Vision and Wi-Fi fusion in probabilistic appearance-based localization.

Mathieu Nowakowski[1, 2], Cyril Joly[1], Sébastien Dalibard[2], Nicolas Garcia[2] and Fabien Moutarde[1]

Abstract

This paper introduces an indoor topological localization algorithm that uses vision and Wi-Fi signals. Its main contribution is a novel way of merging data from these sensors. The designed system does not require to know the building plan or the positions of the Wi-Fi Access Points. By making Wi-Fi signature suited to the FABMAP algorithm, this work develops an early-fusion framework that solves global localization and kidnapped robot problems. The resulting algorithm has been tested and compared to FABMAP visual localization, over data acquired by a Pepper robot in three different environments: an office building, a middle school and a private apartment. Numerous runs of different robots have been realized through several months for a total covered distance of 6.4km. Constraints were applied during acquisitions to make the experiments fitted to real use cases of Pepper robots. Without any tuning, our early-fusion outperforms the performances of visual localization in all testing situations and with a significant margin in environments where vision faces problems such as moving objects or perceptual aliasing. In such conditions, 90.6% of estimated localizations are less than 5m away from ground truth with our early-fusion framework compared to 77.6% with visual localization. Furthermore, compared with other classical fusion strategies, the early-fusion produces the best localization results since in all tested situations, it improves visual localization results without damaging them where Wi-Fi signals carry little information.

Keywords

Topological localization, kidnapped robot, low-cost sensors, visual appearance, Wi-Fi fingerprints, data fusion.

1 INTRODUCTION

1.1 Problem statement

This paper addresses the problem of indoor localization for mobile service robotics. The current market trend consists in a mass deployment of affordable mobile robots interacting with humans. This raises the need for low-cost solutions enabling those robots to map their environment, and constantly know where they are when they move in it. Numerous projects have been proposed to solve the problem of localization. However, most of these solutions are based on the use of expensive sensors, such as laser range finders, and are designed for specific platforms (Kummerle et al. 2009).

The need for low-cost localization solutions has focused some research on the use of visual sensors. One investigation field aims attention at solving the problem of place recognition by using visual appearance (Ulrich and Nourbakhsh 2000; Angeli et al. 2008; Sünderhauf and Protzel 2011; Lowry et al. 2014; Lynen et al. 2014). Such algorithms try to associate query images to already mapped places, represented by their visual appearance. Lowry et al. (2016) presents a comprehensive survey of appearance-based approaches.

Our work is built upon the Fast Appearance-Based Mapping algorithm (FABMAP), introduced in Cummins and Newman (2008), that uses visual appearance to detect loop closures. This algorithm achieves robust localization with a low rate of false loop closure detection and can manage big maps by employing an inverted-index (Cummins and Newman 2011).

However, place recognition algorithms using visual appearance have to face the well-known problem of *perceptual aliasing*. Perceptual aliasing happens when two different locations share similar visual appearances (see example on Figure 1). This problem is inherent in repetitive environments. A solution is the use of a multi-sensors localization for disambiguating such situations.

For example, using Wi-Fi helps to disambiguate cases where several locations have similar visual appearances. For example, corridors on opposite sides of a building, or at different floors, have different Wi-Fi signatures but can share comparable appearances. Recent work has introduced a way of including Wi-Fi data in the FABMAP algorithm (Wietrzykowski et al. 2017), but it does not benefit from advantages of both sensors since it only considers Wi-Fi signals.

[1] MINES ParisTech, PSL - Research University, Centre de Robotique, 60 Bd St Michel, 75006 Paris, France.

[2] SoftBank Robotics Europe, 43 Rue du Colonel Pierre Avia, 75015 Paris, France.

Corresponding author:

Mathieu Nowakowski, SoftBank Robotics Europe, 43 Rue du Colonel Pierre Avia, 75015 Paris, France.

Email: mnowakowski@softbankrobotics.com

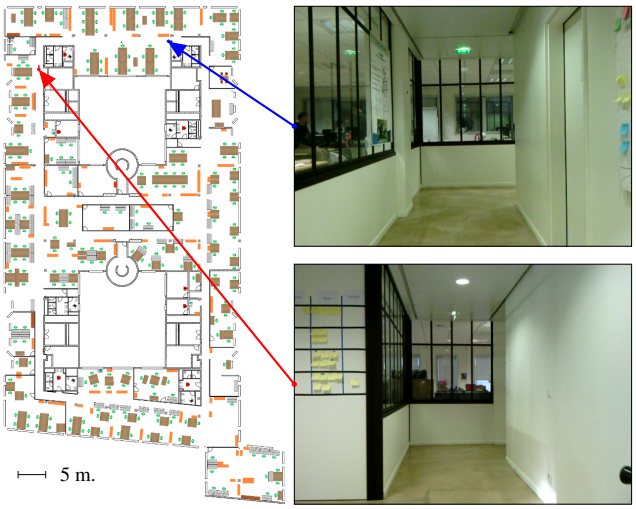


Figure 1. Perceptual aliasing. Two distant locations in a building share similar visual appearances. However the perceived Wi-Fi signatures in these locations are different enough to distinguish them.

In this paper, a novel way of merging visual and Wi-Fi data is introduced in order to solve the global localization and the kidnapped robot problems.

1.2 Use case and social robots

Our novel merging scheme has been tested on Pepper robots, visible on Figure 2. Pepper is a social robot designed by SoftBank Robotics. It can be found mainly in Japanese shops, where it holds receptionist or demonstrator functions. One of Pepper's key strengths directly comes from the interaction it has with users. These interactions are reinforced and simplified thanks to the human behaviours of the robot. For this reason, practical uses of Pepper result in special constraints taken into consideration in this work.

First, the localization process must not hinder human-robot interaction. As such, it cannot take control of the joints of the robot, for example to make the robot look away from the human it is interacting with. Second, some Pepper robots are deployed in environments with little or no internet access. This implies that the localization function must be able to run, at a reasonable framerate, with the - limited - robot computing power.

Figure 2. Pepper robot: designed for making its interaction with human being as natural and intuitive as possible. It has been first presented in Lafaye et al. (2014).

A challenging use case comes from the initialization of the robot pose in its environment. When a Pepper robot is switched on to start its work day, it has to immediately localize itself without any help. This known problem is commonly called the global localization problem. It happens when the robot has no information about its previous pose. A very similar issue is often referred to as the kidnapped robot problem. This problem corresponds to situations where the robot has a false prior on its pose. Both problems are very similar, and methods solving global localization problem can be adapted to solve the other. Our work aims at solving these problems in indoor situations that are typically the operation environments of Pepper robots.

1.3 Related Work

In recent years, several indoor localization algorithms based on Wi-Fi sensors have been introduced (Howard et al. 2003; Olivera et al. 2006; Rohrig and Kiinemund 2007; Biswas and Veloso 2010; Huang et al. 2011; Boonsriwai and Apavatjrut 2013; Jirku et al. 2016). This popularity can be explained by two reasons. First, the Wi-Fi coverage in urban environment is dense enough for being used in localization task. Second, it is easy to equip mobile robots with Wi-Fi sensors.

Because the Wi-Fi localization error is bounded (Olivera et al. 2006), some algorithms choose to fuse Wi-Fi with other sensors (Aparicio et al. 2008; Mirowski et al. 2013; Ocaña et al. 2005). Visual and Wi-Fi localizations are particularly complementary. Even if Wi-Fi localization is less accurate, (Liu et al. 2012), it does not suffer from perceptual aliasing, visually dynamic or repetitive environments. A lot of strategies take advantage of this synergy and use visual and Wi-Fi sensors to create a low-cost localization.

Most of work focusing on solving the localization problem from these sensors uses particle filters for fusion as in Schwiegelshohn et al. (2013); Quigley et al. (2010); Liu et al. (2017). However, the hypothesis converge if there is enough motion. Other approaches are sequential, and usually Wi-Fi guided. They define a set of possible locations from Wi-Fi data, and perform visual localization over it (Ruiz-Ruiz et al. 2011; Werner et al. 2011; Nowicki 2014; Jiang and Yin 2015). Finally, some methods consist in choosing which sensor is the most reliable for current estimation (Biswas and Veloso 2013). Yet, these two last approaches are both likely to suffer from one misled sensor.

1.4 Contribution

Our main contribution is a novel way of merging Wi-Fi and vision for localization tasks. We propose an *early-fusion* process for combining visual and Wi-Fi data that takes the same inputs as the classical FABMAP. In comparison with related work, our approach looks for a compromise on the current estimation by considering data from both sensors together.

The core of our algorithm was first presented in Nowakowski et al. (2017). In this paper, our method deeper considers the Wi-Fi signal. **Experiments from Nowakowski et al. (2017) are increased with new data making the results presented in this paper more reliable.** New tests are specifically realized in multiple environments to evaluate the generality of our solution.

1.5 Paper organization

To introduce our localization solution using Wi-Fi and vision, the FABMAP algorithm is first briefly presented in section 2. The readers familiar with FABMAP can start with section 3 that explains how Wi-Fi data is made compatible with FABMAP formalism. Our early-fusion process is then introduced in section 4 with other merging strategies for comparison. Finally, our experimental acquisitions and localization results are presented and discussed in section 5.

2 Fast Apperance-Based Mapping

In Cummins and Newman (2008, 2011), the authors introduce FABMAP, for Fast Appearance-Based Mapping, a localization algorithm based on the visual appearance. FABMAP discretizes the environment into a succession of topological nodes. Each node constitutes a location L_i, and is associated with one or several visual observations. When processing a query image, FABMAP uses the results of an offline learning stage. This learning is achieved before the map creation and the localization phase over external data.

