer à travailler pour améliorer les performances (taux de détection et de fausse alarme) de la détection visuelle d'objets, en particulier de piétons compte tenu de l'ajout probable à assez court terme d'un item « Evitement de collision piéton » dans les critères d'évaluation euroNcap de sécurité des voitures. De plus, j'oriente actuellement mes recherches dans ce domaine vers un niveau d'abstraction plus élevé, à savoir la déduction en temps-réel par inférence probabiliste, d'informations de plus haut niveau (type de route, zone de travaux, approche d'intersection, etc…) en utilisant les modules plus élémentaires type détection et reconnaissance de certains panneaux, de marquages au sol, de feux tricolores, etc…

Dans le deuxième domaine, après les travaux menés sur la ré-identification de personnes entre caméras (pour la vidéo-protection), et ceux sur l'identification d'objets à l'aide de capteurs 3D (type scanner laser ou Kinect), mes recherches se tournent maintenant vers l'identification de *gestes*, avec des caméras de profondeur et/ou des capteurs inertiels mesurant les mouvements du corps humain. Le principal domaine applicatif visé est lié à la Chaire « PSA Robotique et Réalité Virtuelle » dont le laboratoire est titulaire, et dont un des objectifs est de pouvoir donner aux futurs robots manufacturiers un minimum de compréhension des actions des opérateurs humains afin de permettre d'aller vers une véritable collaboration homme-robot. Est aussi concerné, via le gros projet européen i-Treasure (http://www.i-treasures.eu/), le domaine de la sauvegarde et transmission du patrimoine immatériel, dont le savoir-faire gestuel expert, par exemple d'artisans.

Enfin, concernant le dernier volet, outre l'approfondissement des techniques d'analyse et prédiction de trafic routier, mes travaux évoluent naturellement vers la question difficile du re-routage collaboratif pour permettre aux usagers de la route de contourner les bouchons en se ré-orientant vers des itinéraires alternatifs sans pour autant créer de congestion secondaire qui les bloqueraient à nouveau dans le même « itinéraire bis ».

BIBLIOGRAPHIE

Abramson, Y., B. Steux, and H. Ghorayeb. "YEF (Yet Even Faster) Real-Time Object Detection." *Proceedings of International Workshop on Automatic Learning and Real-Time (ALART'05).* Siegen (Germany), 2005.

Abramson, Y., F. Moutarde, B. Steux, and B. Stanciulescu. "Combining adaBoost with a Hill-Climbing evolutionary feature search for efficient training of performant visual object detectors." *proceedings of FLINS2006 on Applied Computational Intelligence.* Genoa (Italy), 2006.

Agarwal, S., A. Aatitf Awan, and D. Roth. "Learning to Detect Objects in Images via a Sparse, Part-Based Representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, no. 11 (2004).

Akagündüz, E. *3D Object recognition using scale space of curvatures.* Graduate School Of Natural And Applied Sciences Of Middle East Technical University, 2011.

Albu, A. B., et al. "MONNET: Monitoring Pedestrians with a Network of Loosely-Coupled Cameras." *Proc. IEEE International Conference on Pattern Recognition (ICPR).* 2006. 924-928.

Arth, C., C. Leistner, and H. Bischof. "Object Reacquisition and Tracking in Large-Scale Smart Camera Networks." *Proc. 1st ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC'07).* 2007. 156-163.

Ashbrook, A.P, R.B. Fisher, C. Robertson, and N. Werghi. "Finding surface correspondence for object recognition and registration using pairwise geometric histograms." *European Conference on Computer Vision (ECCV'98)*, 1998: 674.

Bahlmann, C., Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler. "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information." *in Proceedings. IEEE Intelligent Vehicles Symposium*, 2005.

Bak, S., E. Corvée, F. Brémond, and M. Thonnat. "Person Re-identification Using Haar-based and DCD-based Signature." *proc. of 7-th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS'2010).* 2010.

Bargeton, A., F. Moutarde, F. Nashashibi, and Bradai B. "Improving pan-European speed-limit signs recognition with a new 'global number segmentation' before digit recognition ." *Proc. IEEE Intelligent Vehicle symposium (IV'2008).* Eindhoven (Netherlands), 2008.

Barnes, N., and A. Zelinsky. "Real-time radial symmetry for speed sign detection." *Proc. IEEE Intelligent Vehicles Symposium.* 2004.

Bay, H., T. Tuytelaars, and L. van Gool. "SURF:Speeded Up Robust Features." *Proceedings of the 9th European Conference on Computer Vision (ECCV'2006).* 2006.

Bdiri, T., F. Moutarde, and B. Steux. "Visual object categorization with new keypoint-based adaBoost features." *Proc. of IEEE Symposium on Intelligent Vehicles (IV'2009).* Xi'An (China), 2009.

Bdiri, T., F. Moutarde, N. Bourdis, and B. Steux. "adaBoost with 'keypoint presence features' for real-time vehicle visual detection ." *Proc. of 16th World Congress on Intelligent Transport Systems (ITSwc'2009).* Stockholm (Sweden), 2009.

Belhumeur, P.N., J.P. Hespanha, and D.J. Kriegman. "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 19 (1997): 711-720.

Blandin, S., D. Work, P. Goatin, B. Piccoli, and A. Bayen. "A general phase transition model for vehicular traffic." *SIAM Journal of Appl. Math.* 71, no. 1 (2011): 107–127.

Borgefors, G. "Distance Transformations in Digital Images." *Computer vision, graphics, and image processing* 34, no. 3 (1986): 344-371.

Bosch, A., A. Zisserman, and X. Munoz. "Representing shape with a spatial pyramid kernel." *Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR '07).* Amsterdam (Netherlands), 2007.

Broggi, A., P. Cerri, P. Medici, P.P. Porta, and G. Ghisio. "Real Time Road Signs Recognition." *Proc. of IEEE Intelligent Vehicles symposium (IV'2007).* Istanbul (Turkey), 2007.

Cecotti, H., C. Choisy, and A. Belaid. "Réseau de neurones à topologie dynamique, comparaison avec des invariants pour la reconnaissance de caractères multi-orientés multi-échelles." *Huitième Colloque International Francophone sur l'Ecrit et le Document*, 2004.

Chang, K., K. Bowyer, and P. Flynn. "Face recognition using 2D and 3D facial data." *ACM Workshop on Multimodal User Authentication*, 2003: 25-32.

Chang, Y.-L., and X. Li. "Adaptive image region-growing." *IEEE Transactions on Image Processing* 3, no. 6 (1994): 868-872.

Chaouch, M., and A. Verroust-Blondet. "A new descriptor for 2D depth image indexing and 3D model retrieval." *Proceedings of the International Conference on Image Processing (ICIP 07)*, 2007: 373-376.

Chen, H., and B. Bhanu. "3D free-form object recognition in range images using local surface patches." *Pattern Recognition Letters* 28, no. 10 (2007): 1252-1262.

Cong, D.N.T., C. Achard, L. Khoudour, and L. Douadi. "Video Sequences Association for People Re-identification across Multiple Non-overlapping Cameras." *International Conference on Image Analysis and Processing (ICIAP'2009).* Springer, 2009. 179.

Csurka, G., C. Dance, L. Fan, J. Williamowski, and C. Bray. "Visual categorization with bags of keypoints." *ECCV Workshop on Statistical Learning in Computer Vision.* 2004.

Cyr, C., and B. Kimia. "A similarity-based aspect-graph approach to 3D object recognition." *International Journal of Computer Vision (IJCV)* 57 (2004): 5-22.

Dalal, N., and B. Triggs. "Histograms of oriented gradients for human detection." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2005).* 2005.

de la Escalera, A., J.M. Armingol, and M. Mata. "Traffic sign recognition and analysis for intelligent vehicles." *Image and vision computing* 21, no. 3 (2003): 247-258.

de la Escalera, A., L. Moreno, M.A. Salichs, and J.M. Armingol. "Road Traffic Sign Detection and Classification." *IEEE Transactions on Industrial Electronics* 44, no. 6 (1997): 848-859.

De Palma, A., and F. Marchal. "Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale macroscopic transportation systems." *Networks and Spatial Economics* 2, no. 4 (2002): 347–369.

Diego, I.M. de, A. Serrano, C. Conde, and E. Cabello. "Face verification with a kernel fusion method." *Pattern Recognition Letters*, 2010.

Dollar, P., Z. Tu, P. Perona, and S. Belongie. "Integral Channel Features." *British Machine Vision Conference (BMVC).* London (England), 2009. 1-11.

Dorai, C., and A.K. Jain. "Cosmos - a representation scheme for 3D free-form objects." *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)* 19, no. 10 (1997): 1115–1130.

Draper, Bruce A., Kyungim Baek, Marian Stewart Bartlett, and J. Ross Beveridge. "Recognizing faces with PCA and ICA." *Computer Vision and Image Understanding* 91 (2003): 115-137.

Escalera, S., and P. Radeva. "Fast Greyscale road sign model matching and recognition." *Frontiers in Artificial Intelligence and Applications / Recent Advances in Artificial intelligence Research and Development*, 2004.

Fang, C.Y., C.S. Fuh, S.W. Chen, and P.S. Yen. "A road sign recognition system based on dynamic visual model." *in Proc IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

Fang, C.Y., S.W. Chen, and C.S. Fuh. "Road-sign detection and tracking." *IEEE Trans. Vehicular Technology* 52, no. 5 (2003): 1329-1341.

Farenzena, M., L. Bazzani, A. Perina, V. Murino, and M. Cristani. "Person Re-Identification by Symmetry-Driven Accumulation of Local Features." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2010).* 2010. 2360-2367.

Finlayson, G., S. Hordley, G. Schaefer, and G. Yun Tian. "Illuminant and device invariant colour using histogram equalisation." *Pattern Recognition* (Elsevier) 38 (2005): 179-190.

Flint, A., A. Dick, and A. van den Hengel. "Thrift: Local 3D Structure Recognition." *Proc. Digital Image Computing: Techniques and Applications (DICTA'07).* 2007. 182-187.

Friedman, J. H., J. L. Bentley, and R. A. Finkel. "An Algorithm for Finding Best Matches in Logarithmic Expected Time." *ACM Trans. Math. Softw.* (ACM) 3 (1977): 209-226.

Frome, A., D. Huber, R. Kolluri, T. Bulow, and J. Malik. "Recognizing objects in range data using regional point descriptors." *European Conference on Computer Vision(ECCV'04)*, 2004: 224–237.

Furtlehner, C., J. Lasgouttes, and A. de La Fortelle. "A belief propagation approach to traffic prediction using probe vehicles." *Proc. 10th Int. Conf. Transportation Systems (ITSC'2007).* 2007. 1022–1027.

Furtlehner, C., Y. Han, J.-M. Lasgouttes, V. Martin, F. Marchal, and F. Moutarde. "Spatial and Temporal Analysis of Traffic States on Large Scale Networks." *Proc. 13th International IEEE Conference on Intelligent Transportation Systems (ITSC'2010).* 2010.

Garcia, M.A., M.A. Sotelo, and E. Martin-Gorostiza. "Fast Traffic Sign Detection and Recognition Under Changing Lighting Conditions." *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*, 2006.

Gavrila, D.M. "Traffic sign Recognition Revisited." *Proceding of the 21st DAGM Symposium fur Musterekennung*, 1999.

Geroliminis, N., and C. F. Daganzo. "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings." *Transportation Research Part B: Methodological* 42, no. 9 (2008): 759-770.

Geroliminis, N., and C. F. Daganzo. "Macroscopic Modeling of Traffic in Cities." *Transportation Research Board 86th Annual Meeting.* 2007.

Geroliminis, N., and J. Sun. *Transportation Research Part B: Methodological* 45, no. 3 (2011): 605-617.

Gheissari, Niloofar, Thomas B. Sebastian, and Richard Hartley. "Person Reidentification Using Spatiotemporal Appearance." *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* {IEEE} Computer Society, 2006. 1528-1535.

Ghosh, B., B. Basu, and M. O'Mahony. "Multivariate short-term traffic flow forecasting using time-series analysis." *IEEE Transaction on Intelligence Transportation Systems* 10, no. 2 (2009): 246 – 254.

Hamdoun, O., A. Bargeton, F. Moutarde, and B. Bradai. "Detection and Recognition of End-of-Speed-Limit and Supplementary Signs for Improved European Speed Limit Support." *proc. 15th World Congress on Intelligent Transport Systems (ITSwc'2008).* New-York (USA), 2008.

Hamdoun, O., and F. Moutarde. "Keypoints-based background model and foreground pedestrians extraction for future smart cameras." *Proc. of 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC'2009).* Como (Italy), 2009.

Hamdoun, O., F. Moutarde, B. Stanciulescu, and B. Steux. "Person Re-identification in Multi-camera System by Signature based on Interest Point Descriptors Collected on Short Video Sequences." *Proceedings of ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC'2008).* 2008.

Hamdoun, O., F. Moutarde, Stanciulescu B., and B. Steux. ""Interest points harvesting in video sequences for efficient person identification",." *Proc. of '8th international workshop on Visual Surveillance (VS2008)' of "10th European Conference on .* Marseille (France), 2008.

Han, Y., and F. Moutarde. "Analysis of Large-scale Traffic Dynamics using Non-negative Tensor Factorization." *Proc. 19th World Congress on Intelligent Transport Systems (ITSwc'2012).* Vienna (Austria), 2012.

Han, Y., and F. Moutarde. "Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization." *Proc. 14th IEEE Intelligent Transport Systems Conference (ITSC'2011).* Washington DC (USA), 2011.

Han, Y., and F. Moutarde. "Statistical Traffic State Analysis in Large-scale Transportation Networks Using Locality-Preserving Non-negative Matrix Factorization." *IET Intelligent Transport Systems journal* sous presse (2013).

Hebert, M., K. Ikeuchi, and H. Delingette. "A spherical representation for recognition of free-form surfaces." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17 (1995): 681-690.

Herring, R., et al. "Using mobile phones to forecast arterial traffic through statistical learning." *Proc. 89th Transportation Research Board Annual Meeting (TRB'2010).* Washington D.C. (USA), 2010.

Herty, M., A. Klar, and L. Pareschi. "General kinetic models for vehicular traffic flow and Monte Carlo methods." *Computational methods in applied mathematics* 5, no. 2 (2005): 155-169.

Hetzel, G., B. Leibe, P. Levi, and B. Schiele. "3D Object Recognition from Range Images using Local Feature Histograms." *Proceedings of Computer Vision and Pattern Recognition (CVPR'01).* 2001. 394-399.

Hilaga, M., Y. Shinagawa, T. Kohmura, and T.L. Kunii. "Topology matching for fully automatic similarity estimation of 3D shapes." *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001: 203-212.

Ho, H.T., and D. Gibbins. "A Curvature-based Approach for Multi-scale Feature Extraction from 3D Meshes and Unstructured Point Clouds." *IET Computer Vision* 3, no. 4 (2009): 201-212.

Hofleitner, A., A. Herring, A. Bayen, Y. Han, F. Moutarde, and A. de la Fortelle. "Large scale estimation of arterial traffic and structural analysis of traffic patterns using probe vehicles." *Proc. of Transportation Research Board 91st annual meeting (TRB'2012).* Washington DC (USA), 2012.

Hofleitner, A., R. Herring, and A. Bayen. "Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning." *Transportation Research Part B* 46, no. 9 (2011): 1097-1122.

Hofleitner, A., R. Herring, and A. Bayen. "Probability distributions of travel times on arterial networks: a traffic flow and horizontal queuing theory approach." *91st Transportation Research Board Annual Meeting.* 2012.

Horn, K. P. "Extended gaussian images." *Proceedings of the IEEE* 72, no. 12 (1984): 1671-1686.

Huang, J., S. Ravi Kumar, M. Mitra, W.J. Zhu, and R. Zabih. "Spatial Color Indexing and Applications." *International Journal of Computer Vision (IJCV)* (Springer) 35 (1999): 245-268.

Huttenlocher, D. P, G. A Klanderman, and W. A. Rucklidge. "Comparing images using the Hausdorff distance." *IEEE Transactions on pattern analysis and machine intelligence*, 1993: 850–863.

Jaffré, G., and P. Joly. "Costume: A new feature for automatic video content indexing." *Proc. RIAO.* 2004. 314-325.

Javed, O., Z. Rasheed, K. Shafique, and M. Shah. "Tracking across multiple cameras with disjoint views." *Proceedings 9th IEEE International Conference on Computer Vision (ICCV'2003).* , 2003: 952-957.

Javed, O.S., and K.M. Shah. "Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras." *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2005).* 2005.

Johnson, AE, and M. Hebert. "Using spin images for efficient object recognition in cluttered 3D scenes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21 (1999): 433-449.

Jones, M., and P. Viola. "Face recognition using boosted local features." *Proceedings of International Conference on Computer Vision (ICCV'03)*, 2003.

Kang, D.S., N.C. Griswold, and N. Kehtarnavaz. "An invariant traffic sign recognition system based on sequential color processing and geometrical transformation." *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, 1994.

Kass, M., A. Witkin, and G. Terzopoulos. "Snakes: Active contour." *International Conference on Computer Vision (ICCV).* 1987.

Kazhdan, M., and T. Funkhouser. "Harmonic 3d shape matching." *ACM SIGGRAPH 2002 Technical Sketch*, 2002: 191.

Kesorn, K., S. Chimlek, S. Poslad, and P. Piamsa-nga. "Visual content representation using semantically similar visual words." *Expert Systems with Applications* 38, no. 9 (2011): 11472-11481.

Lantagne, M., M. Parizeau, and R. Bergevin. "VIP: Vision tool for comparing Images of People." *16th IEEE Conf. on Vision Interface.* 2003. 35-42.

Lee, D.D., and H.S. Seung. "Algorithms for non-negative matrix factorization." *Proc. 13th Neural Information Processing Systems (NIPS).* Denver (USA), 2000. 556-562.

Leibe, B., A. Leonardis, and B. Schiele. "Robust object detection with interleaved categorization and segmentation." *Int. J. Comput. Vision* 77, no. 1-3 (2008): 259-289.

Li, X., and I. Guskov. "Multi-scale features for approximate alignment of point-based surfaces." *Proceedings of the third Eurographics symposium on Geometry processing*, 2005: 217.

Lin, C.J. "On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization." *IEEE Transactions on Neural Networks* 18, no. 6 (2007): 1589 - 1596.

Lipton, A.J., H. Fujiyoshi, and R.S. Patil. "Moving target classification and tracking from real-time video." 1998. 8-14.

Liu, C.H., and H. Fujisawa. "Classification and Learning Methods for Character Recognition: Advances and Remaining Problems." *Machine Learning in Document Analysis and Recognition*, 2008.

Liu, W., J. Lv, H. Gao, B. Duan, H. Yuan, and H. Zhao. "An Efficient Real-Time Speed Limit Signs Recognition Based on Rotation Invariant Feature." *proc. of IEEE Intelligent Vehicles Symposium (IV'2011).* Baden-Baden (Germany), 2011.

Lowe, D. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60 (2004): 91-110.

Lowe, D.G. "Local feature view clustering for 3D object recognition." *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR).* 2001. I-682 - I-688 vol.1.

Loy, G., and N. Barnes. "Fast shape-based road sign detection for a driver assistance system." *in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

Madden, C., E. Cheng, and M. Piccardi. "Tracking people across disjoint camera views by an illumination-tolerant appearance representation." *Machine Vision and Applications* 18 (2007): 233-247.

Marchal, F. "'Contribution to dynamic transportation models." PhD Thesis, University of Cergy-Pontoise, 2001.

Medici, P., C. Caraffi, E. Cardarelli, P.P. Porta, and G. Ghisio. "Real Time Road Signs Classification." *Proceedings of the 2008 IEEE International Conference on Vehicular Electronics and Safety (ICVES'2008).* Colombus (USA), 2008.

Medioni, G.G., and A.R.J. François. "3-D structures for generic object recognition." *Computer Vision and Image Analysis*, 2000: 30-37.

Mian, A., M. Bennamoun, and R. Owens. "Face recognition using 2D and 3D multimodal local features." *Advances in Visual Computing*, 2006: 860-870.

Mian, A., M. Bennamoun, and R. Owens. "On the Repeatability and Quality of Keypoints for Local Feature-based 3D Object Retrieval from Cluttered Scenes." *International Journal of Computer Vision (IJCV)*, 2009.

Mian, A.S., M. Bennamoun, and R. Owens. "Keypoint detection and local feature matching for textured 3D face recognition." *International Journal of Computer Vision (IJCV)* 79, no. 1 (2008): 1-12.

Min, X., J. Hu, Q. Chen, T. Zhang, and Y. Zhang. " Short-term traffic flow forecasting of urban network based on dynamic STARIMA model." *Proc.12th Int. Conf. Intelligent Transportation Systems (ITSC'2009).* 2009.

Miura, J., T. Kanda, and Y. Shirai. "An Active Vision System for Real Time Traffic Sign Recognition." *Proceeding 2000 IEEE Int. Conf. On Intelligent Transportation Systems*, 2000.

Moeslund, T. B., and E. Granum. "A Survey of Computer Vision-Based Human Motion Capture." *Computer Vision and Image Understanding* 81, no. 3 (2001): 231-268.

Moutarde, F., A. Bargeton, A. Herbin, and L. Chanussot. "Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system." *proc. of IEEE Intelligent Vehicles Symposium (IV'2007).* 2007.

Moutarde, F., and Y. Han. "A new traffic-mining approach for unveiling typical global evolutions of large-scale road networks." *Proc. 18th World Congress on Intelligent Transport Systems (ITSwc'2011).* Orlando (USA), 2011.

Moutarde, F., B. Stanciulescu, and A. Breheret. "Real-time visual detection of vehicles and pedestrians with new efficient adaBoost features." *Proc. of 'Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)' of "2008 International conference on Intelligent RObots ans Systems (IROS'2008)".* Nice (France), 2008.

Munder, S., and D.M. Gavrila. "An Experimental Study on Pedestrian Classification." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 28, no. 11 (2006): 1863-1868.

Nagel, K., and M. Schreckenberg. "A cellular automaton model for freeway traffic." *Journal of Physics* 2 (1992): 2221–2229.

Nakajima, C., M. Pontil, B. Heisele, and T. Poggio. "Full-body person recognition system." *Pattern Recognition* 36 (2003): 1997-2006.

Nienhüser, D., T. Gumpp, J.M. Zöllner, and K. Natroshvili. "Fast and Reliable Recognition of Supplementary Traffic Signs." *proc. IEEE Intelligent Vehicles Symposium (IV'2010).* San Diego (USA), 2010.

Ning, H., T. Tan, L. Wang, and W. Hu. "People tracking based on motion model and motion constraints with automatic initialization." *Pattern Recognition* 37 (2004): 1423-1440.

Osada, R., T. Funkhouser, B. Chazelle, and D. Dobkin. "Matching 3D Models with Shape Distributions." *Proceedings of the International Conference on Shape Modeling and Applications (ICSMA)*, 2001: 154.

Park, U., AK Jain, I. Kitahara, K. Kogure, and N. Hagita. "ViSE: Visual Search Engine Using Multiple Networked Cameras." *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06).* 2006. 1204-1207.

Perronnin, F., C. Dance, G. Csurka, and M. Bressan. "Adapted vocabularies for generic visual categorization." *proc. of European Conference on Computer Vision (ECCV).* 2006.

Pettersson, N., L. Petersson, and L. Andersson. "The histogram feature - a resource-efficient Weak Classifier." *IEEE Intelligent Vehicles Symposium (IV'2008).* Eindhoven (Netherlands), 2008.

Pham, T.V., M. Worring, and AWM Smeulders. "A multi-camera visual surveillance system for tracking of reoccurrences of people." *Proc. 1st ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC'07).* 2007. 164-169.

Piccioli, G., E. De Micheli, P. Parodi, and M. Campani. "Robust method for road sign detection and recognition." *Image and Vision Computing* 14, no. 3 (1996): 209-223.

Porikli, F. "Inter-camera color calibration by correlation model function." *Proceedings of IEEE International Conference on Image Processing (ICIP).* 2003. II - 133-6 vol.3.

Prosser, B., S. Gong, and T. Xiang. "Multi-camera Matching under Illumination Change Over Time." *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008.* 2008.

Puthon, A.-S., F. Moutarde, and F. Nashashibi. "Recognition of supplementary signs for correct interpretation of traffic signs." *Workshop on Environment Perception and Navigation for Intelligent Vehicles, Intelligent Vehicle symposium (IV'2013).* 2013.

Puthon, A.-S., F. Moutarde, and F. Nashashibi. "Subsign detection with region-growing from contrasted seeds." *Proc. Intelligent Transportation Systems Conference (ITSC'2012).* 2012.

Quek, Y., M. Pasquier, and B. Lim. "POP-TRAFFIC: A Novel Fuzzy Neural Approach to Road Traffic Analysis and Prediction." *IEEE Transactions on Intelligent Transportation Systems* 7, no. 2 (2006): 133–146.

Rakha, H. "Validation of Van Aerde's Simplified Steady-state Car-following and Traffic Stream Model." *Transportation Letters: The International Journal of Transportation Research* 1, no. 3 (2009): 227-244.

Salti, S., F. Tombari, and L. Di Stefano. "Unique Signatures of Histograms for Local Surface Description." *Proc. of European Conference on Computer Vision (ECCV'2010).* 2010.

Samir, C., M. Daoudi, and A. Srivastava. "Reconnaissance de Visages 3D Utilisant l'Analyse de Formes des Courbes Faciales." *10èmes Journées CORESA (COmpression et REprésentation des Signaux Audiovisuels)*, 2006: 9-10.

Schmid, C. "A Structured Probabilistic Model for Recognition." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).* 1999. 2485.

Schwartz, W.R., and L.S. Davis. "Learning Discriminative Appearance-Based Models Using Partial Least Squares." 2009.

Shaiek, A., and F. Moutarde. "3D keypoint detectors and descriptors for 3D objects recognition with TOF camera." *Proc. of IS&T/SPIE Electronic Imaging conference on 3D Image Processing (3DIP) and applications.* San Francisco (USA), 2011.

Shaiek, A., and F. Moutarde. "3D Keypoints Detection for Objects Recognition." *Proc. 16th Int. Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'2012).* Las Vegas (USA), 2012.

Shaiek, A., and F. Moutarde. "Fast 3D keypoints detector and descriptor for view-based 3D objects recognition." *Proc. International Workshop on Depth Image Analysis (WDIA'2012) of 21st International Conference on Pattern Recognition (ICPR'2012).* Tsukuba (Japan), 2012.

Sivic, J., and A. Zisserman. "Efficient Visual Search for Objects in Videos." *Proceedings of the IEEE* 96 (2008): 548-566.

Sivic, J., and A. Zisserman. "Video Google: A Text Retrieval Approach to Object Matching in Videos." *Proceedings 9th IEEE International Conference on Computer Vision (ICCV'2003).* 2003. 1470-1477.

Soetedjo, A., and K. Yamada. "Traffic sign classification using ring partitioned method." *IEICE Trans. Fundamentals*, 2005.

Song, W., L. Liu, and X. Yang. "Fast and Robust Ridge-Ravine Detection on Triangular Meshes by Smooth Parametric Functions." *Proc. Pacific Graphics*, 2005: 109-111.

Stanciulescu, B., A. Breheret, and F. Moutarde. "Introducing New AdaBoost Features for Real-Time Vehicle Detection." *Proceedings of COGIS'07 conference on COGnitive systems with Interactive Sensors.* Stanford (USA), 2007.

Stauffer, C., and E. Grimson. "Similarity templates for detection and recognition." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2001. 221.

Swadzba, A., and S. Wachsmuth. "Categorizing Perceptions of Indoor Rooms Using 3D Features." *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, 2008: 734-744.

Tombari, F., S. Salti, and L. Di Stefano. "A combined texture-shape descriptor for enhanced 3D feature matching." *IEEE International Conference on Image Processing (ICIP 2011).* 2011.

Torresen, J., J.W. Bakke, and L. Sekanina. "Efficient recognition of speed limit signs." *in Proceedings The 7th International IEEE Conference on Intelligent Transportation Systems*, 2004.

Tsalakanidou, F., D. Tzovaras, and MG Strintzis. "Use of depth and colour eigenfaces for face recognition." *Pattern Recognition Letters (PRL)* 24, no. 9-10 (2003): 1427-1435.

Tu, P., et al. "An intelligent video framework for homeland protection." *In Defense and Security Symposium.* 2007.

Vincent, L. "Morphological Grayscale Reconstruction in Image Analysis: applications and efficient algorithms." *Transactions on Image Processing* 2, no. 2 (2003): 176-201.

Viola, P., and M. Jones. "Rapid Object Detection using a Boosted Cascade of Simple Features." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2001).* Kauai, Hawaï (USA), 2001.

Vranic, D.V., and D. Saupe. "3D model retrieval." *Proc. Spring Conference on Computer Graphics and its Applications (SCCG2000).* 2000. 3-6.

Vranic, D.V., and D. Saupe. "3D shape descriptor based on 3D Fourier transform." *Proceedings of the EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services (ECMCS 2001).* 2001. 271-274.

Wang, Y., and M. Papageorgiou. "Real-time freeway traffic state estimation based on extended kalman filter: a general approach." *Transportation Research Part B* 39 (2005): 141–167.

Winn, J., A. Criminisi, and T. Minka. "Object categorization by learned universal visual dictionary." *proc. of Int. Conf. on Computer Vision (ICCV).* 2005.

Zaharia, T., and F. Prêteux. "Indexation de maillages 3D par descripteurs de forme." *Actes 13eme Congres Francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'02)*, 2002: 48-57.

Zhu, Q., M.-C. Yeh, K.-T. Kwang-Ting, and S. Avidan. "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients." *proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006).* 2006.

REFERENCES DES PUBLICATIONS ET TRAVAUX ENCADRES

I. Thèse

Thèse de doctorat de l'Université Paris VII : "***Dynamique gravitationnelle non-linéaire dans un Univers en expansion et formation des grandes structures : simulations numériques***", Fabien Moutarde (mai 1991).

II. Chapitres de livre

Geo-positionning and Mobility, chapter 3 « *Information, Modeling and Traffic Reconstruction* », Arnaud de La Fortelle, Jean-Marc Lasgoutte et Fabien Moutarde (sous presse, 2013).

III. Brevets

[B.1] "*Method of circle detection in images for round traffic sign identification and vehicle driving assistance device*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi and Benazouz Bradaï, Brevet PCT/EP2010/007569 déposé par VALEO le 11/12/2010.

[B.2] "*Procédé de détection d'un objet cible*", Herbin Anne, Chanussot Lowik et Moutarde Fabien, Brevet FR2917869 déposé par VALEO, éd. INPI. - France, 25 06 2007.

IV. Articles dans des revues scientifiques à comité de lecture

[R.1] "*SKIN, a new family of performant 3D keypoints for view-based 3D objects recognition* ***",*** Ayet Shaiek et Fabien Moutarde, soumis à un journal, en cours de relecture.

[R.2] "*Statistical Traffic State Analysis in Large-scale Transportation Networks Using Locality-Preserving Non-negative Matrix Factorization*", Yufei Han et Fabien Moutarde, IET Intelligent Transport Systems journal, Vol. 7, issue 3, pages 283-295 (2013).

[R.3] "*Scale invariance and self-similar evolution of dark matter halos*", F. Moutarde, J.-M. Alimi, F. Bouchet, et R. Pellat, The Astrophysical Journal vol. 441, pages 10-17 (1995).

[R.4] "*Precollapse scale invariance in gravitationnal instability*", F. Moutarde, J.-M. Alimi, F.Bouchet, R. Pellat, et A. Ramani, The Astrophysical Journal vol. 382, pages 377-381 (1991).

[R.5] "*Collisionless formation of filaments in an expanding Universe*", J.-M. Alimi, F. Bouchet, R. Pellat, J.-F. Sygnet, et F. Moutarde, The Astrophysical Journal, vol. 354, pages 3-12 (1990).

V. Articles dans des conférences internationales à comité de lecture

[CI.1] "*Fast 3D keypoints detector and descriptor for view-based 3D objects recognition*", Ayet Shaiek et Fabien Moutarde, proc. International Workshop on Depth Image Analysis (WDIA'2012) of 21st International Conference on Pattern Recognition (ICPR'2012), Tsukuba (Japan), 11 nov. 2012.

[CI.2] "*Analysis of Large-scale Traffic Dynamics using Non-negative Tensor Factorization*", Yufei Han and Fabien Moutarde, proc. 19th World Congress on Intelligent Transport Systems (ITSwc'2012), Vienna (Austria), 22-26 octobre 2012.

[CI.3] "*Subsign detection with region-growing from contrasted seeds*", Anne-Sophie Puthon, Fabien Moutarde et Fawzi Nashashibi, proc. 15th IEEE Intelligent Transportation Systems Conference (ITSC'2012), Anchorage (USA), 16-19 septembre 2012.

[CI.4] "*3D Keypoints Detection for Objects Recognition*", Ayet Shaiek et Fabien Moutarde, proc. 16th Int. Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'2012), Las Vegas (USA), 16-19 juillet 2012.

[CI.5] "*Large scale estimation of arterial traffic and structural analysis of traffic patterns using probe vehicles*", Aude Hofleitner, Ryan Herring, Alexandre Bayen, Yufei Han, Fabien Moutarde et Arnaud de la Fortelle, proc. of Transportation Research Board 91st Annual Meeting (TRB'2012), Washington DC (USA), 22-26 janvier 2012.

[CI.6] "*A new traffic-mining approach for unveiling typical global evolutions of large-scale road networks*", Fabien Moutarde et Yufei Han, proc. 18th World Congress on Intelligent Transport Systems, Orlando (USA), 16-20 octobre 2011.

[CI.7] "*Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization*", Yufei Han et Fabien Moutarde, proc. 14th IEEE Intelligent Transport Systems Conference (ITSC'2011), Washington DC (USA), 5-7 octobre 2011.

[CI.8] "*Clustering and Modeling of Network level Traffic States based on Locality Preservative Non-negative Matrix Factorization*", Yufei Han and Fabien Moutarde, proc. 8th Intelligent Transport Systems (ITS) European congress, Lyon (France), 6-9 juin 2011.

[CI.9] "*3D keypoint detectors and descriptors for 3D objects recognition with TOF camera*", Ayet Shaiek et Fabien Moutarde, proc. of IS&T/SPIE Electronic Imaging conference on 3D Image Processing (3DIP) and applications, San Francisco, USA, 26-27 janvier 2011.

[CI.10] "*Joint interpretation of on-board vision and static GPS cartography for determination of correct speed limit*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi et Anne-Sophie Puthon, proc. 17th ITS world congress (ITSwc'2010), Busan, Korea, 25-29 octobre 2010.

[CI.11] "*Spatial and Temporal Analysis of Traffic States on Large Scale Networks*", Cyril Furtlehner, Yufei Han, Jean-Marc Lasgouttes, Victorin Martin, Fabrice Marchal et Fabien Moutarde, proc. 13th International IEEE Conference on Intelligent Transportation Systems (ITSC'2010), Madeira Island, Portugal, 19-22 septembre 2010.

[CI.12] "*3D Object Recognition and Facial Identification Using Time-averaged Single-views from Time-of-flight 3D Depth-Camera*", Hui Ding, Fabien Moutarde, et Ayet Shaiek, proc. 3rd Eurographics Workshop on "3D Object Retrieval" (3DOR2010), Norrköping, Sweden, mai 2010.

[CI.13] "*adaBoost with 'keypoint presence features' for real-time vehicle visual detection*", Taoufik Bdiri, Fabien Moutarde, Nicolas Bourdis et Bruno Steux, proc. of 16th World Congress on Intelligent Transport Systems (ITSwc'2009), Stockholm, Sweden, septembre 2009.

[CI.14] "*Keypoints-based background model and foreground pedestrians extraction for future smart cameras*", Omar Hamdoun et Fabien Moutarde, proc. of 3rd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2009), Como, Italy, 30 août - 3 septembre 2009.

[CI.15] "*Visual object categorization with new keypoint-based adaBoost features*", Taoufik Bdiri, Fabien Moutarde, et Bruno Steux, proc. of IEEE Symposium on Intelligent Vehicles (IV'2009), held in XiAn, China, juin 2009.

[CI.16] "*A robot behavior-learning experiment using Particle Swarm Optimization for training a neural-based Animat*", Fabien Moutarde, proc. of 10th International Conference on Control, Automation, Robotics and Vision (ICARCV 2008), Hanoï, Vietnam, 17-20 décembre 2008.

[CI.17] "*Detection and Recognition of End-of-speed-limit and Supplementary Signs For Improved European Speed Limit Support*", Omar Hamdoun, Alexandre Bargeton, Fabien Moutarde, Benazouz Bradai et Lowik Chanussot, proc. of 15th World Congress on Intelligent Transport Systems (ITSwc'2008), New York City, USA, 16-20 novembre 2008.

[CI.18] "*Interest points harvesting in video sequences for efficient person identification*", Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu et Bruno Steux, proc. of '8th international workshop on Visual Surveillance (VS2008)' of "10th European Conference on Computer Vision (ECCV'2008)", Marseille, France, 17 octobre 2008.

[CI.19] "*Real-time visual detection of vehicles and pedestrians with new efficient adaBoost features*", F. Moutarde, B. Stanciulescu, et A. Breheret, proc. of 'Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)' of "2008 International Conference on Intelligent RObots and Systems (IROS'2008)", Nice, France, 26 septembre 2008.

[CI.20] "*Person Re-identification in Multi-camera System by Signature based on Interest Point Descriptors Collected on Short Video Sequences*" Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu et Bruno Steux, proceedings of ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC-08), Stanford University, California, USA, September 7-11, 2008.

[CI.21] "*Improving pan-European speed-limit signs recognition with a new 'global number segmentation' before digit recognition*", Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi et Benazouz Bradai, proc. of IEEE Intelligent Vehicles Symposium (IV'2008), Eindhoven, Netherlands, 4-6 juin 2008.

[CI.22] "*Introducing New AdaBoost Features for Real-Time Vehicle Detection*", B. Stanciulescu, A. Breheret et F. Moutarde, proceedings of COGIS'07 conference on COGnitive systems with Interactive Sensors, held in Stanford University, California, USA, November 26-27, 2007.

[CI.23] "*Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system*", Fabien Moutarde, Alexandre Bargeton, Anne Herbin et Lowik Chanussot, proc. of IEEE Intelligent Vehicles Symposium (IV'2007), Istanbul, 13-15 juin 2007.

[CI.24] "*Combining adaBoost with a Hill-Climbing evolutionary feature search for efficient training of performant visual object detectors*", Y. Abramson, F. Moutarde, B. Steux et B. Stanciulescu,

proceedings of FLINS2006 on Applied Computational Intelligence, Gênes, Italie, 29-31 août 2006 (pages 737-744).

[CI.25] "*U*F clustering: a new performant "cluster-mining" method based on segmentation of Self-Organizing Maps*", F. Moutarde & A. Ultsch, proceedings of WSOM'05, Paris, 5-8 sept. 2005 (pages 25-32).

[CI.26] "*Fast semi-automatic segmentation algorithm for Self-Organizing Maps*", D.Opolon & F.Moutarde, proc. of ESANN'2004 conference, Bruges, 28-30 avril 2004 (pages 507-512).

[CI.27] "*Intelligent Compression of Still Images*", F. Gilbert & F.Moutarde, proc. of 6th ACM International Multimedia Conference, Bristol, 12-16 septembre 1998.

[CI.28] "*Adaptive compression of still images: automating the choice of algorithm and parameters*", F. Gilbert & F. Moutarde, proceedings of "Electronic Image Capture and Publishing" symposium (Europto series), Zurich, 18-22 Mai 1998.

[CI.29] "*Automatic recognition of numerical modulation: comparison of various neural networks and other classifiers*", C. Louis, P. Séhier & F. Moutarde, proceedings of ICANN 95, Paris, 5-10 octobre 1995 (pages 161-170).

VI. Articles dans des conférences nationales à comité de lecture

[CN.1] "*Détecteurs de points d'intérêt 3D basés sur la courbure*", Ayet Shaiek et Fabien Moutarde, 15° colloque 'COmpression et REprésentation des Signaux Audiovisuels' (CORESA'2012), Lille (France), 24-25 mai 2012.

[CN.2] "*Vidéosurveillance intelligente : ré-identification de personnes par signature utilisant des descripteurs de points d'intérêt collectés sur des séquences*" Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu and Bruno Steux, proc. of Workshop on « Surveillance, Sûreté, Sécurité des Grands Systèmes » (3SGS'08), Troyes, 4-5 juin 2008.

[CN.3] "*Apprentissage de détecteurs visuels d'objets par dopage utilisant un algorithme hybride évolution-escalade*", Y. Abramson, F. Moutarde, B. Steux and B. Stanciulescu, proceedings of 8° conférence francophone sur l'apprentissage automatique (CAp'2006), Trégastel, France, 22-24 mai 2006.

[CN.4] "*Un algorithme de dimensionnement pour atteindre un reseau d'architecture minimal*", B. Gittler, C. Louis & F. Moutarde, proc. of "Neural Networks and their applications", Marseille, 15-16 décembre 1994.

[CN.5] "*Réseaux de neurones pour l'identification de procédés industriels*", F. Moutarde, P. Delesalle & M. Menahem, proc. of Neuro-Nimes 92, Nimes, 2-6 novembre 1992 (pages 443-454).

VII. Thèses soutenues

[T.1] Omar HAMDOUN (déc. 2010), "*Détection et ré-identification de piétons par points d'intérêt entre caméras disjointes*", Thèse de doctorat Ecole des Mines de Paris, 2010.

[T.2] Alexandre BARGETON (déc. 2009, co-encadrement avec Fawzi Nashashibi), "*Fusion multi sources pour l'interprétation de l'environnement routier*", Thèse de doctorat Ecole des Mines de Paris, 2009.

[T.3] Ayet SHAIEK (mars 2013), « *Reconnaissance d'objets 3D par points d'intérêt* », Thèse de doctorat Ecole des Mines de Paris, 2013.

[T.4] Anne-Sophie PUTHON (avril 2013, co-encadrement avec Fawzi Nashashibi), « *Détermination de la vitesse limite par fusion de données cartographiques et vision temps-réel embarquée* », Thèse de doctorat Ecole des Mines de Paris, 2013.

VIII. Thèses en cours

[T.5] Emilie WIRBEL (2° année thèse CIFRE avec Aldebaran Robotics), *« Localisation et navigation multi-modales d'un robot humanoïde en environnement domestique* », Thèse de doctorat Ecole des Mines de Paris.

[T.6] Tao-Jin YAO (1° année, thèse financée par VeDeCom), « *Apprentissage statistique et fouille de données pour la caractérisation des environnements et situations de conduite* », Thèse de doctorat Ecole des Mines de Paris.

Annexe : copie des principales publications

Improving pan-European speed-limit signs recognition with a new "global number segmentation" before digit recognition

Alexandre Bargeton, Fabien Moutarde, Fawzi Nashashibi, and Benazouz Bradai

***Abstract*— In this paper, we present an improved European speed-limit sign recognition system based on an original "global number segmentation" (inside detected circles) before digit segmentation and recognition. The global speed-limit sign detection and correct recognition rate, currently evaluated on videos recorded on a mix of French and German roads, is around 94 %, with a misclassification rate below 1%, and not a single validated false alarm in several hours of recorded videos. Our greyscale-based system is intrinsically insensitive to colour variability and quite robust to illumination variations, as shown by an on-road evaluation under bad weather conditions (cloudy and rainy) which yielded 84% good detection and recognition rate, and by a first night-time on-road evaluation with 75% correct detection rate. Due to recognition occurring at digit level, our system has the potential to be very easily extended to handle properly all variants of speed-limit signs from various European countries. Regarding computation load, videos with images of 640x480 pixels can be processed in real-time at ~20frames/s on a standard 2.13GHz dual-core laptop.**

I. Introduction

Car automation increases by progressive integration of more and more advanced driving assistance systems. For instance, most current GPS navigators now include a function to inform the driver of the supposed current speed-limit, a feature increasingly motivating drivers as automated speed-limit enforcement gets more common. Furthermore a desired evolution for Adaptive Cruise Control (ACC) would be the development of smarter ACC able to automatically tune target cruising speed depending on current speed-limit.

However speed-limit information extracted from GPS cartographic data is neither always complete nor systematically up-to-date. Moreover, temporary speed limits for road works (see example on figure 1), as well as variable speed limits, are by definition not included in pre-defined digital cartographic data. Therefore a *visual* real-time speed-limit sign detection and recognition system is a mandatory complement to GPS systems for designing high-level advanced driving assistance systems such as Speed Limit Support (SLS) and smart ACC.

