Image and WLAN Bimodal Integration for Indoor User Localization

Milan D. Redžić, Christos Laoudias, Ioannis Kyriakides

Abstract—Recently, we experience the increasing prevalence of wearable cameras, some of which feature Wireless Local Area Network (WLAN) connectivity, and the abundance of mobile devices equipped with on-board camera and WLAN modules. Motivated by this fact, this work presents an indoor localization system that leverages both imagery and WLAN data for enabling and supporting a wide variety of envisaged location-aware applications ranging from ambient and assisted living to indoor mobile gaming and retail analytics. The proposed solution integrates two complementary localization approaches, i.e., one based on WLAN and another one based on image location-dependent data, using a fusion engine. Two fusion strategies are developed and investigated to meet different requirements in terms of accuracy, run time, and power consumption. The one is a light-weight threshold-based approach that combines the location outputs of two localization algorithms, namely a WLAN-based algorithm that processes signal strength readings from the surrounding wireless infrastructure using an extended Naive Bayes approach and an image-based algorithm that follows a novel approach based on hierarchical vocabulary tree of SURF (Speeded Up Robust Features) descriptors. The second fusion strategy employs a particle filter algorithm that operates directly on the WLAN and image readings and also includes prior position estimation information in the localization process. Extensive experimental results using real-life data from an indoor office environment indicate that the proposed fusion strategies perform well and are competitive against standalone WLAN and imaged-based algorithms, as well as alternative fusion localization solutions.

Index Terms—Indoor user localization, WLAN, Images, Fusion, Hybrid, Time efficiency

1 Introduction

Different measurement types currently available from modern commercial portable hardware (cellphones, tablets) are able, if properly fused, to offer diverse information and lead to improved accuracy in indoor environment localization [1]–[3]. This work addresses indoor localization using Wireless Local Area Network (WLAN) technology in conjunction with image sensing. Whilst Global Navigation Satellite Systems (GNSS), such as the Global Positioning System (GPS), have become synonymous with user localization, their robustness and availability under certain conditions is questionable. For instance outdoors, satellite signals can be affected by obstacles, multipath propagation and tall buildings that inevitably lead to high location errors. In addition, satellite signals are weak or totally blocked inside buildings.

WLAN technology has demonstrated promising performance in indoor localization; however, it requires accurate modeling of the complex indoor multipath propagation environment and varying signal obstructions or reflections due to motion [4]–[6]. Researchers have investigated image-based localization, for example in [7], also associated with challenges such as occlusion, changes in lighting, noise and blur. Many localization methods in the literature are based on the hybridization or fusion of Ultra-Wide Band (UWB) and WLAN, WLAN and Radio Frequency (RF) tags (indoors) and GPS and WLAN (outdoors) [8]–[11]. Presently, there is a limited number of localization solutions based on the fusion of RF and image sensing methods [5], [12]–[15].

The motivation for combining WLAN and image data to infer user location indoors is that these are fundamentally different and complementary sensor modalities, which in combination may provide rich information on the observed scene and mitigate errors associated with each individual modality. In fact, there is a number of dynamic adjustments during system operation that can be made to meet diverse application-specific requirements in terms of positioning error and computational complexity, which is directly linked to battery depletion on mobile devices. Moreover, nowadays, modern sensor-rich smartphones can be easily employed as WLAN and image data acquisition hubs, e.g., see the *Campaignr*[1] micropublishing platform [16]. Such a localization system can be orientated towards the context-aware needs and capabilities of a user and becomes extremely useful for a multitude of applications including ambient assisted living, i.e., assistive technologies for memory and visually impaired individuals, tourist-oriented services that enhance user experience in museums and galleries,

- *Milan D. Redžić is with Huawei Ireland Research Center, Dublin, Ireland. E-mail: milan.redzic@huawei.com*
- *Christos Laoudias is with KIOS Research and Innovation Center of Excellence, University of Cyprus, Cyprus. E-mail: laoudias@ucy.ac.cy*

The work of C. Laoudias was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant 739551 (KIOS CoE) and the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

- *Ioannis Kyriakides is with the Engineering Department, University of Nicosia, Cyprus. E-mail: kyriakides.i@unic.ac.cy*

1. http://www.campaignr.com

indoor gaming, in-shop advertisement and coupon distribution, as well as health and daily life monitoring. Thus, for example, for memory impaired people, taking and using images in the localization framework works as a memory prosthesis. These images can be automatically segmented and clustered into specific events during a particular time-frame or an activity, thus allowing people to recall different aspects of their daily lives. Other possible uses of the proposed system include the navigation assistance for visually impaired people, tourist-oriented guidance applications, and health and daily life monitoring especially for the elderly, and indoor localization for enhancing indoor vehicle/robot autonomy.