Given a query image, the goal of FABMAP is to compute the following value for each place L_i in a topological map:

$$p(L_i|\mathcal{Z}^k) = \frac{p(Z_k|L_i, \mathcal{Z}^{k-1})p(L_i|\mathcal{Z}^{k-1})}{p(Z_k|\mathcal{Z}^{k-1})} \quad (1)$$

where Z_i is the i^{th} observation and $\mathcal{Z}^i$ is the set of all observations, up to i. Three terms can be identified in (1): the likelihood $p(Z_k|L_i, \mathcal{Z}^{k-1})$, the normalization term $p(Z_k|\mathcal{Z}^{k-1})$, and $p(L_i|\mathcal{Z}^{k-1})$ that can be considered as a prior knowledge on the current pose because the approach assumes independence between current pose and past observations. Note that in our work, this last term is not used because our intention is to solve the global localization problem.

The three next sub-sections respectively introduce the computations of the observation Z_k, the likelihood and the normalization term.

2.1 Visual Appearance Description

The first step of FABMAP is to transform a query image into a compact image descriptor that is suited to the localization context. This compact descriptor is called observation vector, or visual appearance, and is noted Z in (1). To do this, FABMAP uses the bag-of-words approach introduced in computer vision in Sivic and Zisserman (2003). Keypoints are extracted in the image, and their descriptors are then associated with words of a vocabulary. In FABMAP, the observation Z indicates which words of the vocabulary are present on the query image.

For a vocabulary of N words, Z thus contains N binary values indicating the presence or absence of the corresponding word in the query.

The vocabulary used comes from an offline learning. It is usually built thanks to a clustering method like the k-means performed over a lot of keypoints descriptors extracted from learning images. Learning images can be chosen in databases according to the operating environment of the algorithm (indoor, outdoor, etc.).

2.2 Observation Likelihood

The second step constitutes the core of the algorithm and computes the likelihood term $p(Z_k|L_i, \mathcal{Z}^{k-1})$ conditioned on the location. This term is simplified into $p(Z_k|L_i)$ in Cummins and Newman (2008), assuming independence between current pose and past observations. Approaches using the bag-of-words framework can compute similarity scores between queries and references thanks to methods like the *Term Frequency - Inverse Document Frequency* (Salton and Buckley 1988), or hierarchical vocabularies (Nister and Stewenius 2006). The main contribution of FABMAP is the use of a Chow Liu tree, (Chow and Liu 1968), that captures correlations between the different words of the vocabulary. This approach is motivated by the fact that certain visual words are often detected on specific objects and thus, tend to co-occur. Generally, these correlations are learnt offline, with the database used during the vocabulary creation.

Experiences realized in Cummins and Newman (2008) show that learning these correlations helps to avoid false associations due to perceptual aliasing. It also helps to achieve correct associations between images, even if they share few words in common.

2.3 Normalization

The normalization step allows to detect unknown locations. In Cummins and Newman (2008), the authors split $p(Z_k|\mathcal{Z}^{k-1})$ into two sums, one representing the visited locations M, the other the unknown world $\bar{M}$:

$$\begin{aligned} p(Z_k|\mathcal{Z}^{k-1}) &= \sum_{m\in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) \\ &+ \sum_{u\in \bar{M}} p(Z_k|L_u)p(L_u|\mathcal{Z}^{k-1}) \end{aligned} \quad (2)$$

The second summation cannot be evaluated directly. The authors of FABMAP propose to approximate (2) by:

$$\begin{aligned} p(Z_k|\mathcal{Z}^{k-1}) &\approx \sum_{m\in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) \\ &+ p(L_{new}|\mathcal{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k|L_u)}{n_s} \end{aligned} \quad (3)$$

where, $p(L_{new}|\mathcal{Z}^{k-1})$ corresponds to the probability of being in a new location, and is a user-specified input of the algorithm (set to 0.9 in Cummins and Newman (2008)). The second sum of equation (3) consists in sampling an observation Z to create a place model associated with unknown location. The sampling of Z is realized from training set of n_s images.

In addition to the vocabulary, the Chow Liu tree and the n_s samples, the authors of Cummins and Newman (2008) list some user-specified inputs. In our work, these parameters are set to the values specified in Cummins and Newman (2008). Figure 3 summarizes the successive steps of the algorithm.

3 Including Wi-Fi data in FABMAP

In the related literature (Biswas and Veloso 2010; Liu et al. 2012; He and Chan 2016), a Wi-Fi signature - sometimes referred to as a *fingerprint* - consists of a list of visible Access Points (APs), each one being characterized by its

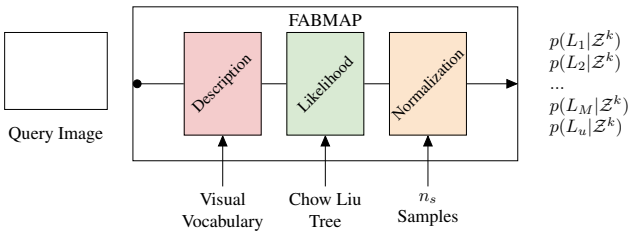


Figure 3. Inputs of each steps of the FABMAP algorithm.

MAC address and its signal strength (Received Signal Strength Indication - RSSI). Most of Wi-Fi localization algorithms collect Wi-Fi signatures during an exploration stage and then generate a map modeling the distribution of Wi-Fi signals in the environment. Such approaches have the advantage of not requiring to know the positions of the APs in the environment. These strategies particularly suit the topological localization of FABMAP once they are made compatible with it (Wietrzykowski et al. 2017; Nowakowski et al. 2017). This section introduces how our work integrates Wi-Fi information into FABMAP, following the steps of Figure 3.

3.1 Defining a Wi-Fi vocabulary

The first step of visual FABMAP consists in turning a query image into a vector of binary values: the visual appearance. This vector is built thanks to a vocabulary of visual words. To do the same thing with Wi-Fi signatures, a correct definition of what is a Wi-Fi word is needed.

In previous work (Nowakowski et al. 2017), each MAC address characterizes one Wi-Fi word. In contrast with classical visual vocabulary, Wi-Fi vocabulary can therefore only be built online or after an exploration phase. Indeed, it is useless to define a global and complete Wi-Fi vocabulary since it is not possible to know the APs a robot is going to encounter in an environment before exploring it. After an exploration phase, the encountered MAC addresses constitute the vocabulary. The values of Wi-Fi observation vector $Z_{\text{Wi-Fi}}$ indicate which known APs are visible in a Wi-Fi signature. However, this definition does not take advantage of the information carried by the RSSI.

In Wietrzykowski et al. (2017), the authors discretize the Wi-Fi signal strength over 10 bits in range $]-110dBm, -10dBm]$. Each bit is associated with a threshold and is set to 1 if the perceived signal strength exceeds this threshold. Consequently, the Wi-Fi signal identified by a MAC address is described by 10 words in the observation vector.

That last method can be seen as an *incremental* representation. Even if such representation is expected to manage small temporal variations of RSSI, it is possible that the strongest correlations learnt between words of the vocabulary come from Wi-Fi words associated with the same APs. Our work thus introduces another technique that we call the *exclusive* representation. A signal coming from a MAC address is still described by multiple words, but only one word is set to 1 following its RSSI. A range for usable RSSI is defined and split into multiple b bins. Each bin is associated with one binary value. When the perceived signal strentgh is between the upper and lower limits of a bin, its value is set to one. Otherwise it is set to zero. Thus, it is possible to compute a vector of b words for each perceived signal in a Wi-Fi signature. A comparison of the incremental and exclusive representations is shown on Figure 4 for a Wi-Fi signal identified by a MAC address and described by $b = 5$ Wi-Fi words.

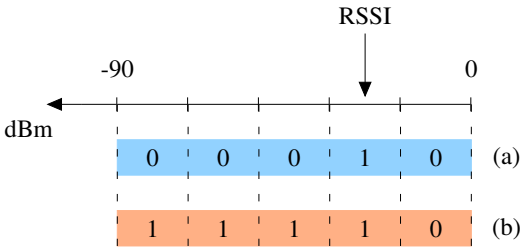


Figure 4. Example of Wi-Fi signal perceived from an AP being encoded over $b = 5$ words with the exclusive (a) and the incremental (b) methods.

With both approaches, $Z_{\text{Wi-Fi}}$ is the ordered concatenation of the vectors describing the signals of all known APs in the current signature. Each unperceived AP of the vocabulary in current signature is associated with a vector full of zeros. So with K_W known MAC addresses, a Wi-Fi signature is transformed into an observation vector $Z_{\text{Wi-Fi}}$ of $b \times K_W$ binary values.

At a given location, Wi-Fi signal strength varies significantly. To make their inputs more reliable, most Wi-Fi localization solutions stay motionless for a while, and compute mean and standard deviation of the RSSI coming from each AP (Biswas and Veloso 2010; Jirku et al. 2016). This approach is not possible in our use case because of Pepper's motion behaviours. Therefore, the issue of how many words to use for encoding Wi-Fi intensity arises without the possibility of averaging. If the use of a lot of Wi-Fi words is expected to increase the accuracy, a too precise representation can mislead our system. This issue is investigated in 5.8.

3.2 Tree Structure

Previous sub-section explained that Wi-Fi vocabulary can only be built and completed online or after an exploration phase. It is thus possible to build a Chow Liu tree that catches correlations between Wi-Fi words from the collected Wi-Fi signatures. However, in order to avoid the learning of redundant correlations, it is necessary to make sure that observation vectors come from different places. In our system, spatial difference is ensured thanks to odometry data and temporal difference is ensured by timestamps.

3.3 Normalization and virtual Wi-Fi locations

In the visual world, sampling an observation for normalization is easy. To do this, training images needed during the offline learning and coming from other environments are used. In the Wi-Fi world, employing this method is not so simple. One Wi-Fi vocabulary is specific to one environment. For this reason, using real Wi-Fi signatures collected in training environments does not make a lot of sense since the computed observation vectors would only be composed

of zeros. A solution is to simulate virtual Wi-Fi signatures according to the collected data.