Fig. 1. Example of roadwork temporary speed-limit sign, whose information can obviously not be present in GPS cartographic data, and therefore has to be visually detected and recognized for a Speed Limit Support system.

II. Related and previous work

A Traffic Sign Recognition (TSR) system usually involves two main steps: 1/ detection of potential traffic signs in the image, based on the common shape/color design of sought traffic signs; 2/ classification of the selected regions of interest (ROIs) for identifying the exact type of sign, or rejecting the ROI.

Most recently published TSR approaches make use of color information for the detection step (see e.g. [1], [2], [3], [4] or [10]), which makes it easier, but less robust. In contrast with that, our TSR system, whose initial version was already presented in [9] and [13], uses a shape-based detection working on grayscale images. As was already advocated by Gavrila in [5], by Barnes and Zelinsky in [6], and confirmed by García-Garrido et al. in [8], using a shape-based detection increases robustness for detection of signs with colors faded away by time, and makes it possible to work properly even with difficult illumination conditions, such as glare by background sun or light, in the dark, or even at night.

For the classification step, nearly all published works on speed-limit sign recognition use a "holistic" approach in which the whole sign (in fact most of the time a set of features extracted from it) is fed into a classifier. Various kinds of classifiers are used: Bayesian Maximum Likelihood

Manuscript received Januar 7, 2008.

Alexandre Bargeton, Fabien Moutarde and Fawzi Nashashibi are with the Robotics Laboratory (CAOR), Ecole des Mines de Paris (ParisTech), 60 Bd Saint-Michel, F-75272 Paris cedex06, FRANCE, Phone: (33) 1-40.51.92.92, Fax: (33) 1-43.26.10.51 (e-mail: Alexandre.Bargeton@ensmp.fr, Fabien.Moutarde@ensmp.fr, Fawzi.Nashashibi@ensmp.fr)

Benazouz Bradai is with Valeo Driving Assistance Domain, 34 rue St-André, ZI des Vignes, F-93012 Bobigny, FRANCE, Phone: (33) 1-49.42.60.95 (e-mail: benazouz.bradai@valeo.com).

approach after Linear Discriminant Analysis (LDA) [1], Uncorrelated Fisher Discriminant Analysis (UFDA) [12] or after Image matrix based Discriminant analysis (IFDA) [12], ART1 (Adaptive Resonance Theory) neural network in [2], normalized correlation-based pattern matching in [3] and something similar in [6], Radial Basis Function (RBF) network in [5], a fuzzy set approach applied to color and shape features in [7], backpropagation neural network applied to global normalized sign image in [8] and [10], and more recently new learning methods such as ensemble learning in [11].

Contrary to all those works, and as already described in more details in [9] and [13], our Speed Limit Support system prototype relies on a digit extraction and recognition scheme, as in Torresen et al. [4], but more general and robust because we use separate recognition of each of the 2 or 3 digits, and employ a digit segmentation that is "orientation insensitive" so as to be able to recognize properly slightly tilted signs.

The recognition of extracted digits can be made, as described in [14], by any of commonly used algorithms from pattern recognition domain (kNN, MLP, RBF, SVM,…), which all can reach more than 98.5% in character recognition accuracy. In our algorithm, digits are recognized by a multi-layer neural network (MLP) trained on digit examples extracted from real videos.

In our system, a temporal integration on several frames is also used to confirm (or infirm) the recognition of the sign into the video, and provide a confidence level.

Fig. 2. Example of E.U. speed-limit sign detection and recognition with our initial system. All detected circles are shown (in red or green), and the candidate digits segmented inside them are outlined in red. The speed signs recognized on the current image are shown on the top black zone, with their associated confidence, and the currently validated speed limit is superimposed on top-center of image.

In our first version, the digit segmentation consisted in connected components extraction inside binarized detected circles. This technique proved its robustness and computation efficiency. However, it suffers in some cases of non-segmentation when two digits "touch" each other on the binarized image (see examples on fig. 3, 4, 6 or 7), thus resulting in non-recognition of the speed limit sign. These cases, which mostly happen for small and/or distant signs, 3-digits signs whose digits are not well-separated, or tagged signs, are not so common but significantly impaired performance, notably on German signs whose 3-digits signs are often small. We therefore decided to design a complementary technique to solve this problem, which is presented in this paper.

We also present here a new systematic evaluation of our improved system, on a mix of French and German roads, which are very encouraging for a future pan-European Speed Limit Support system, and results of recent on-road evaluations under bad weather conditions, and at night.

III. IMPROVED DIGIT CHARACTER SEGMENTATION

The Optical Character Recognition (OCR) domain provides three mains techniques in order to segment characters [15]: (1) via the image histogram, (2) via over-segmentation, (3) and via word by word segmentation. The first one fails with overlapped characters and depends on the precision of circles detection and image quality (see figure 3).

Fig. 3. Histograms of binarized speed limit signs

The second one is not applicable for real-time detection and can imply a lot of false recognition. We therefore turned to some hybrid of the third kind of method and our initial technique: we first try to globally segment the "word" inside the sign, which is in fact a 2- or 3-digits number in this case. The new proposed algorithm to segment digits inside a speed limit sign thus consists in two steps:

- Find the number / word into the circular sign
- Segment digits into the obtained rectangular zone

A. *"Global number segmentation" into the sign*

We further divide the number segmentation in two successive steps:

- finding upper/lower limits around the number
- determining left/right limits of the number

The two algorithms are detailed below.

A.1. Search upper/lower limits: The general idea is some kind of pixel-by-pixel "guided propagation". For each sub-quarter of the image, starting from each pixel on a

small central horizontal segment (Fig 4a – blue segment), browse the image pixel by pixel choosing one of the 3 neighbours pixels (Fig 4c) firstly attracted by the vertical center and secondly by the top (resp. bottom) of the image (Fig 4a). The vertical limits are the highest (resp. lowest) end point (Fig 4b). To prevent the propagation on the surrounding black circle, a rule is added to block the way exceeding a vertical segment / limit (i.e. stop the propagation to the outside if it's to too far from the vertical center). A last rule is that if the height of a propagation from a starting pixel is too close to the height of the sign (ie propagated on the outer circle) or too small (ie no propagation), this propagation is considered as failed, and not included in the computation of the upper / lower limit.

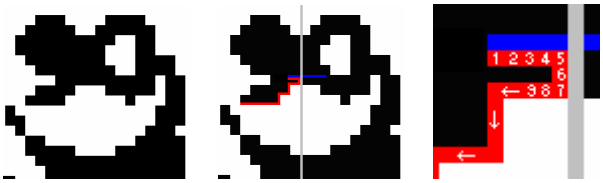


Fig. 4a. Searching vertical limits

(left) Initial binarized image – (center) a bottom left subquarter propagation way in red starting from the blue centered segment – (right) a zoomed part of this way from start with indexed pixels and direction of the propagation

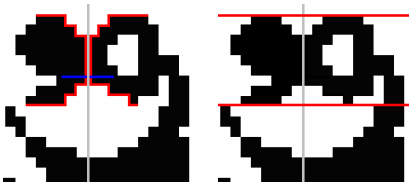


Fig. 4b. Searching vertical limits

(left) The propagation in red for each sub-quarter – (right) the result is the highest way for each half part (top – bottom) of the image.

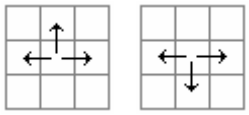


Fig. 4c. Searching vertical limits

The three neighbors used for top propagation (left) and bottom one (right)

A.2. Search left/right limits: Browse column by column from the vertical center to left and right, inside the previously found region. The main idea is that the first column (on each side) from which a black pixel is vertically outside the region or which is a white column (no more digit) is the searched vertical limit (see figure 5). In order to prevent premature stop between two separated digits, we temporally save the first horizontal index and continue our search by browsing columns with the same stop criteria. When there is no more way to stop, we use the last saved index as the limit.

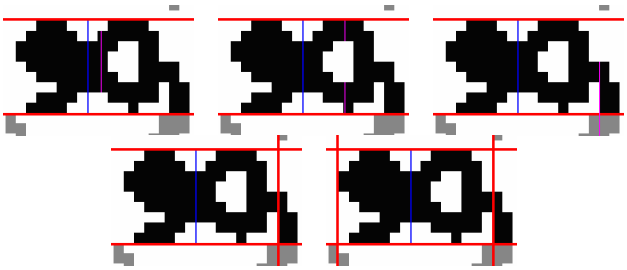


Fig. 5. Searching horizontal limits: exploration column by column (purple segment) for each half part (left – right) and stop on a white column (left) or exceed black (right).

This algorithm has been tested and validated under various conditions: day and night (see figures 6 and 7), with backlight or front light illumination resulting with some poor quality binarized images (see figure 6).

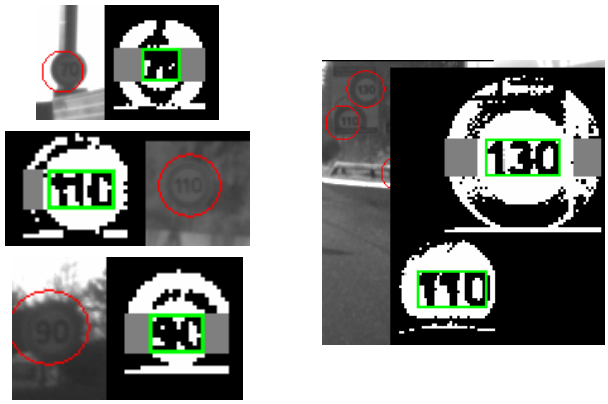


Fig. 6. Global number segmentation

B. Digit segmentation inside the segmented number

We use the same technique as in our initial method, but instead of binarizing the speed limit sign image, we binarize only the rectangular region found by our "number segmentation" algorithm. The main idea is that the binarization of the sub-image is much easier and provides a greater result than in the circle / speed image (see figure 7). In this new image we can easily extract connected components.

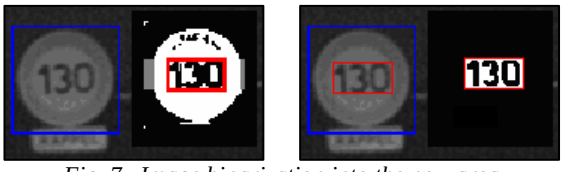


Fig. 7. Image binarization into the new area

There are still some digit-overlap cases, but the practice showed that overlap is done by only one or two pixels between digits in this new type of image. So in parallel we

provide a histogram segmentation (see figure 3) which now can work very well.

IV. Experiments and Results

A traffic sign detection and recognition system can be evaluated with various criteria. What is important in our application context is that all sought signs are validated *at least once* during the video segment between its first appearance and its final disappearance. We therefore evaluate only the *global* system performance by comparing type/time/position of all *validated* signs issued by our system to a ground-truth indicating a space-time visibility interval and type for each potentially detectable speed sign. The main resulting measure is the percentage of speed signs correctly detected and validated within their space-time visibility interval. This comparison and measure is of course done on video recordings independent from those used for extracting digits for training.

A. Recognition improvement on "difficult cases"

A first evaluation of the improvement brought by our new "number segmentation" approach has been made by considering only a set of speed-limit signs which were NOT correctly recognized with our initial digit-segmentation-based method. The percentage of those "difficult signs" that are now properly identified with our algorithm including "number segmentation" turned out to be 11/18 = 61%. This means our approach does not solve *all* recognition problems, but nevertheless provides a solution for a quite significant proportion of them, such as the tagged sign on figure 8.

B. Recognition improvement global evaluation

A more thorough evaluation has been done for the E.U. system, using recordings *on a mix of French and German roads and streets*, under various daytime illumination conditions, and containing ~140 speed-limit signs covering 11 different limit values (30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130). The corresponding videos have been manually annotated to indicate on each frame the position and type of speed-limit sign, so that an automated comparative evaluation is possible. The outcome is that the global system correct detection rate (SCDR) is 94 % with our improved "global number segmentation" approach, compared to 85% with our initial "digit segmentation" algorithm.

As can be seen on table 1, our new "number segmentation" quite significantly improves the correct detection rate. Also, the misclassification rate (signs for which a wrong speed value has been validated) remains below 1%. Nearly all of the 6% remaining non-correct-detections in our current improved system are just *missed* signs (most of the time because of not contrasted enough edges of the sign preventing a correct circle detection). Most importantly, not a single *validated* false alarm has been noticed in several hours of daytime recording: all spurious signs are efficiently filtered by our tracking and validation module.

Sign recognition method	Signs detected, recognized and validated with correct type	Misclassified signs
Initial digit-segmentation	**85 %**	**0,7%**
New "global number segmentation" before digit segmentation	**94 %**	**0,7%**

Table 1. Global evaluation of European speed limit sign detection on French+German roads.

Fig. 8. Example of a tagged speed-limit sign correctly recognized thanks to our new global number segmentation

C. On-road evaluation under bad weather conditions

As a complement to our in-lab evaluation done on pre-recorded videos, we hereafter present results of a recent on-road test conducted by Valeo with one of their experimental cars. The on-road evaluation (Table 2) is done *under bad weather condition (cloudy and rainy)* on 50km long run in France. Out of the 13 missed signs, 3 signs are partially occluded by others vehicles or dirty and thus nearly impossible to detect/recognize, 4 signs are with very low contrast inside tunnels, and 3 signs were correctly recognized but not validated as confidence did not pass validation threshold.

D. Preliminary night on-road evaluation

Because our algorithm is greyscale-base, and intrinsicly rather insensitive to luminosity variations, it shows very

promising results for night-time operation. This was recently quantified by a first on-road night evaluation done by Valeo: during this one hour live test, 78% of 60 speed limit signs encountered have been detected and correctly recognized.

Speed limit	**Good validated detection and recognition**	**Correctly recognized but not validated**	**Missed signs**
30	1		1
50	10		3
70	21		3
80	6		
90	28	3	3
110	4		
Total	70 **84 %**	3 **4 %**	10 **12 %**

Table 2. On-road evaluation in France, under bad weather conditions (cloudy and rainy).

V. CONCLUSIONS, DISCUSSION AND PERSPECTIVES

We have presented an improved robust and efficient visual speed-limit signs detection and recognition with ~94 % global correct sign detection rate on French+German roads. According to our evaluation, our new "global number segmentation" approach brings a quite significant 9% increase in detection rate, compared to our initial already performant system. The system requires only greyscale videos, and is able to process in real-time a video flow of images with 640x480 pixels at ~20 frames/s on a standard 2.13GHz dual-core laptop. It has a remarkably low false alarm rate (less than 1 spurious sign in several hours of operation).

The quantitative E.U. system evaluation reported in the present paper was restricted to videos recorded on French and German streets and roads. Evaluations in other E.U countries (in which sign digits are sometimes slightly different), are currently in progress, with very promising results (see example of successful detection of an Italian speed-limit sign on figure 9), as long as the ODR neural network is trained on a database with examples of all required variants of digits.

Our use of digit extraction and recognition instead of global sign recognition clearly facilitates the handling of aspect variability of the same sign across different E.U. countries, and even inside a single country such as France for instance. Even though we presently train a specific neural network for speed sign digits recognition, a final really pan-European system (i.e. properly recognizing all variants of speed-limit signs in all European countries) could probably work with a "universal" digit recognition module, therefore not requiring the previous collect of all European variants of each speed-limit sign digits.

Fig. 9. Correct recognition of a Italian speed-limit sign illustrating promising results for pan-European speed-limit recognition.

Fig. 10. Successful LED speed-limit sign recognition obtained with our system without specific re-training, by just inverting pixel scale inside region of detected circles.

Another illustration of the advantage of our speed-sign recognition being based on grayscale and digit recognition is the ease of adaptation to operate properly on LED signs, which are more and more common, especially for variable speed limits, for which the visual recognition is a mandatory complement to GPS. According to a first quick experiment, by simply inversing the image pixel scale inside the circles of potential signs, and then applying our algorithm, we can obtain, without any specific re-training, correct recognition of LED signs, as illustrated on figure 10.

Other work done in parallel and presented elsewhere includes recognition of end-of-speed-limit signs, and of some supplementary sign placed under signs, those two features

being essential for a fully pertinent Speed Limit Support system.

Finally, the final system really makes sense when integrated with GPS information, which can provide "baseline" information for the unavoidable cases of signs occulted by other vehicles. We therefore have begun to develop a framework for fusion of the output of visually-detected speed limits with GPS cartographic speed-limit data, as presented in [16]. Preliminary experiments show quite promising results for a final system that could take into account those two complementary speed-limit information sources.

REFERENCES

[1] Bahlmann C., Zhu Y., Ramesh V., Pellkofer M. and Koehler T., "A System for Traffic Sign Detection, Tracking, and Recognition Using Color, Shape, and Motion Information", IEEE Intelligent Vehicles Symposium (IV 2005), Las Vegas, June 2005.

[2] de la Escalera A., Armingol J.M. and Mata M., "Traffic sign recognition and analysis for intelligent vehicles", Image and Vision Computing, 21:247-258, 2003.

[3] Miura J., Kanda T. and Shirai Y., "An active vision for real-time traffic sign recognition", Proc. IEEE Conf. on Intelligent Transportation Systems, pages 52-57, Dearbom, MI, 2000.

[4] Torresen J., Bakke J.W., and Sekania L., "Efficient recognition of speed limit signs", Proc. IEEE Conf. on Intelligent Transportation Systems (ITS), Washington DC, 2004.

[5] Gavrila D.M., "Traffic sign recognition revisited", Proc of 21st DAGM symposium fur Musterekennung, pp. 86-93, Springer-Verlag, 1999

[6] Barnes N. and Zelinsky A., "Real-time radial symmetry for speed sign detection", Proc. IEEE Intelligent Vehicle Symposium, pages 566-571, Parma, Italy, 2004.

[7] Fang C.-Y., Chen S.-W and Fuh C.-S., "Road-sign detection and tracking", IEEE Trans. On Vehicular Technology, Vol.52, N°5, September 2003.

[8] García-Garrido M. A., Sotelo M. A., and Martín-Gorostiza E., "Fast traffic sign detection and recognition under changing lighting conditions", Proc. IEEE Intelligent Transportation Systems conference, pages 811-816, Toronto, Canada, 2006.

[9] Moutarde F., Bargeton B., Herbin A. and Chanussot C., "Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system", proc. of IEEE Intelligent Vehicles Symposium, Istanbul, 13-15 juin 2007.

[10] Broggi A., Cerri P., Medici P., Porta P. and Ghisio G., "Real Time Road Signs Recognition", IEEE-IV2007 , Istanbul, Turkey, 13-15 June, 2007

[11] Kouzani A., "Road-Sign Identification Using Ensemble Learning", IEEE-IV2007 , Istanbul, Turkey, 13-15 June, 2007

[12] Zhang C., Yang Q., Zhang W and Wang M., "Traffic sign recognition using Discriminant analysis based Algorithms", proc. of 14th World congress on Intelligent Transportation Systems (ITS), Beijing, China, 9-13 October, 2007

[13] Moutarde F., Bargeton B., Herbin A. and Chanussot C., "Modular Traffic Sign Recognition applied to on-vehicle real-time visual detection of American and European speed limit signs", proc. of 14th World congress on Intelligent Transportation Systems (ITS), Beijing, 9-13 octobre 2007.

[14] Liu C.-L., Fujisawa H., "Classification and Learning for Character Recognition: Comparison of Methods and Remaining Problems in Neural Networks and Learning in Document Analysis and Recognition", First IAPR TC3 NNLDAR Workshop, p1-7, Seoul, Korea, August 29, 2005.

[15] Casey R. G., Lecolinet E., "A Survey of Methods and Strategies in Character Segmentation", IEEE Transaction on Pattern Analysis and Machine Intelligence, Volume 18, Issue 7, P690-706, July 1996.

[16] Lauffenburger J. Ph., Bradai B., Basset M. and Nashashibi F., "Navigation and Speed Signs Recognition Fusion for Enhanced Vehicle Location", Proc. 17th IFAC (International Federation of Automatic Control) World Congress, Seoul, Corea, 6-11 July, 2008.

Author manuscript, published in "ITSC (2012)"

Subsign detection with region-growing from contrasted seeds

Anne-Sophie Puthon, Fabien Moutarde and Fawzi Nashashibi

***Abstract*— Speed limit determination systems for cars based on vision are more and more developed. Roadsign detection is nowadays a well managed problem. However, in some situations this information is not sufficient to know the speed limitation. Restrictions are sometimes applicable and specified by subsigns. These small rectangles often provide essential information about the applicability scope (vehicle type, condition, lane, etc.) of speed limits. We present an approach of subsign localization based on region growing with an initial step of seed selection using morphological reconstruction. A comparison is also performed with three other techniques based on edge, color and graph on two databases gathering French and German subsigns. The obtained subsign correct detection is above 65%.**

Fig. 1. Example of German (top) and French (bottom) subsigns.

I. Introduction

With the increasing need of making car a safe mean of transport comes the development of more and more sophisticated ADAS (Advanced Driving Assistance Systems). To help the driver having the most complete overview of its surrounding environment, manufacturers equip their cars with lots of sensors and dedicated systems, like pedestrian detection or lane keeping assistant. Concerning the speed management, TSR (Traffic Sign Recognition) systems aim at detecting and recognizing speed limit signs located on the side (or above) of the road by using an embedded camera. However, in some situations, these limitations are restricted to a particular category of vehicle, circumstance (weather condition, date or time, etc.) or lane (in highway exits). This additional information is specified by supplementary rectangular signs located under the corresponding speed limit. The main challenges of the task are the variability of the signs in size and ratio as well as in the type of information they provide (arrow, text, pictogram, etc.). Figure 1 illustrates some various subsigns encountered on road. Moreover, intelligent vehicle systems require high performance and must work in unsupervised environment with a moving camera.

Unfortunately, nowadays only very few dedicated methods have been implemented for the localization of these subsigns ([1], [2] and [3]). By extension, subsigns can be seen as rectangular homogeneous regions with contrasted symbols. Thus, related subjects are detection of U.S. speed limit signs, license plate or direction indicators localization. They are all designed for being seen by drivers at a relative great distance, *i.e.* on clean background with a readable font. Among the different approaches a distinction can be made between edge, color and spatial based methods. Edge-based techniques rely on the border detection and the rectangular shape of the objects. Jung *et al.* [4] use the Hough transform to retrieve the rectangles in the image with their parameters (size, ratio, orientation). Keller *et al.* [5] and Hamdoun *et al.* [1] search for horizontal and vertical lines through a voting scheme based on either a gradient map or a Canny filtered image. Color-based methods perform histogram thresholding [6] based on the assumption that the searched object all contain a significant proportion of uniformly colored background (usually a "dirty" yellowish white or grey depending on illumination and contrast). The last approach of Wu *et al.* [7] consists in detecting and clustering keypoints in the image. Then groups verifying a vertical plane hypothesis are kept and assume to belong to roadsigns.

Considering that without contrasted foreground no subsign can be detected, we combine a hole search with a region growing approach. The former consists in selecting pixels with a high contrast with the surrounding region while the latter allow us to delineate homogeneous regions. The developed approach is therefore not limited to the subsign detection but is generalizable to the detection of different types of informative regions like the U.S. speed signs. To validate our method we compared it to several techniques using different types of characteristics. The paper is organized as follows. Section II describes our region growing approach. The three comparative algorithms are introduced in section III as well as the evaluation protocol and the results. Finally, section IV presents the conclusion and future works.

Manuscript received July 12th, 2012.

A.-S.P. and F.M. are from the Robotics Centre of Mines ParisTech, 60 boulevard Saint-Michel F-75272 Paris Cedex 06, FRANCE (phone: (33)1-40-51-94-54, email: {anne-sophie.puthon, fabien.moutarde}@ensmp.fr)

F.N. is both from the Robotics Centre of Mines ParisTech and from INRIA, IMARA Team, BP 105 F-78153 Le Chesnay Cedex, FRANCE (phone: (33)1-39-63-52-56, email: fawzi.nashashibi@inria.fr)

hal-00741617, version 1 - 14 Oct 2012

hal-00741617, version 1 - 14 Oct 2012

II. Region growing from contrasted seeds

A. Region growing

Region-based approaches aim at segmenting the image into connected and homogeneous groups of pixels. Introduced by Zucker [8], rgion growing consists in selecting a set of "seeds", or initial regions R_0, to which neighbouring pixels are agglomerated if they fulfill homogeneity predicates. It is a bottom-up approach contrary to the splitting technique which starts from a set of regions which are iteratively split if the criteria are not verified. The principle is the following:

1) **Selection of the seeds**
   The first critical issue is the choice of the initial sets of pixels from which regions will grow. They actually need to belong to the objects to segment and must be a representative part of them. In section II-B we introduce our specific method for extracting seeds belonging to the subsign. It is based on a hole search.

2) **Selection of candidates for the aggregation**
   At each iteration a set of pixels is added to a region R until the equilibrium is reached. To be eligible, a pixel p must be adjacent to the current region R (in the sense of the 4- or 8-connexity) and verify a *local* homogeneity criterion κ_L such that:

$$\exists q \in N_v(p) \cap R, \mid \frac{I(p)}{I(q)} - 1 \mid \leq \kappa_L \qquad (1)$$

   with $N_v(p)$ the set of neighboring pixels of p and $I(p)$ the grey value of pixel p in the image I.

3) **Agglomeration**
   A candidate pixel p is finally merged to the region R if the *global* homogeneity predicate κ_G is verified:

$$\mid \frac{I(p)}{\mu_{R_0}} - 1 \mid \leq \kappa_G \qquad (2)$$

   where μ_{R_0} is the mean value of the initial region R_0. This criteria ensures that the global variance of the region is limited and less than $2\mu_{R_0}$.

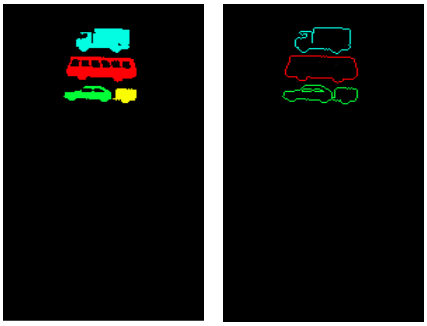
The values experimentally chosen for our implementation as giving the best results are $N_v = 4$, $\kappa_L = 0.1$ and $\kappa_R = 0.1$.

B. Selection of seeds

To draw the drivers' attention, road information are assumed to be highly contrasted, generally black symbols on white background. To segment these local extrema, we use the morphological reconstruction of an image I with marker M of Vincent [9]. Basically, it helps to extract the peaks of the image marked by M. This operator can be expressed as:

$$\rho_I(M) = \bigvee_{n \geq 1} \delta_I^{(n)}(M) \qquad (3)$$

(a) Original image. (b) Reconstruction result. (c) Resulting holes.



(d) Connected components selected. (e) Sets of seeds used for the region growing.

Fig. 2. Illustration of our method of seed selection. The reconstruction 2(a) has filled all the holes of the original image. By subtracting 2(b) from 2(a) we recover the contrasted pixels 2(c). Then we proceed to a connected component extraction 2(d) of the thresholded image applying 5. Finally, the seeds used by the region growing correspond to the pixels located at a distance $d = 1$ of these components 2(e).

where $\delta_I^{(n)}(M)$ is the geodesic dilatation of M of size n and $\bigvee_n X(n)$ means the maximum value taken by $X(n)$ for all n.

The seed selection procedure is depicted in figure 2 and consists in:

1) **Filling the holes**
   As we are searching for the holes in the image, we perform the morphological reconstruction of the complemented image $\tilde{I} = 255 - I$ of the source image I. The marker is designed to extract only the connected components which do not touch the image borders.

2) **Extracting the holes**
   By subtracting the resulting image $\rho_{\tilde{I}}(M)$ from the source I, we get the only pixels belonging to the dark connected components surrounded by light ones.

$$I_H = I - \rho_{\tilde{I}}(M) \qquad (4)$$

3) **Filtering**
   In order to filter out the slightly contrasted pixels, we threshold the image depending on the mean μ and the standard deviation σ of I_H. The remaining pixels are then grouped into connected components. Only the

hal-00741617, version 1 - 14 Oct 2012

ones containing at least 3 highly contrasted pixels are kept. This ensures to eliminate some noisy elements. Thus, the remaining components CC must verify:

$$\begin{cases} \forall p \in C, I_H(p) \geq \mu + \sigma \\ Card(p \in C | I_H(p) \geq \mu + 3\sigma) \geq 3 \end{cases} \quad (5)$$

4) **Generating the seeds**
In our approach, the regions to grow are the light homogeneous and rectangular subsigns surrounding these dark pixels. We therefore consider all the pixels located at a distance $dist = 1$ from the previous black connected components. Each connected set of these pixels represents an initial region for subsigns. By choosing directly a set as seed rather than a single pixel, the process is speeded up and made robuster as more information about the region is available.

III. Evaluation

A. A comparative study

As mentioned earlier only few techniques were implemented for the subsign detection, mainly using edges. To evaluate our approach, we thus implemented an existing one. Moreover we developed two other methods based on image segmentation and relying on different features such as color distribution and spatial relationships. The objectives was to compare our technique of region growing to diverse alternatives.

- **Edge-based**
This approach aims at detecting the subsign borders by the use of a Canny filter [1]. It works similarly to a template-matching. Horizontal and vertical lines are first searched for in the image. Then they are grouped by pairs and finally in rectangles if they meet certain conditions.

- **Color-based**
We implemented a technique based on the histogram distribution from Zhang *et al.* [10]. This choice is motivated by the assumption that a subsign and its background must have different color mode to be detected. The objective is to fit the histogram to a mixture of Gaussians without knowing *a priori* the number of modes. It consists in iteratively splitting and merging the Gaussians previously evaluated as the worst fitting the current estimation.

- **Graph-based**
Felzenszwalb *et al.* [11] proposed a segmentation lying on a compromise between the internal homogeneity of regions Int and the dissimilarity Dif between them. Two regions R_1 and R_2 are separable if

$$Dif(R_1, R_2) > min(Int(R_1) + \tau(R_1), Int(R_2) + \tau(R_2)$$

where $\tau(R) = \kappa / \mid R \mid$ is the threshold function.
We modify the threshold to take into consideration the rectangular shape of the region and its global homogeneity.

$$\tau(R) = \frac{\kappa}{\mid R \mid}(\beta_{size} + \beta_{rect}.f_{rect}(R) + \beta_{hom}.f_{hom}(R)) \quad (6)$$

$f_{rect} \to 0$ when the shape is close to a rectangle.
$f_{hom}(R) \to 0$ when $var(R) \to 0$

For every approach, a further step of region filtering is performed. It aims at keeping only those centered in the image, with a ratio, width and height within a given interval.

B. Evaluation procedure

- **Databases**
To our knowledge, no database is currently available for the specific task of subsign detection/recognition. Hence, we compared the four techniques on our own databases. We dispose of subsigns of two countries, France and Germany. This allows us to perform a first evaluation in an European context. Table I gives an overview of the two databases of the study. We consider the total number of subsigns in the sense of a frame per frame evaluation.

- **Comparison criteria**
The rectangles output by the different approaches are compared with several criteria. Firstly, in many TSR applications segmented regions are evaluated with the Jaccard's measure J (see figure 3).

$$J = \frac{GT \cap ALGO}{GT \cup ALGO} = \frac{I}{U} \quad (7)$$

A high value of J shows a good overlapping surface I between the ground truth (GT) and the algorithm result (ALGO) with a low disjoint area. However, the use of this measure only can be misleading. Figure 4 shows three different configurations of Ground Truth and Algorithm giving the same value of J. Using this single measure to validate or not a rectangle can result in a loss of information, as the detection is the early stage of the whole system.

To compare our detection techniques, we secondly introduce two other criteria, the overlap O and the centering

Database	Country	Subsigns (fr)
dbF	France	1040
dbG	Germany	12546

TABLE I
Database statistics.

hal-00741617, version 1 - 14 Oct 2012

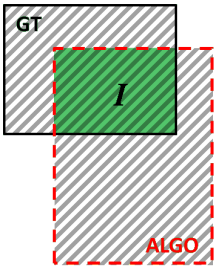


Fig. 3. Illustration of the Jaccard's measure J. The intersection $I = GT \cap ALGO$ corresponds to the overlap between the Ground Truth (GT) and the region output by the algorithm (ALGO). $U = GT \cup ALGO$ (in stripes) is the union of both regions.

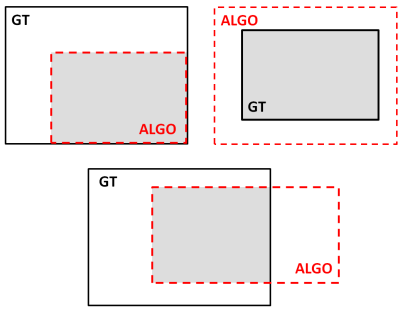


Fig. 4. Examples of three configurations of GT and ALGO resulting in $J = 0.5$.

C.

$$O = \frac{I}{Area(GT)} \quad (8)$$

$$D = \frac{ALGO \setminus GT}{Area(GT)} \quad (9)$$

$$C = \frac{dist(center_{GT}, center_{ALGO})}{d_{GT}/2} \quad (10)$$

The overlap has an interest in the context of a further classification allowing to control the amount of overlapping area the recognition needs. The last criteria C corresponds to the ratio of the distance between the region centres and the half of the diagonal d_{GT} of GT.

- **Results**

The comparison is performed for the criteria values:

$$J \geq 0.5 \quad (11)$$

$$O \geq 0.5 \quad (12)$$

$$O \geq 0.5 \,\&\, D \leq 1.5 \quad (13)$$

$$C \leq 0.2 \quad (14)$$

Results are shown in table II. The Jaccard's measure defined in equation 11 corresponds to the value selected in most of the paper about subsign detection. An overlap of 0.5 ensures that at least the half of the ground truth is covered by the algorithm. Combined to a disjoint measure less than 1.5, output regions are limited and avoid a further classification step to deal with noisy regions.

Generally speaking, the color-based approach appears to give the worst results. An explanation can be that our grayscale source images do not give enough color information to be able to efficiently detect the subsigns. Moreover, the process of splitting and merging until equilibrium is quite long and fastidious making this technique not very applicable as is. The two best techniques are the edge-based and ours regarding J or the combination of O and D. Regarding the centering measure, the best technique is the region growing as expected. The contrasted pixels used to generate the initial seeds are indeed mainly located in the middle of subsigns. Figure 7 shows the feature image and the resulting rectangles for each method. Finally, some results obtained with our region-growing technique are given in figures 5 and 6 for French and German subsigns. Actually the main sources of errors come from:

- a lack of contrast, appearing for instance in rainy conditions or when blur moving occurs;
- the absence of seeds in the region growing process because of an insufficient contrast of the symbols.

IV. Conclusion and Future Works

We presented a new technique of subsign detection based on region growing with a search of contrasted pixels as initial seeds. This idea was motivated by the assumption that roadsign in general are specifically designed to be seen by drivers at great distances. We thus implemented an approach based on a morphological reconstruction to get highly contrasted holes in the image. A comparison was then performed with three other image processing-based approaches in order to validate our method. For this evaluation, we proposed three different criteria adapted to the detection task. The final results highlight two algorithms, ours and the modified graph-based, with correct subsign

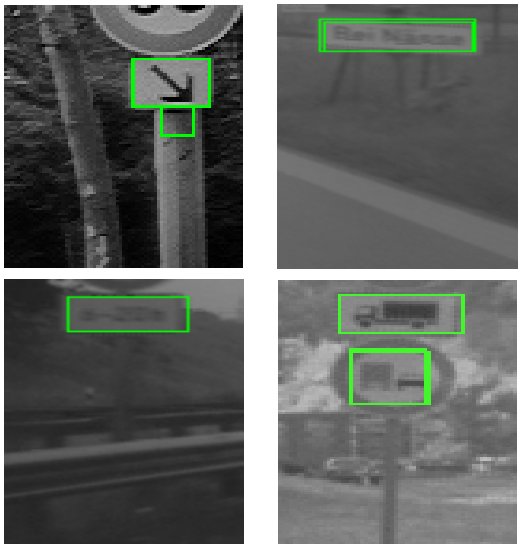


Fig. 5. Some good results obtained with the region-growing method for French and German subsigns.

hal-00741617, version 1 - 14 Oct 2012

	Method	$J \geq 0.5$		$O \geq 0.5$		$O \geq 0.5 \& D \leq 1.5$		$C \leq 0.2$	
		fr	rank	fr	rank	fr	rank	fr	rank
dbF	Edge	0.60	2	0.80	3	0.73	2	0.51	3
	Color	0.48	3	0.73	4	0.53	4	0.56	2
	Graph	0.47	4	0.81	2	0.58	3	0.49	4
	Region	0.66	**1**	0.89	**1**	0.80	**1**	0.64	**1**
dbG	Edge	0.69	2	0.85	3	0.77	2	0.58	2
	Color	0.40	4	0.76	4	0.46	4	0.45	4
	Graph	0.68	3	0.90	**1**	0.72	3	0.58	2
	Region	0.75	**1**	0.90	**1**	0.78	**1**	0.73	**1**

TABLE II

RESULTS OBTAINED ON THE TWO DATABASES FOR THE FOUR IMPLEMENTED TECHNIQUES. NUMBER OF SUBSIGNS CORRECT IN THE SENSE OF EACH CRITERIA PER FRAME AND RANKING.

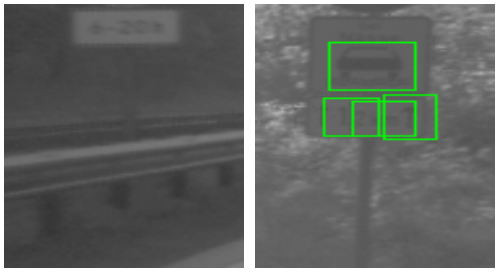


Fig. 6. Some bad results obtained with the region-growing method due to no seed selection (left) or a low contrast (right).

detection above 70%.

As future work, we aim at developing the recognition stage in order to validate the complete process. A research axis is to split the subsigns into different meta categories, for instance text, arrows and vehicle pictograms to perform a first coarse classification. Then a finer recognition will give the final subsign type and eventually the written message. Finally, the complete system has to be tested under real-time and various weather conditions.

V. ACKNOWLEDGMENTS

This work was conducted in the framework of the Speedcam project (ANR-09-VTT-11), funded by the ANR (Agence Nationale de la Recherche).

REFERENCES

[1] O. Hamdoun, A. Bargeton, F. Moutarde, B. Bradai, and L. Chanussot, "Recognition of End-of-Speed-Limit and Supplementary Signs for Improved Speed Limit Support," in *World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting*, 2008.

[2] W. Liu, J. Lv, H. Gao, B. Duan, H. Yuan, and H. Zhao, "An Efficient Real-Time Speed Limit Signs Recognition Based on Rotation Invariant Features," in *IEEE Intelligent Vehicles Symposium*. IEEE, 2011, pp. 1000–1005.

[3] D. Nienhueser, T. Gumpp, J. Zollner, and K. Natroshvili, "Fast and reliable recognition of supplementary traffic signs," in *IEEE Intelligent Vehicles Symposium*, 2010, pp. 896–901.

[4] C. Jung and R. Schramm, "Rectangle detection based on a windowed hough transform," in *Brazilian Symposium on Computer Graphics and Image Processing*, 2004, pp. 113–120.

[5] C. Keller, C. Sprunk, C. Bahlmann, J. Giebel, and G. Baratoff, "Real-Time Recognition of US Speed Signs," in *IEEE Intelligent Vehicles Symposium*. IEEE, 2008, pp. 518–523.

[6] S. Chang, L. Chen, Y. Chung, and S. Chen, "Automatic License Plate Recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 1, pp. 42–53, 2004.

[7] W. Wu, X. Chen, and J. Yang, "Detection of Text on Road Signs from Video," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 4, pp. 378–390, 2005.

[8]

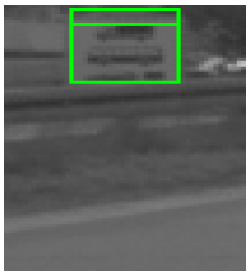
[9] L. Vincent, "Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[10] Z. Zhang, C. Chen, J. Sun, and K. Luk Chan, "E.M. Algorithms for Gaussian Mixtures with Split-and-merge Operation," *Pattern Recognition*, vol. 36, no. 9, pp. 1973–1983, 2003.

[11] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

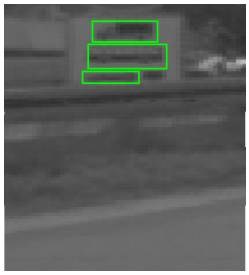
hal-00741617, version 1 - 14 Oct 2012

(a) Canny image.



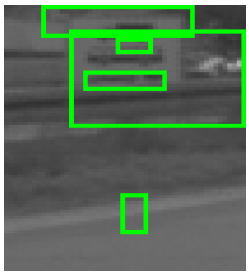
(b) Rectangle found by the edge-base approach.

(c) Final regions output by graph-based.

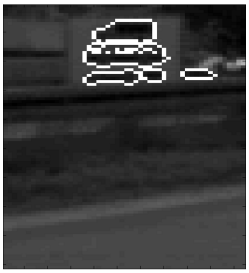


(d) Corresponding rectangles.

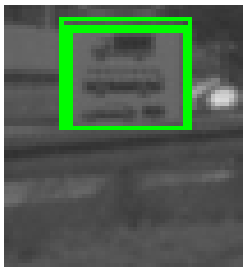
(e) Segmentation obtained from the color-based approach.



(f) Filtered rectangles.



(g) Initial seeds of the region growing.



(h) Output of the region growing.

Fig. 7. Examples of segmentations obtained for the different implemented techniques.

Real-time visual detection of vehicles and pedestrians with new efficient adaBoost features

Fabien Moutarde, Bogdan Stanciulescu and Amaury Breheret

***Abstract*— This paper deals with real-time visual detection, by mono-camera, of objects categories such as cars and pedestrians. We report on improvements that can be obtained for this task, in complex applications such as advanced driving assistance systems, by using new visual features as adaBoost weak classifiers. These new features, the "connected control-points" have recently been shown to give very good results on real-time visual rear car detection. We here report on results obtained by applying these new features to a public lateral car images dataset, and a public pedestrian images database. We show that our new features consistently outperform previously published results on these databases, while still operating fast enough for real-time pedestrians and vehicles detection.**

I. Introduction and related work

AUTONOMOUS vehicles, as well as most Advanced Driving Assistance System (ADAS) functions, require real-time perception analysis. This environment perception can be done using various sensors such as lidars, radars, ultrasonic devices, etc… However, compared to other sensors, visual perception can provide very rich information for very low equipment costs, if an abstract enough scene analysis can be conducted in real-time.

One of the key bricks required for such an automated scene analysis is efficient visual detection of most common moving objects in car environment: vehicles and pedestrians. Many techniques have been proposed for visual object detection and classification (see eg [10] for a review of some of the state-of-the-art methods for pedestrian detection, which is the most challenging). Of the various machine-learning approaches applied to this problem, only few are able to process videos in real-time. Among those last ones, the boosting algorithm with feature selection was successfully extended to machine-vision by Viola & Jones [4][5]. The adaBoost algorithm was introduced in 1995 by Y. Freund and R. Shapire [1][2], and its principle is to build a *strong classifier*, assembling weighted weak classifiers, those being obtained iteratively by using successive weighting of the examples in the training set.

Most published works using adaBoost for visual object class detection are using the Haar-like features initially proposed by Viola & Jones for face and pedestrian detection.

Manuscript received June 10, 2008.

F. Moutarde, B. Stanciulescu and A. Breheret are all with the Robotics Laboratory of Mines ParisTech, 60 Bd St Michel, 75006 Paris, FRANCE (33-1-40.51.92.92, {Fabien.Moutarde,Bogdan.Stanciulescu}@ensmp.fr).

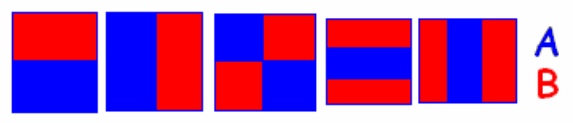


Fig.1: Viola &Jones Haar-like features

These weak classifiers compute the absolute difference between the sum of pixel values in red and blue areas (see figure 1), with the respect of the following rule:

if $\left|Area(A) - Area(B)\right| > Threshold$ **then True else False**

Fig. 2: Some examples of adaBoost-selected Viola-Jones features for car detection (top) and pedestrian detection (bottom).

However, the adaBoost outcome may strongly depend on the family of features from which the weak classifiers are drawn. But rather few investigations have been done on using other kinds of features with adaBoost: Zhu et al. in [13] defined and successfully applied adaBoost features directly inspired from the Histogram of Oriented Gradient (HOG) approach initially proposed (combined with SVM) by Dalal [12]; Baluja et al. in [14] and Leyrit et al. in [15] both use pixel-comparison-based feature very similar, although simplified, to our lab's control-points approach ([6][7][8][9]); very recently Pettersson et al. in [16] proposed efficient gradient-histogram-based features inspired from HOG.

II. CONTROL-POINTS ADABOOST FEATURES

Several years ago, Abramson & Steux [6][7] proposed an original set of features, the control-points, for faster and more illumination-independant adaBoost classifiers.

These features operate directly at pixel level (at one among 3 different possible resolutions) and are illumination-independent. Each of these features can be computed by only a few pixel comparisons, which makes them extremely fast, thus providing very good real-time performances for the resulting detector. Arbitrary points are divided in two groups, one called the positive set and the second called the negative set. Examples are classified as positive, if the following condition applies:

$$\min(\{P_i^+, i=1,...,N_+\}) - \max(\{P_j^-, j=1,...,N_-\}) > V$$

OR

$$\min(\{P_j^-, j=1,...,N_-\}) - \max(\{P_i^+, i=1,...,N_+\}) > V$$

V is the minimum separation threshold between the two point groups, P_i^+ a point from the positive group, P_j^- a point from the negative group, and N_+ and N_- the number of points in the respective groups.

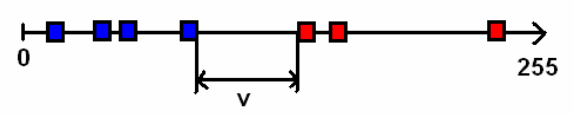


Fig. 3a: Positive-classified example with respect to the threshold V.

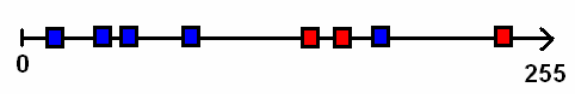


Fig. 3b: Negative-classified example.

In a linear representation of the pixel values, an example is classified as positive if the two point groups are separated by at least the value of threshold V (see figure 3a). Negative examples are those that do not respect this characteristic: values of the control-points of the two groups are interleaved (see figure 3b).

One can see on the figure 4 some examples of control-points features acting on vehicle or pedestrian detection. Each feature operates at either full-, half- or quarter-resolution of the minimal detection window size (80x32 for the lateral car case, and 18x36 for the pedestrian case). An examined image or sub-window is thus resized to those 3 resolutions before the features are applied.

On the upper-left example of figure 4, the feature will respond positively if the 2 pixels values (on the correctly resized image) corresponding to the 2 white squares *all* have higher luminance (with margin ≥ V) than *all* 3 pixels values corresponding to the 3 red squares (or opposite). This particular feature can therefore be interpreted as detecting some usual contrast between the car itself and region just below, with shadow and dark tyres. Similarly, the lower-left feature seems to detect some contrast between pedestrian center and the background. Such interpretation of selected control-points features is not always very clear, however.

AdaBoost requires a ”weak learner”, i.e. an algorithm which will select and provide, for each adaBoost step, a ”good” feature (i.e. with a ”low-enough” weighted error measured on the training set). The weak learner used by Viola and Jones is just an exhaustive search of all ~180,000 possible features in their set of features. But as our control-point family features is absolutely huge (there are more than 10^{35} of them for a 36 × 36 detection window size), a systematic full search is definitely not possible. We therefore use as weak learner a genetic-like heuristic search in feature space: an evolutionary hill-climbing described in more details in [8].

The core of our heuristic search weak-learner is to define specific mutations adapted to the feature-type, and apply them to a population of initially random features. A single mutation of one control-points feature typically consists in adding, moving, or removing one of the points, changing working resolution, or modifying the value of threshold V. When evolution provides no more improvement, the best feature of the population is selected and the weak-learner returns it to be added as the next adaBoost feature.

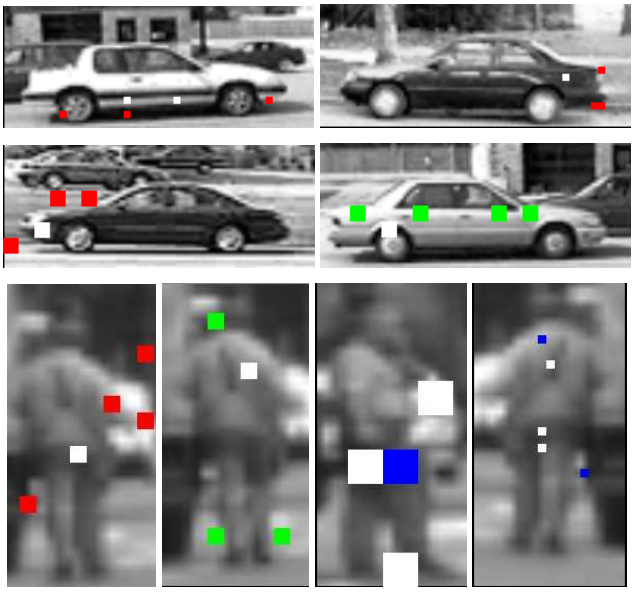


Fig. 4: Some examples of adaBoost-selected Control-Points features for car detection (top) and pedestrian detection (bottom line). Some features operate at full resolution of detection window (eg rightmost bottom), while others work on half-resolution (eg leftmost bottom), or even at quarter-resolution (third on bottom line).

III. NEW "CONNECTED-CONTROL-POINTS" FEATURES

As presented in [9], we have recently explored new types of adaBoost features in the context of rear car detection. It turned out that among those, the new "connected control-points" significantly outperformed all others. This feature is a particular form of the control-points feature. It contains 2 up to 12 points, and the principle is exactly the same as described in II. The difference is that the "control-points" of a given feature are constrained to remain connected with 8-connectivity, which implies each point must touch another one *at least by a corner*.

As mentioned in section II, the classical control-points features family is extremely large, and therefore difficult to search efficiently by the weak-learner. By imposing the 8-connectivity constraint, the search-space size decreases to $\sim 3 \times 10^{19}$ possible combinations instead of $\sim 10^{35}$, which makes it easier to explore efficiently for our heuristic. Besides, the connectedness constraint will force each feature to focus on a more localized part of the detection window.

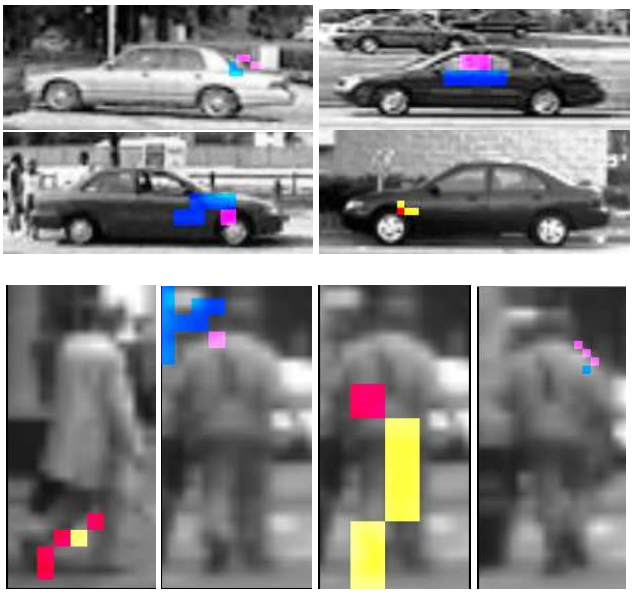


Fig. 5: Some examples of adaBoost-selected new "connected control-points" features for lateral car detection (top) and pedestrian detection (bottom line).

In figure 5 are shown some examples of the "connected control-points" features resulting from the adaBoost training process for cars and pedestrians. The evolutionary heuristic weak-learner we use is exactly the same as for standard control-points, except that the mutation operator has been modified to maintain the connectedness constraint. As can be seen by comparison to figure 4, because of the connectedness constraint, each of the new features tend to operate on a particular region (as can readily be seen on figure 5), contrary to basic control-points features whose points positions are sometimes disseminated throughout the detection window (see eg bottom right on figure 4). As a result, our connected-control-points features are in some way a kind of generalization of Haar-like features, but much more flexible in shape so that they can adapt themselves to detect any particular contour or contrast geometry. Note on examples of figure 5 that the features we obtain are even more general and flexible than the generalization of Viola&Jones type features proposed by Treptow and Zell in [17], with which they had obtained better detection performances than with standard Viola&Jones Haar-like features.

IV. EXPERIMENTS AND RESULTS

Encouraged by our good results on rear car detection [9], we decided to test our new "connected control-points" features for other kind of objects encountered in vehicle environment: lateral cars, and pedestrians. In order to allow comparisons with other published methods, we have chosen to work on publicly available databases: the "lateral cars" by UIUC [11], and the pedestrian database collected by Munder and Gavrila [10].

A. *Lateral cars database*

The lateral cars database contains 500 positive examples and 500 negative examples, all of size 100x40 pixels. For evaluation, we use, as in [11], the set of 108 wider field independent images containing 139 lateral cars at various scales, ranging from roughly 0.8 to 2 times the size of cars in the training images. This test set comes with an associated ground truth allowing automated computation of correct detection and false alarm rates.

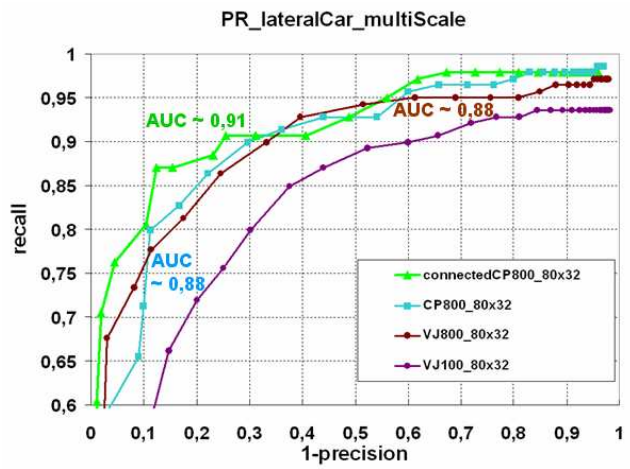


Fig. 6: Precision-recall curves for adaBoost lateral car detectors obtained with connected-control-points features (upper green curve), standard control-points (cyan curve), and ViolaJones Haar-like features (maroon)

All trainings were done for 800 boosting steps. Figure 6 shows precision-recall curve of resulting detectors obtained with various feature families. We use precision-recall metrics in order to allow easy comparison with (rather poor) results of the method presented in [11] on the same database. Our new "connected control-points" features (upper curve, and

best Area Under Curve with 0.91, instead of 0.88) outperforms both our usual simple control-points, and Haar-like features.

Figure 7 shows some detection results on test wider-field images by our connected-control-points adaBoost classifier. These illustrate the robustness to at least moderate occlusion, of classifiers built with our new features.

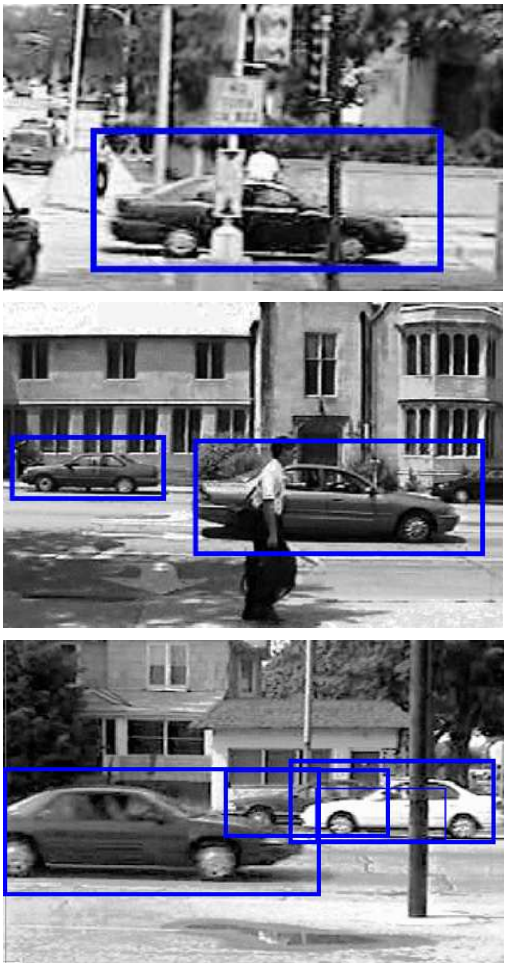


Fig. 7: Some detection results with our connected-control-points adaBoost classifier, which illustrates its robustness to at least moderate occlusion.

If we compare detectors with similar computation loads (in this particular setup, control-points features operate ~ 8 times faster than our implementation of ViolaJones Haar-like features), then the superiority of our new connected control-points features over Haar-like features is even clearer (see figure 7). It should be noted however that our ViolaJones classifiers were obtained using the same heuristic weak-learner as for control-points (with adapted mutation operator), rather than usual full-search which would anyway have been prohibitively long for a 80x32 detection window size.

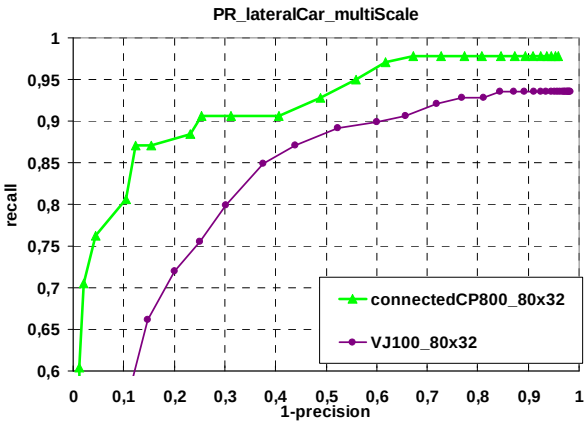


Fig. 7: Precision-recall for adaBoost lateral car detection, when comparing detectors with similar computation loads. At equivalent computation time, our new connected-control-points features clearly outperform ViolaJone Haar-like features.

B. Pedestrians database

The pedestrian database comprises 3 training sets and 2 test sets (each one of the 5 sets with 4800 positive examples and 5000 negative ones). As suggested in [10], 3 independent trainings were conducted on unions of 2 of the 3 training sets, and the evaluation was done on the 2 test sets, producing a total of 6 evaluations, to be averaged, for each feature type. In each training, 2000 boosting steps were allowed, therefore producing adaBoost detectors assembling 2000 weak-classifiers.

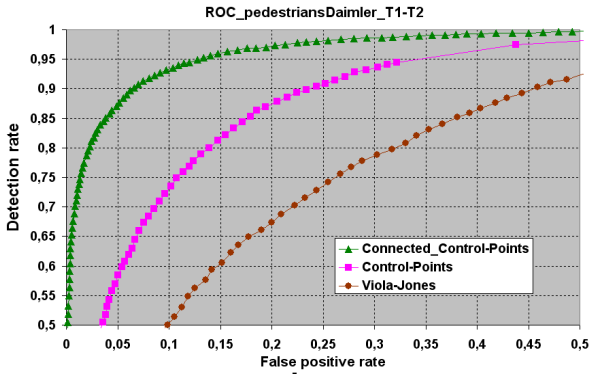


Fig. 9: Averaged ROC curves for adaBoost pedestrian classifiers obtained with various feature families

As one can see in figure 9, the classifiers obtained with the new “connected control-points” features have by far the best classification results. The Viola-Jones performs rather poorly, even when compared to “ordinary control-points”.

We also compared the performance of our new classifier to the Viola-Jones classifier performance reported in [10], which was obtained with openCV implementation. As can be seen on figure 10, our “connected control-points” pedestrian classifier has a significantly better performance, which confirms the results obtained with our own implementation (with which we did not use cascade for our comparisons).

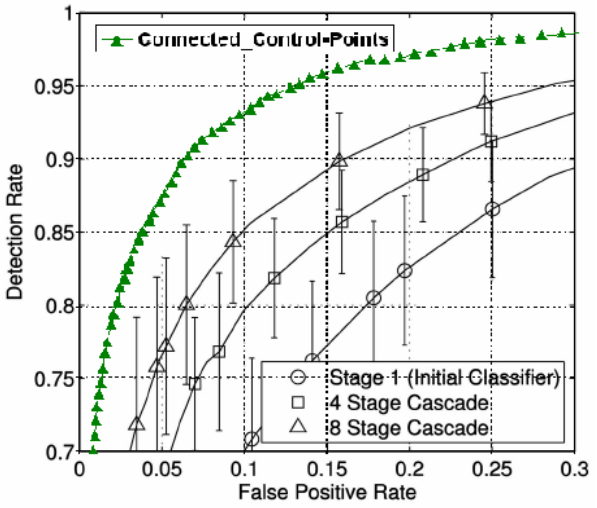


Fig. 10: ROC curves comparing our boosted "connected control-points" (upper curve, green) to boosted ViolaJones cascade result reported in [10].

Moreover, we finally compare to the best methods reported in [10] on figure 11, where one can see that boosting with our new features seems to be even better than the best algorithms (namely quadratic and RBF SVM, and NN-LRF) reported in [10].

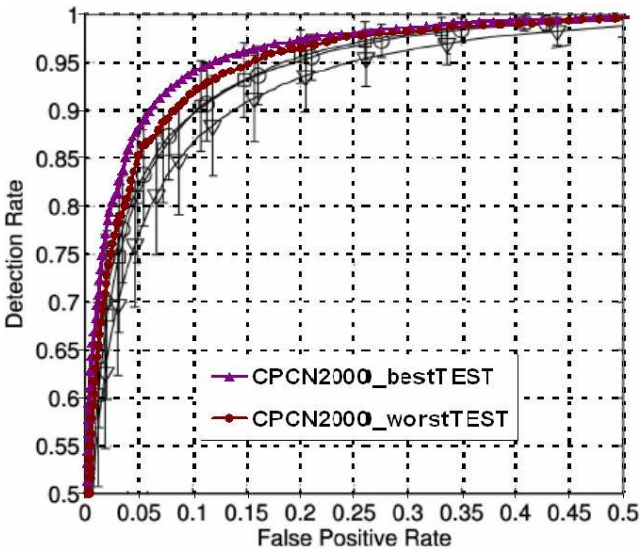


Fig. 11: ROC curves comparing our boosted connected control-points (two upper curves) to best algorithms results reported in [10].

It should be noted that the best algorithms from [10], to which we compare on figure 11, are reported in [10] to operate at ~ 250 ms per test sample on a 3.2 Ghz Pentium IV PC, while our boosted classifier containing 2000 "connected control-points" features requires only ~0.4ms per test image from the database, on a 2 GHz Intel Core2 laptop.

V. Conclusions and perspectives

We have presented a new feature type, which we call "connected control-points", for adaBoost training of visual object classifiers.

We report here on test of these new features on two publicly available databases: one for lateral cars, and one for pedestrians on which many classification algorithms have already been tested and results published. It turns out that the adaBoost strong classifiers obtained with our new features, while being extremely fast (~0.4ms per pedestrian image classification on a 2Ghz laptop), clearly outperform both standard Viola-Jones boosted cascade and even the most powerful (but very slow) classification algorithms reported so far on the pedestrian database.

Given previous tests conducted by us on real-time visual rear car detection application [9] that have also shown these new "connected control-points" features to provide better results than other features used in boosting, we think these new features have a very promising potential for improving real-time detection performance of visual object classes in general, and particularly the kind of objects that should be efficiently detected and tracked in intelligent vehicle applications.

References

[1] Yoav Freund, Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", '95 European Conference on Computational Learning Theory, pages 23–37, 1995.

[2] Yoav Freund Robert E. Schapire, "A short introduction to boosting", Journal of Japanese Society for Artificial Intelligence 14(5), pages 771–780, 1999.

[3] R. Meir and G. Ratsch, "An introduction to boosting and leveraging", S. Mendelson and A. Smola, Editors, Advanced Lectures on Machine Learning, LNCS. Springer Verlag, pages 119–184, 2003.

[4] Paul Viola, Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2001), Volume 1, page 511, Kauai , Hawai, USA, 2001.

[5] Paul Viola, Michael J. Jones, Daniel Snow. « Detecting pedestrians using patterns of motion and appearance", IEEE International Conference on Computer Vision (ICCV'2003), pages 734–741, Nice, France, 2003.

[6] Yotam Abramson, Bruno Steux, "Illumination-independent pedestrian detection in real-time", internal report, CAOR, Mines ParisTech, 2003.

[7] Yotam Abramson, Bruno Steux, Hicham Ghorayeb, "YEF (Yet Even Faster) Real-Time Object Detection ", 2005 Proceedings of International Workshop on Automatic Learning and Real-Time (ALART'05), Siegen, Germany, page 5, 2005.

[8] Yotam Abramson, Fabien Moutarde, Bruno Steux, Bogdan Stanciulescu, "Combining adaBoost with a Hill-Climbing evolutionary feature search for efficient training of performant visual object detectors", Proceedings of FLINS2006 Conference on Applied Computational Intelligence, Genova, Italy, pages 737-744, 2006.

[9] Bogdan Stanciulescu, Amaury Breheret, Fabien Moutarde, "Introducing New AdaBoost Features for Real-Time Vehicle Detection", Proceedings of COGIS2007 Cognitive Systems with Interactive Sensors, Stanford University, USA, 2007.

[10] Stefan Munder and Dariu M. Gavrila. "An Experimental Study on Pedestrian Classification". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, nr 11, pp. 1863-1868, 2006.

[11] Shivani Agarwal, Aatif Awan, and Dan Roth, "Learning to Detect

Objects in Images via a Sparse, Part-Based Representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.26, nr 11, pp. 1475-1490, 2004.

[12] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection", proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), held in San Diego, CA, USA, 20-26 June 2005.

[13] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, Shai Avidan, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), held in New York, NY, USA, 17-22 June 2006.

[14] Shumeet Baluja, Mehran Sahami, and Henry A. Rowley, "Efficient face orientation discrimination", proc. *International Conference on Image Processing (ICIP '04)*, pp. 589-592 Vol. 1, 24-27 Oct. 2004.

[15] Laetitia Leyrit, Thierry Chateau, Christophe Tournayre and Jean-Thierry Lapresté, "Association of AdaBoost and Kernel Based Machine Learning Methods for Visual Pedestrian Recognition", proc. *IEEE* Intelligent Vehicles Symposium (IV 2008), Eindhoven (Nederland), June 2008.

[16] Niklas Pettersson, Lars Petersson and Lars Andersson, "The Histogram Feature – A Resource-Efficient Weak Classifier", proc. *IEEE* Intelligent Vehicles Symposium (IV 2008), Eindhoven (Nederland), June 2008.

[17] A. Treptow and A. Zell, "Combining adaBoost learning and evolutionary search to select features for real-time object detection", In Proc. of IEEE Congress on Evolutionary Computation (CEC 2004), Portland, Oregon, vol. 2, pp. 2107-2113, IEEE Press (2004).

ADABOOST WITH "KEYPOINT PRESENCE FEATURES" FOR REAL-TIME VEHICLE VISUAL DETECTION

Taoufik Bdiri, Fabien Moutarde, Nicolas Bourdis and Bruno Steux

Robotics Laboratory (CAOR), Mines ParisTech
60 Bd Saint-Michel, F-75006 PARIS, FRANCE
Tel.: (33) 1-40.51.92.92, Fax: (33) 1.43.26.10.51
Fabien.Moutarde@mines-paristech.fr

ABSTRACT

We present promising results for real-time vehicle visual detection, obtained with adaBoost using new original "keypoints presence features". These weak-classifiers produce a boolean response based on presence or absence in the tested image of a "keypoint" (~ a SURF interest point) with a descriptor sufficiently similar (i.e. within a given distance) to a reference descriptor characterizing the feature. A first experiment was conducted on a public image dataset containing lateral-viewed cars, yielding 95% recall with 95% precision on test set. Moreover, analysis of the positions of adaBoost-selected keypoints show that they correspond to a specific part of the object category (such as "wheel" or "side skirt") and thus have a "semantic" meaning.

KEYWORDS: vehicle visual detection, image categorization, boosting, interest points

INTRODUCTION AND RELATED WORKS

Efficient and reliable detection of surrounding moving objects (such as pedestrians and vehicles) is one of the key features for enhancing safety in intelligent vehicles. It is particularly interesting to be able to properly detect laterally incoming cars that could lead to lateral collisions. A module implementing reliable (particularly with very low false alarm rate) visual detection of laterally incoming cars could for instance be used for developing useful Advanced Driving Assistance Systems (ADAS) such as a Lateral Collision Warning (LCW) system.

We present here a new method for the visual recognition and detection of a given object type, with a first test application for laterally-viewed cars. Many techniques have been proposed for visual object detection and classification (see e.g. [3] for a review of some of the state-of-the-art methods for pedestrian detection, which is the most challenging). Of the various machine-learning approaches applied to this problem, only few are able to process videos in real-time. Among those, the boosting algorithm with feature selection was successfully extended to machine-vision by Viola & Jones [2]. The adaBoost algorithm was introduced in 1995 by Y. Freund and R. Shapire [1], and its principle is to build a *strong classifier*, assembling weighted

weak classifiers, those being obtained iteratively by using successive weighting of the examples in the training set. Most published works using adaBoost for visual object class detection are using the Haar-like features initially proposed by Viola & Jones for face and pedestrian detection.

However, adaBoost outcome may strongly depend on the family of features from which the weak classifiers are drawn. Recently, several teams [4][5] have reported interesting results with boosting using other kinds of features directly inspired from the Histogram of Oriented Gradient (HOG) approach. Our lab has been successfully investigating boosting with pixel-comparison-based features named “control-points” (see [6] for original proposal, and [7] for recent results with a new variant).

To our knowledge, the idea of using interest point descriptors as boosting features was first proposed by Opelt et al. in [8], but it was in a more general framework, and they were considering SIFT points and descriptors [9] which are quite slow to compute, compared to the SURF points and descriptors [10]. In the present work we investigate boosting of “keypoint presence features”, where “keypoint” are a variant of SURF points implemented in our lab, and already successfully applied to real-time person re-identification [11].

CAMELLIA “KEYPOINTS ”

The interest point detection and descriptor computation is performed using “key-points” functions available in the Camellia image processing library (http://camellia.sourceforge.net). These Camellia key-points detection and descriptor functions – named CamKeypoints - implement a variant of SURF [10]. SURF itself is an extremely efficient method (thanks to the use of integral images) inspired from the more classic and widely used interest point detector and descriptor SIFT [9].

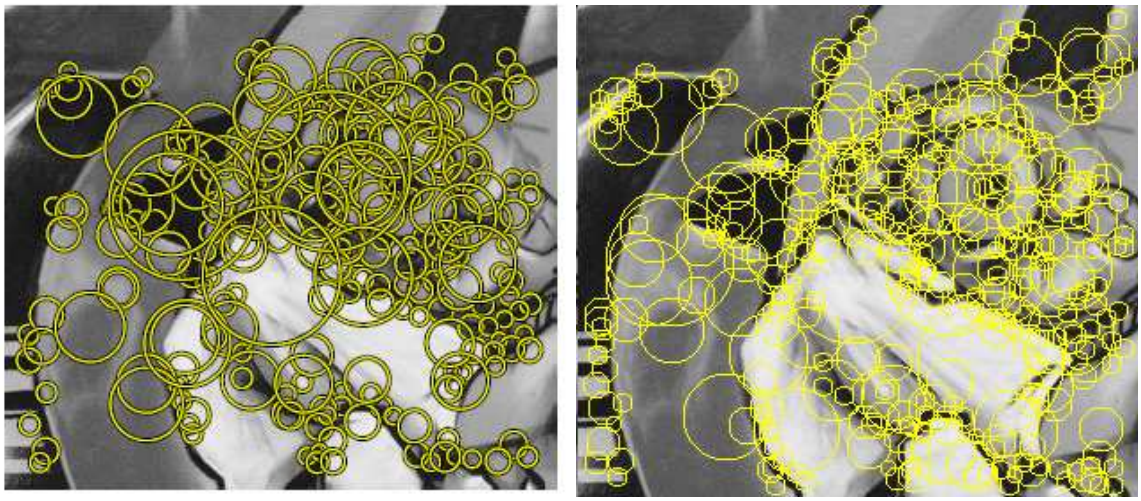


Figure 1 - SURF interest points (left) v.s. Camellia keypoints (right); they are very similar except for voluntary suppression of multiple imbricated blobs at different scales (cf. upper left).

As for SURF interest points, the detection of Camellia keypoints is a “blob detector” based on finding local Hessian maxima, those being efficiently obtained by approximating second order derivatives with box filters computed with integral image. Our keypoints are however not exactly the same as SURF points (as can be seen on figure 1), in particular because multiple imbricated blobs at various scales are voluntarily avoided. In contrary to SURF and SIFT, CamKeypoint scale selection is not based on overlapping octaves, but on a set of discrete scales from which the scale of a keypoint is derived by quadratic interpolation. This speeds up the keypoints detection compared to SURF by a factor of 2, without sacrificing the quality of scale information, as was shown by some experiments.

The descriptor used for each Camellia keypoint is similar to the SURF descriptor : an image patch corresponding to the keypoint location and scale is divided in 4x4=16 sub-regions, on each of which are efficiently computed (by using integral image approach) the following 4 quantities:

$$\sum \boldsymbol{dx}$$
$$\sum |\boldsymbol{dx}|$$
$$\sum \boldsymbol{dy}$$
$$\sum |\boldsymbol{dy}|$$

where *dx* and *dy* are respectively the horizontal and vertical gradient. The total descriptor size is therefore 16 x 4 = 64. In order to avoid all boundary effects in which the descriptor abruptly changes when a keypoint physically changes, bi-linear extrapolation is used to distribute each of the 4 quantities above into 4 sub-regions. Experiments have shown that this really improves the quality of the descriptor wrt. SURF. In addition to this, CamKeypoints support color images by adding 32 elements of gradient information by color channel (U and V) to the signature, resulting in a 128 descriptor size for color descriptors.

Another main difference between Camellia Keypoints and SURF lies in that the Camellia implementation uses integer-only computations – even for the scale interpolation –, which makes it even faster than SURF, and particularly well-suited for potential embedding in camera hardware. SIFT and SURF make extensive use of floating point computations, which makes these algorithms power hungry.

BOOSTING “KEYPOINTS PRESENCE ” FOR CATEGORIZATION

The rationale of boosting “keypoint presence features” for image categorization is that it should be possible, for a given object category, to determine a set of characteristic interest points whose *simultaneous* presence would be representative of that particular category. This is similar in spirit, but with a completely different algorithm, to the “part based” approach proposed by [12].

Our object recognition approach uses the same general feature-selecting boosting framework as pioneered by Viola&Jones in [2]. The originality of our work is to define and use as weak classifiers a new original feature family, instead of Haar features. This new feature type is a weak classifier that answers positively on an image if and only if, among all the Camellia keypoints detected in the image, there is at least one

of them whose descriptor is similar enough to the “reference keypoint descriptor” associated with the weak-classifier.

More formally, each “keypoint presence” weak-classifier is defined by a keypoint SURF descriptor D (in $\Re^{64}$) and a descriptor difference threshold scalar value d. This weak-classifier h(D,d,I) answers positively on an image I if and only if I contains at least one keypoint whose descriptor D’ is such that $|D-D'|<d$, where the “sum of absolute difference” (SAD) L1-distance is used: if the descriptors of two keypoints K_1 and K_2 are respectively $\{D_1[i], i = 1…64\}$ and $\{D_2[i], i = 1…64\}$, then, the distance between K1 and K2 is given by equation 1 below:

$$\mathrm{Dist}(K_1, K_2) = \sum_{i=1}^{64} \mathrm{abs}(D_1[i] - D_2[i]) \quad \text{(Eq. 1)}$$

The training method is the standard feature-selecting adaBoost algorithm, in which the descriptor D is chosen among all descriptors found in *positive* example images. The threshold value d is chosen by considering the matrix M of distances Mij between positive keypoint Ki and all images Ij, this distance being defined as the smallest descriptor difference $|D(Ki)-D(Kpj)|$ for all keypoints Kpj found in image j.

More formally, let the training set be composed of positive images Ip1, Ip2, …., and of negative images In1, In2, …We first apply the Camellia keypoint detector on all *positive* images Ip1, Ip2, …, and build the “positive keypoints set” $Spk = \{K_1, K_2, K_3, …, K_Q\}$ as the union of all Camellia keypoints detected on any positive examples of the training set. The adaBoost feature-selection has to select, at each boosting step, a particular “keypoint presence” weak classifier defined by a 64D descriptor and a scalar threshold. The descriptor will be chosen among those of positive keypoints collected in Spk.

In order to choose a threshold value, we apply keypoints detection on all negative images as well, so that we can compare descriptors of the positive keypoints in Spk to descriptors of all keypoints found in training images. We define the “distance” between any given keypoint K and any given image I as the smallest descriptor difference between K and all keypoints KIj found in image I:

$$\mathrm{dist}(K,I) = \min_{KIj \text{ keypoint found in image } I} \{ \mathrm{dist}(K, KI_j) \} \quad \text{(Eq. 2)}$$

where dist(K,KIj) is the SAD of descriptors as defined in equation (1). This allows us to build a matrix M of distances between positive keypoints and all training images, where Mij = dist(Ki , Ij). As illustrated on figure 2, this QxN matrix (with Q the number of positive keypoints and N the number of training images) has at least one zero on each line, on the column corresponding to the positive image in which the keypoints was found.

We execute a growing sorting of the distance matrix M, row by row, and then we take the *middle of each two successive distances* in the sorted matrix to build the set $\{T_{ik}, k=1,…,N\}$ of candidate threshold values for a feature testing presence of the corresponding positive keypoint K_i.

I_{p1}	I_{p2}	I_{p3}	I_{n1}	I_{n2}	I_{n3}
0	x	x	x	x	x
0	x	x	x	x	x
x	0	x	x	x	x
x	0	x	x	x	x
x	0	x	x	x	x
x	x	0	x	x	x
x	x	0	x	x	x
x	x	0	x	x	x
x	x	0	x	x	x

Figure 2 - Matrix of distances between keypoints found on positive image examples (one for each row) and all N training images (positives and negatives, one image for each column)

At each boosting step, we choose among all (Ki,Tik) couples the one that gives the lowest weighted error on the training set: (i*,k*) = argmin_{ik} ($\sum_{j=1}^{N} w_j |h(Ki,T_{ik},I_j) - l_j|$), and the selected weak classifier is $h(K_{i^*},T_{i^*k^*}, .)$.

EXPERIMENT ON LATERAL-VIEWED CAR DATASET

For a first evaluation of our approach, we used the publicly available (http://l2r.cs.uiuc.edu/~cogcomp/Data/Car/) lateral-car dataset collected by Agarwal et al. [12]. This database contains 550 positive images and 500 negative images. For training, we use 352 positive images, and 322 negative images, the rest being used as a test set for evaluation. Note that the partition between training and testing subset is random. Some examples from the training set are shown on figure 3.

Figure 3 - Some positive (2 left columns) and negative (right column) examples from the training set

Figure 4 shows the typical error evolution during adaBoost training: as is usual with boosting, the training error quickly falls to zero, and the error on test set continues to diminish afterwards. This shows that boosting by assembling features extracted from our new “keypoint presence” family does work and allow to build a strong classifier able to discriminate a given object category. On this particular case, there seems to be no clear improvement on test dataset for boosting steps T>150.

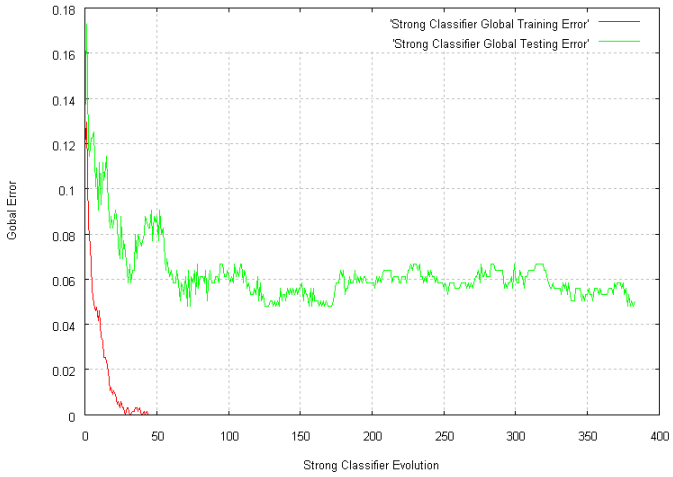


Figure 4 - Typical evolution, during successive boosting steps, of errors on training and test

Figure 5 shows the precision-recall curve, computed on the independent test set, for boosted strong classifiers with respectively 10 and 300 “keypoint presence” weak-classifiers assembled. The classification result is very good, with a recall (percentage of lateral cars recognized) of ~95%, for a precision (percentage of true lateral cars among all test images declared positive by the classifier) of ~95%.

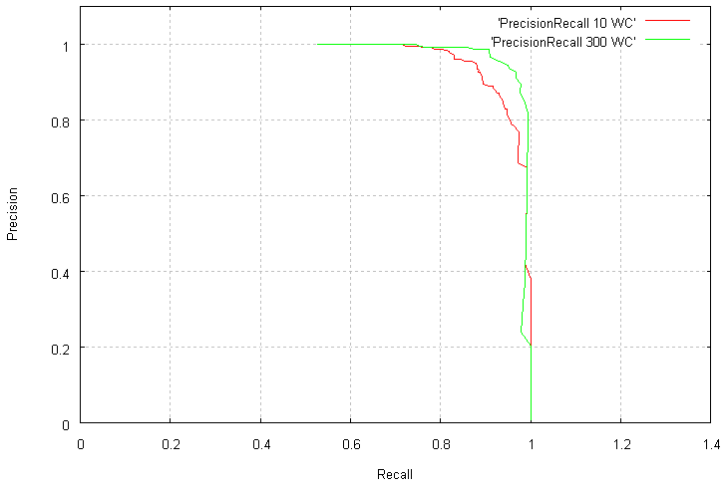


Figure 5 - Precision-recall curve computed on test set, for strong boosted classifier assembling 10 and 300 weak classifiers.

There are several motivations for our new feature type. One is that a classifier based on the simultaneous presence of several characteristic keypoints matches the intuition we can have on how human do categorize image by spotting some characteristic parts. In order to check if our adaBoost-selected keypoints make sense from this point of view, we decided to check on positive images where are located the "positively responding keypoints" for a given feature of the strong classifier. Figure 6 illustrates the positions of all keypoints, cumulated on all positive example images, that are within the descriptor distance threshold of one given adaBoost-selected keypoints. This clearly shows that the keypoints selected correspond to specific parts of the object category, such as the wheels or the side skirt, which means they have a semantic signification relative to the object category.

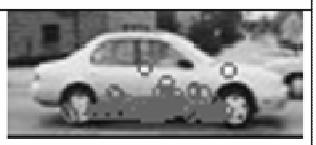
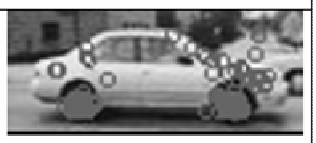
Weak Classifier	Image 1	Specificity
40		Keypoints concentrated on the rectangular side of the car
230		Keypoints concentrated under the car between the two wheels
300		Keypoints concentrated on wheels

Figure 6 - Position of adaBoost positively responding keypoints, cumulated on all positive example images: each selected keypoint seem to correspond to a specific part of the car.

Another motivation for these new kind of adaBoost features is that, by nature of the features, it is possible to derive the localizations in the image of objects of the searched category quite straightforwardly by some kind of Hough-like method applied to the positions of object-category-specific keypoints, thus making costly window-scanning unnecessary.
A preliminary result of keypoints filtering followed by object detection and localization, applied on a real on-board video, is illustrated on figure 7. The keypoint filtering uses a specific keypoint classifier trained to discriminate between "lateral car keypoints" and "background keypoints". From these remaining keypoints, those that are compatible with one of our adaBoost-selected weak-classifiers are used to derive candidate bounding-boxes by applying a Hough-like method

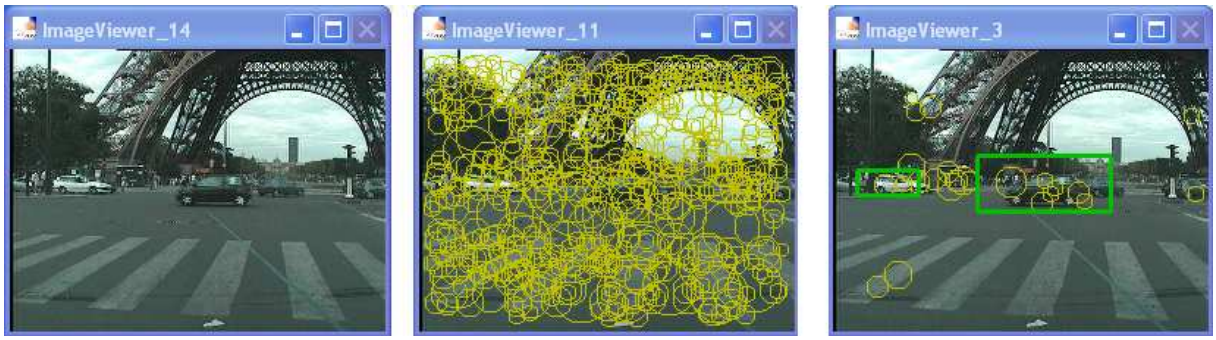


Figure 7 - First detection test on a video: the processed frame on left side, all keypoints on middle, and on the right side only keypoints classified as "lateral-car keypoints" and the bounding boxes obtained with our Hough-like localization method.

CONCLUSIONS, DISCUSSION AND PERPECTIVES

We have presented a first successful test of boosting "keypoint presence features", applied to lateral car recognition, yielding 95% recall with 95% precision on test set. Moreover, analysis of the positions of adaBoost-selected keypoints show that they correspond to a specific part of the object category (such as "wheel" or "side skirt") and thus have a "semantic" meaning.

Regarding the performance attained, it is important to note that in the potential application (a Lateral Collision Warning system), it is not too problematic to miss some of the cars, but what is most important is to ensure a low false alarm rate. Therefore a 95% recognition rate is quite sufficient. The 95% precision could seem to imply too many false alarms, but one should keep in mind that we report here only rate computed on a still images database, in which the negative examples can be more confusing than background on an empty road or street. Also, the target application would analyze a video, so that a temporal filtering would most probably get rid of most false alarms, which would probably not arise on all successive frames, contrary to true laterally incoming cars.

Perspectives include tests on various other datasets, including for other object categories such as pedestrians (for which a preliminary test was encouraging). If tests on other dataset for other categories are also successful, then this would imply that our method is quite general, and could be used for recognition of various types of potential obstacle.

Another important work underway is the improvement and optimization of our object-localization method based on the analysis of positions of positively-responding keypoints, that allows the detection step to be done without any tedious window-scanning, contrary to most existing detection object detection algorithms. Finally, we think the recognition performance of our method could be further improved by exploiting the relative positions of keypoints, instead of only their presence.

REFERENCES

[1] Freund Y., and Schapire R.E., “A decision-theoretic generalization of on-line learning and an application to boosting”, *1995 European Conference on Computational Learning Theory*, pages 23–37, 1995.

[2] Viola P., and Jones M., "Rapid Object Detection using a Boosted Cascade of Simple Features", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’2001)*, Volume 1, page 511, Kauai , Hawai, USA, 2001.

[3] Munder S. and Gavrila D. M., “An Experimental Study on Pedestrian Classification”. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.28, nr 11, pp. 1863-1868, 2006.

[4] Zhu Q., Yeh M.-C., Kwang-Ting Cheng K.-T., and Avidan S., “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients”, proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), held in New York, NY, USA, 17-22 June 2006.

[5] Pettersson N., Petersson L. and Andersson L., “The Histogram Feature – A Resource-Efficient Weak Classifier”, proc. *IEEE* Intelligent Vehicles Symposium (IV 2008), Eindhoven (Nederland), June 2008.

[6] Abramson Y., Steux B., and Ghorayeb H., “YEF (Yet Even Faster) Real-Time Object Detection “, *2005 Proceedings of International Workshop on Automatic Learning and Real-Time (ALART’05)*, Siegen, Germany, page 5, 2005.

[7] Moutarde F., Stanciulescu B., and Breheret A., "Real-time visual detection of vehicles and pedestrians with new efficient adaBoost features", proceedings of *'Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)' of "2008 International Conference on Intelligent RObots and Systems (IROS'2008)"*, Nice, France, 26 sept 2008.