Multiple variations in Wi-Fi signatures can be identified for unknown locations considering the propagation of Wi-Fi signals. An unknown location can:

- share the same Wi-Fi signature as a mapped place,
- have unknown APs in its Wi-Fi signature,
- bring up new combinations of known APs.

Virtual Wi-Fi signatures are simulated considering these changes.

In our measurements, the number of APs in a Wi-Fi signature follows a normal distribution for a specific environment. The mean μ and the standard deviation σ on the number of APs perceived in the Wi-Fi signatures collected during the exploration are thus identified. To randomly generate a virtual Wi-Fi signature, our method first selects a number of perceived APs following the normal distribution $\mathcal{N}(\mu, \sigma)$. Each of these simulated APs is then randomly associated with a known or unknown MAC address and an RSSI in the usable range of strength defined (in our case $]-90dBm, 0dBm]$).

With this formalism, FABMAP localization results based on Wi-Fi data can be computed. However, as shown in section 5, using Wi-Fi data alone shows poorer accuracy than visual-based localization. Next section explains how to take advantage of both visual and Wi-Fi sensors.

4 Merging Visual and Wi-Fi data

This section introduces various fusion strategies, and among them, our early-fusion process. To our knowledge, this approach has never been studied for solving the global localization problem using Wi-Fi and visual data. Our methodology is shown on Figure 5. The following subsections present more classical ways of merging vision and Wi-Fi, discuss the interest of the early-fusion and our choices concerning the inputs of this algorithm.

4.1 Sequential fusion

In multi-sensors system, sequential fusions try to take advantage of the different levels of accuracy of the different sensors. Two methodologies can be identified when using sequential fusion with Wi-Fi and visual data:

1. Wi-Fi-guided fusion, in which a visual localization is realized over possible locations determined from Wi-Fi data;
2. Wi-Fi check fusion, where the result from visual localization must be confirmed by Wi-Fi data.

These approaches use the fact that Wi-Fi localization is less accurate than the visual one, but never produces aberrant results. However, FABMAP normalization style enables the algorithm to detect loop closures with the assumption that no one has been missed. In practice, FABMAP detects at most one loop-closure for each query. However, in order to make the presented sequential fusions work, Wi-Fi localization has to furnish a set of several possible locations. Work presented in Stumm et al. (2013) tackles this issue by introducing another normalization approach. The same technique is used in our evaluated sequential merging strategies.

4.2 Late-fusion and Early-fusion

An intuitive way of merging localization results coming from multiple sensors can be called the *late-fusion*. Each sensor s provides a probability $p_s(L_i|\mathcal{Z}_s^k)$ of being in a location L_i knowing its observations $\mathcal{Z}_s^k$. For a multi-sensor platform composed of two sensors s_1 and s_2, the result p_{lf} from the late-fusion can be written as:

$$p_{lf}(L_i|\mathcal{Z}_{s_1,s_2}^k) = \alpha \times \frac{p_{s_1}(L_i|\mathcal{Z}_{s_1}^k)\, p_{s_2}(L_i|\mathcal{Z}_{s_2}^k)}{p(L_i)} \quad (4)$$

Where α ensures that $\sum_i p_{lf}(L_i|\mathcal{Z}_{s_1,s_2}^k) = 1$. Focusing on the global localization problem and considering that n_M locations have been mapped, plus one location associated with the unknown world, note that for every i, $p(L_i) = \frac{1}{n_M+1}$.

With the late fusion, each localizer gives a result according to the perceptual area of its sensor. Therefore, a sensor misled by perceptual aliasing can clearly pervert the system. In this work, we propose to merge the data before the steps of likelihood computation and normalization. We thus realize an *early-fusion*. The output probabilities of the algorithm are noted $p_{ef}(L_i|\mathcal{Z}^k)$. The idea is to concatenate the observation vectors obtained from visual and Wi-Fi sensors **as presented in Figure 5.** This resulting observation vector then becomes the query for a single localization algorithm. The early-fusion asset is the computation of a compromise between the different sensors.

The overall process of early-fusion thus corresponds to the classical fabmap using as input the concatenation of visual and Wi-Fi appearance vectors. Computing $p_{ef}(L_i|\mathcal{Z}^k)$ is done in the same way as described in subsections 2.2 and 2.3. The learning of correlations between vocabulary words driving the computation of likelihood and normalization terms is presented in the following subsection.

4.3 Which correlations to learn for Early-fusion?

When merging Wi-Fi and visual data by early-fusion, the question of which correlations to learn remains. We choose to split the correlations learning. Instead of learning one tree, two trees capture words that respectively co-occur in the visual vocabulary and in the Wi-Fi vocabulary. Both trees are passed together as a single data structure to the localization algorithm. This split learning can be explained by two reasons:

1. First, learning new correlations between all Wi-Fi and visual words is not obvious. For instance, to learn new visual correlations from nodes collected during exploration, all visual words of the vocabulary must have been seen at least once during the exploration. In practice, it is not possible to ensure that this assumption is verified.
2. Second, the normalization step encourages this choice. Indeed the visual and Wi-Fi unknown world is too difficult to simulate whereas using split learning allows to simply concatenate unknown visual and Wi-Fi samples.

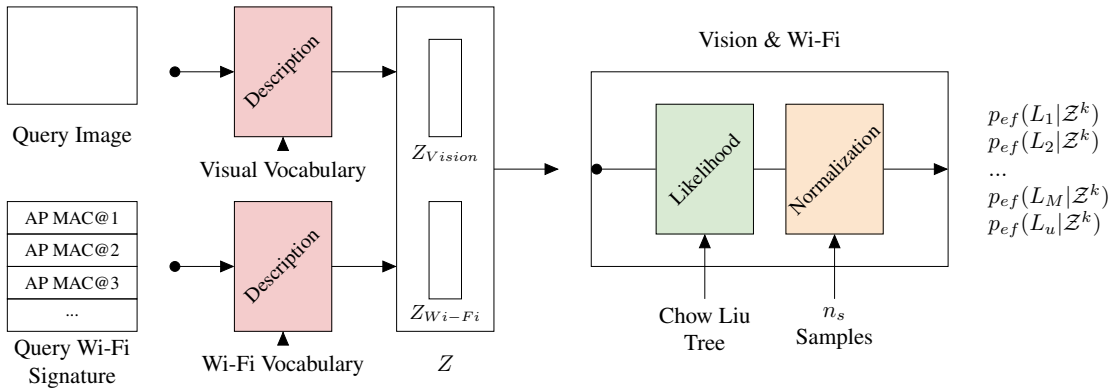


Figure 5. Early-fusion framework, using visual and Wi-Fi data.

Even if the correlations learning is split, section 5 shows that early and late fusions generate different results. Both fusions are also compared to sequential fusions (4.1) that are more classical Wi-Fi and vision merging styles.

5 Evaluation

5.1 Experimental conditions

Our algorithm was evaluated on data acquired by several Pepper robots. **Each robot used is equipped with the same camera model and the same Wi-Fi device. In these conditions, the differences in localization performances are not significant from one robot to another.** Acquisitions were done by driving a Pepper thanks to a remote control. During the acquisitions, the robot autonomously acquired images every 2s, and Wi-Fi signatures every 10s, while moving at an average speed of 0.35 m/s. Diverse constraints came from the need of a visually natural localization (1.2). For instance, blurry images could result from the fact that we did not want the robot to stop for image acquisitions. Moreover, motion behaviours of the robot were kept. When navigating, Pepper looks in the direction of its motion. So in straight lines, images were not taken in discriminative perpendicular directions (left or right), but in the direction of movement. The attached video illustrates the data acquisition and localization process.

In order to see the performances of our system, multiple environments were used in this paper. During our acquisitions, the environments were not specifically equipped and the tested localization algorithms only used the existing visual and Wi-Fi landmarks. Thus, data has been collected in:

- the open-spaces of an office building on two different floors (5.4),
- corridors and classrooms of a junior high school spread over three different floors (5.5),
- the different rooms of a private appartment (5.6).

Each of these environments has some specific aspects. However, because Pepper is intended to be handled by non-expert users, the same acquisition scheme was kept in all these situations.

The paths we ran constitute a set of 10120 images and 6110 Wi-Fi signatures. The total covered distance is 6.4km long. We also paid attention to the diversity and reality of our acquisition scenarios: occlusions, dynamic environment, realistic velocity, user interactions, blur, various times of day, etc.

Finally, the visual vocabulary used was learnt from 3000 images of indoor scenes, extracted from database presented in Quattoni and Torralba (2009). To satisfy the real-time constraint, FAST keypoints detection was used (Rosten and Drummond 2006; Rosten et al. 2010) combined with the ORB binary descriptors (Rublee et al. 2011).

5.2 Annotations: initial exploration and localization

Our formalism defines topological nodes by an associated couple (image ; Wi-Fi signature). In practice, images were collected faster than Wi-Fi signatures. Therefore, in order to associate acquired images with Wi-Fi signatures, two scenarios were chosen respectively for mapping and localization phases:

1. During the initial exploration phase, images were associated with the estimated temporally closest Wi-Fi signature.
2. During the localization tests, images were linked to the last Wi-Fi signature acquired.

The positions of all collected images were manually annotated, resulting in ground truth of node positions. For each environment, the different acquisitions were split in two: 40% were used for map creation and 60% constituted our queries. Each query corresponded to a mapped place. Global localization was thus realized in an entirely mapped environment.

Note that all Wi-Fi correlations were learnt over examples used for the mapping.