[8] Opelt A., Pinz A., Fussenegger M., and Auer P., ”Generic Object Recognition with Boosting”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 28, No. 3, March 2006.

[9] Lowe, D. “Distinctive Image Features from Scale-Invariant Keypoints” *International Journal of Computer Vision, Vol. 60*, pp. 91-110*, Springer,* 2004*.*

[10] Bay H., Tuytelaars T. & Gool L. V., “SURF:Speeded Up Robust Features”, *Proceedings of the 9th European Conference on Computer Vision (ECCV’2006)*, Springer LNCS volume 3951, part 1, pp 404--417, 2006.

[11] Hamdoun O., Moutarde F., Stanciulescu B. and Steux B., "Interest points harvesting in video sequences for efficient person identification", *proceedings of '8th international workshop on Visual Surveillance (VS2008)' of "10th European Conference on Computer Vision (ECCV'2008)"*, Marseille, France, 17 oct. 2008.

[12] Agarwal S., Aatif Awan A., and Roth D., “Learning to Detect Objects in Images via a Sparse, Part-Based Representation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.26, nr 11, pp. 1475-1490, 2004.

Interest points harvesting in video sequences for efficient person identification

Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu, and Bruno Steux
Robotics Laboratory (CAOR), Mines ParisTech, 60 Bd St Michel, F-75006 Paris, FRANCE
{Omar.Hamdoun,Fabien.Moutarde,Bogdan.Stanciulescu,Bruno.Steux}@ensmp.fr

Abstract

We propose and evaluate a new approach for identification of persons, based on harvesting of interest point descriptors in video sequences. By accumulating interest points on several sufficiently time-spaced images during person silhouette or face tracking within each camera, the collected interest points capture appearance variability.
Our method can in particular be applied to global person re-identification in a network of cameras. We present a first experimental evaluation conducted on a publicly available set of videos in a commercial mall, with very promising inter-camera pedestrian re-identification performances (a precision of 82% for a recall of 78%). Our matching method is very fast: ~ 1/8s for re-identification of one target person among 10 previously seen persons, and a logarithmic dependence with the number of stored person models, making re-identification among hundreds of persons computationally feasible in less than ~ 1/5s second.
Finally, we also present a first feasibility test for on-the-fly face recognition, with encouraging results.

1. Introduction and related work

In many video-surveillance applications, it is desirable to determine if a presently visible person, vehicle, or object, has already been observed somewhere else in the network of cameras. This kind of problem is commonly known as "re-identification", and a general presentation of this field for the particular case of person tracking can be found for instance in §7 of [1]. Re-identification algorithms have to be robust even in challenging situations caused by differences in camera viewpoints and orientations, varying lighting conditions, pose variability, and also, for global persons, rapid change in clothes appearance.

A first category of person re-identification methods rely on biometric techniques (such as face or gait recognition). Face identification in "cooperative" context on high-resolution images with well-controlled pose and illumination can now be done with very good performance (see eg [12] or [13]). However, on-the-fly face matching in wide-field videos is still a very challenging problem.

A second group of methods try to perform person re-identification using no biometrics but only global appearance. Among these, various approaches have been proposed: signature based on color histograms (such as in [2], [3] or [14]), texture characteristics (see eg [4]). More recently some works have proposed the use of matching of interest points for establishing correspondance between objects, like cars in [5], and also for person re-identification as for instance in [6].

We here propose and evaluate a re-identification scheme using matching of interest points harvested in several images during short video sequences. The central point of our algorithm lies in the exploitation of image sequences, contrary to the method proposed in [6] where matches are done on image-to-image basis. This allows to get a more "dynamic" and multi-view descriptor than when using single image, and is a bit similar in spirit to the "averaging of interest-point descriptors throughout time sequence" used in the work by Sivic and Zisserman in [7]. However, contrary to them, we do not use SIFT [8] detector and descriptor, but a locally-developped (see §2) and particularly efficient variant of SURF [9]. This is also in contrast with Gheissari et al. in [6] who use a color-histogram of the region around interest points for their matching. Note also that we do not use a "vocabulary" approach as in [5] or [7], but use a direct matching between interest point descriptors.

We hereafter describe our method in more details, present a first performance evaluation on publicly available real-world videos, and a preliminary feasibility test for application to on-the-fly face identification.

2. Description of our re-identification scheme

In this section we detail the algorithmic choices made in our re-identification approach. Our method can be separated in two main stages: a learning phase, and a recognition phase. The learning consists in taking advantage of tracking of a given person, vehicle or object in a sequence from one camera, in order to extract interest points and their descriptors necessary to build the model.

The interest point detection and descriptor computation is done using "key-points" functions available in the Camellia image processing library (http://camellia.sourceforge.net). These Camellia key-points detection and characterization functions are implementing a very quick variant, which shall be described elsewhere in more details, inspired from SURF [9] but even faster. SURF itself is an extremely efficient method (thanks to use of integral images) inspired from the more classic and widely used interest point detector and descriptor SIFT [8].

Apart from some technical details, the main difference between Camellia keypoints and SURF lies in optimization of Camellia implementation, using integer-only computations, which makes it even faster than SURF, and particularly well-suited for embedding in camera hardware.

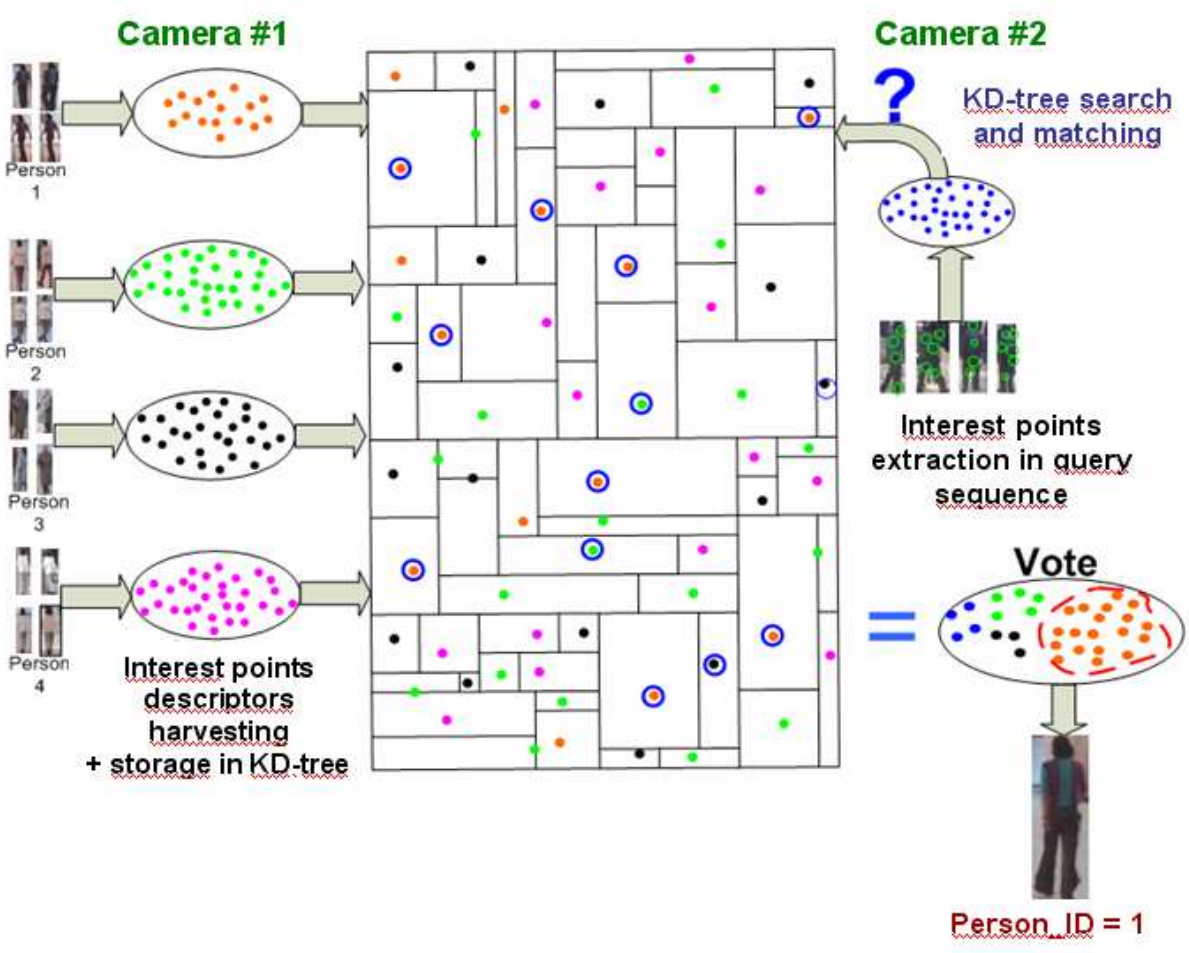


Figure 1: schematic view of model building (left), and of re-identification of a query (right).

Our recognition step uses tracking of the to-be-identified-person, and models built during learning stage, in order to determine if the signature of the currently analyzed person is similar enough to one of those already "registered" for which signatures have been stored from other cameras. Our method can be detailed in the following 5 steps:

1. **Model building**
   A model is built for each detected and tracked person. In order to maximize the quantity of non-redundant information, we do not use every successive frame, but instead images sampled every half-second. The person model is obtained as the accumulation of interest point descriptors obtained on those images.
2. **Query building**
   The query for the target persons is built on several evenly time-spaced images, exactly in the same way as the models, but with a smaller number of images (therefore collected in a shorter time interval).
3. **Descriptor comparison**
   The metric used for measuring the similarity of interest point descriptors is the Sum of Absolute Differences (SAD).
4. **Robust fast matching**
   A robust and very fast matching between descriptors is done by the employed Camellia function, which implements a Best Bin First (BBF) search in a KD-tree [10] containing all models.
5. **Identification**
   The association of the query to one of the models is done by a voting approach: every interest point extracted from the query is compared to all models points stored in the KD-tree, and a vote is added for each model containing a close enough descriptor; finally the identification is made with the highest voted-for model.

3. Experimental evaluation for application to re-identification of persons

For the person re-identification application, a first experimental evaluation of our method has been conducted, on a publicly available series of videos (http://homepages.inf.ed.ac.uk/rbf/CAVIAR)showing persons recorded in corridors of a commercial mall. These videos (collected in the context of European project CAVIAR [IST 2001 37540]) are of relatively low resolution, and include images of the same 10 persons seen by two cameras with very different viewpoints.

The model for each person was built with 21 evenly time-spaced images (separated by half-second interval), and each query was built with 6 images representing a 3 second video sequence (see figure 2). Camera color potential variability is avoided by working in grayscale. Illumination invariance is ensured by histogram equalization of each person's bounding-box.

The re-identification performance evaluation is done with the precision-recall metric:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\,call = \frac{TP}{T\,\mathrm{arg}\,et\ number}$$

with TP (True Positives) = number of correct query-model re-identification matching, and FP (False Positives) = number of erroneous query-model matching.

Table 1: precision and recall, as a function of the score threshold for query-model matching (ie minimum number of similar interest points).

Score threshold for query-model matching (number of matched points)	Precision (%)	Recall (%)
40	99	49
35	97	56
30	95	64
25	90	71
20	85	75
15	82	78
10	80	79
5	80	80

The resulting performance, computed on a total of 760 query video sequences of 3 seconds, is presented in table 1, and illustrated on a precision-recall curve on figure 3. The main tuning parameter of our method is the "score threshold", which is the minimum number of matched points between query and model required to validate a re-identification. As expected, it can be seen that increasing the matching score threshold, increases the precision but at the same time lowers the recall. Taking into account the relatively low image resolution, our person re-identification performances are good, with for instance 82% precision and 78% recall when the score threshold is set to a minimum of 15 matching interest points between query and model.

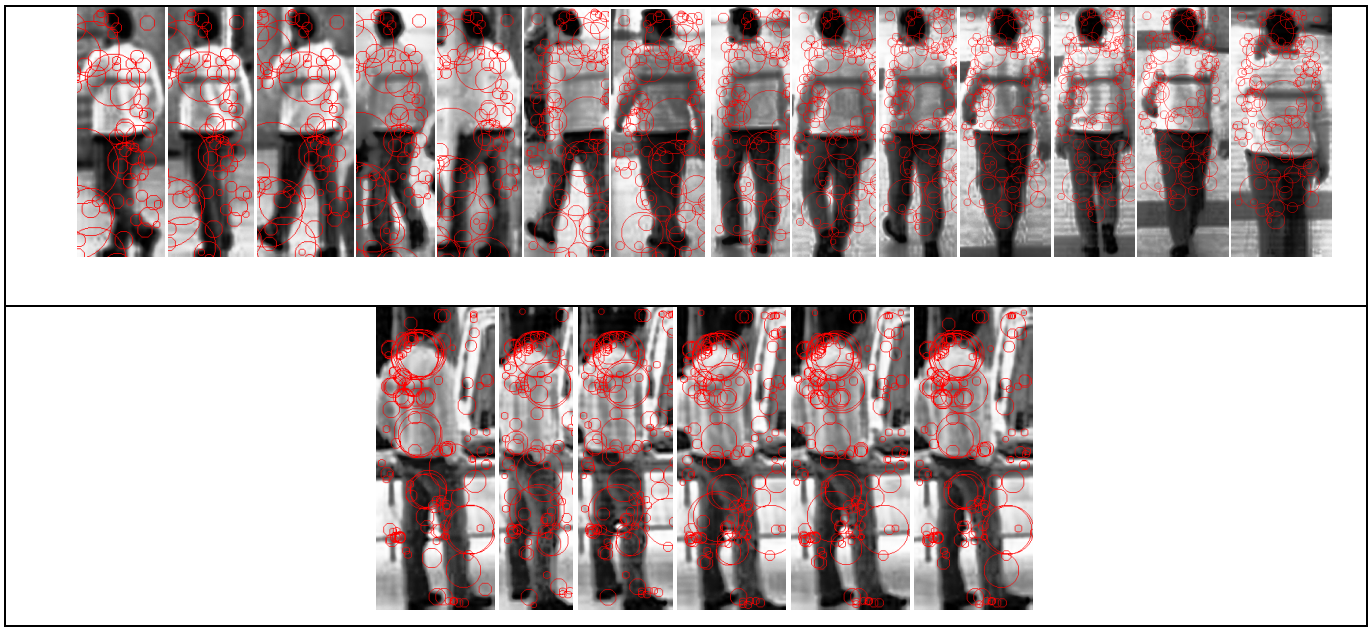


Figure 2: Visualization of detected key-points on 14 of the 21 images for one person's model (top line), and on the 6 images of a successfully matched re-identification query for the same person (bottom-line).

For comparison, [14] which use color histograms for re-identification on another video set including 15 different persons report best results of ~80% positive rate for ~80% true negative proportion among negatives. Also, the best result reported in [6] on a set of videos including 44 different persons is 60% correct best match (and they achieve 80% only when considering if true match is included in the 4 best matches). It is of course difficult to draw any strong conclusion from these comparisons, as the video sets are completely different, with different numbers of persons, which may be more or less similar, but it seems that the order of magnitude of our first evaluation of re-identification performances is rather encouraging.

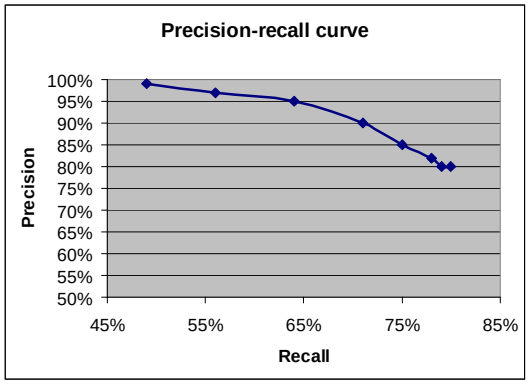


Figure 3: Precision-recall curve in our first person "global" re-identification experiment

In order to quantify the advantage of harvesting interest points on several images, we tried to use various numbers of images per model of person, and checked the impact on obtained precision and recall. As can be seen on figure 4, there is a very clear and significant *simultaneous* improvement of both precision *and* recall when the number of images per model is increased. This validates the interest of the concept of harvesting interest points for the same person on various images.

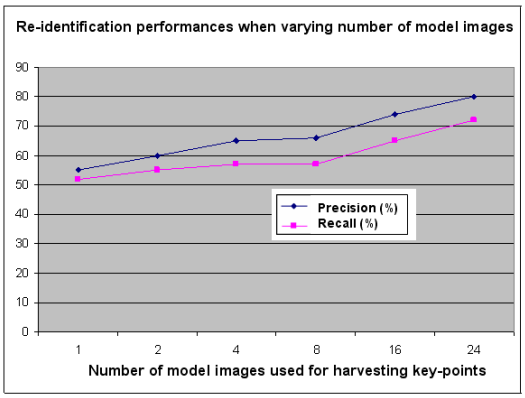


Figure 4: Influence of the number of images used per model on the re-identification precision and recall

It is also important to emphasize the high execution speed of our re-identification method: the computing time is less than 1/8 s per query, which is negligible compared to the 3 seconds necessary to collect the six images separated by ½ s.

Table 2: Total number of interest points, and re-identification computation time as a function of the number of images used for each person model

Number of images used in model sequences	Total number of stored interest points	Computation time for re-identification (ms)
1	1117	123
2	2315	132
4	5154	141
8	11100	149
16	22419	157
24	32854	161

More importantly, due to the logarithmic complexity of the KD-tree search with respect to the number of stored descriptors, the query processing time should remain very low even if large number of person models were stored. In order to verify this, we compared the re-identification computation time when varying the number of images used in model sequences, as reported in table 2.

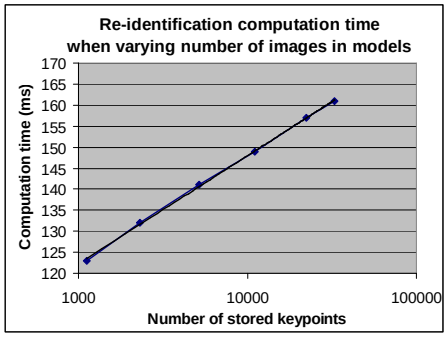


Figure 5: Re-identification computation time as a function of number of stored keypoint descriptors; the dependence is clearly logarithmic

Indeed figure 5 shows that the re-identification computation time scales logarithmically with number of stored descriptors. Since the number of stored descriptors is roughly proportional to the number of images used, if 100 to 1000 person models were stored instead of 10 (with ~ 20 images for each), the KD-tree would contain 10 to 100 times more key-points, i.e. ~ 0.25 to 2.5 millions of descriptors. Extrapolating from figure 5, we therefore expect a query computation time ≤ 1/5 s for a re-identification query among hundreds of registered person models. However, the *reliability* of re-identification among such a high number of persons of course remains to be verified.

4. Feasibility study of application to on-the-fly face identification

To evaluate the generality of our approach, we have also begun preliminary tests for a feasibility study of application to on-the-fly face identification face in real time. We have installed four IP cameras in our research lab in order to build up our own experiment with different cameras.

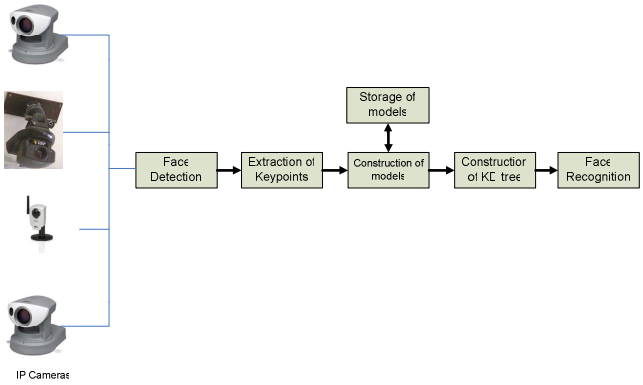


Figure 6: The general diagram of our application to real-time on-the-fly face identification

The general principle is the same as for person global re-identification, except that we need a face detection module to locate face for model harvesting as well as for re-identification by face. Figure 6 shows a general diagram of our on-the-fly face identification system.

For the face detection, we began by using just the standard Viola-Jones face detection algorithm [11] implemented in open source library OpenCV. Figure 7 shows the points of interest extracted by the Camellia keypoints detector on the faces of several persons in our lab. As for global person re-identification, we use several sufficiently different images for building the model of each person.

The quantification of the identification performances of our system is currently still in progress, but we have already qualitatively interesting results, as illustrated on figure 8 where one can see some successful on-the-fly face identification, including some on difficult situations with occlusion, or when a person wears dark sunglasses (while no such image example was present in its model).

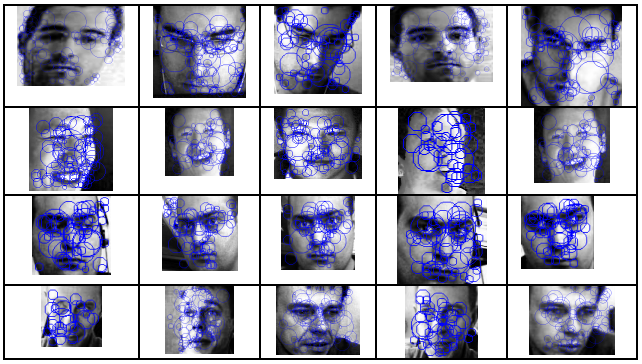


Figure 7: Examples of interest points extracted with Camellia KeyPoint detector inside detected faces used as models

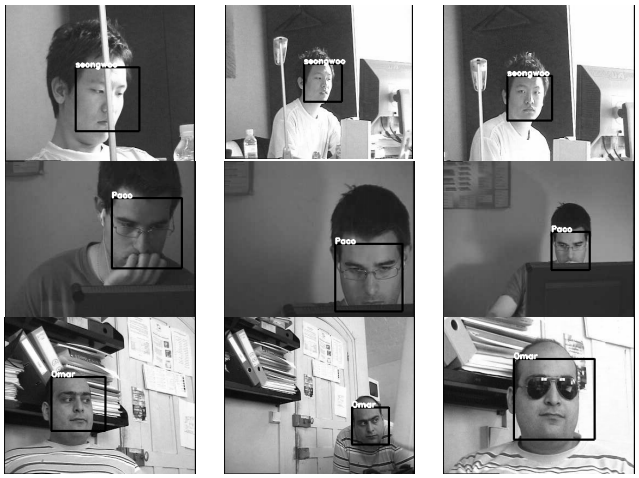


Figure 8: Examples of successful on-the-fly face identification with our interest point harvesting approach

5. Conclusions

We have presented a new re-identification approach based on matching of interest-points collected in query short video sequence with those harvested in longer model videos used for each previously seen and registered "object" (person, face or vehicle).

We have conducted a first evaluation for application to global pedestrian re-identification in multi-camera system on low-resolution videos. This yielded very promising inter-camera person re-identification performances (a precision of 82% for a recall of 78%). It should be noted that our matching method is very fast, with typical computation time of 1/8s for re-identification of one target person among 10 stored signatures for previously seen persons in other cameras. Moreover, this re-identification time scales logarithmically with the number of stored person models, so that the computation time would remain below 1/5 second for a real-world-sized system potentially involving tracking of hundreds of persons.

We have also set up a first feasibility evaluation of application of the same method for on-the-fly face identification, with encouraging successful face identification in difficult situations (occlusion, adding sunglasses, etc…).

More thorough evaluations have to be done for pedestrian re-identification, including on other video corpus, and with more registered persons, which are currently under progress. For on-the-fly face recognition, a quantitative evaluation is underway. Also, our re-identification scheme will soon be integrated in the global video-surveillance processing, which will allow to restrict interest points inside the person area, therefore excluding most background keypoints, which should improve significantly the performance of our system.
Finally, we also hope to further increase performances, either by exploiting relative positions of matched interest points, and/or by applying a machine-learning to built "models".

6. References

[1] Tu, P.; Doretto, G.; Krahnstoever, N.; Perera, A.; Wheeler, F.; Liu, X.; Rittscher, J.; Sebastian, T.; Yu, T. & Harding, K., "An intelligent video framework for homeland protection" *Proceedings of SPIE Defence and Security Symposium - Unattended Ground, Sea, and Air Sensor Technologies and Applications IX*, Orlando, FL, USA, April 9--13, 2007.

[2] Park, U.; Jain, A.; Kitahara, I.; Kogure, K. & Hagita, N., "ViSE: Visual Search Engine Using Multiple Networked Cameras", *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)-Volume 03*, 1204-1207 (2006).

[3] Pham T.; Worring M.; Smeulders A., "A multi-camera visual surveillance system for tracking reoccurrences of people", *Proc. of 1st AC/IEEE Int. Conf. on Smart Distributed Cameras* held in Vienna, Austria, 25-28 sept. 2007.

[4] Lantagne, M.; Parizeau, M. & Bergevin, R. , "VIP : Vision tool for comparing Images of People", *Proceedings of the 16th IEEE Conf. on Vision Interface*, pp. 35-42, 2003

[5] Arth C.; Leistner C.; Bishof H., "Object Reacquisition and Tracking in Large-Scale Smart Camera Networks**"**, *Proc. of 1st AC/IEEE Int. Conf. on Smart*

Distributed Cameras held in Vienna, Austria, 25-28 sept. 2007.

[6] Gheissari, N.; Sebastian, T. & Hartley, R., "Person Reidentification Using Spatiotemporal Appearance", *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006)*-Volume 2, IEEE Computer Society, pp. 1528-1535, New-York, USA, June 17-22, 2006.

[7] Sivic J. and A. Zisserman A., "Video Google: A text retrieval approach to object matching in videos", *Proceedings of 9th IEEE International Conference on Computer Vision (ICCV'2003)*, held in Nice, France, 11-17 october 2003.

[8] Lowe, D. "Distinctive Image Features from Scale-Invariant Keypoints" *International Journal of Computer Vision, Vol. 60*, pp. 91-110, *Springer,* 2004,

[9] Herbert Bay, Tinnr Tuytelaars & Gool, L. V. "SURF:Speeded Up Robust Features", *Proceedings of the 9th European Conference on Computer Vision (ECCV'2006)*, Springer LNCS volume 3951, part 1, pp 404--417, 2006

[10] Beis, J. & Lowe, D., "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces", In *Proc. 1997 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1000-1006, Puerto Rico, 1997.

[11] Viola, P. & Jones, M., 'Rapid object detection using a boosted cascade of simple features', *Proc. CVPR* **1**, 511—518, 2001.

[12] Belhumeur P. N., Hespanha J., and Kriegman D.J., "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection". In *IEEE European Conference on Computer Vision*, pages 45–58, 1996

[13] Draper B., Baek K., Bartlett M.S., and Beveridge R., "Recognizing faces with PCA and ICA". *Computer Vision and Image Understanding: Special issue on Face Recognition*, 91, 2003

[14] J. Orwell, P. Remagnino, G.A. Jones, "Optimal Color Quantization for Real-time Object Recognition" in *'Real Time Imaging*', 7(5) Academic Press, October, pp. 401-414. ISBN/ISSN 1077-2014 (2001).

SKIN, a new family of performant 3D keypoints for view-based 3D objects recognition

Ayet Shaiek, Fabien Moutarde and Arnaud de La Fortelle

Robotics lab (CAOR), Mathématiques et Systèmes, Mines ParisTech, 60 Bd St Michel, F-75006 Paris, France

Tel: (33) 1 40.51.92.92, E-mail: Fabien.Moutarde@mines-paristech.fr

Abstract. We propose, for 3D object recognition from points cloud of single 3D view, a new family of 3D keypoints detectors and descriptors, nicknamed SKIN. Our method makes use of the 2D organization of range data produced by 3D sensor to extract interest points (IPs) based on local shape variation measures. Our novel 3D interest points detectors rely on surface type classification, and combines the Shape Index (SI) – Curvedness (C) map with the Gaussian (H) - Mean (K) map. For each extracted keypoint, a local description is then computed by histogram based signature, in which we join information about the relationship between the reference point normal and its neighbor's normals, with information about the shape index values of this neighborhood. These two new proposed descriptors, IndSHOT and IndThrift, stem from existing descriptors respectively, CSHOT (Color Signature of Histograms of OrienTations) and Thrift. Our surface patch descriptors are used to find matching between query-model view pairs in effective way. Performance evaluation of these detectors and descriptors in terms of stability and repeatability shows the robustness of our new proposed 3D keypoints. Experimental results are presented to validate our recognition approach, which outperforms state-of-the-art 3D keypoints on several public 3D objects databases, e.g. 95% correct recognition (instead of 91%) on Stuttgart benchmark containing 42 objects. Moreover, evaluation on datasets obtained with Kinect™ shows that our SKIN 3D keypoints can reach ~80% correct recognition (among 37 objects) on real noisy and low-resolution depth images.

Keywords: 3D object recognition, Depth images, 3D keypoints detector and descriptor, Mean and Gaussian curvatures, Surface classification, Shape Index.

1 Introduction

There has been strong research interest in 3D object recognition over the last decade, due to the promising reliability of the new 3D acquisition techniques. 3D recognition, however, conveys several issues related to the amount of information, class variability, partial information, as well as scales and viewpoints differences. As previous works in the 2D case have shown, local methods perform better than global features to partially overcome those problems. For extraction of global features, the complete and isolated object shape is needed. These methods are therefore less successful when dealing with partial shape. An example of global 3D feature is volumetric part-based description proposed by Medioni and François (2000). In the case of local approaches, two stages referred to as detection and description, are needed to compute local invariant features. Detection of keypoints deals with the extraction of repeatable and robust salient points from images. Description projects the neighborhood of a keypoint into a proper feature space. Some other tasks that benefit from feature detection are object registration, mesh simplification, and shape segmentation, just to name a few.

The field of 2D Point-of-Interest (PoI) feature has been the source of inspiration for the 3D interest-points detectors. For example, the Harris detector has been extended to three dimensions, first by Scovanner et al. (2007) with two spatial dimensions plus the time dimension, then by Viksten et al. (2008) who discussed variants of the Harris measure and recently in Knopp et al. (2010) where a 3D-SURF adaptation is proposed. Regarding description of 3D shape in a patch around a keypoint, several approaches have already been proposed in the literature. For example, Frome et al. (2004) define a "shape context" around a 3D keypoint, calculated by counting the number of 3D points lying in its neighborhood. Another efficient 3D descriptor, is the SHOT descriptor proposed by Tombari et al. (2010), which is based on the definition of a local, robust and invariant Reference Frame RF, and achieves both state-of-the-art robustness, and descriptiveness. Results demonstrate the higher descriptiveness embedded in SHOT with respect to Spin Images (Johnson and Hebert, 1999), Exponential Mapping (EM) and Point Signatures (PS).

Our application objective is to recognize an isolated 3D object given in a request view, from a training database containing some views of same and other objects. Our idea is to propose a local method that combines some existing approaches in order to improve recognition performance. Our proposed new method aims to detect salient keypoints that are repeatable under moderate viewpoint variations. We propose to use a measure of

curvature in the line of the work of Chen and Bhanu (2007), and construct a patch labeling to classify different surface shapes using both mean-gaussian curvatures (HK) and shape index-curvedness (SC) couples (see Akaganduz et al., 2009). Thus, we select keypoints according to their local surface saliency which leads to four new 3D Interest Point (IP) detectors, named SC_HK_xxx. Furthermore, we define two novel 3D keypoints descriptors: the first one is dubbed IndSHOT and emphasizes the shape description by merging the SHOT descriptor with the Shape Index histogram; our second new descriptor, called IndThrift, extends the 1D Thrift into 2D version by adding the Shape Index information into the final histogram. We name "SKIN" (for Sc_hK points described by shape Index and Normals) the family of 3D keypoints obtained by combining one of our new SC_HK_xxx 3D Interest Point detectors, with one of our new 3D descriptors joining shape-Index with normals information. In this work, we evaluate the effectiveness of the many possible combinations between our approaches and state-of-the-art 3D detectors and descriptors, so as to identify optimal pairs. In §2, we review the state-of-the-art methods for 3D keypoints detection and description considered in our investigation. The complete recognition system with detection, description and matching phases is introduced in §3. Finally, parameters, datasets and evaluation of proposed methods are presented in §4.

2 Related work

In order for the paper to be self-contained, state-of-the-art algorithms, referenced in our work, are briefly discussed in this section.

2.1 Keypoint detectors

The aim of this step is to pick out a repeatable and salient set of 3D points. 3D detectors can be divided into two categories, namely fixed scale and multi-scale detectors. The first kind of detectors finds distinctive keypoints at a specific and constant scale. Multi-scale variants build a scale-space defined on the surface, and a characteristic scale is thus associated to each keypoint: it is found by looking for the maximum of the saliency along the scale dimension. The key stage common to both types of detectors is the selection of keypoints as local extrema of a saliency measure. Principal curvatures correspond to the eigenvalues of the Hessian matrix and are invariant under rotation. The usual pair of Gaussian curvature K and mean curvature H only provides a poor representation, since the values are strongly correlated. Instead, these measures are used in composed form with curvature based

quantities. We now briefly present the principle of some detectors referenced in our work, namely: 3D SURF, 3D Harris, Shape Index, SC map, and HK map.

2.1.1 3D SURF

3D SURF, proposed by Knopp et al. (2010) is a multi-scale detector that constructs scale spaces from the voxelized version of the original mesh. A measure of saliency S is defined on each cubic grid bin over several scales (three octaves). S corresponds to the absolute value of the determinant of Hessian matrix H, encoding second Gaussian derivatives. Preselected IPs are detected only on blob and saddle regions. The final IP selection is performed with a non-maximal suppression algorithm.

2.1.2 3D Harris

Sipiran and Bustos (2011) proposed to extract local maximums of the Harris response calculated on vertices. Two options are suggested to select the final set of interest points:

- selection of points having the biggest Harris response: only a constant portion (ρ) of points of interest are kept. The restriction on extraction only points with big saliency can penalize certain parts of the object. We call this version « Harris_Fract ».
- grouping of points (clustering) which we name « Harris_Clust ». This approach allows to obtain a distribution on the entire surface of the object. The clustering algorithm is recalled below in Algorithm1.

The value of ρ can be considered as a fraction of the diagonal of the object bounding rectangle. These 3D Harris detectors implementations require, basically, two parameters to be set: K used to calculate the Harris response for a given vertex, and the value of the parameter of the final selection.

Algorithm 1 Interest Points Clustering

Require: Set P of pre-selected interest points in decreasing order of Harris operator value
Ensure: Final set of interest points

Let Q be a set of points
$Q \leftarrow \emptyset$
for $i \leftarrow 1$ to $|P|$ **do**
  if $min_{j \in [1,|Q|]} \|P_i - Q_j\|_2 > \rho$ **then**
    $Q \leftarrow Q \cup \{P_i\}$
  end if
end for
Return Q

2.1.3 Shape Index

This detector type was proposed by Chen and Bhanu (2007), and uses the shape index (SI) for feature point extraction. It is a quantitative measure of the surface shape at a point p, and is defined by equation (1):

$$SI_p = \frac{1}{2} - \frac{1}{\pi} \times \text{arctg}\left(\frac{k_p^1 + k_p^2}{k_p^1 - k_p^2}\right) \qquad (1)$$

where k^1_p and k^2_p are respectively maximum and minimum principal curvatures. With this definition, all shapes are mapped onto the [0,1] interval where every distinct surface shape corresponds to a unique value of SI (except for planar surfaces, which will be mapped to the value 0.5, together with saddle shapes). Larger shape index values represent convex surfaces and smaller shape index values represent concave surfaces. The main advantage of this measure is the invariance to orientation and scale. A point is marked as a keypoint if its shape index SI_p satisfies equation (2) within point neighbors:

$$\begin{cases} SI_p = \max(SI_k)\,; k \in neighbors \;\; and \;\; SI_p \geq (1+\alpha) \times \mu \\ \qquad\qquad or \\ SI_p = \min(SI_k)\,; k \in neighbors \;\; and \;\; SI_p \leq (1-\beta) \times \mu \end{cases} \qquad (2)$$

where μ is the mean of Shape Index over neighbors of SI points, and $0 \leq \alpha, \beta \leq 1$. In above expression (2), α and β parameters control the selection of feature points. We denote this detector by « SI ».

2.1.4 HK and SC classification

The idea here (see Cantzler and Fisher, 2001) is to build shape classification space using the pair mean curvature-Gaussian curvature (HK) or the pair shape index-curvedness (SC). Typically, for HK classification, the type function T_p associates, to each couple of H and K values, a unique type value according to equation (3):

$$T_p = 1 + 3\left(1 + sgn_{\varepsilon_H}(H)\right) + \left(1 - sgn_{\varepsilon_K}(K)\right);\ \text{sgn}_{\varepsilon_X}(X) \begin{cases} +1 & if\ X > \varepsilon_X, \\ 0 & if\ |X| \leq \varepsilon_X, \\ -1 & if\ X < \varepsilon_X \end{cases} \qquad (3)$$

where ε_H and ε_K are two thresholds over the H and K. Nine region types are defined.

In the shape index-curvedness (SC) space, S encodes the shape type (SI) and C defines the degree of curvature and is the square-root of the deviation from flatness. The curvedness can be used to indicate how highly or gently curved a surface is. Similarly to HK representation, the continuous graduation of S subdivides surface shapes into 9 types. Planar surfaces are classified using the C value. A type function S_p is defined that associates a unique

type value to each couple of SI and C values (i.e values between 0.8125 and 0.9375 correspond to dome and $S_p = 7$), as described in equation (4):

$$\begin{cases} S_p = 0 \; if \; C \leq \varepsilon_C \\ else \\ S_p \in [1,8] \; ; \; SI \in [\![0,1]\!] . \end{cases} \quad (4)$$

The idea after labeling step is to select vertices in regions with uncommon local structure as opposed to vertices on smooth or nearly planar sections of a surface which have low interest. Thus, for both classifications, salient regions are selected as those of one of the 5 following types: dome, trough, spherical, saddle rut and saddle ridge regions. For instance, in a human-shaped model, those interest point detectors select vertices on the face, hands, and feet. More details can be found in (Akagandunz et al., 2009) and (Cantzler and Fisher, 2001). We denote the two above detectors respectively by «HK» and «SC».

2.2 Keypoint descriptors

After keypoints detection step, a 3D descriptor is computed around each selected interest point. In the case of range data, the dominant orientation at a point is the direction of the surface normal at that point. Histogram-based methods are typically based on the feature point normals. We present briefly the principle of existing descriptors that inspired our own description phase.

2.2.1 LSP

The LSP (Local Surface Patches) descriptor was proposed by Chen and Bhanu (2007) to address recognition of 3D free-form objects from range images. In the original implementation of the LSP, authors calculate local surface properties which are 2D histogram, surface type and the centroid. The 2D histogram consists of shape indexes and angles between the normal of reference point and that of its neighbors. The shape index axis is in the [0,1] range; the second axis is the dot product of surface normal vectors which is in the [-1,1] range.

2.2.2 Thrift

Flint et al. (2007) proposed this 3D descriptor that extends 2D SIFT. The local shape information is described using histogram of the cosine angles between normals. Surface normals are approximated by fitting a least-squares plane to the points in a sphere centered on the point. The descriptor of the interest point z is a histogram

that cumulates, for each neighbor in the support of z, the angle between two normals calculated over two windows of that neighbor.

2.2.3 SHOT

The recently proposed SHOT descriptor achieves computational efficiency, descriptive power and robustness by defining 3D repeatable local Reference Frame (RF). We briefly summarize here the structure of the SHOT descriptor. The reader is referred to (Tombari et al., 2010) for details on the descriptor. Given the local RF, an isotropic spherical grid is defined to encode spatially well localized information. The introduction of geometric information concerning the location of the points within the support is performed by first calculating a set of local histograms of normals over the 3D volumes defined by a 3D grid superimposed on the support, and then grouping together all local histograms to form the final descriptor. The normal estimation is based on the eigenvalue decomposition of a novel scatter matrix defined by a weighted linear combination of neighbour point distances to the feature point, lying within the spherical support. The eigenvectors of this matrix define repeatable, orthogonal directions in presence of noise and clutter. For each sector of the spherical grid a histogram of normals is defined and the overall descriptor SHOT results from the juxtaposition of these histograms. Furthermore, the CSHOT descriptor was proposed by Tombari et al. (2011) as an improvement of the SHOT descriptor by adding texture information to the 3D data. The process of combination succeeds to form more robust and descriptive signature.

3 SKIN: the proposed new family of 3D keypoints

Our detection and description strategy is based on local differential quantities (normals and curvatures) which can be calculated from first and second derivatives, in a local context region. Curvatures can be computed indirectly as the rate of change of normal orientations. In the following, we present the steps of our approach.

3.1 Pre-processing of the 3D points cloud

As we address a recognition scenario wherein only 2.5 views are matched, we deal with some views of the models from specific viewpoints. In the work presented here, we exploit the lattice structure provided by the range image. First, we search the coordinates of the maximum and minimum points at x-axis and y-axis in the sample, and build a bounding box based on the two limit points. Using the (i, j) coordinates of each point in this

box, we determine the two nearest neighbours of each point and we generate a mesh using a triangulation process. The distance between face vertices is thresholded by the mean distance of the points cloud. In the case of Kinect™ databases, we also apply a gaussian bilateral filter (Fleischman et al., 2003) to smooth out noise and quantification artifacts. The main advantage of this filter is to reduce noise while preserving shape details. Afterwards, we construct mesh using the new vertices corresponding to the Gaussian average of a set of nearest neighbour points.

In our approach, neighbour points are given by a spherical region around the point, with a support radius R presenting a proportion r_1 of the bounding box diagonal, so as to make our method robust to different spatial samplings and to scaling. In practice, we adjust a local polynomial surface to the selected neighborhood to calculate differential measures. An advantage of subdividing the point cloud in local regions is to avoid mutual impact between them.

3.2 Proposed SKIN 3D keypoint detectors

Combination of SC and HK criteria

Theoretically, the two classifications HK and SC should provide the same result; however, in practice some areas (like planar regions) are classified differently. Therefore we suggest combining the two criteria to increase reliability. In fact, our result will be validated with two measures of keypoints detection. After labeling points with a pair of value (T_p, S_p), points with salient type pair are selected, in other words, if the two labels values are the same and the surface type corresponds to one of the 5 salient region types previously mentioned. Moreover, in this paradigm, plane surfaces are not taken in account. So we chose to select, in addition to those 5 surface types, planar regions.

Selection of keypoints

Afterwards, the obtained candidate 3D keypoints are ranked according to some criteria, so that we can select only the most relevant ones. We have experimented three different ranking criteria: the curvedness C, a measure of quality factor, and a confidence value of the feature point.

The quality factor was introduced by Mian et al. (2009) and used for ranking keypoints after the detection process. We associate at each point *k* a quality measure Q_k is given by:

$$Q_k = \frac{1000}{r^2} \sum |K| + max(100K) + |min\,(100K)| + max(10\,k_p{}^1) + |min\,(10k_p{}^2)|\,;\;\; K = k_p{}^1 k_p{}^2 \quad (5)$$

where $k_p{}^1$ and $k_p{}^2$ are maximum and minimum principal curvatures, respectively. Summation, maximum and minimum values are calculated over the point neighbours. Absolute values are taken so that positive and negative curvatures do not cancel each other; positive and negative values of curvatures are equally descriptive. In a map of factor quality values of one model, brighter pixels correspond to the highest values of FQ, and are located in descriptive regions within important shape variation.

The confidence value γ was used by Ho and Gibbins (2009) in a curvature based approach for multi-scale feature extraction. They proposed an equivalent 3D version of the 2D Gaussian smoothing kernels in order to estimate the 3D scale space representation. The scale of a point on the surface is defined as the size of the neighbourhood. γ measure computes the deviation measure of curvedness towards the neighborhood. In our case, this confidence value of a feature in point *p* is defined as :

$$\gamma(p) = \frac{\left|c_p - \mu_{N_p}\right|}{\sigma_{N_p}}\;;\; \mu_{N_p} = \frac{\sum_{p_j \in N_p} c_{p_j}}{n};\; \sigma_{N_p} = \sqrt{\frac{\sum_{p_j \in N_p} \left(c_{p_j} - \mu_{N_p}\right)}{n-1}} \quad (6)$$

where N_p is a set of all *n* points in the neighbourhood of p. c_p is the curvedness of p. μ_{N_p} and σ_{N_p} are the mean and standard deviation of the curvedness of all vertices in N_p respectively.

Using different selection criteria and grouping method, we propose two manners for final IPs selection:

a. Using a ranking measure and processing the clustering algorithm (Algorithm 1 is applied by replacing the Harris operator with ranking measure) to filter ordered keypoints. Our ranking measures are: first, curvedness value C, which forms the SC_HK _C detector; second, quality factor FQ which leads to the SC_HK_FQ detector; and third, confidence value γ which forms the SC_HK_Conf detector. The ρ value in Algorithm 1 controls the number of returned interest points.

b. Considering points with the same pair value (T_p, S_p) and group them using the connected- component labeling. Connectivity is carried out by checking the 8-connectivity of each point. Finally, the centers of the connected component are selected as keypoints. The point with the maximum value of curvedness over the

selected keypoints is chosen to represent each connected component. We call the detector combining the two criteria « SC_HK_Conn ».

In the case of planar regions, a big number of points are chosen and are not all really representative of the saliency. In order to have well distributed interest points in the object surface, the proposed idea, here, is to cluster preselected points according to their relative distance and we threshold the distance between final keypoints (as a fraction of the bounding box's diagonal).

Summary of the SC_HK keypoints detection algorithm

The different steps of our combination algorithm are as following:

1. Generate the mesh (delimit the bounding box and compute the mean distance between points)
2. Border points elimination
3. Calculate the couple of salience measures (T_p, S_p) for each patch (size proportional to the diagonal of the bounding box) and select potential IPs
4. Compute values (C or FQ or Conf) for each extracted IPs
5. Regroup and select detected points by one of the two approaches below:
    a. In connected components (T_p, S_p) and select IP with the highest C in the largest components
    b. By ranking and clustering (algorithm 1) according to criterion C, FQ or Conf

Table 1 – Summary of the new proposed 3D keypoints detectors

	Final selection criterion	**Grouping method**
SC_HK_C	Highest curvedness	Ranking and clustering
SC_HK_FQ	Highest Quality factor	Ranking and clustering
SC_HK_Conf	Highest index of confidence	Ranking and clustering
SC_HK_Conn	Highest curvedness	Connected component

3.3 Proposed SKIN 3D keypoint descriptors

In real range data, the density of points cloud on a surface is determined by the camera viewpoint and the relative distance between object surface and the camera. Hence it is important that our descriptor be robust to such changes in sampling density. The most common measure unaffected by the sampling density at one point is the

surface normal at that point. There is a further advantage to using surface normal information which is its relative robustness to changes in viewpoint. Thus, surface normals based descriptor is more robust to changes in scales, viewpoints and sampling density than other descriptors that use only location information such as "Spin images" (Koenderink and Dorn, 1992) for example. Inspired from state-of-the-art descriptors, our contribution is to combine the formalism of some shape-based existing descriptors to form new ones.

3.3.1 IndThrift

The Thrift descriptor is invariant to full 3D rotation due to the use of comparison of surface normals estimated at two different scales. It was shown that combining measures increase descriptor reliability. Thus, we take profit of this advantage and join the LSP and Thrift formalism. We propose a variant of the Thrift descriptor extending the original 1D histogram into a 2D histogram by adding Shape Index information (as illustrated on right of *figure 1*). More clearly, for each interest point z, we define the spherical support with radius R. For each $y \in support(z)$, we define two spherical windows W_1 and W_2 with radius r_{small} and r_{large} respectively:

$$W_1 = \{p \in X : \|p - y\| \leq r_{small}\}$$

$$W_2 = \{p \in X : \|p - y\| \leq r_{large}\} \quad (7)$$

One normal (respectively N_{small} and N_{large}) is associated with each of the above windows. The set of points belonging to each window are fitted to a least-squares plane. The 2D histogram is made as follows: for each point v of the neighborhood of the IP, we increment a compartment of the histogram corresponding to the couple (index shape of v, angle between two normals of v). The cosine axis accumulates the values of the cosine of the angle between the two normals of every point in the support of the IP. The index axis accumulates the values of the index shape of points belonging to the neighborhood. A bilinear interpolation on both axes brings more robustness and stability to this new descriptor. We name this new descriptor IndThrift (to denote Shape Index + Thrift).

3.3.2 IndSHOT

Similarly to the design of CSHOT descriptor, we suggest to concatenate the histogram of shape index values to the histogram of angle values between the reference surface normals at the feature point and the neighbour's ones. First of all, we accumulate point counts into bins according to a cosine function of the angle between the normal

at each point within the matching part of the grid and the normal at the feature point. For each of the local histograms, a coarser binning is created for directions close to the reference normal direction and a finer one for orthogonal directions. In this way, small differences in orthogonal directions to the normal, which are the most informative ones, cause a point to be accumulated in different bins. Secondly, shape index values of the feature point and those of its neighbours relying in the spherical support are grouped into bins. Finally, we merge the shape index values and the cosine values into one descriptor that we call IndSHOT (Shape Index + SHOT). We perform the same process as in the CSHOT to juxtapose the two histograms, where index shape histogram replaces the color histogram (shown in *figure 1*). The final descriptors, composed of (model ID, index shape + cosines histograms, surface type, the 3D coordinates of keypoint), are saved to be used in the matching process.

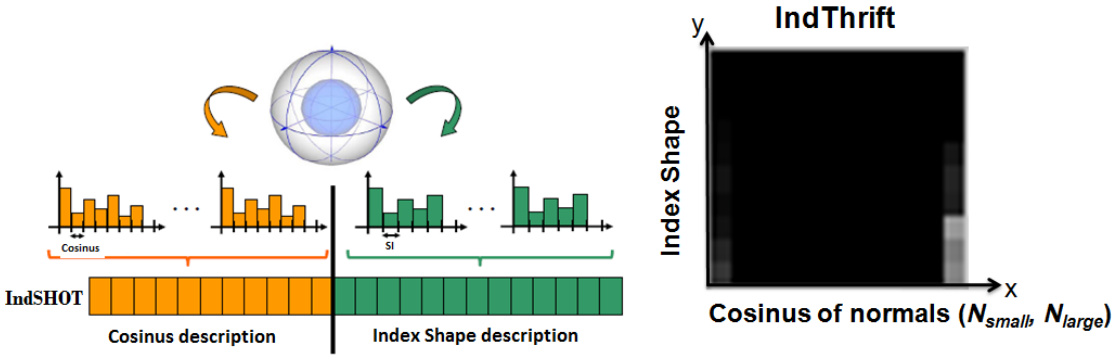


Fig.1. IndSHOT and IndThrift descriptors representation

3.4 Matching process

We are validating the proposed SKIN 3D keypoints detectors and descriptors using a view matching approach. Here, we focus on solving the surface matching problem based on local features, by point-to-point correspondences obtained by matching local invariant descriptors of feature points. Given a test object, we compute a measure of similarity between descriptors extracted on the test view and those of the models in database. The main motivation of histogram matching is its low computational cost. In most cases, similar shape patches are assigned to the same histogram cells. A comparison measure is needed to evaluate the contents of matching cells. We operate as follows:

For each histogram from test view, we find the best matching histogram from database, using the Euclidian distance. To speed up the comparison process, we use a KD-tree structure. Two keypoints are matched according

to their histogram distance and their types of surface. For a test object, a set of nearest neighbors is returned after histogram matching. In the case of multiple correspondences, the potential corresponding pairs are filtered based on the geometric constraint used in (Chen and Bhanu, 2007) and thresholding the Euclidean distance between features coordinates of the two matched surface patches. The closest couple of features in term of coordinates distance is the more likely to form a consistent correspondence. A system of incremental votes for each class gives the final matched class.

4 Experimental results

We evaluate our SKIN 3D keypoints family in two ways: first by assessing separately the repeatability of position of detected keypoints and the stability under rotation and zoom of computed descriptors; and then with evaluation in a specific context of the overall recognition rates of various detector + descriptor combinations, in order to single out the best performing ones. The first part of this section describes data and parameters used in our evaluation. Experimental results on several datasets are then presented to verify the effectiveness of the proposed approach and compare it with state-of-the art.

4.1 Data and Parameters

Datasets

We performed our experiments on four range data sets. 2.5D views for each model constitutes the model libraries, i.e. we focus on 2.5D versus 2.5D object recognition:

a. The first dataset is the public Minolta database from Ohio State University[1] which is captured with a laser scanner laser (Minolta Vivid 900 laser range scanner). For our viewpoint invariance evaluation, we first select 9 objects (top row of Figure 2) then 25 objects (adding 5 objects to the 20 models used in [8]). For scale invariance evaluation, 7 objects are chosen: bluedino (3 scales), brain (3 scales), facesimages (3 scales), gc_bottle (4 scales), reddino (3 scales), valve (4 scales) and yellowhorn (3 scales).

[1] http://cheepnis.cse.nd.edu/~flynn/3DDB/3DDB/RID/index.htm

b. The second dataset is the Stuttgart University Range Image[2], composed of 42 objects. Range images are generated synthetically to have an important angle variation. For each object, 66 poses are taken for training and 258 poses for test which leads to a total of 2772 training poses and 10836 test poses. Angle view differences are estimated to be 23-26° in training and 11.5-13° in test.

c. The third dataset is our own dataset (Lab-dataset) captured with the Microsoft Kinect™ device and composed of 20 objects (Ex. prism, ball, fan, trash can, etc). Depth-maps of 640x480 pixels are generated from several viewpoints (3 to 10 angle views per object). The 20 objects are shown on third row of Figure 2.

d. The last dataset is the public RGB-D Object dataset[3], a set of views of 51 object categories acquired by means of a Kinect™ sensor. There are common household object categories. In our experimentation, we use 37 objects with 25 views per object for only one object per category, which constitute a dataset of 1150 views. The list of the following objects are labelled from 1 to 37 respectively: apple_1, ball_1, banana_1, bell_peper_1, binder_1, calculator_1, camera_1, cap_1, cell_phone_1, cereal_box_3, coffee_mug_1, comb_1, flashlight_1, food_bag_1, food_box_1, food_can_1, food_cup_1, garlic_1, greens_1, hand_towel_1, instant_noodles_1, keyboard_1, Kleenex_1, lemon_1, lightbulb_1, lime_1, marker_1, mushroom_1, notebook_1, onion_1, orange_1, peach_1, pear_1, pitcher_1, plate_1, potato_1, rubber_eraser_1, scissors_1, shampoo_1, soda_can_1, sponge_1, stapler_1, tomato_1, toothbrush_1 and watter_bottle_1.

[2] http://range.informatik.uni-stuttgart.de
[3] http://www.cs.washington.edu/rgbd-dataset/

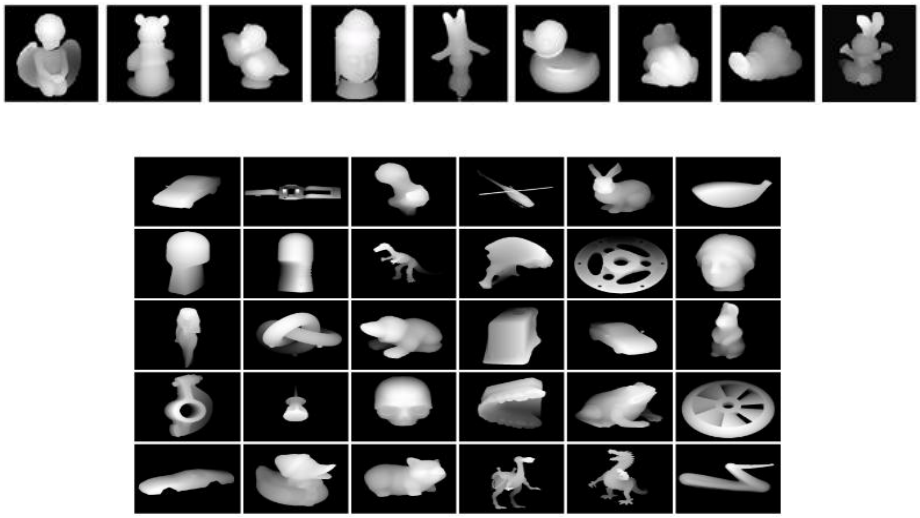


Fig.2: On top line, 9 objects of Minolta public database; on second row, the 42 objects of the Stuttgart public dataset; on third row, the 20 objects of the lab-dataset; on bottom row, 45 objects from the RGBD public dataset

Parameters

The original Spin-image descriptor of Koederink and Doorn (1992) is used in our comparison. As for SURF, Harris and SHOT, we used the publicly available original C++ implementation. For all others tested techniques, we have re- implemented them. In our implementation of the two detectors SC and HK, classification process is

followed by points filtering according to their relative distances using the clustering algorithm1. This post-processing allows the reduction of the final number of keypoints and ensures a correct distribution over the object surface.

Experimental results in which we investigate the effect of different parameters choices in rising methods performance will not be shown in this paper. The numbers of feature points detected from these range images vary from 4 to 250, depending on the viewpoint, the point cloud density and the complexity of input shape. The parameters of our approach are: r_1= 2, $\boldsymbol{\alpha}$ =0.05, β= 0.05. H_{zero}, K_{zero} and C_{zero} are chosen so that $K_{zero} = H^2_{zero}$ and $C_{zero} = K_{zero}$; empirically selected values are 0.01, 0.0001 and 0.01 respectively.

For Harris detector, parameter K = 0.04 and the value of the parameter ρ is equal to 0.01 for Harris_Fract, and to 0.006 for Harris_Clust. Concerning curvature computation and mesh processing, the CGAL[4] and VTK[5] libraries are used.

4.2 Repeatability of detected 3D keypoints

To evaluate detector performance, we first illustrate a visual comparison of keypoint positions detected with SC_HK_FQ, SC_HK_C and SC_HK_Conf detectors, as shown on left 3 views of figure 3. It reveals that the final selected points are quite well localized although the data is noisy (folds of the plastic waste bag, corners of the can). The relative stability of keypoints' positions detected with SC_HK_Conf detector when varying viewpoints for fan object is shown in the last three views of figure 3. Clearly, we recover almost same keypoint positions in the different views, mainly on the fan support and the fan wheel.

[4] http://www.cgal.org/
[5] http://www.vtk.org/

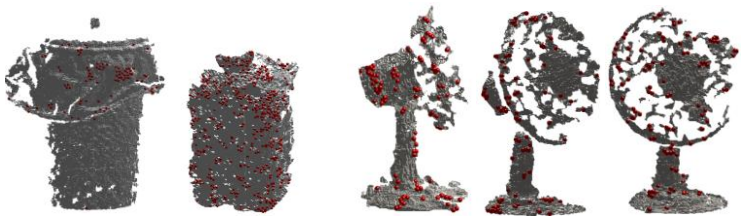


Fig.3. Keypoints detection results on our Lab-dataset for, respectively from left to right: SC_HK_FQ applied to trash, SC_HK_C applied to can, and SC_HK_Conf applied to fan (3 different views).

Keypoint repeatability corresponds to the number of repeatable points between views of the same object. For a quantitative analysis of different detectors repeatability, we operate as follows:

- For each object *O*, real transformation *T* between two views is known. We compute the distance between each detected point on *T*(*O*), and the nearest point (after applying the transformation *T)* detected on *O*. The repeatability of one detector is determined by fixing a threshold on the smallest distance between corresponding pairs keypoints on the two views.
- In order to normalize the comparison between detectors, we normalize distances by dividing them with the maximal distance computed over the whole base and for all detectors.

Evaluation of rotation invariance for our 3D keypoints is performed on the 9 objects of Minolta database. We apply on the 18 views of each object the 10 following detectors: SC_HK_FQ, SC_HK_C, SC_HK_Conf, SC_HK_Conn, Harris_Fract, Harris_Clust, SI, SC, HK and 3D-SURF. A total of 17 transformations are considered. For scale invariance evaluation, we consider the transformation between the original view and the corresponding scaled one of 7 Minolta objects. For each object, 11 views are scaled and the overall repeatability measure is deduced as the mean of the 11 repeatability values. By varying the threshold on the ratio between the distance of the nearest neighbor and a maximal distance, we plot the corresponding repeatability average in figure 4.

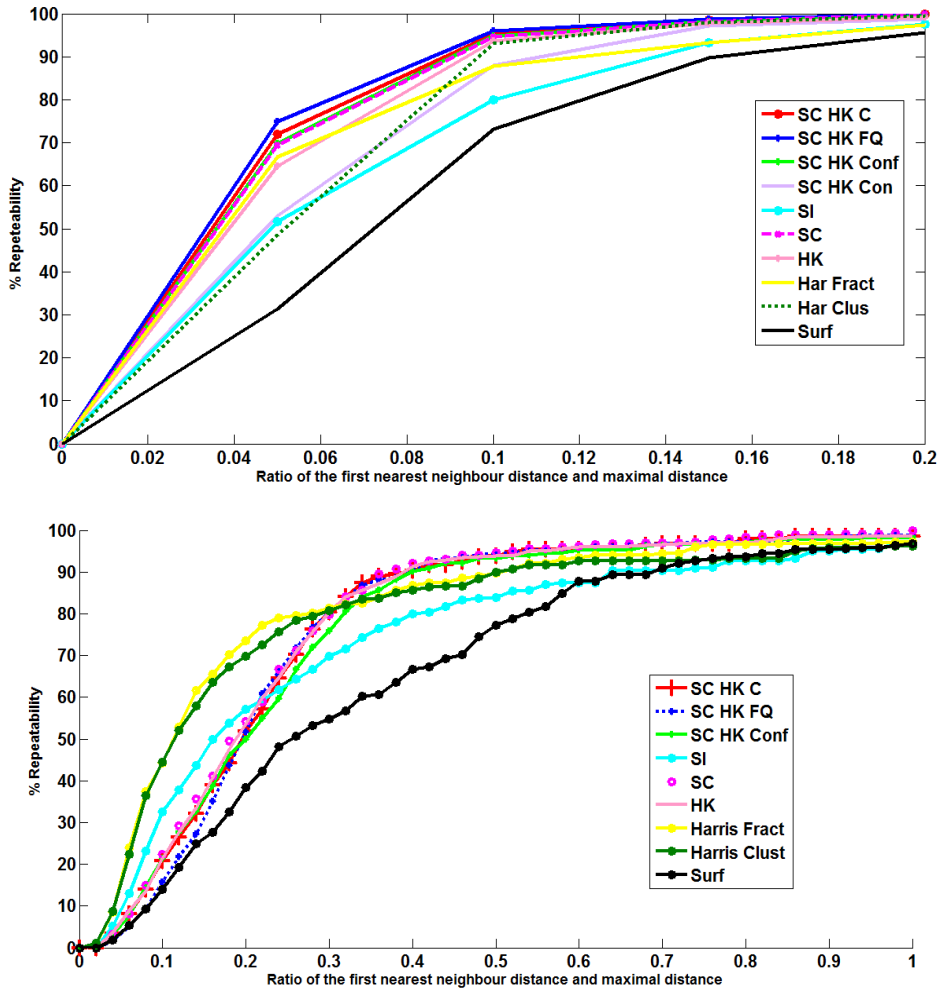


Fig.4. Evaluation and comparison of position repeatability of 3D keypoints detection: on top, for 9 objects of Minolta database under viewpoint variation; on bottom, for 11 views of 7 objects at two different scales.

The three detectors SC_HK_C, SC_HK_Conf and SC_HK_FQ exhibit comparable repeatability performance under viewpoint variation, with SC_HK_FQ curve slightly better than other detectors. We notice that the combining process allows a better feature point filtering than SC or HK alone, as false detected points in both are eliminated, and points with correct surface type remain. The invariance of the measures S, C, H and K to rotation explains the high repeatability of their corresponding based detector. The combination criteria succeeded to raise the performance of single criteria detectors. The 3D-SURF curve is dramatically under the other ones, but it could be explained by the effect of parameters adjustment used to obtain almost the same number of IPs with every detector. The Harris_Fract is better than the Harris_Clust for small threshold values. This behavior is inversed at 0.09 threshold value. For a fixed threshold value equals to 0.16, the five first repeatability performances are:

98.8% for SC_HK_FQ, 98.5% for SC_HK_C, 98.2% for SC, 98.2% for SC_HK_Conf and 97.9% for Harris_Clust. SC_HK_Conn version has the worst repeatability, thus, we will not take into account this detector for the remaining evaluation.

As for repeatability under scale variation, we can observe the clear superiority of Harris_Fract, Harris_Clust and SI detectors for small distance threshold values. The behaviors of SC_HK_C, SC_HK_FQ, SC and HK detector curves are almost equals, with SC_HK_C performing slightly better. The same remark about the Surf exception is valid in our evaluation of the scale impact. In a fixed threshold value equals to 0.4, where the performance of most tested detectors is superior to 80%, we obtain repeatability rates: 89.7% for SC_HK_FQ, 89.7% for SC_HK_C, 87.9% for SC_HK_Conf, 83.9% for Harris_Fract, 81.9% for Harris_Clust,78.2% for SI and 62.0% for 3D-SURF.

These experiments globally show that our new SKIN 3D keypoints detectors are relatively stable under viewpoint and scale variation, and have performances comparable with, or above, state of the art 3D keypoints detectors.

4.3 Stability of 3D keypoints descriptors

The description phase, as well as keypoint detection, has to ensure invariance with respect to geometric transformations. For the purpose of descriptor evaluation, we consider, here again, the relation between descriptors computed in same spatial positions in one original view and in its transformed view.

For a fair comparison, and to avoid detectors errors affecting the descriptors performance assessment, we use a "pseudo-detector" that picks out random set of IPs (about 150 IPs), and compute corresponding descriptors in the original view; then for the same physical positions in the transformed view, we calculate corresponding descriptors. The Euclidian distance is used to match descriptors. We have a matching pair of descriptors if the ratio between distance to first nearest neighbor and distance to second nearest neighbor is inferior to a threshold th_d. A descriptor correspondence is validated (positive) if the spatial positions points are the same. Our descriptor matching experiments relies on Precision-Recall curves, as the total number of negatives is not well-defined in such experiments. Curves are obtained by varying th_d value. In our evaluation of the descriptors rotation stability, we select 5 views, with 20° difference angle, for each object in the 9 objects Minolta dataset, so as to consider 4 rotations around the y axis (for example, a transformation from view 0° to view 20°). Scale invariance is evaluated on two scales of the 7 Minolta models, with averaging over 11 views for each object. The

process described above is applied for both datasets and we show, in figure 5, Precision-Recall curves for angles variation and for scales variation.

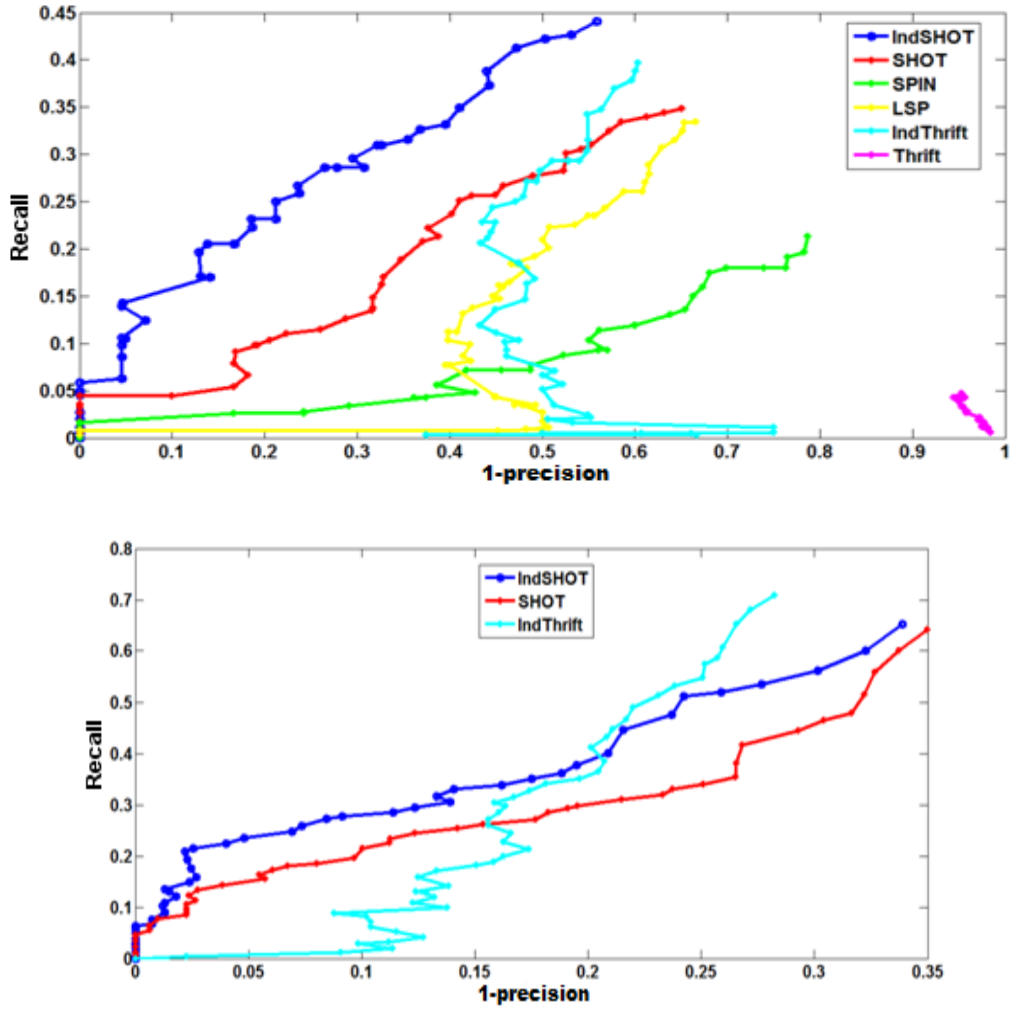


Fig.5. Evaluation and comparison of 3D keypoints descriptors stability: on top Recall vs 1-Precision curves for angles variations on 9 Minolta models; on bottom, same curves for scale variations on 7 Minolta models

In general, the potential of IndSHOT is much higher than that of other tested descriptors. A very significant improvement is apparently made by adding the shape index information to the original version of SHOT. IndThrift has a lower stability in comparison with SHOT and IndSHOT, but succeeds to overcome the SHOT performance starting from a certain value of th_d (>0.65).

Recall that Spin image is a local surface descriptor that encodes two of the three cylindrical coordinates (α, β) of its surrounding points. α is the radial coordinate and β is the elevation coordinate. The Spin images transform isn't invariant to scale, and despite the fact that histograms are calculated relatively to estimated surface normal, it exhibits some robustness to rotation; this descriptor is sensitive to small changes in the computed surface normal, which can explain its low stability in our case.

Rotation invariance is mainly due to the use of local robust frame (RF) in SHOT and IndSHOT formalism. In addition, surface normal is quite invariant to viewpoint and sampling and Shape Index is invariant to orientations. As a result to this assumption, IndThrift and IndSHOT which cumulate normals cosinus and Shape Index values are invariants to rotations. Although it is a normal based descriptor, Thrift has the lowest stability. The absence of RF is the major imperfection of this descriptor. LSP has similar behavior to IndThrift one since both encode almost the same information (normals vs Index Shape). In scale variation results, IndSHOT outperforms the SHOT descriptor, and IndThrift has the best recall-precision values from a certain value of th_d (>0.60). Scale invariance still represents a challenge in our method since we deal with a mono-scale approach.

4.4 Recognition rates

In this section, we carry out experiments to identify the best performing couple of detector + descriptor. Our test protocol for the four databases object recognition is the following:

- For Minolta dataset, we select one test view from the *N* total number of views in the dataset, and the *N-1* views are used as the training set; this process is repeated for the N views of the whole database.
- For the Stuttgart dataset, we train the 66 views and test on 258 views of each object.
- For the Lab-dataset and the RGBD dataset, we select one random view per object as the query and use the remaining views for training. This process is repeated for *n* times.

Recognition results on Minolta Database

We consider the combination of 8 detectors with 5 descriptors. In table 2.a, we show recognition results of those combinations and 3D-SURF detector+descriptor, for angle variation on 25 objects of Minolta database. Overall recognition rates using IndSHOT descriptor are higher than those with other descriptors. The pair SI+IndSHOT is the top-performing couple. The SC_HK_C version of proposed detector combined with IndSHOT performs well

and is in second position. Results of scale variation on 7 objects of Minolta database, in table 2.b, show that the couple (SC_HK_C, IndSHOT) and the pair (SC_HK_FQ, IndThrift) give the same and greatest recognition rate. Besides, we notice that IndThrift seems to be more effective than IndSHOT in scale variation which confirm previous descriptors evaluation result.

Table 2 – Recognition rate (%) results on Minolta public database:
(a) on top, for angle variation on 25 objects; (b) on bottom, for scale variation on 7 objects

Descriptor / Detector	IndSHOT	*SHOT*	IndThrift	*LSP*	*Spin*	*SURF*
SC_HK_FQ	**83.1**	76.4	79.3	69.9	55.9	
SC_HK_conf	**84.1**	74.7	80.2	74.1	57.6	
SC_HK_C	**84.6**	83.7	80.2	67.8	58.2	
SI	**86.6**	*75.4*	81.4	*74.7*	*58.9*	
HK	**83.7**	*72.0*	81.4	*73.7*	*58.7*	
SC	**82.9**	*74.5*	80.2	*71.2*	*56.4*	
Harris_Fract	**83.5**	*83.3*	79.5	*73.3*	*64.5*	
Harris_clust	**79.1**	*73.9*	67.2	*61.0*	*54.9*	
SURF						***82.8***

Descriptor / Detector	IndSHOT	IndThrift	*SURF*
SC_HK_Conf	86.3	**86.7**	
SC_HK_C	**88.6**	86.7	
SC_HK_FQ	86.2	**88.6**	
SI	*79.0*	***80.9***	
Harris_Clust	*67.6*	***68.6***	
Harris_Fract	*79.3*	***82.8***	
SURF			***71.9***

It is very important that keypoints detected in the original surface will be present in the noisy data. In order to evaluate the recognition rate, white Gaussian Noise with four values of the standard deviation was added to the 3D surfaces. Curves of recognition rates, as a function of the 4 noise levels, and corresponding to SC_HK_Conf, SC_HK_C, SC_HK_FQ and SI detectors using IndSHOT descriptor and IndThrift descriptors are displayed in figure 6. When noise is introduced to the point clouds, the details of the shape are less visible and the variation of the local surface patches will increase. As a result, more features points are detected in noisy images compared to the original one. IndThrift curves decrease less rapidly than the IndSHOT curves which indicate that IndSHOT is slightly more robust to noise. Performances of our three detectors are quite similar and show better robustness than SI detector when associated with the IndThrift descriptor.

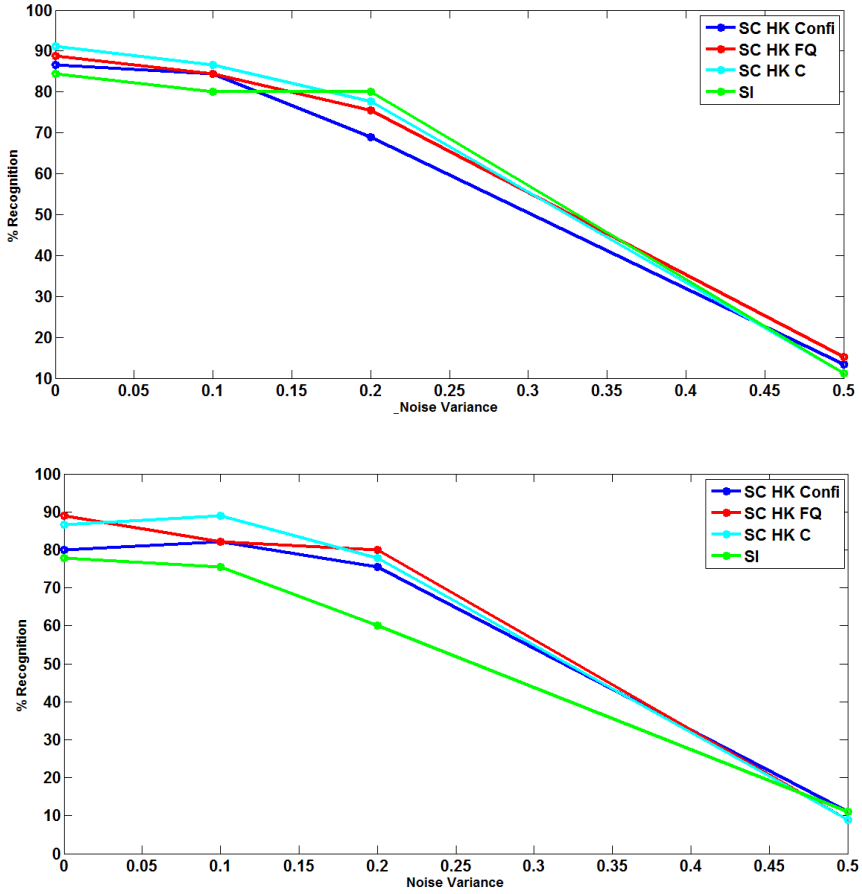


Fig.6. Curves of recognition rates as a function of 4 noise levels, for SC_HK_Conf, SC_HK_C, SC_HK_FQ and SI detectors: on top, combined with IndSHOT descriptor; on bottom, combined with IndThrift descriptor.

Recognition results on Stuttgart Database

Hinton diagram on the 42 objects of Stuttgart database using the couple SC_HK_C+IndSHOT is shown in figure 7. The conjunction of the SC_HK_C detector with the IndSHOT descriptor seems to provide a pertinent description of the local surface typology. In fact, a comparison between our approach and existing methods prove this conclusion: we obtain 94.9% of correct recognition rate with this combination, to be compared with 90.97% obtained with HK method (Eskizara, 2009) and 91.03% with SI approach (Hozatli, 2009). The lowest recognition rates in the diagram correspond to mechanical objects which are planar dominant surface object. The risk, here, is redundant information that leads to confusion in matching process.

It should also be noted that the overall computation time for our recognition process (detection+description+matching features) is similar to fastest existing 3D keypoints: about ~9.62s for the

combination (SC_HK_C+IndSHOT) on Lab-dataset, versus 10.13s for the 3D-SURF detection+description. This is essential since we aim at near real-time applications for object recognition.

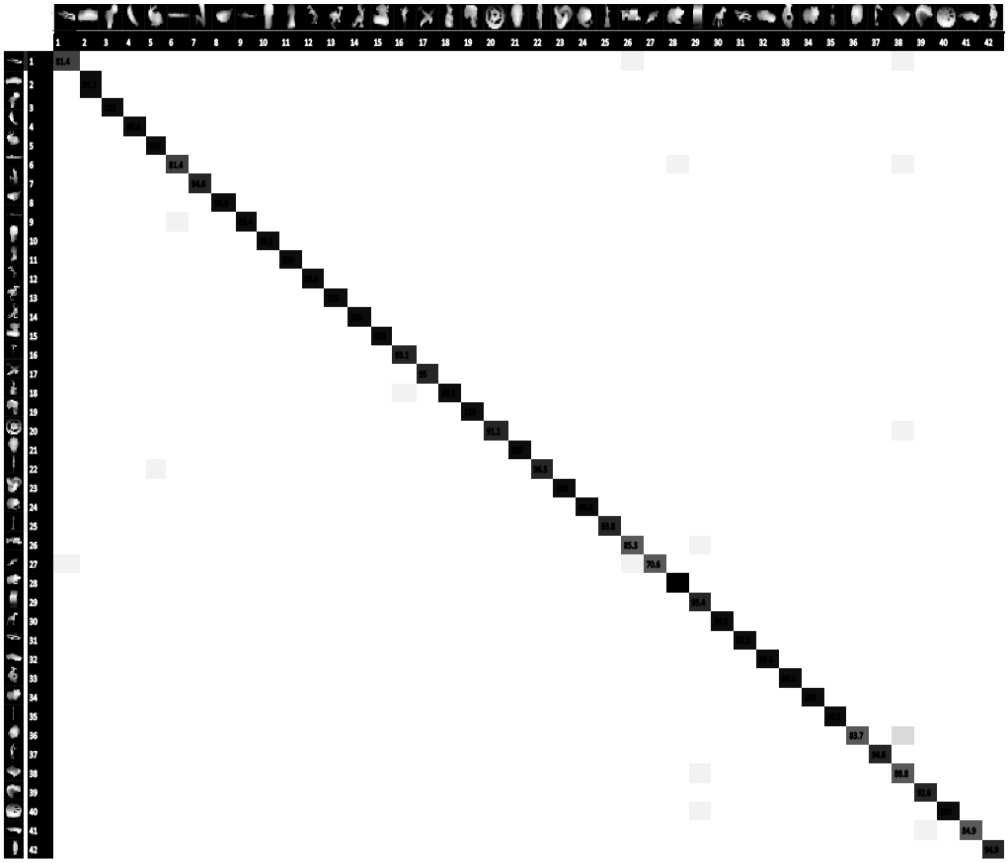


Fig.7. **Hinton diagram for 42 models of Stuttgart database. The global recognition rate is 94.9%, to be compared with ~91% obtained by previously published results on same benchmark**

In conclusion, experiments on the two first databases allowed comparison of our SKIN 3D keypoints to state-of-the-art techniques and published results, and show that they have better performance.

4.5 Applicability to Kinect™ data

A drawback of two evaluations presented above is that they are done either on very high-resolution real data (Minolta scanner dataset), or on synthetic perfect data (Stuttgart benchmark). Since our application aim is to use low-cost real-time 3D sensor for object recognition, it is essential that we also evaluate applicability of our SKIN 3D keypoints to this kind of noisy and low-resolution 3D data.

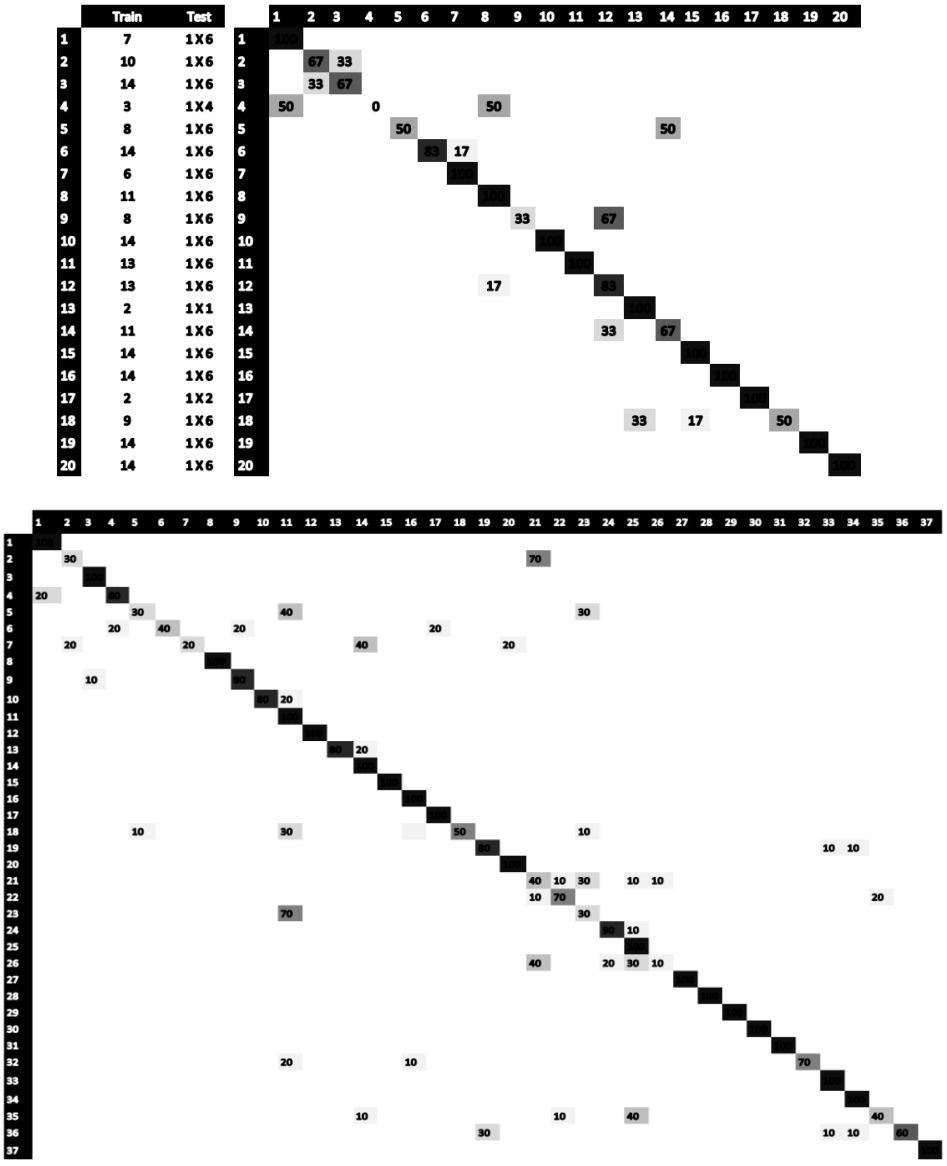


Fig.8. Hinton diagram using SC_HK_FQ detector and IndThrift descriptor: on top, for our Lab-dataset (80%); on bottom, for 37 objects of RGB-D public dataset (78.1%)

We therefore carry out experiments also on our Lab-dataset, and on RGB-D public dataset, using the three versions of proposed detectors combined with the two proposed descriptors IndSHOT and IndThrift. Overall recognition rates using SC_HK_FQ+IndThrift were better, thus, we will present in the following only the best recognition rates. In figure 8, the cross recognition rates between models are displayed in the confusion matrix. Gray level determines the rate of the recognition. Black is for high and white is for low recognition rate. The overall recognition rate is quite promising for our conjunction SC_HK_FQ and IndThrift method with 78.1% on

the RGB-D dataset. The recognition rate in the Lab-dataset is about 80%. The reason behind this lower result is the high similarity between object shapes included in this dataset (two boots objects, parallelepipedic shapes, cylindrical shapes, etc). It should be noted that the recognition rate varies according to the view angle chosen for the query, with higher rates are achieved when the query‘s view angle is given between two view angles in the training base. Finally, the noisy character of those two databases makes harder the detection and description phases.

5 Conclusions and Perspectives

The contribution presented is the proposal of “SKIN”, a new family of 3D keypoints detectors and descriptors. This family consists in the combination of: 1/ original 3D keypoint detectors, SC_HK_C, SC_HK_FQ and SC_HK_Conf, based on the idea of combining 2 criteria of shape type; 2/ new 3D keypoint descriptors, IndSHOT and IndThrift, based solely on shape characteristics.

The proposed detectors combine SC (shape curvedness) and HK criteria with ranking criteria based either on Curvedness, on confidence, or on Factor Quality. It was already shown in our previous work that the selected 3D keypoints are more repeatable than for alternative detectors, and this is confirmed here by the good inter-view matching reached in our experiments. The proposed IndSHOT descriptor encodes the occurrence frequency of shape index values vs. the cosine of the angle between the normal of reference feature point and that of its neighbours. It seems to be significantly more descriptive than original SHOT from which we have crafted it. Similarly IndThrift is a descriptor combining Shape Index information with the original Thrift descriptor.

Finally, our SKIN new 3D keypoints family is evaluated on challenging 3D object recognition scenarios characterized by the presence of viewpoint variations and a few number of views on real-world depth data. Our experiments on public datasets show that SKIN 3D keypoints outperform state-of-the-art: for example the combination of SC_HK_C detector + IndSHOT descriptor allows 94.9% correct recognition (to be compared with ~91% for other published works) on the 42 objects of public Stuttgart benchmark. Our SKIN 3D keypoints are shown to perform well also on real low-resolution and noisy depth images crom Kinect™ sensor: using the combination of SC_HK_FQ detector + IndThrift descriptor, the outcome is very promising, with 78.1% correct

recognition on 37 objects from RGBD public dataset, and 80% on our own lab dataset containing 20 “everyday” objects (some of which are rather similar one to another).

The main drawback of current version of our SKIN 3D keypoints is that detection operates at a fixed scale, chosen as a fraction of the object bounding box. However, we are currently working on a multi-scale version of our detectors to overcome this limitation, and provide a specific scale associated to each 3D keypoint.

References

Akagunduz E., Eskizara O. and Ulusoy I., 2009. "*Scale-space approach for the comparison of HK and SC curvature descriptions as applied to object recognition*", Proc. Int. Conf. on Image Processing (ICIP), 413-416.

Cantzler H. and Fisher R. B., 2001. "*Comparison of HK and SC curvature description methods*", Proc. Conference on 3D Digital Imaging and Modeling, 285-291.

Chen H. and Bhanu B, 2007. "*3D free-form object recognition in range images using local surface patches*," Pattern Recognition Letters, Vol. 28(10), 1252-1262.