5.3 Evaluation metrics

For each query, the highest score computed by the algorithm was considered as the current localization in the map:

$$L_{max} = \underset{L_i \in M \cup \bar{M}}{\operatorname{argmax}} \; p(L_i|\mathcal{Z}^k) \tag{5}$$

To evaluate accuracy, the Euclidean distance between the annotated positions (x, y) of the query and the associated mapped place L_{max} was computed. This distance was set to infinity when the algorithm was mistaken between two floors

or two environments. This choice is explained by the fact that these kind of errors are much more problematic in practical cases. Finally, when $L_{max} = L_u$, corresponding to unknown location, our evaluation considered the result as rejected.

Produced results are discussed in the following. We have used the FABMAP2.0 algorithm (Cummins and Newman 2011), and adapted the open source implementation introduced in Glover et al. (2012) to our use. For the set of queries, localizations were computed using visual data only, Wi-Fi data only, and different merging styles: early and late fusions, and the two sequential fusions presented in sub-section 4.1.

In order to appraise the global localization performances of the different merging approaches, results are presented according to two types of graphs showing:

- the *rate of correct localizations*: measured as the cumulative distribution of distances between estimated localizations and ground truth.
- the *rate of misplaced localizations*: measured as the percentage of queries leading to localizations farther away from ground truth than a given distance.

Thus, both types of graphs show different information. If the first one gives information on the accuracy of the algorithms, the second one highlights its errors. On graphs showing the rate of misplaced localizations, the different rejection rates can be deduced as:

$$1 - Rate(error > d = 0) \quad (6)$$

Appearance-based SLAM algorithms are sometimes evaluated on a topological best-place correct rate. Because of the nature of Wi-Fi based localization, which is not as precise as visual localization, we chose to evaluate our algorithm on its metrical correctness. It also makes sense for the kidnapped-robot problem to know how far, geometrically, the algorithm is from ground truth. If needed, the presented results can be interpreted as a topological best-places correct rate. Given the robot moving speed and data acquisition frequency, references are placed, in average, every 70cm. As stated earlier, queries were made only on an already mapped route, which makes each query at most 35cm away from its closest reference. Our main performance index is the rate of correct localizations within 5m, which corresponds to a match with one of the 10 best topological places.

In the following sub-sections, localization results in various environment are presented. First, signals strengths in Wi-Fi signature will not be used. Only the presence or absence of known access points will be taken into account. The use of RSSI will be then investigated in sub-section 5.8.

5.4 Localization in an office building

The environment where most of our acquisitions were realized is the office floors of SoftBank Robotics Europe that are mainly composed of open spaces. This facet is significant considering the propagation of Wi-Fi signals. In such environments, Wi-Fi signatures are more difficult to distinguish because there are no obstacles creating important changes. Possible uses of Pepper in large indoor environments, like shops, malls or airports, have motivated this choice of testing conditions.

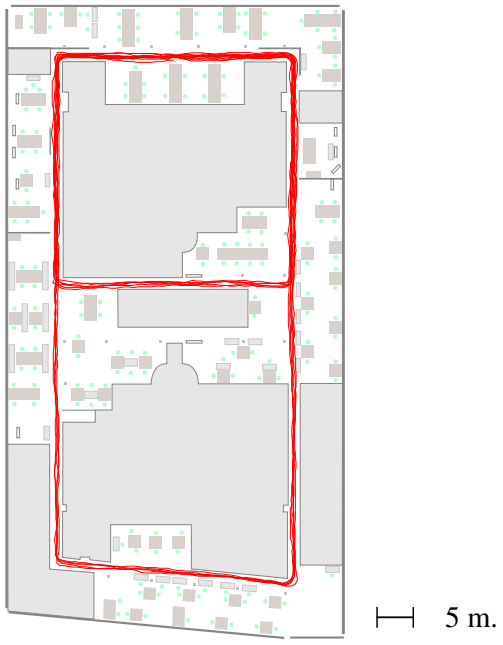
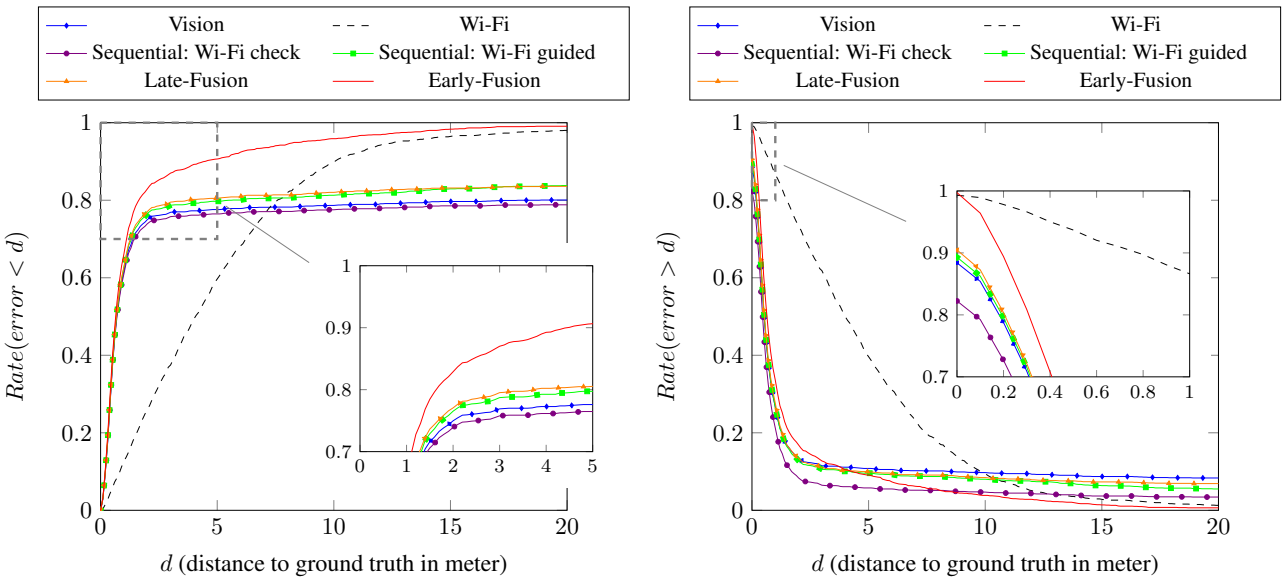


Figure 6. Examples of paths run by Pepper robots on a floor of SoftBank Robotics Europe office building.

Acquisitions were realized with different Pepper robots over several months and at two different floors of the building. Some examples of paths run by Pepper on a floor of the office building are shown on Figure 6.

5.4.1 Evaluation of Early-Fusion: In such testing conditions, the use of Wi-Fi data was expected to decrease the number of aberrant localizations. Indeed, the particularly dynamic and repetitive aspects of the open-space make the Access Points be more stable landmarks than visual keypoints. These expectations were verified by the the results plotted on Figure 7a and Figure 7b. The localization based only on Wi-Fi produced 1.3% of aberrant estimations localized more than 20m away from ground truth whereas classical visual FABMAP generated 8.3% of deviant localizations. However, the visual localization still resulted in more accurate results with 75.0% of the queries localized within 2m away from their true positions.

The interest of combining Wi-Fi and vision for localization tasks is confirmed by the results introduced in Figures 7a and 7b. All presented merging approaches show rates of misplaced localizations lower than visual FABMAP for a distance of 5m. Concerning the rates of correct localization, our early-fusion framework clearly outperformed the other algorithms with 90.6% of the queries localized within 5m away from their true positions, compared to only 77.6% for vision-only. On Figure 7b, the rejection rate of sequential fusion with Wi-Fi check allowed to reduce the rate of errors for this merging style, but the price to pay was high: in about one in five cases (17.8%, *cf.* curve on Figure 7b at $d = 0$), the robot did not succeed in estimating a position.



(a) Rate of correct localizations.

(b) Rate of misplaced localizations.

Figure 7. Localization results computed from the different localization algorithms in the office building of SoftBank Robotics Europe.

Resulting values from early-fusion and Wi-Fi check localizations are extracted in Table 1 for comparison with classical FABMAP.

	Correct localization rate (%) within				Misplaced localization rate (%) away from			
	2m	5m	10m	20m	2m	5m	10m	20m
Vision	75.0	77.6	78.7	80.1	13.4	10.8	9.7	8.3
Wi-Fi	26.0	59.7	89.8	98.0	73.3	39.6	9.5	1.3
Wi-Fi guided	76.5	79.7	81.3	83.8	12.8	9.6	8.0	5.5
Wi-Fi check	73.9	76.5	77.5	78.8	**8.3**	**5.8**	4.7	3.4
Late-Fusion	77.1	80.5	82.0	83.6	13.4	9.9	8.4	6.9
Early-Fusion	**82.9**	**90.6**	**95.9**	**99.1**	16.8	9.1	**3.9**	**0.6**

Table 1. Comparison of classical FABMAP with the presented localization schemes using Wi-Fi data only, and merging visual and Wi-Fi data.

Examples of complicated query images taken in the environment are shown on Figure 8. The images chosen highlight some typical complex situations. This environment was particularly dynamic since all acquisitions were realized on working hours. Between two runs of the robots, furniture could be moved, people were not systematically sitting at the same place, and light conditions could significantly evolve. Furthermore, people were not asked to avoid interaction with Pepper. Most of acquisitions thus contained occlusions due to interaction or people walking in front of the camera (on Figure 8, queries (a) and (b)). In all cases of images with occlusions, early-fusion produced same or better pose estimations than the visual localization. Queries (c) and (d) on Figure 8 indicate some examples of perceptual aliasing. Query (c) was misled by the presence of a houseplant that produced a lot of visual keypoints. A similar houseplant on another floor of the building duped the visual localization. Early-fusion process did not fall in this error thanks to the use of Wi-Fi data. A very hard case of perceptual aliasing can be seen with query (d) of Figure 8. In this case too, the early-fusion generated an acceptable localization (0.70m from true position) in comparison with vision (37.87m from true position).