Eskizara Ö., 2009. "*3D Geometric Hashing Using Transform Invariant Features*", M.Sc. thesis, Electrical and Electronics Engineering Department, Middle East Technical University.

Fleishman S., Drori I. and Cohen-Or D., 2003. "*Bilateral mesh denoising*", ACM Transactions on Graphics (TOG), 3 : vol. 22. - pp. 950-953 (July 2003).

Flint A., Dick A. and van den Hengel A., 2007. "*Thrift: Local 3D Structure Recognition*", Proc. Digital Image Computing: Techniques and Applications (DICTA'07), 182–188.

Frome A., Huber D., Kolluri R., Bulow T. and Malik J., 2004. "*Recognizing objects in range data using regional point descriptors*". In Proc. 8th European Conf. on Computer Vision (ECCV’2004).

Ho H.T. and Gibbins D., 2009. “*A Curvature-based Approach for Multi-scale Feature Extraction from 3D Meshes and Unstructured Point Clouds*”, IET journal on Computer Vision, 4: Vol. 3, 201-212.

Hozatli A., 2009. "*3D Object Recognition by Geometric Hashing for Robotics Applications*", M.Sc. thesis, Electrical and Electronics Engineering Department, Middle East Technical University.

Johnson A.E. and Hebert M., 1999. "*Using spin images for efficient object recognition in cluttered 3D scenes*," IEEE PAMI 21, 433-449.

Knopp J., Prasad M., Willems G., Timofte R. and Van Gool L., 2010. "*Hough Transform and 3D SURF for robust three dimensional classification*", Proceedings of the 11th European Conference on Computer Vision (ECCV'2010).

Koenderink J. and Doorn A. J., 1992. "*Surface shape and curvature scale*", Image Vis. Comput., vol. 10, no. 8, 557–565.

Medioni G.G. and François A.R.J, 2000. "*3-D structures for generic object recognition*", Computer Vision and Image Analysis, Vol. 1, pp. 30-37.

Mian A., Bennamoun M. and Owens R., 2009. "*On the Repeatability and Quality of Keypoints for Local Feature-based 3D Object Retrieval from Cluttered Scenes*", International Journal of Computer Vision, vol 89: 2-3, pp. 348-361.

Sipiran I. and Bustos B., 2011. "*Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes*", Int. J. Vis. Comput., vol. 27, pp. 963–976.

Tombari F., Salti S. and Di Stefano L., 2010. "*Unique Signatures of Histograms for Local Surface Description*", Proc. 11th European Conference on Computer Vision (ECCV'2010).

Tombari F., Salti S. and Di Stefano L., 2011. "*A combined texture-shape descriptor for enhanced 3D feature matching*"; IEEE 18th International Conference on Image Processing (ICIP'2011.

Scovanner P., Ali S. and Shah M., 2007. "*A 3-dimensional SIFT descriptor and its application to action recognition*", Proceedings of the 15th International Conference on Multimedia, 357–360.

Viksten F., Nordberg K. and Kalms M., 2008. "*Point-of-Interest Detection for Range Data*", Proc. of 19th International Conference on Pattern Recognition (ICPR'2008), 1-4.

Statistical Traffic State Analysis in Large-scale Transportation Networks Using Locality-Preserving Non-negative Matrix Factorization

Yufei HAN and Fabien Moutarde

CAOR, MINES-ParisTech, 60 Boulevard Saint-Michel, 75006, Paris

Yufei.Han@mines-paristech.fr, Fabien.Moutarde@mines-paristech.fr

Abstract

Statistical traffic data analysis is a hot topic in traffic management and control. In this field, current research progresses focus on analyzing traffic flows of individual links or local regions in a transportation network. Less attention are paid to the global view of traffic states over the entire network, which is important for modeling large-scale traffic scenes. Our aim is precisely to propose a new methodology for extracting spatio-temporal traffic patterns, ultimately for modeling large-scale traffic dynamics, and long-term traffic forecasting. We attack this issue by utilizing Locality-Preserving Non-negative Matrix Factorization (LPNMF) to derive low-dimensional representation of network-level traffic states. Clustering is performed on the compact LPNMF projections to unveil typical spatial patterns and temporal dynamics of network-level traffic states. We have tested the proposed method on simulated traffic data generated for a large-scale road network, and reported experimental results validate the ability of our approach for extracting meaningful large-scale space-time traffic patterns. Furthermore, the derived clustering results provide an intuitive understanding of spatial-temporal characteristics of traffic flows in the large-scale network, and a basis for potential long-term forecasting.

Keywords: LPNMF, Statistical Analysis, Network-level Traffic State

1. Introduction

Most traffic information systems make use of floating-car data collected from distributed probing vehicles [1][2][3] as a major data feed for quantitative traffic state evaluation. Acquired floating-car data are aggregated over small time periods (5-15 mn) to estimate traveling time of vehicles, in order to identify traffic states (congestion or free-flowing) for each link. For urban transportation network of decent scale, the traffic management department processes real-time traffic information from thousands of links simultaneously, which is an overwhelming task. Therefore, automatic analysis of traffic information, e.g. unveiling characteristics of traffic flow variations [4][5][6][7][8] is necessary for efficient management strategies and adjusting demands of traffic sources.

Most published works on traffic data analysis focus on modeling temporal dynamics for individual links (either in arterial networks or highways) using model-driven [9][10][11][12][13] or data-driven methods [14][15][16][17][18][19][20][21][22][23][24]. The model-driven methods, like Cellular Automata [12] and other underlying physical models [9][10][11][13], are usually equipped with parameters that are calibrated with structural assumptions to simulate temporal evolution of traffic states. Excellent and as they are for modeling free way or arterial links, the model-driven methods present less efficiency in modeling urban traffics. The velocity flow field of urban transportation is easily subject to the fluctuations induced by intersections of links, traffic signals at the crossings and so on. These fluctuations lead to spatio-temporal traffic events. It is thus difficult to find a local stationary regime for the velocity using the physical models. In contrast, data driven approaches describe statistical dependencies using the block-box machine learning methodologies. They are more popular due to relaxation of prior assumptions during modeling traffic dynamics. Kalman filter [25] and ARMA (Autoregressive Moving Average) [26], originated from

state space theory, are used to predict temporal variations of traffic flows [14][15][16][17][18]. It is an extension of these linear models from sequential signal processing to traffic domain. Efficient as they are for short-term temporal prediction, they can not track nonlinear fluctuations of traffic flows. In [19][20][21][22], neural networks [27] and hybrid non-linear dynamic systems are used to approximate short-term non-linear variations of traffic states. Due to the intrinsic multiple-input and multiple-output (MIMO) structures, neural networks can integrate spatial-temporal correlations between local links into a computational framework. In [23][24], spatial correlations between local links are considered in Markov Random Field [28] and Multi-Agent System [29] based traffic models. These inspiring works concatenate global structural information of transportation networks to improve descriptive power of traffic flow models.

Extending both the model-driven and data-driven methods to large-scale urban network, we need to tackle curse-of-dimensionality caused by enlarged scale of the modeling target. Increasingly more parameters are needed to capture details about temporal dynamics of thousands of links, which increases intrinsic complexity of the built traffic model. Therefore, it is necessary to introduce regularization terms, providing prior knowledge about global configuration of traffic states over the entire network and serve as consistency constraints of the spatio-temporal congestion structure, e.g. co-occurrence of congestion in the network during specific time intervals. Furthermore, local regions with independent traffic flow behaviors can be treated separately in a divide-and-conquer manner. Therefore, mining the spatial configuration patterns of congestion and large-scale macroscopic traffic dynamics over the entire network is highly informative for constructing the computationally tractable models to describe local fringes of large-scale traffic scenes. Such macroscopic view of traffic flow configurations can be also used to identify bottleneck of transportation networks, in order to improve traffic management strategies. Besides, drivers can make use of the global traffic state information to optimize their

traveling plan ever before they leave their own garage. Nevertheless, little progress has been reported on analyzing global congestion configurations of large networks. We attack this issue by performing clustering analysis of traffic configurations over the entire large-scale network simultaneously. We define the network-level traffic state as a multi-dimensional vector containing traffic states of all local links. A matrix factorization based dimension reduction method named as Locality-Preserving Non-negative Matrix Factorization (LPNMF) [39], is adopted to derive a compact representation of high-dimensional network-level traffic states. K-means clustering [48] performed on the derived compact LPNMF representation provides intuitive understanding of typical spatial configuration patterns of global traffic states and large-scale traffic dynamics of network-level traffic states contained in the data. The flowchart of this work is illustrated in Figure 1.

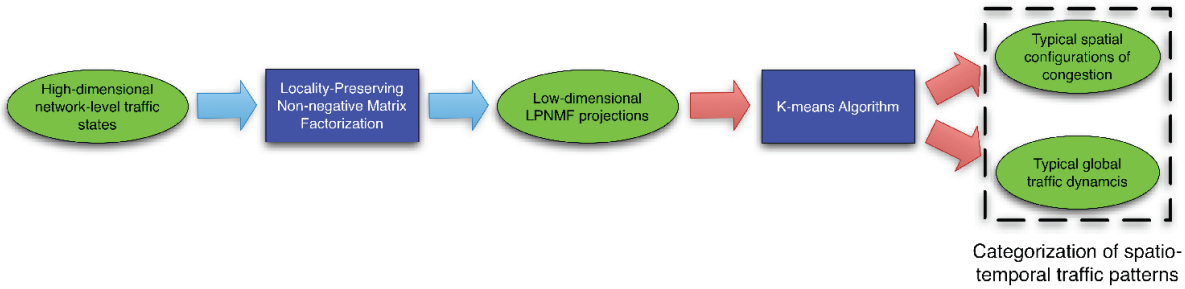


Figure 1. *The flow chart of the proposed clustering methodology*

This article is organized as follows. Section 2 introduces LPNMF employed in the analysis. Section 3 presents the simulated traffic data of a large-scale urban network, used as data source in the following analysis. In Section 4, we illustrate detailed clustering results of spatial configuration patterns obtained through the LPNMF projection. Based on the LPNMF based representation, Section 5 further performs a clustering analysis on temporal behaviors of network-level traffic states. Section 6 draws some conclusions and discusses our future work.

2. Traffic data mining with Locality-Preserving Non-negative Matrix Factorization

2.1. Basic scheme of Non-negative Matrix Factorization (NMF)

LPNMF is an extension of basic NMF [30][31][32][33] by introducing constraints on topological structures of the derived projection subspace. In this section, we introduce basic principles of extracting a flexible representation from original network-level traffic states using the NMF-like scheme. NMF is a particular matrix factorization algorithm, which is in the same family of techniques as the well-known PCA (Principle Component Analysis) [34]. As mentioned in Section 1, i-th entry of a network-level traffic state corresponds to the traffic state on i-th link of the network. In urban traffic sequences, the number of links in any network of decent scale is often over one thousand. Thus, the network-level traffic state has a rather dense data structure. Assuming m samples of n-dimensional network-level states are stored into the column space of a $n \times m$ matrix **X** , NMF factorizes **X** as a product of a $n \times s$ non-negative loading matrix **M** and a $s \times m$ non-negative scoring matrix **V**, in order to minimize the Frobenius norm [35] of the reconstruction error between **X** and the product of **M** and **V**:

$$(M,V) = \arg\min_{M \geq 0, V \geq 0} \|X - MV\|_F \tag{1}$$

Each column of **V** is the NMF projection of the corresponding network-level traffic states. s is the dimensionality of the NMF projection subspace that is spanned by columns of **M**. Normally, s is set to be much less than the row length of **X**, thus **V** forms a low-dimensional representation of network-level traffic states. The specificity of NMF is the non-negativity constraint on **M** and **V**. Each network-level traffic state $\mathbf{X}_j \in \mathbf{R}^n$ is approximated by an additive linear superposition of the column space of **M** in NMF [31][32] as in Eq.2:

$$X_j = \sum_{i=1}^{s} M_i V_{i,j} \quad (2)$$

where $\mathbf{X_j}$ and $\mathbf{M_i}$ is the *j-th* column of **X** and the *i-th* column of **M** respectively. $V_{i,j}$ is the element located at the *j-th* column and *i-th* row of **V**. Treating columns of **M** as the learned base for reconstructing the network-level traffic states, they represent typical structural patterns of traffic configurations. $V_{i,j}$ represents to which degree the *j-th* network-level traffic state vector is associated with the spatial configuration pattern of local traffic states represented by $\mathbf{M_i}$. For example, if $\mathbf{M_i}$ can better represent the *j-th* network-level traffic state, $V_{i,j}$ will take the largest value in the *j-th* column of **V** [36]. The row-wise average $\frac{1}{m}\sum_{j=1}^{m} V_{i,j}$ evaluates the importance of the corresponding NMF basis vectors $\mathbf{M_i}$ in representing the spatial congestion configuration. In this sense, the additive combination shown in Eq.2 leads to a part-based decomposition of the network-level traffic states. Localized groups of entries in each basis vector $\mathbf{M_i}$ with distinctively large magnitudes indicate typical patterns or important components of the original data representation. Benefited from the property, NMF is usually used for extracting semantic components of objects from images [30][32] and latent topics from text documents [36][39]. Motivated by the sounding properties of NMF, we use it rather than PCA to investigate spatial patterns and dynamic properties of network-level traffic states.

An iterative procedure, named as MU (Multiplicative Update) [30][33], is used to solve NMF optimization. In each iteration of MU, either of **M** or **V** is fixed alternatively, the other one is then updated by solving a non-negativity constrained least square problem based on KKT theorem [30][31]. Given the dimensionality of NMF projection as *s*, each iteration of MU has a computation cost in $O(nsm)$. As reported in [30], MU converges to the optimum solution with definite iterations. In our work, MU generally takes 600 iterations before its convergence, much less than the number of samples

contained in the data matrix. Therefore, MU has better computational efficiency than SVD in our case. With finer tuning of matrix multiplication using Strassen's algorithm or Coppersmith-Winograd approach [30] [33], computational efficiency of MU can be improved in a further step.

2.2.Locality-Preserving Penalty for Non-negative Matrix Factorization

For clustering analysis, we want that geometrical structure of the projection space is consistent with the intrinsic characteristics of the traffic data. In particular, we need that projections of network-level traffic states are close to each other, if they are similar in the original high-dimensional space. The consistence of geometrical structures (distance measures between data points) is of utter importance for clustering analysis and dynamic modeling [37][38]. Any artifacts introduced into distance measures in the projection space could change cluster assignments or temporal dynamic patterns. Motivated by this idea, we propose to use a regularized NMF [39][40] to derive the representation of network-level traffic states, named as LPNMF. It aims to minimize the following objective function O, as shown in Eq.3:

$$O = \|X - MV\|_F^2 + \lambda Tr(VLV^T) \tag{3}$$

$$L = D - W \tag{4}$$

$$D_{i,i} = \sum_j W_{i,j} \tag{5}$$

where $Tr$is the trace of a matrix and λ is the regularization parameter. The first term is the Frobenius reconstruction error as illustrated in Eq.1, while the second one is the structural regularization of NMF projections. In this term, **L** is called Graph Laplacian [41][42] as defined in Eq.4. In the matrix **W**, the element $W_{i,j}$ located at i-th row and j-th column, is the pair-wise similarity measure matrix between the i-th and j-th network-level traffic state vectors, corresponding to the i-th and j-th

column of **X**. According to Eq.5**, D** is a diagonal matrix whose entries are column sums of **W**. Graph Laplacian originates from spectral graph theory [43][44][45]. By adding the Graph Laplacian based constraints, the obtained low-dimensional representation **V** is calibrated to have similar geometrical structures as the original data **X** without increasing further computation cost. Based on this property, we can unveil the characteristics of global traffic states more efficiently in the low-dimensional manifold **V** without loss of intrinsic data distribution information.

2.3. Distance measure between network-level traffic states

To perform LPNMF, we need to define a similarity measure between network-level traffic states that evaluates differences between spatial configurations of local traffic states. The traffic state of one link is usually closely correlated with its up-stream or downstream nearest neighbors with the same orientation of traffic flows. For example, the links u_i^j and d_i^m are upstream and down-steam nearest neighbors of the link *i* respectively. Assuming the link *i* fell into heavy traffic congestion, the links u_i^j and d_i^m are more likely to be congested than those far from the link *i*. Motivated by the property, we propose a weighted fusion scheme among traffic states of geometrical neighborhoods to derive the similarity measure. We firstly calculate link-wise differences of traffic states between corresponding links. For each link *i*, we then obtain a weighted sum of the link-wise difference values with respect to the link *i* and its neighbors, which is defined to be local variation v_i of traffic states around the current link, as denoted in Eq.6:

$$v_i = \sum_j w_j^u a(u_i^j) + \sum_m w_m^d a(d_i^m) + w^i a(i) \tag{6}$$

a is the link-wise difference of traffic states between the corresponding link. w_j^u, w_m^d and w^i are the weights respectively attached to up-stream neighbors, downstream neighbors and the current link *i*. After that, we map $\{v_i\}$ into [0,1] using a Gaussian

kernel in Eq.7 as the similarity measure between two network-level traffic states:

$$S = e^{-\frac{\sum_i v_i}{2\delta^2}} \tag{7}$$

To normalize range of the weighted sum, the sum of all weights is required to be 1. The weight w^i corresponding to the link *i* should be the largest one. Weights of the neighboring links can be designed to be proportional to degrees of traffic state correlation between one specific neighboring link and the current link *i*. In this article, all neighboring links are evaluated with the same weight value. By performing weighted fusion of local neighborhoods in the network, the derived similarity measure not only represents traffic state variations between corresponding links but also indicates the spatial correlations between local neighborhoods. We feed this similarity measure into the matrix **W** in Eq.4 and derive the regularized low-dimensional representation of the network-level traffic states.

3. Metropolis simulation and IAURIF database

3.1 Metropolis traffic simulation software

The benchmark IAURIF database used to verify the validity of the proposed LPNMF based method is generated by simulating real-traffic sequences of a large-scale traffic network using Metropolis [46][47]. Metropolis is a planning software designed to model urban transportation systems. It allows the user to study impacts of transportation management policies for metropolitan areas and their fringes in a time-dependent framework. Metropolis simulates commuters' traveling behaviors and congestion in urban areas. The core of the simulation system is a dynamic simulator that integrates joint commuters' departure time and their choices of routes in the transportation network [46]. During simulation of traffic sequences, each commuter is characterized by specific parameter values individually [46][47]. At any moment, locations of all commuters are known. Given traveling plans in Origin-Destination (O-

D) matrix, commuters choose the shortest path dynamically from their current locations to destination. Traveling time of each commuter on a specific link is estimated based on queuing theory, by minimizing a cost function that achieves trade-off between arriving close to the desired arrival time to incur congestion and arriving early/late compared with the desired arrival time to avoid congestion [46]. Congestion in the network is modeled at a macroscopic level. The congestion laws deciding travel delays of each link depend on the setting of incoming traffic flow during a time period and average rate of occupancy [46][47]. To launch simulation, METROPOLIS requires the geometrical structures of the network with static congestion laws and the O-D matrix of all commuters as static simulation settings. By calibrating them, it is easy to introduce traffic events into the network, e.g. congestion of specific spatio-temporal structures.

3.2 Settings of IAURIF database

The network that we focus in IAURIF database covers totally 13627 links in Paris and its suburb region, as shown in Figure 2(a). There are totally 146 simulated traffic sequences in the data set. Each simulated traffic sequence covers 8 hours of traffic data observations, involving congestion in peak hour. Total 48 time sampling steps within each simulation divide the whole 8 hours into 15-minute bins over which the network traffic flows are aggregated. To represent local traffic states, we use traffic index [46][47] in Eq.8:

$$x_{pq} = \frac{\Delta t_p^0}{\Delta t_{pq}} \in [0,1] \tag{8}$$

The denominator is the observed traveling time of link p at the time interval q. It is calculated by averaging all observed traveling time of commuters on the link p within the specific time interval. The numerator is the minimum traveling time among all

commuters on this link within the given time interval. According to Eq.8, the traffic index belongs to [0,1]. As the value of the traffic index decreases, the corresponding link becomes more congested. We store the traffic index of each link at each time sampling step in matrix **X** containing 13627 rows and 7008 columns. Each column is the network-level traffic state observed at the same time interval, represented as a 13627 dimensional vector.

The simulated traffic sequences involve three different configurations of O-D matrix of all 3000 commuters in the network. They start their travel from outskirts of Paris into the central area within 8 hours in each simulation. With this setting, we aim to describe traffic behaviors during morning peak hour of the Paris transportation network. Different configurations of O-D matrix result in variations of global congestion level and different spatial distribution of congestion during peak hour. In the first setting, traffic demands are distributed relatively evenly in the outskirt area near the central Paris, leading to light isotropic congestion inside and around the central Paris. For the second case, we set more travel plans from the northern outskirt area to the central Paris, which produces local congestion patterns in both the northern outskirt and central region. Furthermore, we add random variance to the total amount of travel plans contained in the O-D matrix, covering both globally light and heavy congestion sharing the specific spatial congestion patterns in the network. In the third case, we increase travel paths inside the central and northern area to cause extremely heavy traffic burden in the corresponding areas. As a result, the derived traffic sequences suffer from global congestion ever since the beginning of simulation. They are used to simulate occurrence of extreme accidents in the network. We name the three settings as “Isotropic Traffic Demand” (ITD), “Anisotropic Traffic Demand” (ATD) and “Extreme Traffic Demand” (ETD). Following the three settings, we generate 37, 91, 18 simulations respectively.

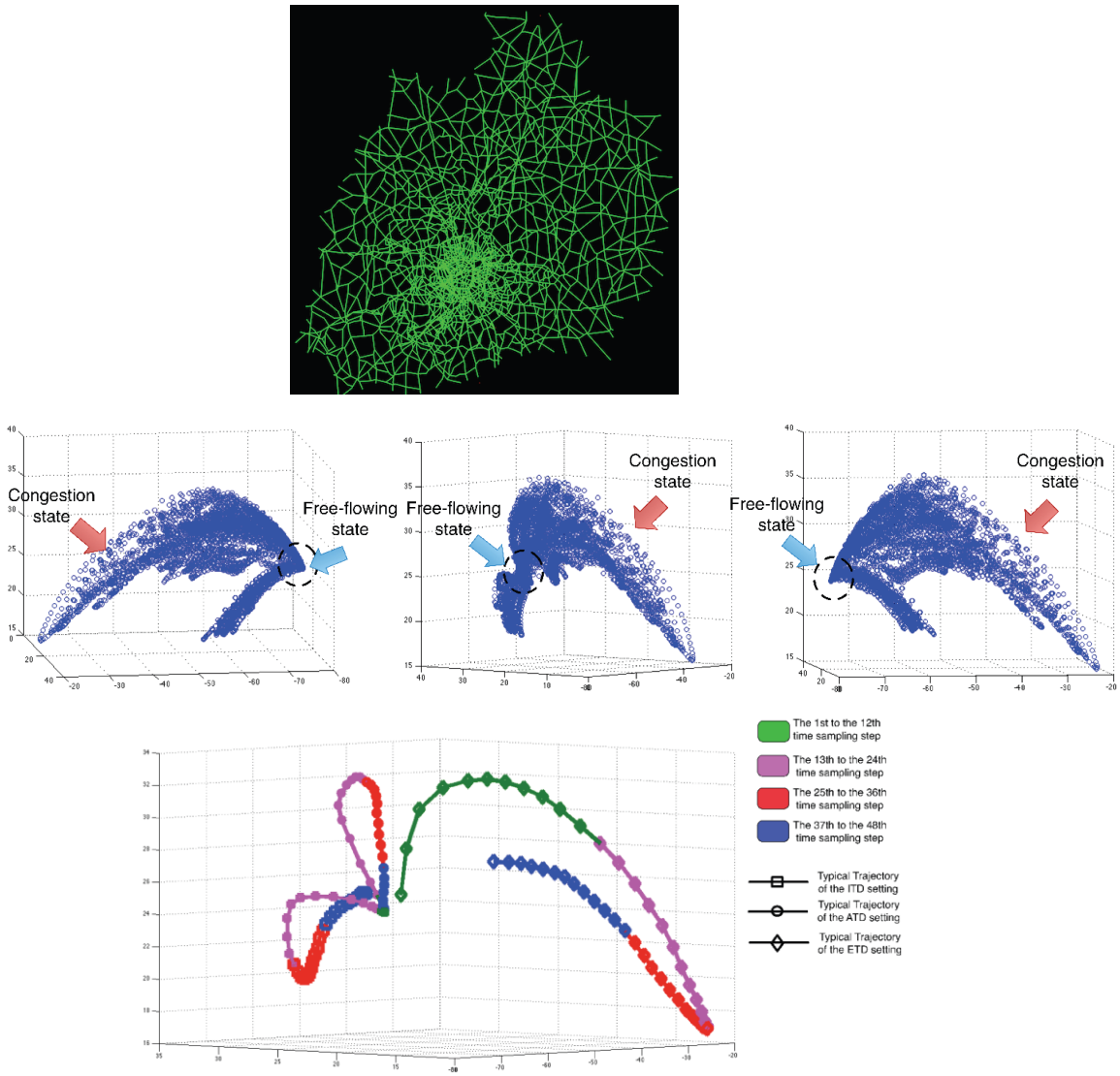


Figure 2. *(a) Traffic network of Paris and suburb regions; (b) A three-views diagram of network-level traffic states in 3 dimensional PCA space; (c) Three typical trajectories corresponding to the three settings of traffic demands*

To visualize distribution of the network-level traffic states, we project all 13627 dimensional vectors into 3 dimensional PCA space, as shown in three different viewpoints in Figure 2(b). In our work, figures illustrating data distribution and clustering results in the PCA space, as Figure 2(b), 2(c), Figure 4, Figure 6 and Figure 8, the three axes correspond to the top three principal components that keep most variances of the original data. The observations of the global free flowing states are distributed within a small region compactly. In contrast, those of medium or

severe congestion are distributed sparsely and biased from the region of the free flowing state. Spatial configurations of local traffic states keep the same if the whole network is global free flowing. On the contrary, congestion occurred at different parts of the network changes the spatial configurations in different ways, increasing variations of global traffic patterns. In Figure 2(c), we link network-level traffic states of the same simulated sequences following their temporal orders. The resultant trajectory represents temporal evolution of network-level traffic states. Different markers on trajectories are used to indicate different traffic demand settings. In each trajectory, color legends are used to indicate successive time intervals. The typical trajectories of the ITD and ATD setting have distinctively different orientations in the PCA space, consistent with difference of spatial congestion patterns. All trajectories start from global free flowing state, as we initialize all simulations with global free flowing state. Trajectories of the ITD and ATD settings converge to the free flowing state, indicating the network restores its fluidity after peak hour. In contrast, the ETD setting leads to much server congestion in the network, with some links congested even at the end of simulations. Thus the corresponding trajectory is quite different from the other two, and converges to the area located far from the free flowing state. Our clustering analysis involves two subjects. Firstly, we perform clustering on network-level traffic states, in order to unveil typical spatial configurations of network-level traffic states, as described in Section 4. After that, we adopt clustering on temporal trajectories of network-level traffic states in Section 5, based on which we could study large-scale traffic dynamics.

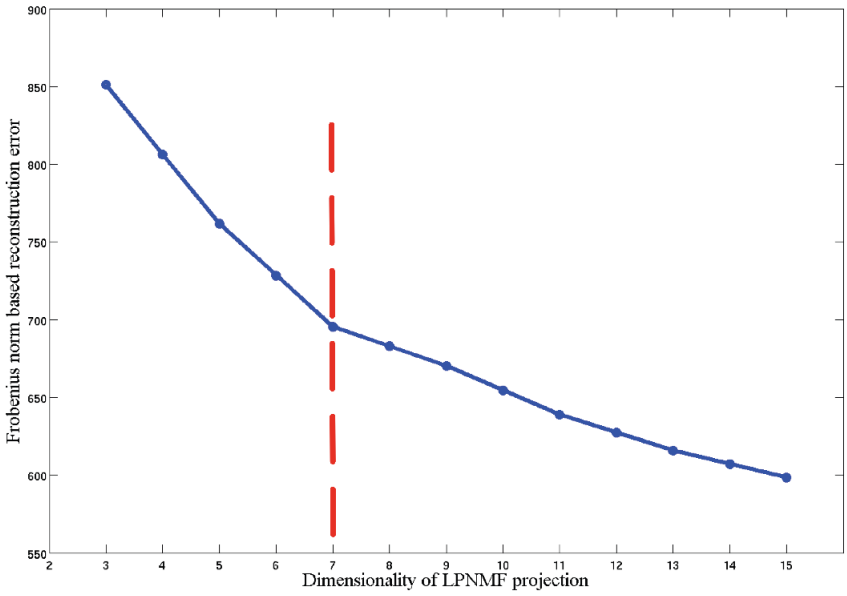


Figure 3 *Frobenius norm based reconstruction errors of different s*

4. Spatial configurations of network-level traffic states

In our work, each LPNMF projection is considered as *s* dimensional signature feature of global traffic configuration. To choose proper dimensionality of LPNMF projection for clustering, we evaluate the Frobenius norm based reconstruction error of the factorization results (shown in Eq.1) with different *s* ranging from 3 to 15, as shown in Figure 3. The reconstruction error declines much slower with *s* larger than 7, indicating 7 dimensional LPNMF projection is competent for describing network-level spatial congestion patterns. Therefore, we set *s* to be 7 in the followings.

In our clustering scheme, the number of the clusters *K* in K-means is decided by following a statistical compactness evaluation of the clusters. Given *p* derived clusters, the compactness c of the clusters is evaluated using the average of sample variances of each cluster, as defined in Eq. 9:

$$c = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{N_i - 1}\sum_{j=1}^{N_i}(nd_j - \frac{1}{N_i}\sum_{j=1}^{N_i} nd_j)\right) \quad (9)$$

N_i is the number of samples assigned into the *i*-th cluster. $\{nd_j\}$ are the network-level traffic observations of the *i*-th cluster. This criterion has been used in ward-

linkage hierarchical clustering [49]. The average sample variance represents general level of compactness of clusters, which is used as the stopping criterion of hierarchical division of the cluster structure. The lower average sample variance is, the more compact clusters we obtain. In our case, we expect network-level traffic states sharing similar spatial configurations of congestion to present a compact cluster structure. To achieve this goal, for each *K* ranging from 3 to 7, we evaluate the compactness measure. Further larger *K* results in cluttered clustering structure, which is difficult to explain the correspondence between the derived clusters and underlying global traffic state patterns. According to Table 1, the compactness measure declines when *K* increases from 3 to 5. *K* larger than 5 introduces little change. It indicates that *K* equaling to 5 is a suitable setting to unveil typical network-level traffic state patterns.

Table 1. *Compactness measure of clustering results based on LPNMF and PCA*

Algorithm	3	4	5	6	7
LPNMF	182.0142	179.0684	171.7805	173.9921	175.6909
PCA	197.1148	182.7953	156.6382	154.2231	152.9623

Figure 4 illustrates the cluster structures in the PCA space. Figure 4(a) illustrates clustering results when *K equals to* 3. The cluster of green legends contains network-level traffic data between the 1st and 20th time sampling step of all 146 simulations, which counts 76% of all samples in the cluster. The left 24% of the cluster come from the interval ranging from the 35th to 48th time sampling step of 97 simulations. All data of the cluster are distributed within the region of the global free flowing state. We therefore name it as “Free Flowing Cluster” (FFC). The cluster of blue legends is composed mainly of traffic state observations collected between the 18th to 40th time sampling step from all 37 simulations of the ITD simulation setting. The cluster of the red legends consists of traffic data between the 20th and 48th time sampling step

from the 105 simulations of the ATD and ETD setting. We thus name these two clusters as “Light Isotropic Congestion Cluster” (LICC) and “Anisotropic Congestion Cluster” (ACC) respectively. Both of them cover peak hour ranging from the 20th until 40th time sampling step. The corresponding three exemplars in the figure indicate typical spatial congestion patterns of the three clusters. In our work, we use the average spatial configuration of traffic indexes as the exemplar of the corresponding cluster, which is calculated by taking link-wise average over the 30% most congested network-level traffic state observations in each cluster. The congested links with traffic indexes lower than 0.79 are labeled using red legends, in order to make the spatial congestion pattern of each cluster visually distinctive. The LICC exemplar contains evenly distributed congestion around the outskirt and central region of Paris. In contrast, congestion concentrates in the central region and the northern outskirt in the ACC exemplar, which is consistent with characteristics of the ATD and ETD setting. In the figure, spatial traffic configurations of the ITD and ATD settings are separated perfectly in the clustering result. By increasing the number of clusters to 5 in Figure 4(b), we can find more details about spatial traffic configurations. The ACC cluster is split into three sub-clusters labeled by pink, purple and black legends respectively. Figure 5 illustrates exemplars of the obtained sub-clusters following the setting in Figure 4(a). Over 90% of the black-labeled cluster are collected between the 10th and 48th step of all 18 simulations of ETD setting. Due to the extremely heavy traffic demands of the ETD setting, congestion appears since the 10th step until the end. The traffic configuration of this cluster presents severe congestion in the network. Consistently, the exemplar of this cluster shown in Figure 4(a) illustrated much severer congestion over the central region and its surroundings than the others. Therefore, we name it as “Heavy Congestion Cluster” (HCC). The pink-labeled cluster corresponds to the peak hour period ranging from the 15th to 35th time sampling step of 88 simulations following the ATD setting. In the purple-labeled cluster, data samples come from the time interval after peak congestion (from the

30th time sampling step to the 48th time sampling step) of 95 simulations. 75 of them are shared with the pink-labeled cluster. It means that the 75 sequences evolve from the pink-labeled cluster to the purple-labeled cluster successively during the peak hour period. Among the left 20 simulations, 3 of them correspond to the ATD setting but with heavier congestion. The other 17 simulations are derived based on the ETD setting and shared with the HCC cluster. The purple-labeled cluster covers the tails of the 17 simulated sequences after the peak of congestion, while the HCC cluster involves the peak hour interval. Therefore, we name the sub-clusters labeled by pink and purple legends as “Peak Congestion Cluster” (PCC) and “After-peak Congestion Cluster” (APCC) respectively. Figures 5(b) and 5(c) illustrate the exemplar of the two sub-clusters. The general congestion level of the PCC exemplar is heavier than the APCC exemplar. Both of them have similar congestion pattern inside the central Paris, while the PCC exemplar contains more congested links in the northern outskirt. The two sub-clusters indicate gradually restoration of traffic conditions during peak hour, representing a spatio-temporal traffic pattern in the network.

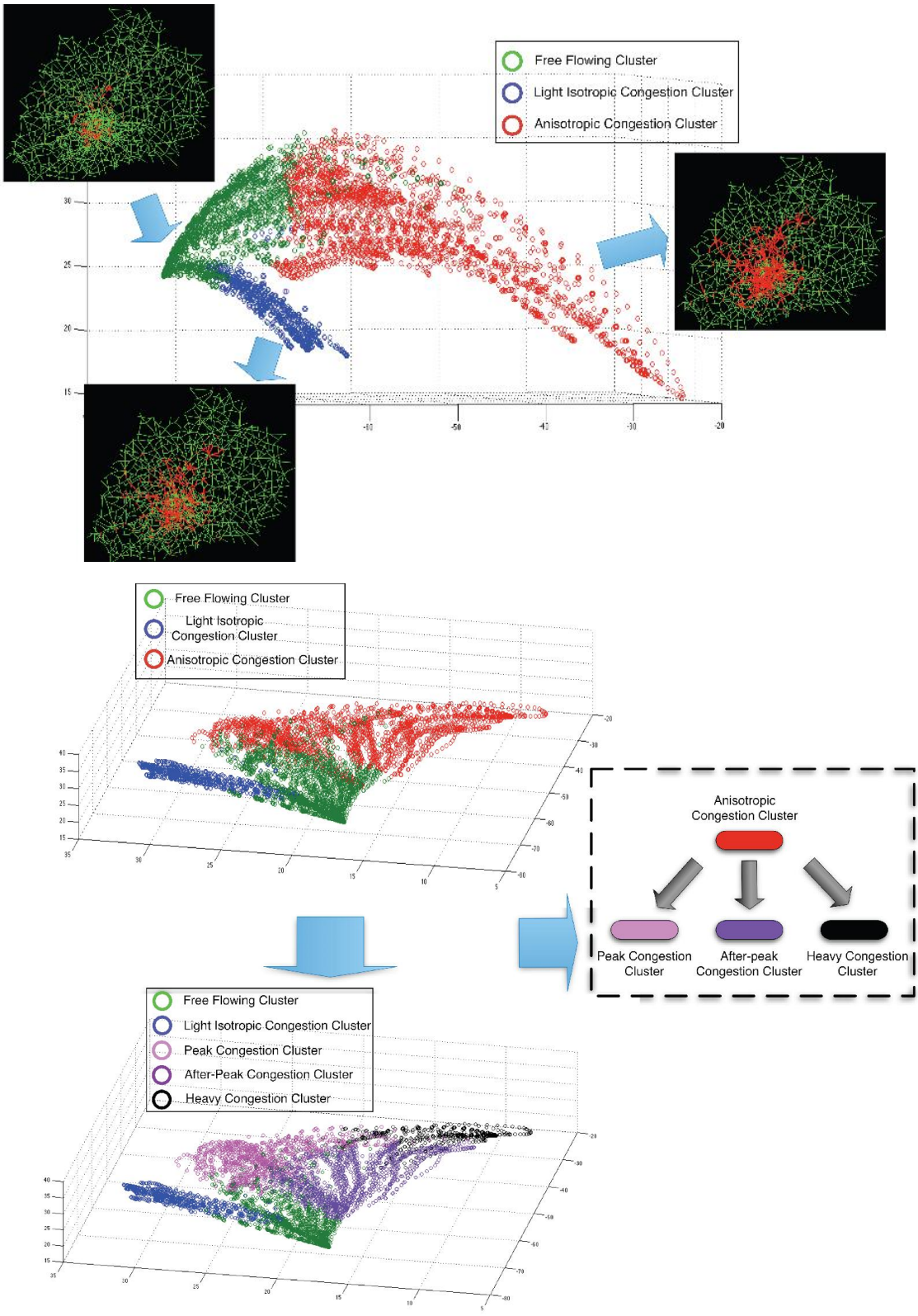


Figure 4. *(a) Three clusters and exemplars of network-level traffic states; (b) Division of clusters after increasing the number of clusters to 5.*

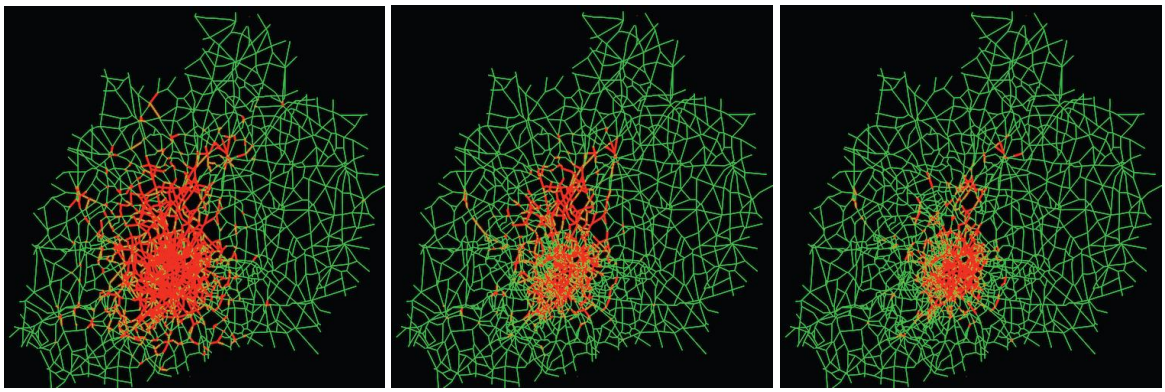
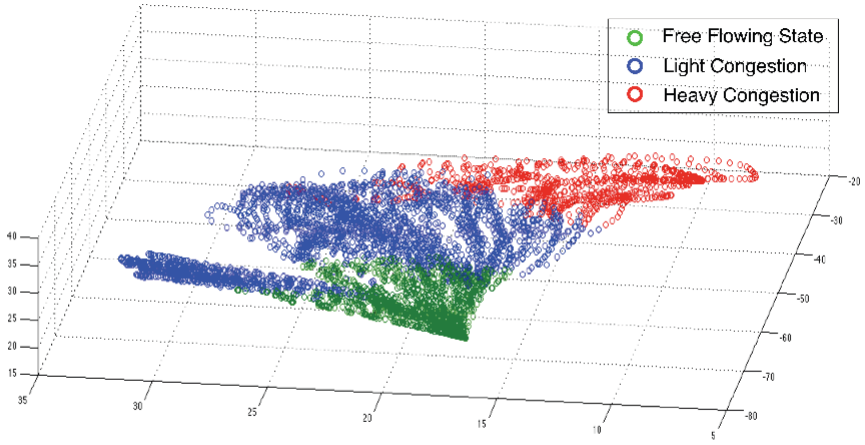
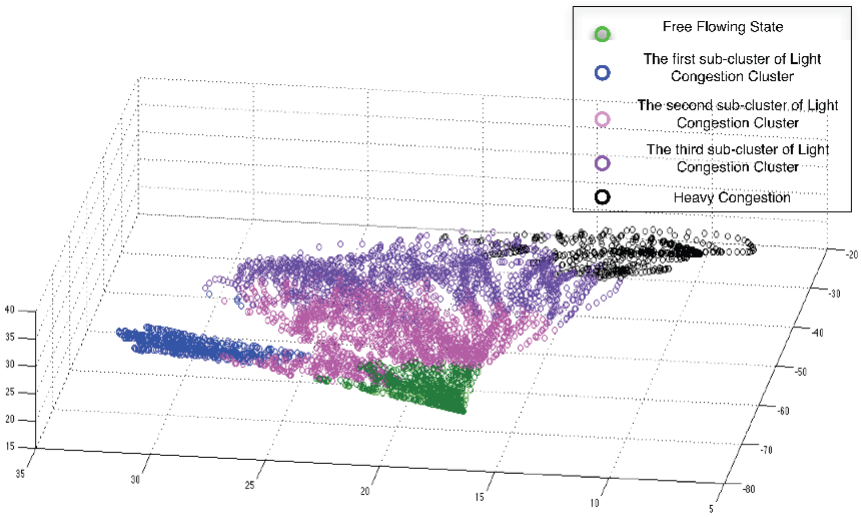


Figure 5. *Exemplars of the sub-clusters: (a) the exemplar of the HCC cluster; (b) the exemplar of the PCC cluster; (c) the exemplar of the APCC cluster.*

To verify capability of LPNMF in unveiling global traffic patterns, we perform K-means on *PCA projections* of the network-level traffic states and compare the derived clustering results. PCA is known as a baseline algorithm in manifold learning. For clustering, we keep the first 15 principal components that contain over 50% variance of the original data. The obtained 15 dimensional PCA projections are then used for clustering. For comparison, we vary *K* in K-means from 3 to 7. According to Table 1. PCA based clustering leads to higher average sample variances when *K* is 3 and 4. Figure 6 shows the derived clustering structures. As shown in the figure, with *K* to be 3, the PCA based clustering only indicates variations of average congestion level over the whole network, ignoring differences between spatial congestion patterns of the ITD and ATD settings. It results in high variances of the cluster structure. By increasing the number of clusters to 5, the PCA based clustering separates global traffic states of the two different simulation setting, thus obtains more compact clusters in statistical sense. However, the derived clusters of pink and purple legends in Figure 6 fail to identify the spatio-temporal structure of traffic patterns as shown in Figure 4(b). Compared with PCA, LPNMF is not only a dimension reduction tool, but also a feature extraction procedure to construct an informative representation of global traffic states.



3 clusters derived by performing K-means to the PCA projections



5 clusters derived by performing K-means to the PCA projections

Figure 6. *Clustering results derived by performing K-means on PCA projections.*

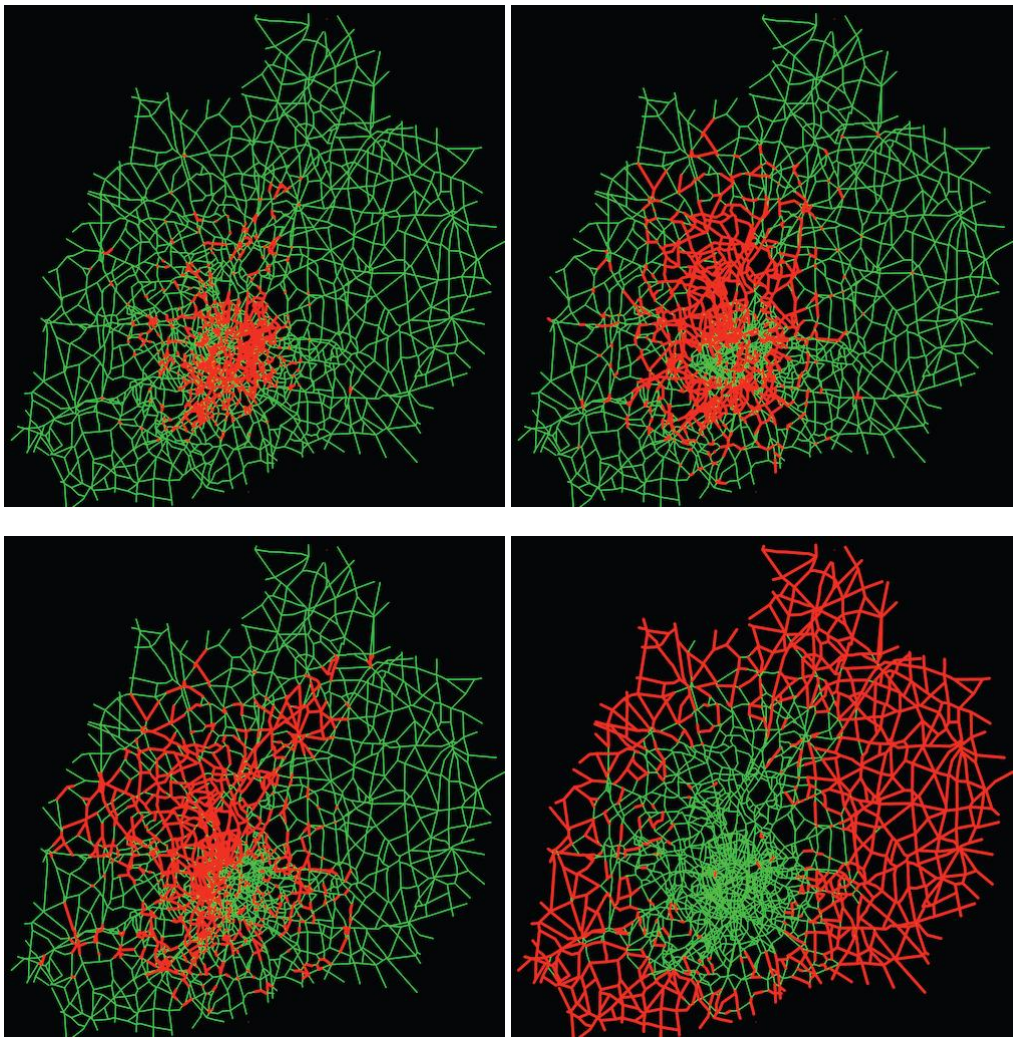


Figure 7. *(a) and (b) are LPNMF basis vectors corresponding to the two highest row-wise average values in* ***V****; (c) and (d) are LPNMF basis vectors corresponding to the two smallest row-wise average values in* ***V***

Besides clustering on the reduced LPNMF representation, we propose to investigate spatial layouts of LPNMF basis vectors that represent important components of global traffic configurations. We calculate the row-wise average $\frac{1}{m}\sum_{j=1}^{m} V_{i,j}$ and sort the *s* average values. The LPNMF basis vectors corresponding to the two largest and two smallest row-wise average values are selected. We analyze the physical significance of the selected basis vectors in the followings. The localized links with the top 20% largest entries (2725 entries) in each basis vector are labeled and their spatial locations are illustrated with red legends. We choose two LPNMF basis

vectors corresponding to the two largest row-wise average in **V** and show their spatial layouts in Figures 7(a) and 7(b). Furthermore, Figures 7(c) and 7(d) show the spatial layouts of the basis vectors corresponding to the two lowest row-wise average values. The links with distinctively large LPNMF magnitudes correspond to the local regions that are highly correlated in forming the spatial traffic configurations. According to Section.2, the LPNMF basis vector with higher row-wise average value of **V** contributes more in representing the spatial congestion patterns. Following this idea, as shown in Figures 7(a) and 7(b), links in central Paris play the most critical role in constituting typical spatial distribution patterns of congestion. Compared with the central region, the outskirt area, especially the northern outskirt, has less but still important contribution in global traffic configurations. Since most travel paths are oriented to the central region in the simulations, Paris center plus its surroundings is expected to be the area that is more likely to be congested during peak hour. Therefore, congestion patterns in this region are the most important factors of global traffic configurations. On the contrary, the circular outskirt regions located far from the central area are free flowing at most times. They contribute the least in global traffic configurations, as shown consistently in Figures 7(c) and 7(d). The spatial layouts of the LPNMF basis vectors represent segmentation of the geographical structure of the network. Different separated regions have effects to different extents on yielding the global traffic state configurations. Links within the same region have homogeneous traffic characteristics. By looking into the basis vectors, we can identify traffic bottleneck of the whole network and extract spatial correlation patterns of the links. Such structural information provides a prior knowledge about how links are correlated with each other when we describe traffic dynamics of the network using graph models [50][51].

5. Temporal analysis of network-level traffic states

In this section, we analyze typical temporal dynamic patterns of network-level traffic states by performing clustering of large-scale traffic temporal behaviors in the 7 dimensional LPNMF projection space. This analysis is important in understanding how the global configuration of traffic states varies throughout a large time period. We also aim to investigate correspondences between the obtained typical temporal dynamic patterns and the underlined simulation settings, which further verifies ability of the LPNMF based method in analyzing global traffic dynamics.

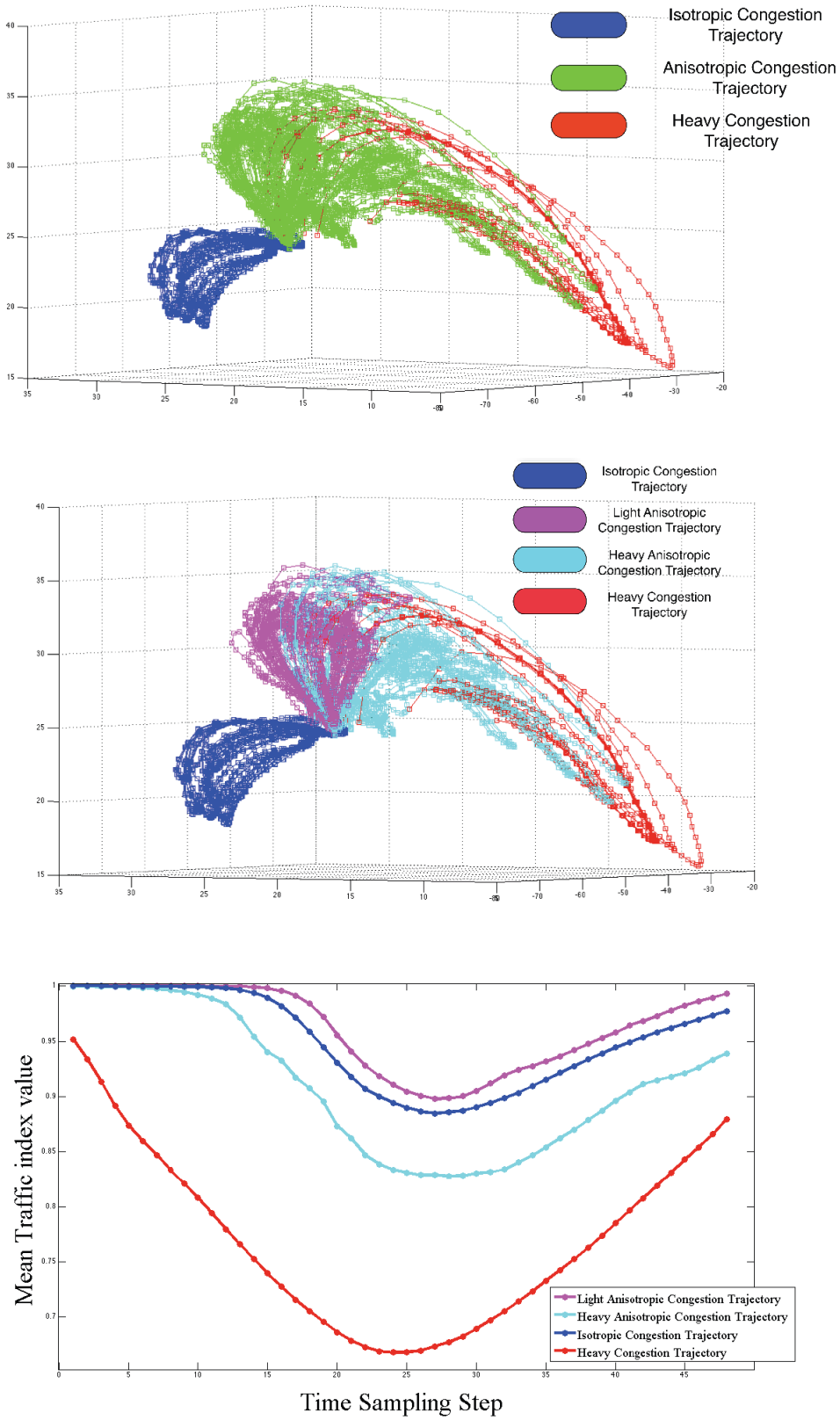


***Figure 8**. (a) Three clusters of the temporal trajectories in 3 dimensional PCA space; (b) Four clusters of the temporal trajectories; (c) Comparison of average temporal dynamic patterns of the four trajectory clusters.*

For each studied traffic sequence, we represent the trajectory of network-level traffic states in the LPNMF space as $\{seq_i\}(i=1,2...48)$, with each seq_i as a 7 dimensional LPNMF projection. To measure similarity between trajectories $\{seq_i^a\}$ and $\{seq_i^b\}$, we compute cosine distance [52] between the LPNMF projections of the corresponding time sampling steps and sum the distances derived in the whole time period, as defined in Eq.10:

$$D=\sum_{i=1}^{48}\frac{seq_i^a\cdot seq_i^b}{\left\|seq_i^a\right\|\left\|seq_i^b\right\|} \tag{10}$$

The cosine distance measures the cosine value of the angle between two corresponding LPNMF projection. It is well normalized into [0,1], which is easy to manipulate in computation. We perform K-means clustering on the temporal trajectories using the distance measure. We set *K* to be 3 firstly, in order to find correspondence between the trajectory clusters and the underlying simulation settings. Figure 8(a) illustrates the derived three clusters of the trajectories. The cluster labeled by blue legends consists of all 37 traffic sequences of the ITD setting, The black-labeled cluster consists of total 11 of 18 trajectories of the ETD setting. The left 7 of 18 ETD simulations are absorbed into the green-labeled cluster. 91 of total 98 trajectories in the green-labeled cluster correspond to the simulated sequences generated following the ATD setting. The obtained three trajectory clusters categorize trajectories of different traffic demand settings accurately. We name them henceforth by “Isotropic Congestion Trajectory” (ICT) and “Anisotropic Congestion Trajectory” (ACT) and “Heavy Congestion Trajectory” (HCT) respectively. By increasing the cluster number from 3 to 4, we can find the ACT cluster is divided further into two sub-clusters, corresponding to different general congestion level during peak hour, named as “Light Anisotropic Congestion Trajectory” (LACT) and “Heavy Anisotropic Congestion Trajectory” (HACT). Figure 8(b) shows the clustering results. For each time sampling step, we treat mean traffic index value (averaged

over all 13627 links in the network) as a crude measure of global traffic state at the current time, the sequence of totally 48 mean index values in one simulation form a general evaluation of large-scale traffic dynamics of the simulation. We further take average of all the 48-D sequences of mean index values in each trajectory cluster. The resultant average sequence represents the general dynamic pattern of the corresponding cluster. Figure 8(c) illustrates average sequences of mean index values corresponding to the trajectory clusters. According to the figure, ICT, LACT and HACT have similar general traffic dynamics with differences in duration of congestion and peak congestion level. In HCT, the network starts to suffer from congestion since the beginning of simulations, which is much different from the others and consistent with the ETD simulation setting. Temporal clustering analysis provides a divide-and-conquer solution to describe underlined large-scale traffic dynamic patterns. Sequences in the same cluster share a common statistical dynamic characteristic of global traffic configurations. By extracting and modeling the typical dynamic process of each cluster using the feasible dynamical models, we are able to improve the controllability and observability of the large-scale traffic dynamic process.

6. Conclusions and perspectives

In this article, we propose and present a new traffic mining methodology for unveiling spatio-temporal traffic patterns, with large-scale modeling and long term forecasting as ultimate goals. Our experiments on large-scale simulated traffic data shows ability of our approach to unveil meaningful congestion patterns and typology of time evolutions; also illustrated is the clear advantage compared to using classical dimension reduction methods such as PCA.

In applications of traffic data analysis, there is still an open issue about developing the on-line NMF training scheme. Since traffic state observations arrive successively

in the form of sequences. It is necessary to update the NMF based model after accumulating a certain number of traffic state observations, which makes the derived model consistent with time-varying traffic configurations of the network. More important from the application point of view, one of our main focuses for future work is exploiting LPNMF low-dimension representation for long-term traffic forecasting.

Acknowledgement

This work was supported by the grant ANR-08-SYSC-017 from the French National Research Agency. The authors specially thank Cyril Furtlehner and Jean-Marc Lasgouttes for providing advice and the benchmark database used in this article.

Reference

[1] Herring, R., Hofleitner, A., Amin, S., Nasr, T., Khalek, A., Abbeel, P., and Bayen, A.: 'Using mobile phones to forecast arterial traffic through statistical learning', Proc. 89th Transportation Research Board Annual Meeting, Washington D.C.,USA, January 2010

[2] 'REACT, Realizing Enhanced Safety and Efficiency in European Road Transport', http://www.react-project.org, accessed March 2010

[3] Work, D., Blandin, S.,Tossavainen, O., Piccoli, B., and Bayen, A: 'A traffic model for velocity data assimilation', Applied Mathematics Research eXpress (AMRX)., 2010,1,pp.1-35

[4] Thiagarajan, A., Sivalingam, L., LaCurts, K., Toledo, S., Eriksson, J., Madden, S., and Balakrishnan, H.: 'VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones', Proc. 7th ACM Conf. Embedded Networked Sensor Systems (SenSys), Berkeley CA,USA, November 2009,pp.85-98.

[5] Krause, A., Horvitz, E., Kansal, A., and Zhao, F.: 'Toward community sensing', Proc. Of Int. Conf. Information Processing in Sensor Networks (IPSN), St.Louis,USA, April 2008, pp.481-492

[6] Liu, H., and Ma, W.: 'A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials', Transportation Research Part C, 2009,17,(1),pp.11-26

[7] Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J., Labrador, M.A., Winters, P.L., Georggi, N.L., and Perez, R.: 'Automating mode detection for travel behavior analysis by using global positioning systems-enabled mobile phones and neural networks', IET Intelligence Transportation System, 2010, 4, (1), pp.37-49

[8] Vanajakshi, L., Subramanian, S.C., and Sivanandan, R.: 'Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses', IET Intelligence Transportation System, 2009, 3, (1), pp.1-9

[9] Chowdhury, D., Santen, L., and Schadschneider, A.: 'Statistical physics of vehicular traffic and some related systems', Physics Report, 2000,329,pp.199-329

[10] Herty M., Klar, A., and Pareschi, L.: 'General kinetic models for vehicular traffic flow and Monte Carlo methods', Computational methods in applied mathematics, 2005, 5,(2), pp.155-169

[11] Rakha, H.: 'Validation of Van Aerde's Simplified Steady-state Car-following and Traffic Stream Model', Transportation Letters: The International Journal of Transportation Research, 2009,1,(3), pp. 227-244

[12] Nagel, K., and Schreckenberg, M.: 'A cellular automaton model for freeway traffic', Journal of Physics,1992,2, pp. 2221–2229

[13] Blandin, S., Work, D., Goatin, P., Piccoli, B., and Bayen, A.: 'A general phase transition model for vehicular traffic', SIAM Journal on Applied Mathematics, to appear, 2011

[14] Statthopoulos, A., and Karlaftis, M.G.: 'A multivariate state space approach for urban traffic flow modeling and predicting', Transportation Research Part C, 2003,11, pp.121–135

[15] Wang, Y., and Papageorgiou, M.: 'Real-time freeway traffic state estimation based on extended kalman filter: a general approach', Transportation Research Part B, 2005, 39, pp.141–167

[16] Min, W., Wynter, L., and Amemiya, Y.: 'Road traffic prediction with spatiotemporal correlations', Technical report, IBM Watson Research Center, 2007

[17] Ghosh, B., Basu, B., and O'Mahony, M.: 'Multivariate short-term traffic flow forecasting using time-series analysis', IEEE Transaction on Intelligence Transportation Systems, 2009, 10, (2), pp.246 – 254

[18] Min, X., Hu, J., Chen, Q., Zhang, T., and Zhang, Y.: 'Short-term traffic flow forecasting of urban network based on dynamic STARIMA model', Proc.12th Int. Conf. Intelligent Transportation Systems (ITSC), 2009, pp.1-6

[19] Yin, H., Wong, S.C., Xu, J., and Wong, C.K.: 'Urban traffic flow prediction using a fuzzy-neural approach', Transportation Research Part C: Emerging Technologies, 2002,10,(2), pp.85–98

[20] Quek, Y., Pasquier, M., and Lim, B.: 'POP-TRAFFIC: A Novel Fuzzy Neural Approach to Road Traffic Analysis and Prediction', IEEE Transactions on Intelligent Transportation Systems, 2006, 7, (2), pp.133–146

[21] Vlahogianni, E. I., Karlaftis, M. G. and Golias, J. C. (2008). Temporal Evolution of Short-Term Urban Traffic Flow: A Non-Linear Dynamics Approach, Computer-Aided Civil and Infrastructure Engineering, 22 (5), 317–325

[22] Vlahogianni, E. I. (2009). Enhancing Predictions in Signalized Arterials with Information on Short-Term Traffic Flow Dynamics, Journal of Intelligent Transportation Systems, 13(2), 73 - 84

[23] Furtlehner, C., Lasgouttes, J., and De la Fortelle, A.: 'A belief propagation approach to traffic prediction using probe vehicles', Proc. 10th Int. Conf. Transportation Systems (ITSC), 2007, pp.1022–1027

[24] Arel, I., Liu, C., Urbanik, T., and Kohis, A.G,: 'Reinforcement learning-based multi-agent system for network traffic signal control', IET Intelligence Transportation System, 2010, 4, (2), pp.128-135

[25] Hamilton, J.: 'Chapter 13 The Kalman Filter', in Hamilton, J.D.(Ed.): 'Time Series Analysis' (Princeton University Press, 1994), pp.372

[26] Hamilton, J.: 'Chapter 3 Stationary ARMA Processes', in Hamilton, J.D.(Ed.): 'Time Series Analysis' (Princeton University Press, 1994), pp.43

[27] MacKay, D.: 'Information Theory, Inference, and Learning Algorithms' (Cambridge University Press, 2003)

[28] Li, S. Z.: 'Markov Random Field Modeling in Image Analysis' (Springer Press, 2009, 3rd edn)

[29] Shoham, Y., and Leyton-Brown, K.: 'Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations' (Cambridge University Press, 2008)

[30] Lee, D.D., and Seung, H.S.: 'Algorithms for non-negative matrix factorization', Proc. 13th Neural Information Processing Systems (NIPS), Denver, USA, 2000, pp.556-562

[31] Lin, C.J.: 'Projected gradient methods for non-negative matrix factorization', Neural Computation, 2007,19, (10), pp.2756–2779

[32] Hoyer, P.O.: 'Non-negative matrix factorization with sparseness constraints',

Journal of Machine Learning Research, 2004, 5, pp.1457–1469

[33] Lin, C.J.: ‘On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization’, IEEE Transactions on Neural Networks, 2007,

18 (6), pp.1589–1596

[34] Jolliffe, I.T.: ‘Principal Component Analysis’ (Springer Press, 2002, 2nd edn)

[35]http://en.wikipedia.org/wiki/Matrix_norm#Frobenius_norm, accessed August 2011

[36] Xu, W., Liu, X., and Gong, Y.H.: ‘Document clustering based on non-negative matrix factorization’, Proc. 26th ACM SIGIR, Toronto, Canada, 2003, pp.267-273

[37] Wang, Y., Jiang, Y., Wu, Y., and Zhou, Z.H.: ‘Local and Structural Consistency for Multi-Manifold Clustering’, Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain, 2002, pp.1559-1564

[38] Chen, G.L., and Lerman, G.: ‘Spectral curvature clustering’, International Journal of Computer Vision, 2009, 81, (3), pp.317–330

[39] Cai, D., He, X., Wu., X., and Han, J.: ‘Non-negative matrix factorization on manifold’, Proc. 8th Int. Conf. Data Mining, Pisa, Italy, 2008, pp.63-72

[40] Cai, D., He, X., Wu., X., Han, J., and Huang, T.: ‘Graph Regularized Nonnegative Matrix Factorization for Data Representation’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, to appear

[41] Cvetković, D. M., Doob, M., and Sachs, H.: ‘Spectra of Graphs: Theory and Applications’ (Wiley New York, 1998, 3rd edn)

[42] Chung, F.R.K.: ‘Spectral Graph Theory’ (CBMS Regional Conference Series in Mathematics, 1997)

[43] Belkin, M.: ‘Problems of Learning on Manifolds’. PhD thesis, University of Chicago, 2003

[44] Belkin, M., and Niyogi, P.: ‘Laplacian eigenmaps and spectral techniques for embedding and clustering’, Proc. 14th Neural Information Processing Systems (NIPS), Vancouver, Canada, 2001, pp.585–591

[45] Belkin, M. ,Niyogi,P., and Sindhwani, V.: ‘Manifold regularization: A geometric framework for learning from examples’, Journal of Machine Learning Research, 2006, 7, pp.2399–2434

[46] Marchal, F.: ‘Contribution to dynamic transportation models’, PhD Thesis, University of Cergy-Pontoise, 2001

[47] De Palma, A., and Marchal, F.: ‘Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale macroscopic transportation systems’, Networks and Spatial Economics, 2002, 2, (4), pp. 347–369

[48] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., and Wu, A.Y., ‘An

efficient k-means clustering algorithm: Analysis and implementation', IEEE Trans Pattern Analysis and Machine Intelligence, 2002, 24, pp.881–892

[49] Hastie, T., Tibshirani, R., and Friedman, J.: '14.3.12 Hierarchical clustering', in Hastie, T., Tibshirani, R., and Friedman, J. (Ed.), 'The Elements of Statistical Learning' (Springer New York, 2009, 2nd edn)

[50] http://en.wikipedia.org/wiki/Cosine_similarity, accessed August 2011

[51] Roweis, S., and Ghahramani, Z.: 'A unifying review of linear Gaussian models', Neural Computation,1999, 11, 2, pp.305-345

[52] Friedman, N., Murphy, K. and Russell, S.: 'Learning the structure of dynamic probabilistic networks', Proc. Uncertainty in Artificial Intelligence (UAI), 1998, pp.139–147

Large scale estimation of arterial traffic and structural analysis of traffic patterns using probe vehicles

Aude Hofleitner* Ryan Herring† Alexandre Bayen‡
Yufei Han§ Fabien Moutarde§ Arnaud de La Fortelle§

91st Annual Meeting of the Transportation Research Board
January 22-26, 2012, Washington D.C

Word Count:

Number of words:	6811
Number of figures:	5 (250 words each)
Number of tables:	0 (250 words each)
Total:	8061

*Corresponding Author, Department of Electrical Engineering and Computer Science, University of California, Berkeley and UPE/IFSTTAR/GRETTIA, France, aude.hofleitner@polytechnique.edu

†Apple Inc. Affiliation during redaction of the paper: California Center for Innovative Transportation, Berkeley CA, ryanherring@berkeley.edu

‡Department of Electrical Engineering and Computer Science and Department of Civil and Environmental Engineering, Systems Engineering, University of California, Berkeley, bayen@berkeley.edu

§Robotics Lab (CAOR), Mines ParisTech, Paris, France, Yufei.Han@mines-paristech.fr
Fabien.Moutarde@mines-paristech.fr

Abstract

Estimating and analyzing traffic conditions on large arterial networks is an inherently difficult task. The first goal of this article is to demonstrate how arterial traffic conditions can be estimated using sparsely sampled GPS probe vehicle data provided by a small percentage of vehicles. Traffic signals, stop signs, and other flow inhibitors make estimating arterial traffic conditions significantly more difficult than estimating highway traffic conditions. To address these challenges, we propose a statistical modeling framework that leverages a large historical database and relies on the fact that traffic conditions tend to follow distinct patterns over the course of a week. This model is operational in North California, as part of the *Mobile Millennium* traffic estimation platform. The second goal of the article is to provide a global network-level analysis of traffic patterns using matrix factorization and clustering methods. These techniques allow us to characterize spatial traffic patterns in the network and to analyze traffic dynamics at a network scale. We identify traffic patterns that indicate intrinsic spatio-temporal characteristics over the entire network and give insight into the traffic dynamics of an entire city. By integrating our estimation technique with our analysis method, we achieve a general framework for extracting, processing and interpreting traffic information using GPS probe vehicle data.

1 Introduction and related work

Traffic congestion has a significant impact on economic activity throughout much of the world. Accurate, reliable traffic monitoring systems, leveraging the latest advances in technology and research are essential for active congestion control. They can also be used to study large scale traffic patterns and to understand specific travel behavior, network bottlenecks, to design long term infrastructure planning and to optimize mobility.

Until recently, traffic monitoring systems have relied exclusively on data feeds from dedicated sensing infrastructure (loop detectors and radars in particular). For highway networks covered by such infrastructure systems, it has become common practice to perform estimation of flow, density or speed at a very fine spatio-temporal scale [3], using traffic flow models developed in the last decades [32, 7, 41]. Probe vehicle data has also been successfully integrated into these models [45, 43, 24, 30]. For arterials, traffic monitoring is substantially more difficult: probe vehicle data is the only significant data source available today with the prospect of global coverage in the future. The lack of ubiquity and reliability, the variety of data types and specifications, and the randomness of its spatio-temporal coverage encourage the use of both historical and real-time data to provide accurate estimates of traffic conditions on large transportation networks. The *Mobile Millennium* project [26] receives probe vehicle data from a dozen of different sources. In Figure 1, we illustrate one of the data source of the *Mobile Millennium* project: it shows a snapshot of probe measurements from San Francisco taxis collected on an arbitrary day from midnight to 7:00am (small dots) as well as a snapshot of the probe locations at 7:00am (large dots). This figure illustrates both the breadth of coverage when aggregating data over long periods of time and the limited information available at a given point in time, limiting the direct estimation of the macroscopic state of traffic at a fine spatio-temporal scale. Note that filtering algorithms are designed to limit the bias of the different sources of data. For example, we filter the measurements during which the hired status of the taxis changes. Aside from less abundant sensing compared to existing highway traffic monitoring systems, the arterial network presents additional modeling and estimation challenges. The underlying flow physics is more complex because of traffic lights (often with unknown cycles), intersections, stop signs, parallel queues, and other phenomena.

We introduce a statistical approach for real-time arterial traffic estimation from probe vehicle data, leveraging massive amounts of historical data. Statistical approaches have been proposed that rely on either a single measurement per time interval or aggregated measurements per time interval [19, 10], neither of which is appropriate in our setting since probe data on arterials is available at random times and random locations. Some researchers have examined the processing of high-frequency probe data (one measurement approximately every 20 seconds or less) [43], which allows for reliable calculation of short distance speeds and travel times. In this article, we specifically address the processing of sparse probe data where this level of granularity is not available. Finally, other approaches based on regression [36], optimization [1], neural networks and pattern matching [8] have all been proposed. None of these approaches addresses the issue of processing sparse probe data on a dense arterial network.

Besides the ability of providing real time traffic estimation, the results produced by the model can be further analyzed to provide a large scale understanding of traffic dynamics both in time and in space. Most of previous research in traffic data analysis focus on temporal dynamics of individual links (either on arterial or highways) using data-driven approaches: in [39, 44, 35], Kalman filter and its extensions, originated from the theory framework of state space linear dynamic model, are used for modeling and tracking temporal variations of traffic flows; [46, 37] use neural networks to achieve short-term non-linear prediction of traffic flows based on historic observations; finally, [40] proposes to perform traffic prediction on individual links based on clustering of temporal patterns of traffic flows, while [11] adopts a time-series

Figure 1: San Francisco taxi measurement locations, observed at a rate of once per minute. Each small dot represents the measurement of the location of a taxi, received between midnight and 7:00am, on March 29th, 2010. The large dots represent the location of taxis visible in the system at 7:00am on that day.

analysis (Autoregressive Moving Average) on traffic flows in order to forecast traffic states. Very little progress has been made in analyzing the temporal dynamics of *global* traffic states of an *entire large-scale road network*. We call *global* traffic state, the aggregation of the congestion states of all the link of the network. Traffic states of neighboring individual roads are often highly correlated (both spatially and temporally) and the identification of specific traffic patterns or traffic configurations is very informative. They can be used to better understand global network-level traffic dynamics and serve as prior knowledge or constraints for the design of traffic estimation and prediction platforms. The analysis of traffic patterns is also useful for traffic management centers and public entities to plan infrastructure developments and to improve the performances of the available network using large-scale control strategies.

This article proposes an algorithm to identify spatial configurations of traffic states over the entire network and analyze large-scale traffic dynamics from traffic state estimates produced and collected over long periods of time. We define the *network-level traffic state* as the vector of traffic states for each link of the network at a given moment in time. It is represented in the form of multi-variate data, where its dimension is proportional to the amount of links in the transportation network. In large networks, this data structure quickly becomes too big to handle, limiting the analysis in the original high-dimensional space. In machine learning, this issue is commonly addressed using dimension reduction techniques (feature extraction) to simplify the representation of the data, remove redundancies and improve the efficiency of analysis techniques such as classification. Important applications of these algorithms include image processing and natural language processing [12, 9]. In this work, we propose to use a dimensionality reduction matrix factorization technique known as *Non-negative Matrix Factorization* (NMF) [6, 25] to obtain a low dimensional representation of network-level traffic states. Both the well-known *Principal Component Analysis* (PCA) method, and *Locality Preserving Projection* (LPP) technique are other examples of matrix factorization [28, 15]. However, in contrast to PCA or LPP, the NMF algorithm imposes strict non-negativity constraints on the decomposition result. This allows NMF to approximate the n-dimensional data vector by an *additive* combination of a set of learned bases. This property also leads to a part-based representation of the original data. The learned bases correspond to *latent components* of the original data so that the original data is approximated by a linear *positive* superposition of the latent components. The properties of the NMF have already been exploited for various applications. In text analysis, the learned bases are used to label different latent topics contained in text documents. In face image representation, the NMF bases indicate important localized components of the face, such as the eyes, the mouth or the cheeks. We expect that the distinctive characteristics of NMF will lead to a low-dimensional representation of network-level traffic states that exhibits global configurations of local traffic states and reflects intrinsic traffic patterns of network-level traffic

states.

The rest of this article is organized as follows. In Section 2 we present the real-time traffic estimation algorithm implemented in the *Mobile Millennium* system. It processes sparsely sampled probe vehicles sending their location at random places and random times and leverages historical data using a Bayesian update. In Section 3, we introduce the NMF algorithm, used in the remainder of the article to perform large scale analysis of the dynamics of traffic. In Section 4 and 5, we illustrate and provide a detailed analysis of typical spatial configuration patterns of network-level traffic states found by NMF projections. Section 6 further analyzes temporal dynamic patterns of the network-level traffic state, which describe evolutions of traffic states in the whole network. In Section 7, we conclude our work and discuss our future plans.

2 Large scale statistical model for arterial traffic estimation

We propose a parametric statistical model for large scale traffic estimation from sparsely sampled probe vehicles. The parameters of the model represent traffic patterns that are learned from massive amounts of probe vehicle travel times collected over long periods of time (section 2.1). The historic patterns are used as prior information in a Bayesian real-time estimation algorithm with streaming data (section 2.2). The statistical model is based on assumptions aimed at limiting the computational complexity of the algorithm while providing an adapted framework for arterial traffic estimation when little data is available in real-time but large quantities of historical data are collected over time.

1. The travel time on a link is a *random variable* (RV) and we assume that travel times on different links are independent RVs. In this article, we assume that travel times are normally distributed. Other distributions can also be used (e.g. Gamma, log-normal, and so on) without modifying the basic concepts of the algorithm. We will specify the equations that require modification under a non-normality assumption.
2. Any given moment in time belongs to exactly one *historic time period*, characterized by a day of the week, a start time and an end time. The set of historic time periods is denoted $\mathcal{T}$.
3. All travel time observations from a specific link l are independent and identically distributed within a given (historic) time period, $t \in \mathcal{T}$.
4. Probe vehicles send their location periodically (typically every minute). A trajectory reconstruction algorithm [27] provides the most likely path p of the vehicle between successive location reports. The path p is defined as a set of consecutive links, L_p, along with the fraction of the first and last link traversed and the total travel time associated with the entire path, y_p (time between successive location reports). The fraction of link l traversed for the pth observation on link l is denoted $w_{p,l}$. The time spent on link l is denoted x_p, l and have the constraint that $\sum_{l \in L_p} x_{p,l} = y_p$. The set of path observations for time period t is denoted $\mathbf{P}_t$. We assume that these path observations constitute the only data available to the model.
5. Each link of the network has a known minimum travel time b_l, found by considering the travel time that results from driving some percentage over the speed limit for the entire link. Note that the maximum travel time is harder to determine as travel times increase with congestion. A statistical analysis of available measurements may provide information about maximum travel times.

2.1 Learning historic traffic patterns

The historical model of arterial traffic estimates the parameters $Q_{l,t}$ of the travel time distribution on each link l for each historic time period t. The corresponding *probability density function* (PDF) of travel times is denoted $g_{l,t}(\cdot)$. In the case of Gaussian distributions, the parameters $Q_{l,t}$ are written $Q_{l,t} = (\mu_{l,t}, \sigma_{l,t})$, where $\mu_{l,t}$ and $\sigma_{l,t}$ represent the mean and the standard deviation of the travel times on link l for the period t.

For the pth path observation, the path travel time distribution is denoted $g_{L_p,t}(\cdot)$. Under assumption 1, the PDF of travel time on a path is computed as the convolution of the PDF of travel times of the links that make up the path.

The historic algorithm determines the values of $Q_{l,t}$ for each link and time period that are most consistent with the probe data received. This is achieved by maximizing the log-likelihood of the data given the parameters, which is written as

$$\arg\max_{\mathbf{Q}_t} \sum_{p \in \mathbf{P}_t} \ln(g_{L_p,t}(y_p)), \tag{1}$$

where $\mathbf{Q}_t$ is the set of $Q_{l,t}$ for all links l of the network. This optimization problem may be challenging due to the high number of variables (number of links times number of parameters per link travel time distribution), coupled through the PDF of path travel times $g_{L_p,t}$. In the case of Gaussian link travel times, solving (1) amounts to simultaneously estimating the mean and the variance of every link in the network and is not formulated as a convex problem. To face this difficulty, we decouple the optimization into two separate subproblems (*travel time allocation* [16, 20] and *parameter optimization*), each of which is easier to solve on its own, and then iterate between these subproblems until converging to an (locally) optimal solution.

If we knew how much time each probe vehicle drove on each link of its path (instead of just the total travel time), it would be easy to estimate the mean and standard deviation for each link in the network (sample mean and standard deviation of the link travel time observations). Since the sampling scheme only provides the total travel time on the path, we determine the most likely amount of time spent on each link (travel time allocation). Unfortunately, the most likely link travel times depend upon the link travel time parameters (μ and σ) that need to be estimated. This would appear to a chicken-and-egg problem, but there is a sound mathematical justification (hard EM) for iterating between these two steps. The link parameters are used to determine the most likely travel times and then the most likely travel times are used to update the parameters.

Travel Time Allocation: To solve the travel time allocation problem, we assume that estimates of the link parameters $Q_{l,t}$ are available (and fixed). We specify *lower bounds* b_l on the travel time allocated for each link l of the network to model the bounded speed of vehicles and ensure that a sensible solution is returned. For each observation $p \in \mathbf{P}_t$, we maximize the log-likelihood of the travel times $x_{l,p}$ spent on each link l of the path, with the constraint that they sum to the path travel time y_p. The optimization problem reads

$$\begin{aligned} \arg\max_{x} \quad & \sum_{l \in L_p} \log \mathcal{N}(x_{p,l};\ w_{p,l}\mu_{l,t}, w_{p,l}\sigma_{l,t}) \\ \text{s.t.} \quad & \sum_{l \in L_p} x_{p,l} = y_p \\ & x_{p,l} \geq w_{p,l} b_l, \forall l \in L_p, \end{aligned} \tag{2}$$

where the minimum travel time $w_{p,l}b_l$ is found using the minimum travel time b_l on link l, scaled by the fraction of link traveled $w_{p,l}$, as introduced in item 5. The notation $\mathcal{N}(x;\ \mu, \sigma)$ represents the PDF of a Gaussian variable with mean μ and standard deviation σ, evaluated at x:

$$\mathcal{N}(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Problem (2) is a (small scale) quadratic program (QP) [4] which can be solved analytically (see algorithm 1 for details). Note that if a vehicle travels faster than the maximum speed, the allocation problem is infeasible. The vehicle is considered as an outlier and the observation is discarded from the set of observations. We call $\mathbf{X}_{l,t}$, the vector of allocated travel times for link l during time period t. Note that the travel times $x_{p,l}$ are scaled by the proportion of the link traveled, $w_{p,l}$, before being added to the set of allocated travel times $\mathbf{X}_{l,t}$.

Parameter Optimization: Given $\mathbf{X}_{l,t}$, the computation of the parameters $Q_{l,t}$ depends on the choice of the class of distribution chosen. In the case of Gaussian distributions, this computation is straightforward as $\mu_{l,t}$ and $\sigma_{l,t}$ respectively represent the sample mean and standard deviation of $\mathbf{X}_{l,t}$.

Full Historic Arterial Traffic Algorithm: After initializing the parameters $Q_{l,t}$ for each link of the network, the algorithm iterates between allocating the travel times for each path in $\mathbf{P}_t$ and optimizing the link parameters given the allocated travel times in $\mathbf{X}_{l,t}$. The convergence of the algorithm is checked by computing the log-likelihood after each iteration, which is guaranteed to increase at each iteration until convergence.

2.2 Bayesian Real-time Traffic Estimation

The parameters $Q_{l,t}$ learned by the historic model are used as prior information to estimate current traffic conditions via a *Bayesian update* (see [38] for more details on Bayesian statistics). In Bayesian statistics, parameters are considered as RVs and thus have a probability distribution. Here, we compute the probability distribution (known as *posterior distribution*) of the mean travel time (seen as a RV) given the allocated link travel times and a prior distribution on the mean travel time denoted f_0.

Let Δ_t represent the duration between successive real time estimates. We run the algorithm at time t_2 using the path data available for the *current time window* $[t_1, t_2]$, with $t_1 = t_2 - \Delta_t$. The duration Δ_t between successive updates depends upon the amount of data available in real-time and should remain inferior to the duration of the historical time intervals. If the data volume is large, the model can be run up to every 5 minutes. Running the model more frequently will likely not increase the performance and may lead to estimates that fluctuate too much due in particular to the periodic dynamics associated with the presence of traffic signals [22, 23].

For generic RVs X and Y with realization x and y, the notation $f(x|y)$ is read "probability that X has the realization x given that Y has the realization y" and denotes the conditional probability of RV X given the observation of the RV Y. Let y_{l,t_2} denote the set of travel times allocated to link l between t_1 and t_2. Using Bayes theorem, the posterior probability on the mean travel time $\hat{\mu}_{l,t_2}$ is proportional to the likelihood of the data times the prior:

$$f(\hat{\mu}_{l,t_2}|y_{l,t_2}, \sigma_{l,t}) \propto f(y_{l,t_2}|\hat{\mu}_{l,t_2}, \sigma_{l,t}) f_0(\hat{\mu}_{l,t_2}), \tag{3}$$

The symbol $\propto$ is read "is proportional to". The proportionality constant is chosen such that the integral of $\hat{\mu}_{l,t_2} \mapsto f(\hat{\mu}_{l,t_2}|y_{l,t_2}, \sigma_{l,t})$ on $\mathbb{R}$ is equal to one. At time t_2, the Bayesian update determines the value of mean travel times $\hat{\mu}_{l,t_2}$ that maximizes the posterior probability.

Assuming Gaussian link travel times, a natural choice for f_0 is a Gaussian distribution (conjugate prior [38]). Since f_0 represents prior information on the mean travel time, its mean is set to the historical mean $\mu_{l,t}$ and its standard deviation $\sigma_{0;\,l,t}$ is chosen to represent how much real-time condition can deviate from the historical values. Typically $\sigma_{0;\,l,t}$ is large to give more weight to real-time data as soon as they are in sufficient quantity. Because of the Gaussian prior, the allocated link travel times y_{l,t_2} and the mean travel time $\hat{\mu}_{l,t_2}$ are jointly Gaussian and we compute the parameters of the posterior (Gaussian) distribution of $\hat{\mu}_{l,t}$. In particular, we update the mean link travel times as the mean of the posterior distribution:

Algorithm 1 Travel time allocation algorithm. The core of the algorithm is contained in lines 11-15, which computes the total expected path variance (V) and the difference between expected and actual travel times (Z). With these two quantities, each link is allocated the expected link travel time adjusted by some proportion of Z, where this proportion is computed using the link variance divided by the total path variance. This procedure can lead to some links being allocated a travel time below the minimum for that link. The set $\mathbf{J}$ is introduced to track the links with initial allocated travel times below the lower bound and the main procedure is repeated by setting the travel times for these links to the lower bound and optimizing with respect to the remaining links. Note that the travel times are scaled by the proportion of the link traveled (line 19) before being added to the set of allocated travel times $\mathbf{X}_{l,t}$.

Require: $t \in \mathcal{T}$ is fixed to some particular time period.

1: **for** $l \in \mathcal{L}$ **do**
2: $\quad \mathbf{X}_{l,t} = \emptyset$ {Initialize allocated travel time sets to be empty.}
3: **end for**
4: **for** $p \in \mathbf{P}_t$ **do** {For all probe path observations.}
5: $\quad$ **if** $\sum_{l \in L_p} w_{p,l} b_l > y_p$ **then**
6: $\quad\quad$ Travel time allocation infeasible for this path. This means that the observation represented travel that is considered faster than realistically possible, so the observation is considered an outlier. Remove p from $\mathbf{P}_t$.
7: $\quad$ **else**
8: $\quad\quad \mathbf{J} = \emptyset$ {$\mathbf{J}$ contains all links for which the travel time allocation is fixed to be equal to the lower bound.}
9: $\quad\quad$ **repeat**
10: $\quad\quad\quad x_{p,l} = w_{p,l} b_l, \forall l \in \mathbf{J}$ {For all links that had an infeasible allocation in the previous pass through this loop, set the allocation to the lower bound.}
11: $\quad\quad\quad V = \sum_{l \in L_p \setminus \mathbf{J}} w_{p,l} \sigma^2_{l,t}$ {Calculate the path variance for the links not fixed to the lower bound.}
12: $\quad\quad\quad Z = y_p - \sum_{l \in \mathbf{J}} w_{p,l} b_l - \sum_{l \in L_p \setminus \mathbf{J}} w_{p,l} \mu_{l,t}$ {Calculate the difference between expected and actual travel time for the links not fixed to the lower bound.}
13: $\quad\quad\quad$ **for** $l \in L_p$ **do** {Allocate excess travel time in proportion of link variance to path variance.}
14: $\quad\quad\quad\quad x_{p,l} = w_{p,l} \mu_{l,t} + \frac{w_{p,l} \sigma^2_{l,t}}{V} Z$
15: $\quad\quad\quad$ **end for**
16: $\quad\quad\quad \mathbf{J} = \mathbf{J} \cup \{l \in L_p : x_{p,l} < w_{p,l} b_l\}$ {Find all links violating the lower bound.}
17: $\quad\quad$ **until** $x_{p,l} >= w_{p,l} b_l, \forall l \in L_p$
18: $\quad\quad$ **for** $l \in L_p$ **do**
19: $\quad\quad\quad \mathbf{X}_{l,t} = \mathbf{X}_{l,t} \cup \left(\frac{x_{p,l}}{w_{p,l}} \right)$ {Add the allocated travel time to $\mathbf{X}_{l,t}$.}
20: $\quad\quad$ **end for**
21: $\quad$ **end if**
22: **end for**
23: **return** $\mathbf{X}_{l,t}, \forall l \in \mathcal{L}$

$$\hat{\mu}_{l,t} = \frac{\sigma^2_{0;\,l,t}}{\frac{\sigma^2_{l,t}}{N_{l,t_2}} + \sigma^2_{0;\,l,t}} \overline{x} + \frac{\sigma^2_{l,t}}{\frac{\sigma^2_{l,t}}{N_{l,t_2}} + \sigma^2_{0;\,l,t}} \mu_{l,t}$$

where N_{l,t_2} is the number of travel times allocated to link l during the current time interval (t_1, t_2) and $\overline{x}$ is the sample mean of the allocated travel times y_{l,t_2}.