Therefore, the early-fusion proved to be able of correctly localize 99.1% of the queries within 20m away from ground truth with the best accuracy. Furthermore, it is worth to notice that even if other strategies had higher rejection rates, they did not succeed in catching up some aberrant mistakes and early-fusion still had the lowest aberrant error rate.

A good example of unacceptable errors is not directly visible on Figure 7a and Figure 7b. It corresponds to the number of queries that were localized at a wrong floor. This kind of errors can be particularly problematic in strategies where the result of a global localization is used to load a sub-map of the environment. Nevertheless, it is still a common mistake in environments where floors are very visually similar, like in hotels or conference centers for example. In this environment, Figure 9 shows the percentages of queries submitted to this type of error for each tested localization algorithm.

Once again, the results on Figure 9 highlight the help provided by the Wi-Fi. The localization algorithm based only on vision confused more often the different floors. It was floor mistaken ten times more often than the localization using Wi-Fi data only. In all fusion schemes, using Wi-Fi helped to reduce these mistakes. However only the early-fusion led to a smaller error rate than the Wi-Fi based localization. These results are particularly interesting because they seem to indicate that the early-fusion process results in a better compromise for pose estimation than the other fusion frameworks.

5.4.2 Long-term localization: All the previous results demonstrate the interest of merging Wi-Fi and vision for localizing in this environment. These good performances encouraged some tests in more challenging situations. So the long-term robustness of our approach was also tested. These tests were realized by spacing initial exploration phase and

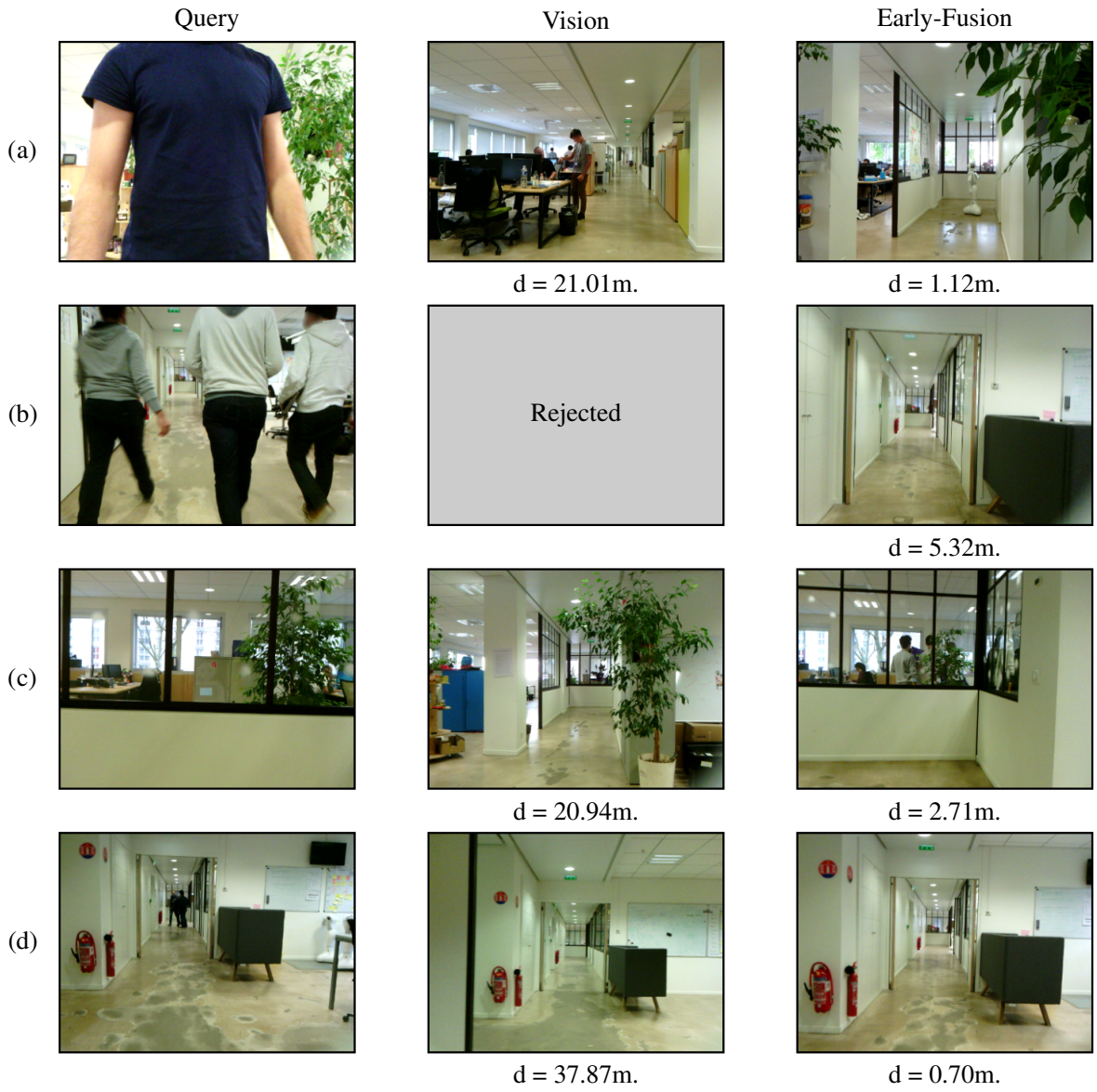


Figure 8. Examples of difficult queries collected in the office building of SoftBank Robotics Europe. Each line shows the query image (left) and the associated location in map computed from classical FABMAP (middle) and early-fusion (right) with the associated distances between estimated position and ground true (below images). The gray image indicates that visual localization considers query (b) as an unknown location.

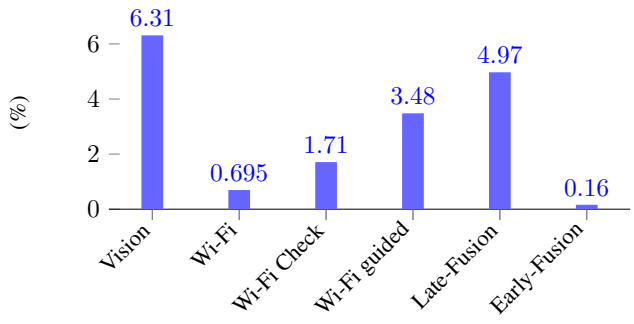
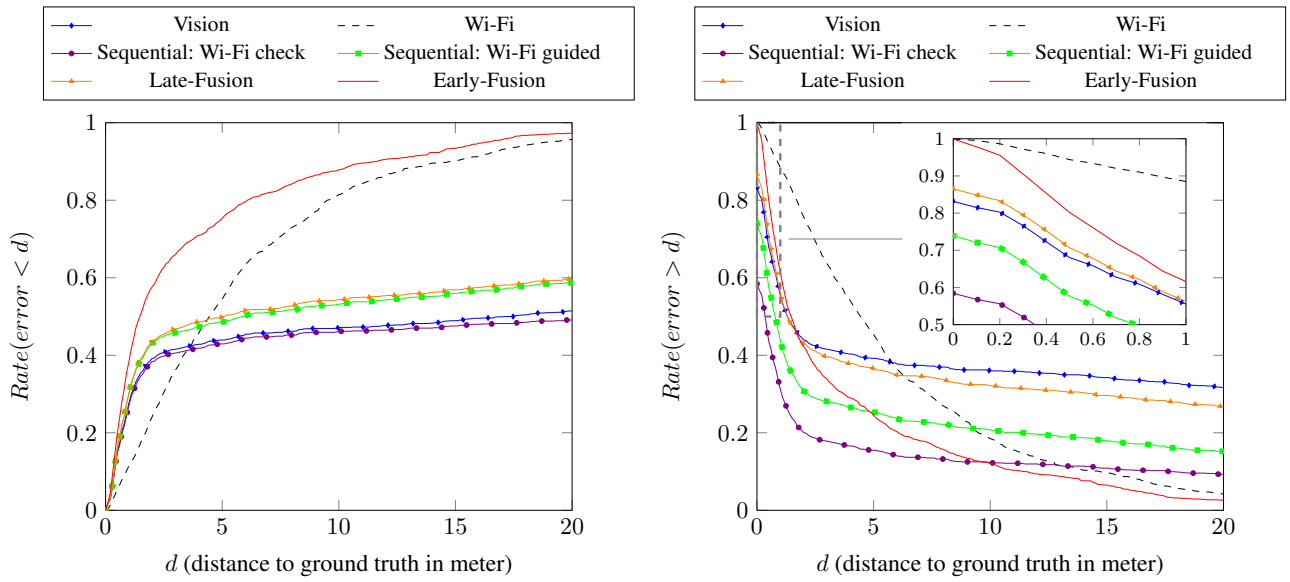


Figure 9. Percentages of queries localized on a wrong floor of the building for each localization algorithms.

localization tests seven months apart. The obtained results plotted on Figures 10a and 10b reveal that visual localization was more deteriorated than Wi-Fi localization. This can be easily explained by all the visual transformations that took place during seven months. These modifications strongly impacted the results of visual and sequential localizations. Conversely, the small changes that occurred in Wi-Fi signals, like new or missing smart-phones mobiles APs, were not significant enough for degrading the outcomes of Wi-Fi localization.

Values in Table 2 show that our proposed early-fusion with Wi-Fi enhanced FABMAP gave the best compromise between visual and Wi-Fi data.

So far, using Wi-Fi data has shown to be helpful for pose estimation. But this can be true because of certain aspects of the environment: its visual repetitiveness, its dynamic objects, its large dimensions and the fact that the Wi-Fi coverage was particularly good with numerous Access Points. To verify the performances of our early-fusion framework in less favorable conditions, the generality of our algorithm was tested in two other environments.