To summarize, the real-time estimation algorithm performs the travel time allocation on each probe observation and then uses the allocated travel times and the historical traffic parameters to perform a Bayesian update of the link parameters.

The precise analysis of the performance of this model is out of the scope of this article. We refer the reader to the following references assessing the results of the *Mobile Millennium* project for more details [2, 17].

3 Non-negative matrix factorization (NMF)

In this section, we present *Non-negative Matrix Factorization* (NMF), which is used for approximating network-level traffic states as positive sums of a limited number of global traffic configurations. NMF [31, 6, 33, 25, 9] is a particular type of matrix factorization, in the same domain as the well-known *Principal Component Analysis* (PCA) method and *Locality Preserving Projection* (LPP). In all cases, given a set of multivariate n-dimensional data vectors placed in m columns of a $n \times m$ matrix X, matrix factorization decomposes the matrix into a product of a $n \times s$ loading matrix M and a $s \times m$ score matrix V, where s represents the dimensionality of the subspace to which we project the original data. Through this matrix decomposition, each n-dimensional data vector is approximated by a linear combination of the s columns of M, weighted by the components in the corresponding column of V. We can regard all s column vectors in loading matrix M as a group of projection bases that are learned optimally to represent the original data. The variable s is typically chosen to be significantly smaller than both n and m so that the obtained score matrix V forms a low-dimensional subspace projection of the network-level traffic states, on which we can perform further data analysis. The specificity of NMF is the enforced positivity of both the weights in V, and of the columns of M forming the NMF decomposition basis. This non-negativity therefore provides an approximation of the n-dimensional data vector by an *additive* combination of a set of learned bases. Furthermore, the NMF components forming the basis tend to be sparse, which leads leads to a part-based representation of the original data.

A network-level traffic state is a vector of size equal to the number of links in the network, where the i^{th} entry corresponds to the traffic state on the i^{th} link of the network. Arterial networks are typically dense (numerous links and intersections) and the number of links in any decent size network is often over a thousand links. Assuming that k samples of n-dimensional network-level traffic states are stored as an $n \times k$ matrix X, NMF factorizes X as a product of a non-negative $n \times s$ matrix M and a non-negative $s \times k$ matrix V which minimizes the Frobenius norm of the reconstruction error between X and its factorized approximation MV. We recall that the Frobenius norm of a matrix $A \in \mathbb{R}^{n \times m}$ with entry on column i and line j denoted $A_{i,j}$ is defined as

$$||A||_F = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} |a_{i,j}|^2}$$

and is equal to the sum of the singular values of A. The matrix factorization problem reads:

$$\arg \min_{(\mathbf{M},\mathbf{V})} \|\mathbf{X} - \mathbf{MV}\|_F \text{ s. t. } \quad M \geq 0,\ V \geq 0, \tag{4}$$

where the inequalities $M \geq 0,\ V \geq 0$ represent the non-negativity constraints (each element of the matrices are non-negative). Training of NMF is implemented using multiplicative updates [31],fixing either M or V and updating the left following the KKT condition. The NMF cost function shown in equation (4) is not convex.However,fixing either M or V leads to a convex subproblem to solve. Multiplicative updates and other gradient based optimization procedure can not guarantee the global optimum of the NMF solution. Nevertheless, in data mining, local minimum is still enough to be useful. Given fixed M or V, the NMF objective is a convex optimization issue. The NMF projects the high-dimensional network-level traffic states on a s-dimensional subspace, which is spanned by the columns of M. According to equation (4), the column space of V corresponds to coordinates of network-level traffic states with respect to the learned set of bases in M. The column space of V forms a low-dimensional representation of the network-level traffic states. As mentioned in the introduction, each network-level traffic state $X_j \in \mathbb{R}^n$ is approximated by an additive linear superposition of the column space of M due to the non-negative constraint. The approximation of X_j is written

$$X_j \approx \sum\nolimits_{i=1}^{k} M_i V_{i,j}, \tag{5}$$

where M_i denotes the i^{th} column of M and $V_{i,j}$ is the element at the i^{th} column and j^{th} row of V. It is important to interpret what the matrices M and V represent in terms of traffic analysis. The column space of M represents typical elements of the spatial configuration patterns with respect to the network-level traffic states. Based on the columns of M, we represent complex spatial arrangements of local traffic states over the entire network. As for V, equation (4) indicates that each element $V_{i,j}$ represents to which degree the j^{th} network-level traffic state observation is associated with the i^{th} expanding basis in matrix M (i^{th} spatial configuration). For example, if the spatial configuration formed by the i^{th} column of M is the best representation of the j^{th} network-level traffic state, then $V_{i,j}$ will take the largest value in the j^{th} row of V [6]. As a result, the derived low-dimensional representation formed by the columns space of V are intuitively consistent with information about spatial distribution patterns of local traffic states. By contrast, the PCA and LPP based projections only aim at best reconstruction of traffic observations with either maximizing data variances or preserving neighboring structures. The projection results of PCA and LPP are thus less likely to be associated with interpretable latent traffic configuration patterns than the NMF. Therefore, we choose NMF to analyze the network-level traffic states in our case.

In this article, the traffic states used for the clustering analysis are *fluidity indices*. A fluidity index is a value in $[0, 1]$ computed as the ratio between the free flow and the estimated travel times. They are provided by the estimation algorithm described in Section 2 and operational in the *Mobile Millennium* [26] traffic platform, which receives data from a dozen of feeds totalling several millions of data points per day for Northern California traffic. The *Mobile Millennium* platform has been operational since November 2008 and has been storing historical data since then, providing a rich database of historical traffic dynamics in the Bay Area. In real-time, the model estimates travel times and fluidity indices from the streaming data and leverages the historical data using the Bayesian update presented in Section 2.2. The estimates are updated on each link of the network every five minutes. We focus our study on a network consisting of 2626 links for a duration of 184 days, from 00:00 May 1st 2010 to 23:55 October 31st 2010, totaling 52292 estimates per link ($12 \times 24 \times 184$). We store the fluidity index of each link at each time sampling step in a matrix X containing 2626 rows and 52292 columns. Our clustering results includes two parts, firstly we perform clustering on network-level traffic states, in order to find some typical spatial configurations of network-level traffic states, as described in Section 4. Secondly, we perform clustering on temporal trajectories of network-level traffic states, from which we can study traffic dynamics, shown in Section 6.

4 Congestion patterns: spatial configurations of global traffic states

An important outcome of dimensionality reduction is to identify typical spatial congestion patterns (i.e. spatial configurations of congestion). While doing this on the original 2626 dimensional data would be rather sloppy and computer intensive, it is much more feasible in the low dimensional space obtained by NMF.

NMF has one essential parameter: the number s of components over which decomposition is done. The parameter s also corresponds to the dimension of the target subspace where we perform clustering. The choice of s is empirical (s is called a *meta parameter*) and is done by analyzing results obtained for increasing values of s from 3 to 30. Our analysis focuses on the reconstruction error (value of the objective function (4) at optimum) and the clustering results. The reconstruction error continually decreases as the dimension s increases. This result is expected as the optimization problem (4) is performed on a larger set and thus the factorization models with higher complexity always leads to better fitting to the original data. Our clustering of global traffic states consists of clustering the traffic data projected in the s-dimension subspace using a *k-means* algorithm [34, 29]. The k-means algorithm is a widely used unsupervised clustering algorithm. It partitions observations into k clusters in which each observation belongs to the cluster with the nearest mean. We represent the clusters obtained in the s-dimensional space in three dimensions, limiting the number of NMF components to three but keeping the clustering results obtained in the s-dimensional space. We notice that values of s inferior to eleven lead to clustering results which seem visually inadequate: the 3D representation of the clusters shows important overlap between the clusters. The clusters become separated for values of s greater than fifteen. Increasing s over 15 does not seem to bring any improvement in the clustering results, while it significantly increases the NMF computation and memory usage costs. Therefore, we set the number of NMF components to $s = 15$ for all subsequent analysis presented in this article. This value achieves a balance between the descriptive power of NMF projection and the computational efficiency.

In clustering analysis, we also need to choose the number of clusters in k-means, denoted by k. The challenge is different from that for the choice of s: the choice of k does not influence the computational costs significantly but changes the interpretability of the results. The number of clusters represents the number of "global congestion patterns" that may arise. Too low values of k may not represent the different congestion patterns whereas too high values of k may decrease the possibilities of interpretation by separating similar congestion states into different clusters. After analyzing the results obtained for increasing values of k, it seems that the most insightful clustering is obtained with $k = 5$ clusters. The average fluidity index value (obtained by averaging index values on all links) are shown for each of the five clusters in the table at the top of figure 2. It appears that two clusters (cyan squares and green stars) correspond to different types of "mostly fluid" states, whereas the remaining three clusters (blue circles, yellow diamonds and red stars) represent "congested states". We study the physical significance of each cluster by constructing histograms of the fluidity index values, counting occurrence frequencies of fluidity index values in each cluster. We find that fluidity index values in the *night and early morning Free-Flow* (NFF) and *Evening Free-Flow* (EFF) cluster are higher as a whole than those in the clusters corresponding to occurrences of congestion (*Morning Increasing Congestion*, MIC, *Mid-Day Congestion*, MDC and *Afternoon Decreasing Congestion* ADC clusters).

Figure 2 shows that the significance of the distributional patterns with respect to evaluating global traffic states is generally consistent with that of average fluidity index values, which implies that the average fluidity index value could also be used as an easy-to-use and efficient indicator of global traffic states in our case.

Marker symbol	Average fluidity	Cluster name
Green stars	0.7757	Night + early morning Free-Flow (NFF)
Blue circles	0.7185	Morning Increasing Congestion (MIC)
Red stars	0.6393	Mid-Day Congestion (MDC)
Yellow diamonds	0.6730	Afternoon Decreasing Congestion (ADC)
Cyan squares	0.7420	Evening Free-Flow (EFF)

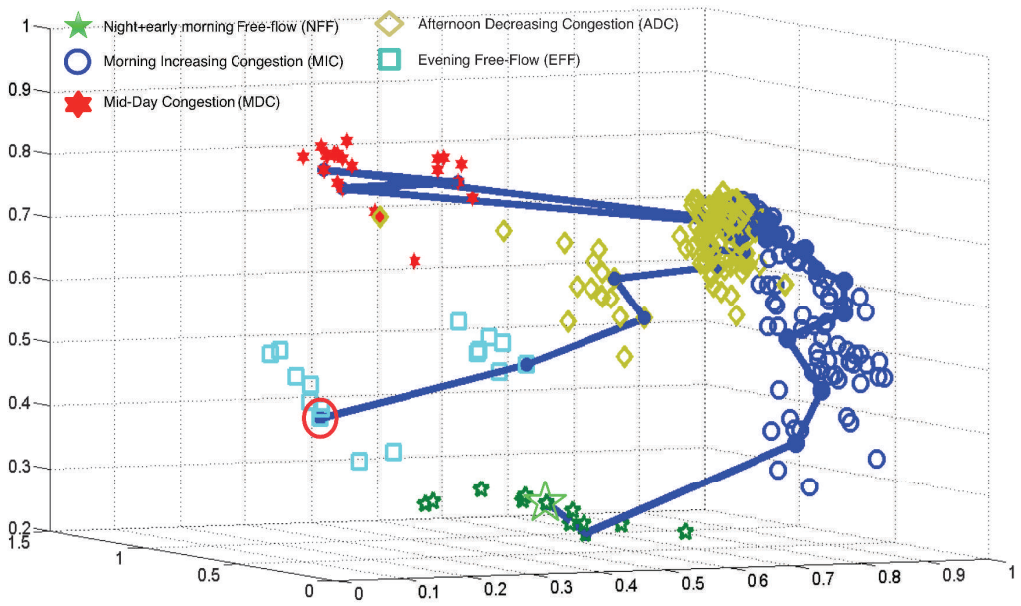
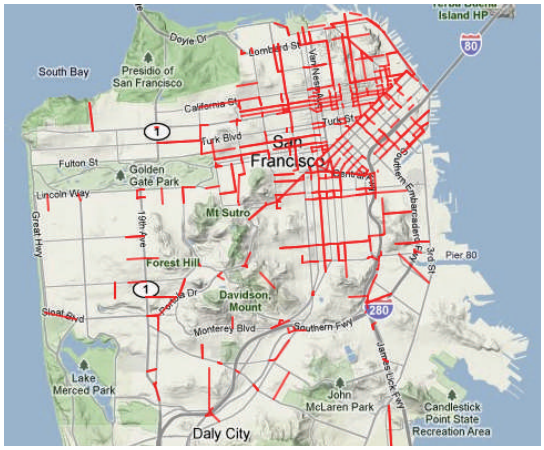
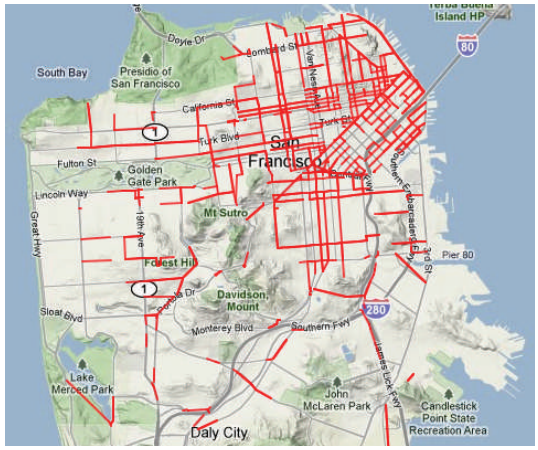


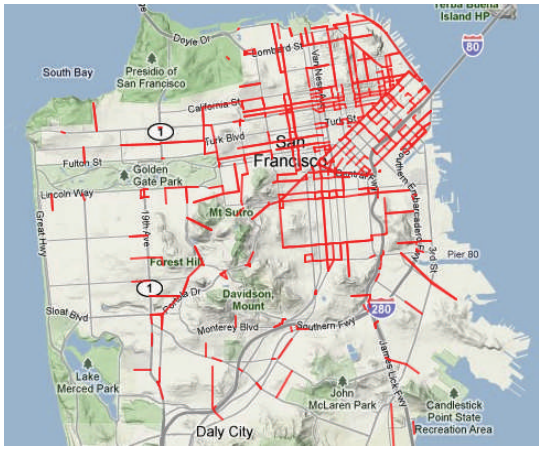
Figure 2: The clustering shows an organization of global congestion states per time of the day. The table shows the average fluidity values of each of the global state clusters. The figure shows the temporal evolution of global congestion states, projected in the 3D-NMF space using different colors and symbols to represent the five different clusters. The first and the last network estimates of the day are represented with a large star and a large circle respectively.



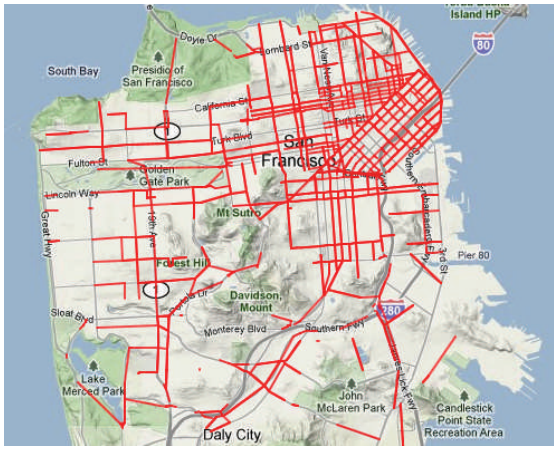
(a) Night Free-Flow (NFF)



(b) Morning Increasing Congestion (MIC)



(c) Evening Free-Flow cluster (EFF)



(d) Afternoon Decreasing Congestion (ADC)

(e) Mid Day Congestion (MDC)

Figure 3: Typical spatial configurations of traffic states for each of the five clusters. On each figure, we display the links with fluid index values less than 0.7 (congested links). Most of the congestion occurs within the region highlighted by the dashed circle, which is the downtown region of San Francisco. The NFF and EFF clusters have a smaller number of links highlighted than the MIC, ADC and MDC clusters indicating the difference in congestion levels.

As done in the primary analysis for the choice of s and for visualization purposes, we illustrate spatial layouts of the global traffic state distribution in 3D-NMF space (obtained by requesting 3 components only instead of 15), but we apply the k-means clustering algorithm in the larger 15-D NMF space. The physical interpretation of the five clusters is clear in Figure 2 in which we show all states projected in 3D-NMF, together with a typical temporal evolution trajectory of a single day. The whole trajectory is indicated by the blue line in Figure 2. The green star and red circle are the starting point and ending point of the trajectory,corresponding to traffic observations at 00:00 and 23:55 respectively. The temporal arrangements of the network-level traffic states along the trajectory underline the temporal interpretation of the five clusters: the green-star cluster corresponds to night and early morning free-flowing, from which typical day evolution goes into morning intermediate states (before 10:00) corresponding to the blue-circle cluster; mid-day congestion (red-star cluster) generally occurs between 10:05 and 15:00, and represents a clearly different congestion state in 3D-NMF space, with a sudden jump of traffic states from the blue-circle cluster to the red-star one, and sudden jump back into the afternoon intermediate state (yellow-diamond cluster) around 15:00. The traffic settles to a specific evening near-free-flow state from 18:00 to 23:55 (cyan-square cluster). Interestingly, both the projection of the global congestion states in 3D-NMF space and the clustering results in 15D-NMF show a clear distinction between morning and afternoon intermediate congestion states, and also between late evening and night/early-morning near-free-flow states.

In Figure 3, we show traffic patterns corresponding to spatial configurations of congestion for centers of each of the five identified clusters. Each cluster center is derived by averaging all elements of the corresponding cluster, so as to indicate a representative spatial configuration of traffic states of each cluster. In this figure, we display the links with fluidity index values less than 0.7 (congested links) on the Google Map screenshots. Generally, most of congestion occurs within the regions highlighted by the dashed circle in figure 3(e). This region corresponds to the downtown region of San Francisco. Compared with the downtown region, the western and southern region of San Francisco are less likely to suffer from congestion (left and bottom part in the San Francisco road network). This analysis is very useful for traffic management centers and public entities to understand the most important bottlenecks that cause heavy traffic conditions. Moreover, our results show that some of the major bottlenecks remain constant throughout the day whereas others evolve with the different traffic patterns of the day. This dynamical analysis can lead to specific management strategies to address this recurring congestion. As a matter of fact, in [13], we constructed a regression model to predict the global traffic dynamics based on the analysis results of the spatial congestion patterns. This work indicates the promising potentials of spatial congestion patterns in forecasting congestion and improving traffic management.

5 Spatial decomposition of the road network

Another motivation for using NMF in dimensionality reduction is its property to approximate original data by an additive linear combination of a limited set of "components" (a.k.a. NMF "basis"). Due to the non-negative constraints, spatial arrangements of the components are usually sparse, which means that values in most regions of each basis are (close to) zero except several localized regions. These localized regions with large values correspond to typical patterns or representative components of the original signals (the global congestion states), and typically correspond to independent "parts" of the data. Therefore, NMF is often used to extract part-based representation or latent semantic topics from the data in image processing or text classification. For example, when NMF is applied to image datasets, it automatically extracts some part-based representation of the type of objects present in the images [31, 25, 9]. We study this "part-based" representation of global congestion states to analyze the physical

significance of NMF components obtained on traffic data.

For arterial traffic, the localized regions with distinctively large values in each NMF basis correspond to a group of links with highly correlated traffic states. In this section, we construct the localized components by selecting the links which represent the top 20% largest values in each basis and indicate their spatial locations using red legends in the road network. Figure 4 shows several typical spatial arrangements of localized components, out of the fifteen arrangements learned during the NMF training. We notice that a component corresponds to streets in a localized West region (Figure 4(a)), and another to streets in the central region (Figure 4(b)), which could indicate that the traffic within each of these regions is highly correlated with each other whereas the traffic between distinct regions exhibits relatively independent behaviors. Such a characterization of independent regions of traffic dynamics is important to significantly reduce the computational costs of a large variety of estimation models, in particular estimation models based on graphical models [10, 18]. We could leverage this characterization in approximate inference algorithms to reduce the computational costs while maintaining an accurate representation of traffic dynamics and limiting the estimation error [5].

Other NMF components highlight correlations of traffic in parallel directions: in Figure 4(c), a majority of the links of the NMF component are horizontally-oriented, wheras in Figure 4(d) a majority of the links are vertically-oriented. As we highlight in Figure 4(c) and Figure 4(d),the links concentrated within the downtown tend to be more consistent with the orientational patterns. These links with similar orientations are likely to have correlated traffic dynamical behaviors, whereas traffic flows with orthogonal orientations have a less important impact on each other. These correlations properties can be used to learn the structure of the graphical model representing conditional independences between traffic states on the network (both spatially and temporarily).

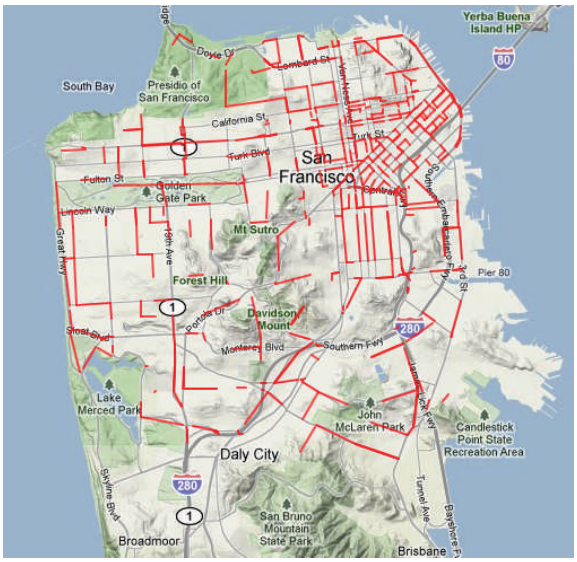
According to the physical representation of the NMF components, it seems that different NMF bases focus on different localized connected regions of the network. This could imply that NMF detects both strong correlation of traffic dynamics within each localized region and relative independence between these regions. However, this connectivity and localization of the components could be improved. Standard NMF does not guarantee connected nor localized components and the above promising results motivate us to investigate this physical representation of spatial configuration of traffic states further. A possible approach is to modify the NMF algorithm in order to favor *localized* sparsity, which should help to unveil more distinct part-based network decomposition.

6 Temporal analysis of global traffic states

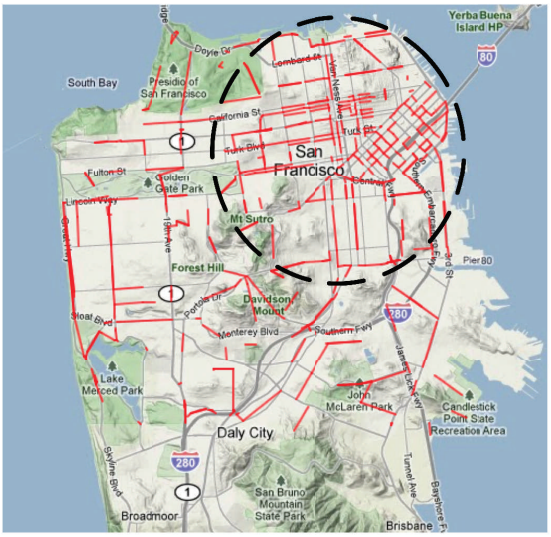
In this section, we analyze the daily dynamics of network-level congestion states projected in the NMF space. This analysis is important to understand how congestion forms and dissipates throughout the day. For each day in the studied period, we represent the trajectory of the network-level traffic states in the NMF space as the projection of the temporal sequence of the network-level traffic states from the beginning to the end of the day. The projections are linked together to form a solid curve representing the trajectory and we notice that trajectories are nearly closed in the NMF space. Note that for visualization purposes, the projection is done on the 3D-NMF space. Figure 5 (top) shows a typical day trajectory with successive temporal intervals along the trajectory plotted using different colors, to give an idea of the dynamics along the curve.

It is noteworthy that over the 184 days of reconstructed traffic data, there are only, in 3D-NMF projection, exactly seven different typical trajectories, as shown in figure 5 (center). Furthermore, our analysis shows that each one of these seven typical trajectories corresponds to

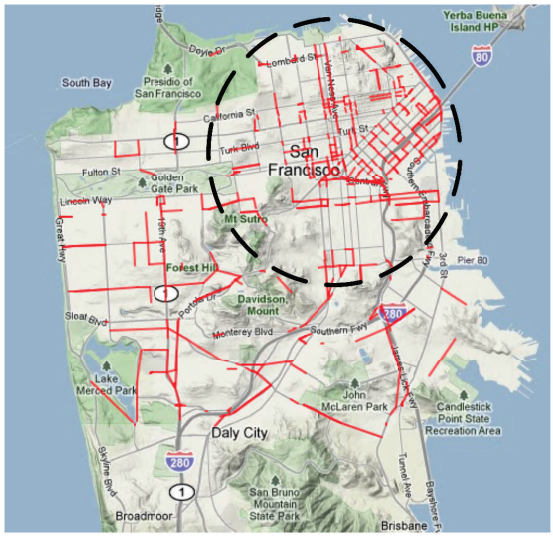
(a) "West Part" NMF component



(b) "Central" NMF component



(c) "East-West transit" NMF component



(d) "North-South transit" NMF component

Figure 4: Examples of interesting NMF components, either highlighting localized behavior (a and b), or flow-direction correlations (c and d).

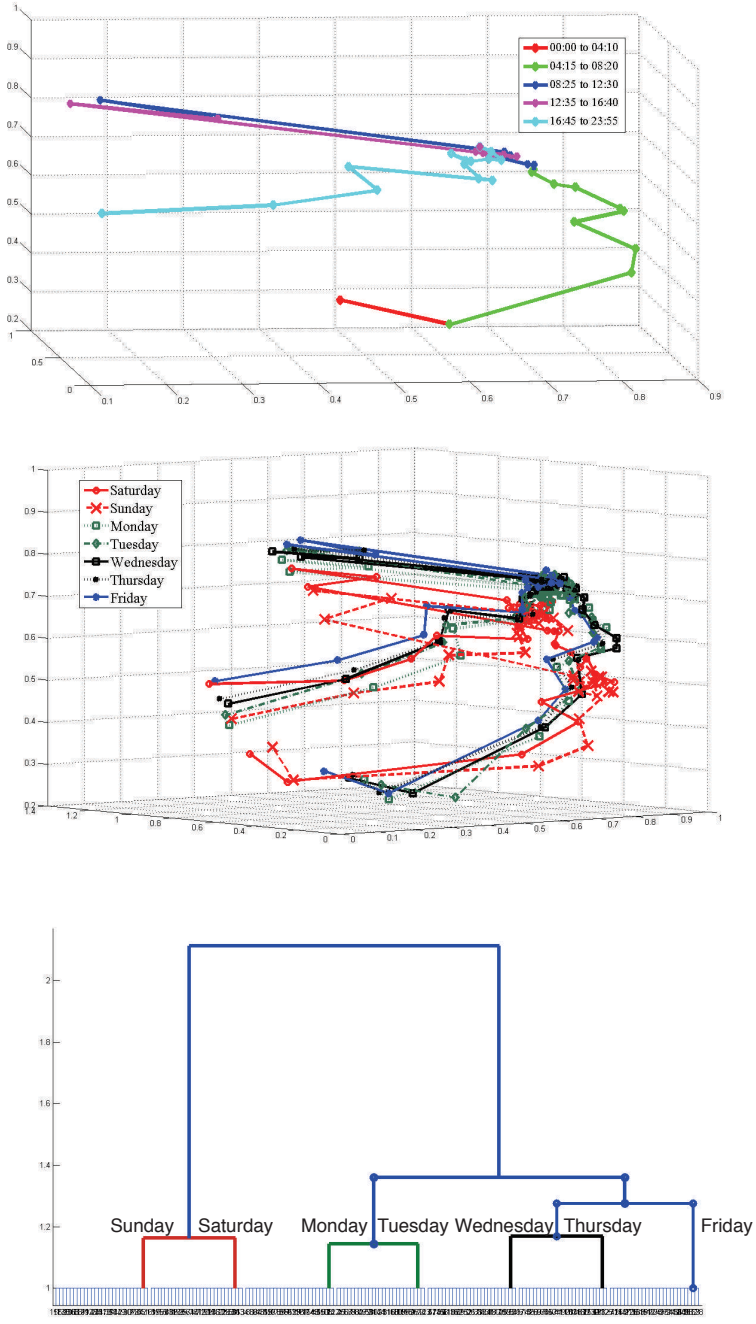


Figure 5: Daily trajectories of network fluidity indices projected in 3D-NMF space exhibit seven different typical trajectories, representing the days of the week. **Top**: Example of a daily trajectory with coloring representing the five different times of the day. **Center**: The seven different trajectories, representing a typical daily dynamic for each day of the week. **Bottom**: Dendrogram representing the hierarchical clustering analysis of the daily trajectories.

a particular day of the week and are thus called *day trajectory patterns*. Note that individual day trajectories for same day-of-week, although superposed in 3D-NMF, are slightly different one from another in 15D-NMF space, in which we perform clustering. Also, there are only two weekday holidays within the analyzed period, and they exhibit only small deviation from the ordinary same day-of-week. This may be a consequence of the estimation algorithms which does not use a holiday-specific model to process the historic data and fuse it with the real-time data.

Differences between the different day trajectory patterns concentrate within the time interval corresponding to transitions between congestion states, in particular between the morning increasing congestion and the mid-day congestion and between the mid-day congestion and the evening decreasing congestion. Characterizing these specific time intervals that represent the differences in daily dynamics allows us to identify and/or predict different evolution patterns of traffic states and to develop mid-term or long-term traffic forecast [14, 13].

In this data set, one complete evolution trajectory contains 288 sampling steps (estimations are performed every five minutes), which is represented by a 2626×288 matrix (the network has 2626 links). As for the previous sections, our analysis is done in 15-D NMF space (3-D space is only used for visualization purposes). Each trajectory is represented by a sequence of 288 network-level traffic state projected on the 15-D NMF space and denoted $\{h_1, h_2, \cdots, h_{288}\}$, where $h_i \in \mathbb{R}^{15}$. To measure similarity between trajectories $\{h_1^a, h_2^a, \ldots, h_{288}^a\}$ and $\{h_1^b, h_2^b, \ldots, h_{288}^b\}$, representing days a and b respectively, we calculate *cosine distances* between the NMF projections at corresponding times of the day and sum the cosine distances over the different estimation times $k = 1 \ldots 288$:

$$D = \sum_{k=1}^{288} \text{cosdis}(h_k^a, h_k^b), \tag{6}$$

$$\text{where cosdis}(h_k^a, h_k^b) = 1 - \frac{h_k^a \cdot h_k^b}{\| h_k^a \| \, \| h_k^b \|}. \tag{7}$$

The function *cosdis* is the cosine distance between two vectors and is defined in (7). It evaluates the cosine value of the angle between the two data vectors h_k^a and h_k^a in the 15-D NMF projection space. Larger cosine distance values indicate more important differences between the two vectors. Due to the mathematical definition of the cosine function, the derived cosine distance is normalized into the range [0,1]. Based on the defined distance measure between sequences, we can perform hierarchical clustering of daily traffic observation sequences in 15D-NMF space [29, 42]. The successive similarity-based groupings are shown on the dendrogram in Figure 5 (bottom) following the same color legends as in the middle figure. In the dendrogram, daily sequences of network-level traffic states are grouped gradually into clusters in the form of U-shaped trees. The height of each U-shaped tree (vertical axis) represents the distance between the sets of daily sequences being connected. Leaf nodes along the horizontal axis correspond to all daily sequences of network-level traffic states. We notice that at the bottom level of the hierarchical tree, daily sequences are first aggregated with respect to each day of the week. It underlines the intuition that each day of the week has a particular temporal dynamic pattern in terms of network-level traffic states. By increasing thresholds of distance settings, we trace back along the U-shaped trees towards its root. The seven days of the week are further clustered into four different groups indicating the days that tend to follow similar dynamic patterns. Weekend (Saturday and Sunday) are clustered together. As for the week days, Monday and Tuesday, representing the beginning of a week, appear to have a different temporal dynamic pattern from Wednesday and Thursday (middle of the week). Traffic dynamics on Fridays also tend to deviate slightly from that of the other days and is assigned to a separate group. As the distance threshold increases, Friday is added to the Wednesday and Thursday cluster. Therefore, we can say that there are generally three kinds of temporal dynamic patterns of network-level traffic states in the data, corresponding to the beginning of the week (Monday and Tuesday), the end

of the week (Wednesday, Thursday and Friday) and weekends (Saturday and Sunday). If we increase the threshold even more, the two clusters of week days merge leading to two clusters representing the week-end days on one side and the week days on the other side. The distance thresholds need to be increased significantly more for these two clusters to merge, which indicates the importance in the differences in daily dynamics between week days and weekends. It is expected for Monday and Friday to have different dynamics (coming back or leaving for the week-end). However, it is slightly surprising that Monday and Tuesday are clustered together while Wednesday and Thursday (and then Friday) form another cluster. The data seems to indicate a beginning of the week vs. end of the week clustering, with Friday being the most different of the other days.

7 Conclusion and discussion

In this article, we have proposed and presented: (1) a probabilistic modeling framework for efficient estimation of arterial traffic conditions from sparse probe data; (2) a novel traffic data mining approach to analyze large-scale traffic patterns and dynamics.

The proposed estimation method leverages massive amounts of historical data to learn statistical distributions of travel times and fuses them with streaming data to produce real-time estimates of traffic conditions using a Bayesian update. The Bayesian framework allows us to properly weight the relative importance of the real time and the historical data (depending on the amount of data available in real time) to produce robust estimates, even when little data is available in real-time. This model is operational in the *Mobile Millennium* system and has been producing travel time and fluidity indices since March 2010 [2].

The output of this estimation model is used as a first real-world platform for a new traffic data mining method using Non-negative Matrix Factorization to allow large-scale analysis of spatial and temporal traffic patterns. The principle is to perform dimensionality reduction, which allows for clustering of spatial congestion patterns, and easy analysis/categorization of temporal daily dynamics. Furthermore, the part-based decomposition feature of Non-negative Matrix Factorization automatically unveils areas of the road network with strong correlations.

Current and future research focus on: (1) integrating traffic flow theory and statistical models to have a more accurate modeling of traffic dynamics, both at the link [22, 23] and at the network [21] level in order to improve the estimation capabilities of the system; (2) modifications of Non-negative Matrix Factorization sparsity constraint to favor geographically-localized components; (3) taking advantage of low-dimensional Non-negative Matrix Factorization representation for performing long-term traffic prediction [13].

Acknowledgements

The authors wish to thank Timothy Hunter from UC Berkeley for providing filtered probe trajectories from the raw measurements of the probe vehicles. We thank the *California Center for Innovative Transportation* (CCIT) staff for their contributions to develop, build, and deploy the system infrastructure of *Mobile Millennium* on which this article relies. This research was supported by the Federal and California DOTs, Nokia, the Center for Information Technology Research in the Interest of Society (CITRIS), the French National Research funding Agency ANR (part of this work was supported by grant ANR-08-SYSC-017 for the "TRAVESTI" project), and French Department of Sustainable Development and Transports MEEDDM. This collaboration between CAOR of Mines ParisTech and Berkeley was initiated partly thanks to funding MEEDDM-09-SDI-003 granted by French authorities within the CalFrance franco-californian cooperation framework.

References

[1] S.J. Agbolosu-Amison, B. Park, and I. Yun. Comparative evaluation of heuristic optimization methods in urban arterial network optimization. In *12th Intelligent Transportation Systems Conference (ITSC '09)*, 2009.

[2] A. Bayen, J. Butler, and A. Patire et al. Mobile Millennium final report. Technical report, University of California, Berkeley, CCIT Research Report UCB-ITS-CWP-2011-6, To appear in 2011.

[3] P. Bickel, C. Chen, J. Kwon, J. Rice, E. Van Zwet, and P. Varaiya. Measuring traffic. *Statistical Science*, 22(4):581–597, 2007.

[4] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[5] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proc. UAI*, volume 98, 1998.

[6] D. Cai, X. He, X. Wu, and J. Han. Non-negative matrix factorization on manifold. In *The proceedings of ICDM 2008*, 2008.

[7] C. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research B*, 28(4):269–287, 1994.

[8] C. de Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 197–203, 2008.

[9] T. Feng, S.Z. Li, H.Y. Shum, and H.J. Zhang. Local nonnegative matrix factorization as a visual representation. In *Proceedings of the 2nd International Conference On Development and Learning*, 2002.

[10] C. Furtlehner, J. Lasgouttes, and A. de la Fortelle. A belief propagation approach to traffic prediction using probe vehicles. In *10th Intelligent Transportation Systems Conference (ITSC '07)*, pages 1022–1027, 2007.

[11] B. Ghosh, B. Basu, and M. O'Mahony. Multivariate short-term traffic flow forecasting using time-series analysis. *IEEE Transaction on Intelligence Transportation Systems*, 10(2):246–254, 2009.

[12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] Y. Han and F. Moutarde. Analysis of network-level traffic states using locality preservative non-negative matrix factorization. In *14th IEEE Intelligent Transport Systems Conference (ITSC'2011)*, 2011.

[14] Y. Han and F. Moutarde. Clustering and modeling of network-level traffic states based on locality preservative non-negative matrix factorization. In *8th Intelligent Transport Systems (ITS) European Congress*, 2011.

[15] X. He and P. Niyogi. Locality preserving projections. In *Proceedings of 17th Neural Information Processing Systems*, 2003.

[16] B. Hellinga, P. Izadpanah, H. Takada, and L. Fu. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C*, 16(6):768 – 782, 2008.

[17] R. Herring. *Real-Time Traffic Modeling and Estimation with Streaming Probe Data using Machine Learning*. PhD thesis, UC Berkeley, Departement of Industrial Engineering and Operations Research, 2010.

[18] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen. Estimating arterial traffic conditions using sparse probe data. In *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, Madeira, Portugal, September 2010.

[19] R. Herring, A. Hofleitner, S. Amin, T. Abou Nasr, A. Abdel Khalek, P. Abbeel, and A. Bayen. Using mobile phones to forecast arterial traffic through statistical learning. In *Proceedings of the 89th Annual Meeting of the Transportation Research Board*, Washington D.C., 2010.

[20] A. Hofleitner and A. Bayen. Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model. *IEEE Intelligent Transportation System Conference (IEEE ITSC '11)*, 2011.

[21] A. Hofleitner, R. Herring, and A. Bayen. Arterial travel time forecast with streaming data: a hybrid flow model - machine learning approach. *submitted, Transportation Research Part B*, 2011.

[22] A. Hofleitner, R. Herring, and A. Bayen. A hydrodynamic theory based statistical model of arterial traffic. *Technical Report UC Berkeley, UCB-ITS-CWP-2011-2*, January 2011.

[23] A. Hofleitner, R. Herring, and A. Bayen. Probability distributions of travel times on arterial networks: a traffic flow and horizontal queuing theory approach. *91st Transportation Research Board Annual Meeting*, January 2012.

[24] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005.

[25] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, vol.5:1457–1469, 2004.

[26] The *Mobile Millennium* Project. http://traffic.berkeley.edu.

[27] T. Hunter, R. Herring, A. Hofleitner, A. Bayen, and P. Abbeel. Trajectory reconstruction of noisy GPS probe vehicles in arterial traffic. *In preparation for IEEE Transactions on Intelligent Transport Systems.*

[28] I.T. Jolliffe. *Principal component analysis*. Springer, 2002.

[29] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, and A.Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans Pattern Analysis and Machine Intelligence*, vol.24:881–892, 2002.

[30] A. Krause, E. Horvitz, A. Kansal, and F. Zhao. Toward community sensing. In *Proceedings of ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, St. Louis, MO, April 2008.

[31] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS 2000, pp.556-562*, 2000.

[32] M. Lighthill and G. Whitham. On kinematic waves. II. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345, May 1955.

[33] C.J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, vol.19,no.10:2756–2779, 2007.

[34] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[35] W. Min, L. Wynter, and Y. Amemiya. Road traffic prediction with spatio-temporal correlations. Technical report, IBM Watson Research Center, 2007.

[36] X. Min, J. Hu, Q. Chen, T. Zhang, and Y. Zhang. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems (ITSC '09)*, 2009.

[37] C. Quek, M. Pasquier, and B. Lim. POP-TRAFFIC:A Novel Fuzzy Neural Approach to Road Traffic Analysis and Prediction. *IEEE Transactions on Inteliigent Transporation Systems*, 7(2):133–146, 2006.

[38] C. Robert. *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[39] A. Statthopoulos and M. G. Karlaftis. A multivariate state space approach for urban traffic flow modeling and predicting. *Journal of Transportation Research Part C*, 11:121–135, 2003.

[40] C. Stutz and T.A. Runkler. Classification and Prediction of Road Traffic Using Application-Specific Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*, 10(3):297–308, 2002.

[41] X. Sun, L. Munoz, and R. Horowitz. Mixture Kalman filter based highway congestion mode and vehicle density estimator and its application. In *Proceedings of the 2004 American Control Conference*, pages 2098–2103, Boston, MA, 2004.

[42] G.J. Szekely and M.L. Rizzo. Hierarchical clustering via joint between-within distances:extending ward's minimum variance method. *Journal of Classification*, vol.22:151–183, 2005.

[43] A. Thiagarajan, L. Sivalingam, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, and H. Balakrishnan. VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones. In *7th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Berkeley, CA, November 2009.

[44] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended kalman filter: a general approach. *Transportation Research Part B*, 39:141–167, 2005.

[45] D. Work, S. Blandin, O. Tossavainen, B. Piccoli, and A. Bayen. A traffic model for velocity data assimilation. *Applied Research Mathematics eXpress (ARMX)*, April 2010.

[46] H. Yin, S.C. Wong, J. Xu, and C.K. Wong. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85–98, 2002.

Résumé

Ce mémoire présente un panorama des travaux de recherche que j'ai menés et dirigés de 2006 à 2013. Ceux-ci couvrent diverses applications des techniques d'apprentissage statistique, en analyse temps-réel d'images et en fouille de données. Le domaine principal est celui des systèmes avancés d'aide à la conduite (Advanced Driving Assistance Systems, ADAS) : détection et reconnaissance de panneaux routiers et de catégories d'objets telles que voitures et piétons, ainsi qu'analyse et prédiction de trafic routier. Des travaux ont aussi concerné l'identification de personnes dans un contexte vidéo-protection, ainsi que l'identification d'objets en 3D pour des applications en robotique. La présentation se veut synthétique, en renvoyant pour les détails aux publications jointes en annexe, et en récapitulant pour chaque sous-domaine les contributions apportées et les expertises acquises.

Quatre thèses ont déjà été soutenues et un post-doctorat encadré sur les recherches présentées. Actuellement les travaux continuent avec deux thèses en cours et un nouveau post-doctorant supervisé, en évoluant d'une part vers un plus haut niveau d'abstraction dans l'analyse d'images (inférence de contexte), et d'autre part vers la localisation par vision en environnement intérieur et surtout la reconnaissance de gestes, dans des contextes automobiles et robotiques.

Summary

This document presents an overview of the research works I have conducted and supervised between 2006 à 2013. These span several applications of statistical machine-learning techniques, in real-time video analysis and data-mining. The main domain is Advanced Driving Assistance Systems (ADAS): detection and recognition of traffic signs and object categories such as cars and pedestrians, and road traffic mining and prediction. Some works were also related to person identification in video-protection context, as well as objects identification in 3D for robotics applications. The presentation tries to be synthetic, pointing for the details to publications attached in appendix, and recapitulating for each field the contributions brought and acquired expertise.

Four PhD thesis have already been defended, and one post-doc supervised, on the research presented. Works are still undergoing on the same line, with two on-going PhD, and an evolution towards: on one hand, higher abstraction level of image analysis (context inference), and on the other side diversification to indoor localization with vision and gestures recognition, in automotive and robotics context.