(a) Rate of correct localizations.

(b) Rate of misplaced localizations.

Figure 10. Long-term localization results in the office building of SoftBank Robotics Europe: initial exploration phase and localization tests are acquired seven months apart.

	Correct localization rate (%) within				Misplaced localization rate (%) away from			
	2m	5m	10m	20m	2m	5m	10m	20m
Vision	39.2	43.9	47.1	51.4	43.9	39.2	36.0	31.7
Wi-Fi	24.2	54.5	81.4	95.7	75.7	45.4	18.5	4.2
Wi-Fi guided	43.2	48.6	53.2	58.7	30.7	25.3	20.7	15.2
Wi-Fi check	38.4	42.9	46.1	49.1	**20.0**	**15.5**	12.3	9.3
Late-Fusion	43.6	49.9	54.3	59.7	42.9	36.6	32.2	26.8
Early-Fusion	**57.7**	**75.4**	**87.9**	**97.3**	42.2	24.5	**12.0**	**2.6**

Table 2. Comparison of the different localization algorithms tested: initial exploration phase and localization tests are acquired seven months apart.

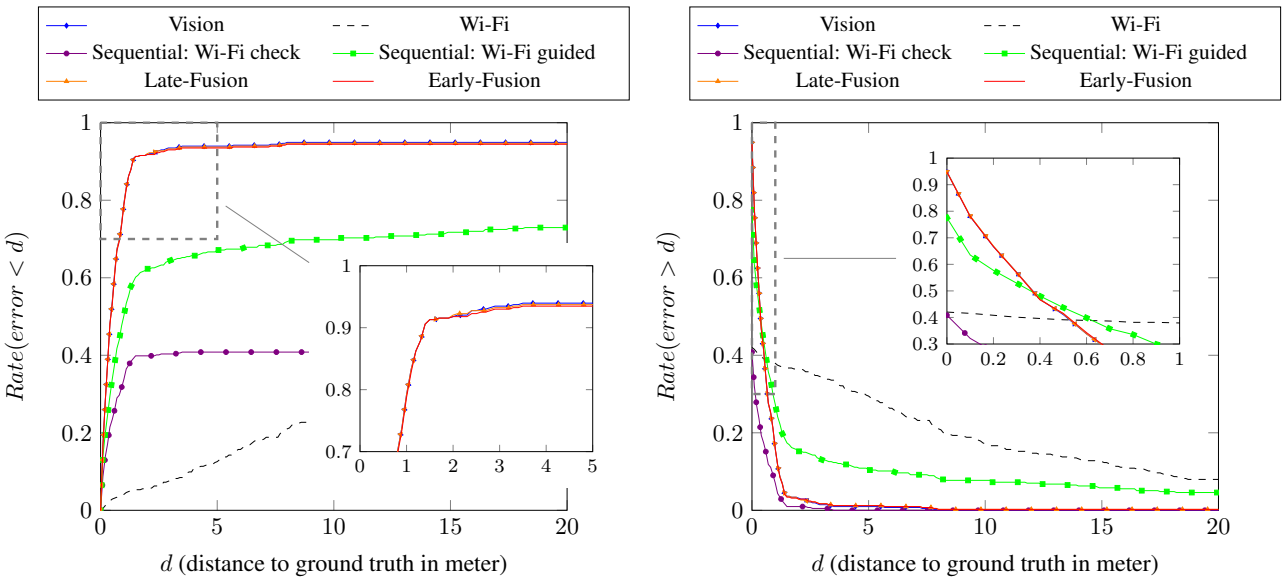
5.5 *Localization with bad Wi-Fi coverage*

Some tests were realized in a junior high school. This environment was chosen for several reasons. First, visual localization produced good estimations **there**. Then, Wi-Fi coverage was not optimal. Only 25 Access Points were visible compared to 544 in the office building of SoftBank Robotics Europe, for both environments with similar sizes. 876m were run over three different floors with a Pepper robot in the corridors and classrooms of a junior high school. Some examples of images acquired in this environment are shown on Figure 11.

Figure 11. Examples of images acquired in the junior high school tested environment.

Graphics on Figure 12 show that Wi-Fi localization performances were much worse than the ones from visual localization. Nevertheless, the good performances of visual FABMAP were not really surprising since the offline learning of our visual vocabulary was realized over the image database presented in Quattoni and Torralba (2009) that contains similar images taken in schools. **Collected signatures in this environment were very redundant and empty in some areas. That damaged the Wi-Fi localization. In subsection 5.8, performances of Wi-Fi localization are investigated taking into account the strengths of perceived signals.**

Despite the poor accuracy of Wi-Fi localization, early and late fusions did not damage the results of visual localization in a significant way. The rates of correct and misplaced localization presented on Figure 12a and Figure 12b are very similar. However, sequential approaches notably deteriorated the performances of the visual localization. Sequential localization with Wi-Fi check became very restrictive and rejects 59.2% of the queries, which strongly degraded its



(a) Rate of correct localizations.

(b) Rate of misplaced localizations.

Figure 12. Localization results computed from data collected in a junior high school with bad Wi-Fi coverage.

rates of correct localizations (**Table 3**). For instance, visual localization correctly localized 94.0% of the queries within 5m away from their true positions compared to only 40.8% with the Wi-Fi check sequential approach.

These results highlight the fact that early-fusion from Wi-Fi and visual data seems to not damage visual localization in environments where Wi-Fi coverage is not optimal.

	Correct localization rate (%) within				Misplaced localization rate (%) away from			
	2m	5m	10m	20m	2m	5m	10m	20m
Vision	91,8	**94,0**	**94,9**	**94,9**	3,0	1,0	**0**	**0**
Wi-Fi	5,6	12,6	24,9	34,1	36,5	29,5	17,1	7,8
Wi-Fi guided	62,3	67,1	69,8	72,4	15,2	10,4	7,7	4,6
Wi-Fi check	39,9	40,8	40,8	40,8	**1,0**	**0**	**0**	**0**
Late-Fusion	**92,0**	93,7	94,7	94,7	2,9	1,2	0,2	0,2
Early-Fusion	91,8	93,5	94,4	94,4	2,9	1,2	0,2	0,2

Table 3. Comparison of classical FABMAP with the presented localization schemes using Wi-Fi data only, and merging visual and Wi-Fi data for operating in a junior high school.

5.6 *Localization in a private apartment*

The last environment used during our tests was a private apartment. Because of its smaller size, testing our algorithm in such environment was interesting to see if the use Wi-Fi would damage visual localization.

Acquisitions were realized on a single floor and the maximal distance between all the recorded poses was 16m long compared to 76m in the office building and 65m in the middle school. In these testing conditions, Wi-Fi coverage was acceptable, with 76 visible Access Points. Several runs were made and they constituted a total covered distance of 440m. Some of the paths run by the robot are plotted on Figure 13.

Figure 13. Examples of paths run by Pepper robot in the different rooms of a private apartment.

Even if this environment was less affected by visual perceptual aliasing, it contained sources of potential errors such as strong variations of brightness and a smaller range of view due to smaller spaces. Some images taken in the apartment are presented on Figure 14.

Without strong visual perceptual aliasing, the localization based on vision produced good results as visible in Figure 15. Thanks to the walls of the different rooms, Wi-Fi signatures were different and localization based only on Wi-Fi data was more accurate in this environment than the one achieved in open-spaces (5.4.1). Some values of Figures 15a and 15b are extracted in Table 4.

Again, the rate of correct localizations was higher for the early-fusion process. The early-fusion also helped to reduce the number of errors. Even if sequential strategy with Wi-Fi check had smaller error rates (2.0% of misplaced localization further than 2m of their true positions, against 4.2% for early-fusion), it rejected 25.0% of the queries.

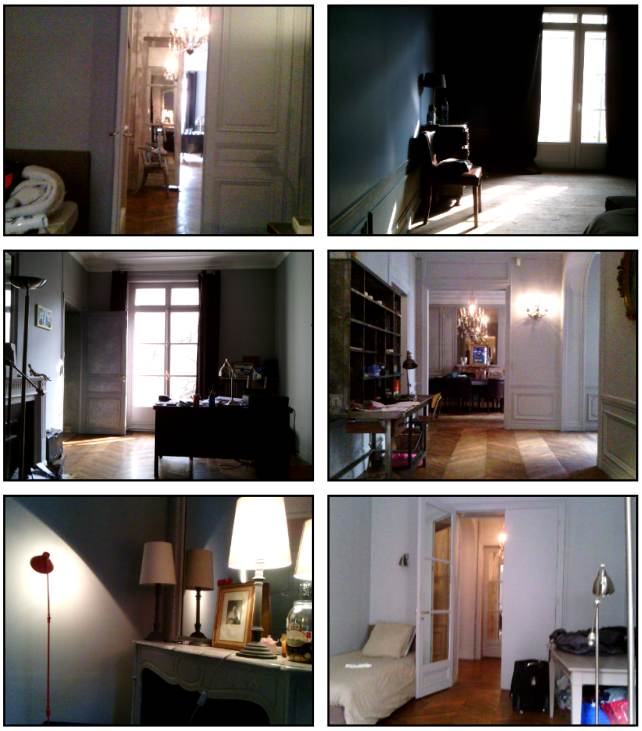


Figure 14. Examples of images acquired in the private apartment tested environment.

	Correct localization rate (%) within				Misplaced localization rate (%) away from			
	2m	4m	6m	8m	2m	4m	6m	8m
Vision	91.8	92.9	93.6	94.4	4.6	3.5	2.8	2.0
Wi-Fi	35.3	70.6	88.5	97.4	64.7	29.4	11.5	2.6
Wi-Fi guided	81.2	82.6	84.8	86.6	6.4	5.0	2.9	1.1
Wi-Fi check	73.0	73.7	74.4	74.8	**2.0**	**1.3**	**0.6**	**0.2**
Late-Fusion	92.9	94.3	94.9	95.6	4.0	2.6	2.0	1.3
Early-Fusion	**94.7**	**96.7**	**97.8**	**98.2**	4.2	2.2	1.1	0.7

Table 4. Comparison of classical FABMAP with the presented localization schemes using Wi-Fi data only, and merging visual and Wi-Fi data for operating in a private apartment.

5.7 Processing time

Finally, the processing time of a query was computed on Pepper robots. Such robots have a quad-core processor Atom E3845 with a CPU clock rate of 1.91GHz. The computation times retrieved were quite similar and around 117ms $\pm$ 59ms for all localization algorithms using visual data. Localization based only on Wi-Fi processed a query in about 0.31ms $\pm$ 0.095ms. This difference can be explained by the fact that generation of the visual appearance descriptor was longer: 99% of the processing time. This aspect encourages the use of fast keypoints detector and descriptor to operate in real-time. Note that our implementation was not parallelized.

In these experiments, our system was able to operate in real-time on board of Pepper robot when acquiring images every 2s. However, it is still possible to go up to 5 images per second.

5.8 The application of RSSI

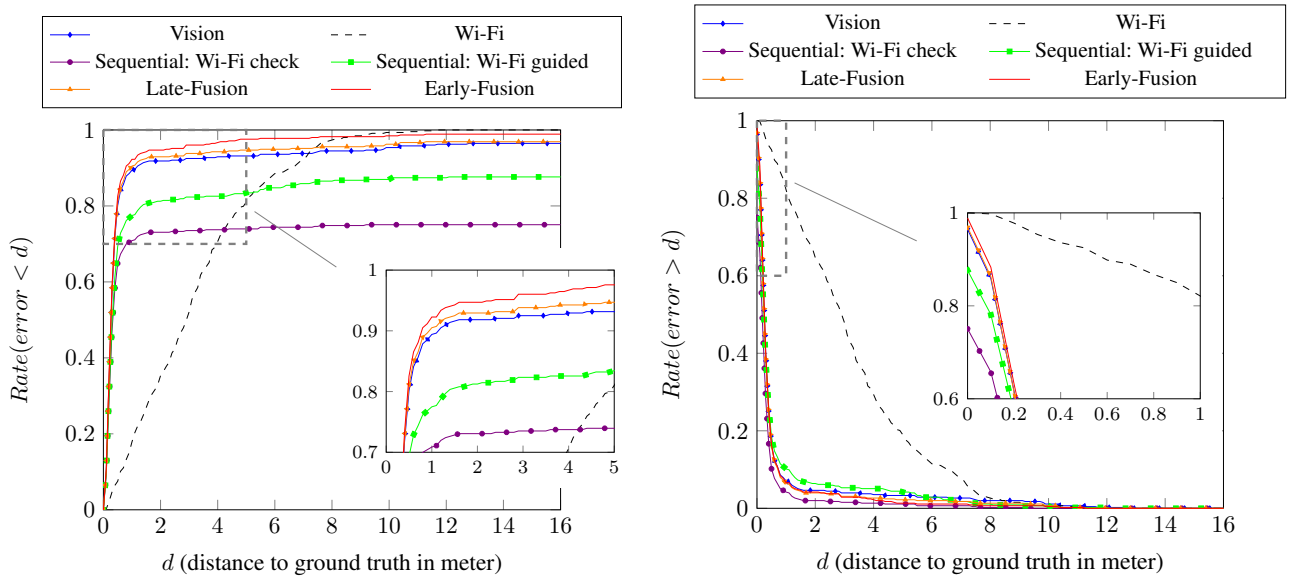
So far, the performances of different localization strategies have been studied without taking into account the intensity of the perceived Wi-Fi signals. As mentioned in sub-section 3.1, the information carried by the RSSI can be used to improve the accuracy of the localization algorithms using Wi-Fi data. This is done by implementing a kind of discretization of the signal strength that fits the FABMAP formalism. For each known Access Point, the RSSI is described by multiple binary words. In order to decide how many words to use for encoding the information from the RSSI, Wi-Fi localization results were computed from data collected by multiple Pepper robots in the office building of SoftBank Robotics Europe. This choice can be explained by the good Wi-Fi coverage in this environment with many visible APs, but was still challenging due to the similarity between the collected signatures (see sub-section 5.4 for more details).

The *incremental* and *exclusive* representations presented in Section 3.1 were investigated. Please note that the chosen range of usable intensities was in our case $]-90dBm, 0dBm]$ (as presented in Figure 4) because of Pepper's hardware and software limitations. Results for localization based on Wi-Fi data only were extracted and some are presented in Table 5. Different numbers of Wi-Fi words were tested to encode the RSSI. The first line of Table 5, for $b = 1$ Wi-Fi word, is the same for both approaches and is used as a reference in the following.

Number of Wi-Fi words to encode RSSI	Correct localization rate (%) within				Rejection rate (%)
	2m	5m	10m	20m	
1	26.0	59.7	89.8	98.0	0.7
4	31.3	68.1	92.6	99.1	0.2
	26 .1	61.5	90.4	98.9	0.4
8	**32.0**	**70.9**	92.9	99.1	0.6
	27 .4	67.5	92.9	99.6	0.4
12	29.8	68.9	90.9	97.8	1.0
	28 .0	68.8	**93 .3**	99.5	0.4
16	31.2	70.1	91.4	97.5	0.6
	26.6	66.5	91.2	**99.7**	0.1
20	29.7	65.8	88.0	96.4	1.2
	28.0	65.7	91.5	**99.7**	0.1

Table 5. Weight of the number of Wi-Fi words used to describe RSSI coming from each Access Point for Wi-Fi localization. Results computed with exclusive representation are plotted over gray background and compared to results from the incremental approach proposed in Wietrzykowski et al. (2017).

The values presented in Table 5 show that taking into account the information provided by the RSSI improves the accuracy of Wi-Fi localization. Both approaches presented in sub-section 3.1 were tested. The most accurate localization results were obtained when the RSSI was encoded over 8 words. For instance, with each perceived strength described by 8 Wi-Fi words, 32.0% of the queries were localized within 5m away from their true positions for the exclusive representation scheme, compared to 26.0%



(a) Rate of correct localizations.

(b) Rate of misplaced localizations.

Figure 15. Localization results computed from data collected in a private apartment.

without the use of the RSSI. However, for incremental and exclusive representations, the localization performances were not significantly different.

The use of the RSSI is tested in all presented environments and for all localization strategies using the Wi-Fi. The obtained results are presented in tables of performance gains, considering a distance to the ground truth of $d = 5$m. These gains are expressed in relation with the approach that does not take into account the RSSI; *i.e.* the method that only considers the presence or absence of known access points in the signature. The values of these gains for the exclusive representation are reported in:

- **Table 6 for experiment in SoftBank Robotics Europe office building;**
- **Table 7 for experiment in explored junior high school;**
- **Table 8 for experiment in tested private appartment;**

For the experiments carried out in the office building (Table 6) and in the junior high school (Table 7), the use of the RSSI improved the localization performances. The best results are reached by encoding the Wi-Fi signals strenghts on $b = 8$ words. However, this improvement is especially significant for the localization using only Wi-Fi data. Approches merging vision and Wi-Fi are less impacted by the use of RSSI, and some of them get worse results. For instance, sequential fusion schemes become more restrictive and the performance gains associated with the rate of correct localization are negative.

Regarding the localization results obtained in the apartment, the impact of the RSSI is less favorable (Table 8). In this environment, the closed rooms create enough variations in the Wi-Fi signals so that Wi-Fi localization can be based on a strategy that only takes into account the presence or absence of known access points in a signature. The use of the RSSI adds quantification noise to our system, since Wi-Fi signals are disrupted by the many obstacles they cross, and the measured gains do not point to any clear improvement.

Considering the strenghts of perceived Wi-Fi signals did not significantly improve the performances of the presented fusion strategies. However, encoding RSSI with b Wi-Fi words multiplies the size of the Wi-Fi vocabulary by b. Even if the use of FABMAP2.0 is compatible with a big vocabulary size, note that the building of Chow Liu tree is a polynomial time algorithm (Chow and Liu 1968). Thus, the learning of correlations between Wi-Fi words would be impacted by the size of the vocabulary.

6 Conclusion

This paper has introduced our early-fusion framework. It is a novel approach for merging data from visual and Wi-Fi sensors in order to solve indoor localization tasks for mobile robots.

The presented method has been tested over data collected by multiple Pepper robots with acquisitions schemes following real use cases of these robots. A total distance of 6.4km has been covered in three different environments: a building office, a junior high school and a private apartment. In all of these various situations, early-fusion has improved the visual localization results. For instance, in an environment where vision faces different problems such as perceptual aliasing or dynamic objects, the improvement of the localization is significant: 90.6% of the queries are correctly localized within 5m from their true positions, compared with 77.6% with visual localization.

Furthermore, compared with other classical fusion approaches, the early-fusion has produced the best results since it improves visual localization results without significantly damaging them even where Wi-Fi signals carry little information. The presented results show that in all our tests, early-fusion is the best compromise when merging visual and Wi-Fi data for solving global localization

	Gains on correct localization rates within 5m (%), for different values of b					Gains on misplaced localization rates away from 5m (%), for different values of b				
	$b = 4$	$b = 8$	$b = 12$	$b = 16$	$b = 20$	$b = 4$	$b = 8$	$b = 12$	$b = 16$	$b = 20$
Wi-Fi	+8,6	+11,2	+9,3	+10,4	+6,1	-8,6	-11,2	-9,3	-10,4	-6,1
Wi-Fi guided	-0,3	+0,3	+0,1	-2,3	-3,6	+0,3	-2,5	-3,2	-1,3	-1,3
Wi-Fi check	+0,4	-3,2	-5,7	-7,6	-11,8	-0,2	-3,0	-4,1	-3,6	-4,3
Late-Fusion	-0,1	+0,1	-0,7	-0,4	-0,1	+0,2	-0,3	+0,4	+0,3	-0,1
Early-Fusion	+0,9	+2,9	+1,1	+1,0	+1,2	-0,9	-3,0	-1,2	-1,2	-1,6

Table 6. Impact of exclusive Wi-Fi discretization in the office of SoftBank Robotics Europe. The performance gains are calculated for different numbers b of Wi-Fi words encoding the RSSI of the perceived Wi-Fi signals.

	Gains on correct localization rates within 5m (%), for different values of b					Gains on misplaced localization rates away from 5m (%), for different values of b				
	$b = 4$	$b = 8$	$b = 12$	$b = 16$	$b = 20$	$b = 4$	$b = 8$	$b = 12$	$b = 16$	$b = 20$
Wi-Fi	+8,9	+15,7	+9,4	+13,0	+11,8	-8,9	-15,7	-9,4	-13,0	-11,8
Wi-Fi guided	+4,1	-6,3	-9,2	-7,2	-13,5	-4,1	+1,4	+3,6	+1,4	+3,6
Wi-Fi check	+8,2	-4,3	-10,0	-4,1	-10,0	+0,5	-0,2	0	0	-0,2
Late-Fusion	+0,5	+0,2	0	+0,2	+0,7	-0,2	-0,5	0	-0,2	-0,7
Early-Fusion	-0,2	+0,2	-0,5	-0,2	0	0	-0,7	0	-0,2	-0,2

Table 7. Impact of exclusive Wi-Fi discretization in the explored junior high school. The performance gains are calculated for different numbers b of Wi-Fi words encoding the RSSI of the perceived Wi-Fi signals.

	Gains on correct localization rates within 5m (%), for different values of b					Gains on misplaced localization rates away from 5m (%), for different values of b				
	$b = 4$	$b = 8$	$b = 12$	$b = 16$	$b = 20$	$b = 4$	$b = 8$	$b = 12$	$b = 16$	$b = 20$
Wi-Fi	+3.1	-2.9	+4.0	-0.2	-1.1	-3.1	+2.9	-4.0	+0.2	-1.1
Wi-Fi guided	-6.2	-5.7	+0.2	-8.4	-9.7	-0.4	0	+0.7	+2.9	+3.5
Wi-Fi check	-5.5	-7.7	-2.0	-13.0	-16.3	+0.7	+0.2	+0.4	+0.7	0
Late-Fusion	+0.7	+0.2	-0.2	+0.4	+0.4	-0.2	0	+0.2	-0.4	0
Early-Fusion	+0.7	+0.2	-0.4	+1.1	-0.9	-0.2	0	+0.2	-0.7	+0.7

Table 8. Impact of exclusive Wi-Fi discretization in the tested private apartment. The performance gains are calculated for different numbers b of Wi-Fi words encoding the RSSI of the perceived Wi-Fi signals.

problem. **Some future work could look into qualifying what are the minimal requirements on the environment to get good Wi-Fi localization performances.**

7 Acknowledgements

These results have received funding from Bpifrance through the PSPC Project ROMEO 2 and from European Union Horizon 2020 program through the MuMMER project under grant agreement No 688147. The authors also thank Scarlett Fres for her precious help during experimental acquisitions.

References

Angeli A, Filliat D, Doncieux S and Meyer JA (2008) Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics* 24(5): 1027–1037.

Aparicio S, Pérez J, Bernardos AM and Casar JR (2008) A fusion method based on bluetooth and wlan technologies for indoor location. In: *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*. IEEE, pp. 487–491.

Biswas J and Veloso M (2010) Wifi localization and navigation for autonomous indoor mobile robots. In: *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, pp. 4379–4384.

Biswas J and Veloso M (2013) Multi-sensor mobile robot localization for diverse environments. In: *Robot Soccer World Cup*. Springer, pp. 468–479.

Boonsriwai S and Apavatjrut A (2013) Indoor wifi localization on mobile devices. In: *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference on*. IEEE, pp. 1–5.

Chow C and Liu C (1968) Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory* 14(3): 462–467.

Cummins M and Newman P (2008) Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research* 27(6): 647–665.

Cummins M and Newman P (2011) Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research* 30(9): 1100–1123.

Glover A, Maddern W, Warren M, Reid S, Milford M and Wyeth G (2012) Openfabmap: An open source toolbox for appearance-based loop closure detection. In: *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, pp. 4730–4735.

He S and Chan SHG (2016) Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons. *IEEE Communications Surveys & Tutorials* 18(1): 466–490.

Howard A, Siddiqi S and Sukhatme GS (2003) An experimental study of localization using wireless ethernet. In: *Field and Service Robotics*. Springer, pp. 145–153.

Huang J, Millman D, Quigley M, Stavens D, Thrun S and Aggarwal A (2011) Efficient, generalized indoor wifi graphslam. In: *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, pp. 1038–1043.

Jiang W and Yin Z (2015) Indoor localization by signal fusion. In: *18th International Conference on Information Fusion*.

Jirku M, Kubelka V and Reinstein M (2016) Wifi localization in 3d. In: *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, pp. 4551–4557.

Kummerle R, Hahnel D, Dolgov D, Thrun S and Burgard W (2009) Autonomous driving in a multi-level parking structure. In: *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, pp. 3395–3400.

Lafaye J, Gouaillier D and Wieber PB (2014) Linear model predictive control of the locomotion of pepper, a humanoid robot with omnidirectional wheels. In: *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*. IEEE, pp. 336–341.

Liu H, Gan Y, Yang J, Sidhom S, Wang Y, Chen Y and Ye F (2012) Push the limit of wifi based localization for smartphones. In: *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, pp. 305–316.

Liu M, Chen R, Li D, Chen Y, Guo G, Cao Z and Pan Y (2017) Scene recognition for indoor localization using a multi-sensor fusion approach. *Sensors* 17(12): 2847.

Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, Corke P and Milford MJ (2016) Visual place recognition: A survey. *IEEE Transactions on Robotics* 32(1): 1–19.

Lowry SM, Wyeth GF and Milford MJ (2014) Towards training-free appearance-based localization: probabilistic models for whole-image descriptors. In: *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, pp. 711–717.

Lynen S, Bosse M, Furgale P and Siegwart R (2014) Placeless place-recognition. In: *3D Vision (3DV), 2014 2nd International Conference on*, volume 1. IEEE, pp. 303–310.

Mirowski P, Ho TK, Yi S and MacDonald M (2013) Signalslam: Simultaneous localization and mapping with mixed wifi, bluetooth, lte and magnetic signals. In: *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on*. IEEE, pp. 1–10.

Nister D and Stewenius H (2006) Scalable recognition with a vocabulary tree. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2. Ieee, pp. 2161–2168.

Nowakowski M, Joly C, Dalibard S, Garcia N and Moutarde F (2017) Topological localization using wi-fi and vision merged into fabmap framework. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 3339–3344.

Nowicki M (2014) Wifi-guided visual loop closure for indoor navigation using mobile devices. *Journal of Automation Mobile Robotics and Intelligent Systems* 8(3): 10–18.

Ocaña M, Bergasa LM, Sotelo M and Flores R (2005) Indoor robot navigation using a pomdp based on wifi and ultrasound observations. In: *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*. IEEE, pp. 2592–2597.

Olivera VM, Plaza JMC and Serrano OS (2006) Wifi localization methods for autonomous robots. *Robotica* 24(4): 455–461.

Quattoni A and Torralba A (2009) Recognizing indoor scenes. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 413–420.

Quigley M, Stavens D, Coates A and Thrun S (2010) Sub-meter indoor localization in unmodified environments with inexpensive sensors. In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, pp. 2039–2046.

Rohrig C and Kiinemund F (2007) Mobile robot localization using wlan signal strengths. In: *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2007. IDAACS 2007. 4th IEEE Workshop on*. IEEE, pp. 704–709.

Rosten E and Drummond T (2006) Machine learning for high-speed corner detection. In: *European conference on computer vision*. Springer, pp. 430–443.

Rosten E, Porter R and Drummond T (2010) Faster and better: A machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence* 32(1): 105–119.

Rublee E, Rabaud V, Konolige K and Bradski G (2011) Orb: An efficient alternative to sift or surf. In: *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, pp. 2564–2571.

Ruiz-Ruiz AJ, Canovas O, Munoz RAR and Alcolea PELdT (2011) Using sift and wifi signals to provide location-based services for smartphones. In: *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, pp. 37–48.

Salton G and Buckley C (1988) Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5): 513–523.

Schwiegelshohn F, Nick T and Götze J (2013) Localization based on fusion of rfid and stereo image data. In: *Positioning Navigation and Communication (WPNC), 2013 10th Workshop on*. IEEE, pp. 1–6.

Sivic J and Zisserman A (2003) Video google: A text retrieval approach to object matching in videos. In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, pp. 1470–1477.

Stumm E, Mei C and Lacroix S (2013) Probabilistic place recognition with covisibility maps. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 4158–4163.

Sünderhauf N and Protzel P (2011) Brief-gist-closing the loop by simple means. In: *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, pp. 1234–1241.

Ulrich I and Nourbakhsh I (2000) Appearance-based place recognition for topological localization. In: *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, volume 2. Ieee, pp. 1023–1029.

Werner M, Kessel M and Marouane C (2011) Indoor positioning using smartphone camera. In: *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*. IEEE, pp. 1–6.

Wietrzykowski J, Nowicki M and Skrzypczyński P (2017) Adopting the fab-map algorithm for indoor localization with wifi fingerprints. In: *International Conference Automation*. Springer, pp. 585–594.