In this work, the problem of efficiently integrating WLAN Received Signal Strength (RSS) and image information for indoor localization is addressed by two fusion strategies. The high-level block diagram of the proposed system architecture is depicted in Figure 1. All the algorithms for WLAN localization, image localization and fusion are calculated by a central unit in a Location Server that resides on the network side, for example a standard laptop used in our experimental setup in Section 7. The WLAN-image equipped Mobile Device, e.g. smartphone, robot, etc., collects the measurements and forwards them to the Location Server. We used such device-assisted approach to avoid heavy computation on the device that may drain battery quickly, although the proposed algorithms could run on the mobile device, in a fully device-based architecture, as long as the battery and storage space (for storing the fingerprint and image databases) are not critical. During localization, an image of the surrounding environment (e.g., captured by a smartphone's camera) and the RSS values from WLAN Access Points (APs) in the vicinity, referred to as fingerprint[2], are provided as inputs to the system.

In the *late fusion* approach (flow shown in solid lines), the *WLAN Localization* component computes a location by matching the input RSS fingerprint against the location-tagged fingerprints that have been collected in advance and stored in the fingerprint database. Similarly, the *Image Localization* component compares the input image with the location-tagged images that span the entire area of interest and are stored in the image database. Then, the *Fusion Engine* employs the *Threshold-based* component that combines the WLAN-based and image-based locations to output the final user location.

Alternatively, in the *early fusion* approach (flow shown in dashed lines), the WLAN and image readings are directly fed into the *Fusion Engine* that employs the *Particle filter* component to fuse the location-dependent data with the aid of an underlying user mobility model that introduces prior location information in the localization process. Thus, contrary to the *late fusion* approach, the computation of intermediate locations by dedicated localization components is avoided. In this work, we focus on the combination of WLAN and image data without exploiting other data (e.g., inertial and magnetic) that are available on modern sensor-rich smartphones. We demonstrate that reasonable accuracy can be achieved only with these two modalities, while outperforming other similar solutions. In particular, the proposed system can be significantly enhanced by incorporating sensor data into the particle filter to improve the underlying kinematic model with more accurate information for the displacement and orientation of the particles.

Beside extending the Naive Bayes approach of [17] to build our WLAN localization algorithm, the main contributions of this work are the following.

- For image-based localization, we introduce a novel algorithm that follows an interest point-based approach and employs a variation of a hierarchical vocabulary tree to efficiently match query images with training images.
- For fusion, two design options are considered to optimally combine WLAN and camera sensory data, namely a light-weight threshold-based scheme and a flexible particle filter algorithm. The use of location quality indicators is also explored for dynamically enabling/disabling a modality acquisition and location computation path in a *hybrid* fashion to meet different requirements. For instance, if the number of sensed WLAN APs in the measured RSS fingerprint is small (indicating that the WLAN-based location might be inaccurate), then the image sampling and localization path could be enabled to deliver the desired accuracy (otherwise it is disabled to extend battery life-time).
- The trade-offs between the WLAN and image modalities, as well as the fusion options are investigated and compared in terms of positioning error, computational complexity, and power consumption. Thus, many insights come up that lead to useful guidelines and best practices for optimizing the operation of such fusion localization solution.

This paper is structured as follows. Section 2 overviews the related work on indoor localization. Section 3 describes our WLAN-based localization method, while Section 4 introduces the novel image-based localization approach. Threshold-based fusion is presented in Section 5, while fusion by means of particle filter is described in Section 6. Section 7 describes the experimental setup and data collection process. A selection of results is presented and compared in Section 8. Time efficiency pertaining to different options is analyzed in Section 9, followed by a comparison with other methods in Section 10. Finally, conclusions and directions for future work are outlined in Section 11.

2 Related work

2.1 WLAN-based localization

WLAN has been a very popular technology for indoor location determination, mainly due to the ease of col-

2. The terms fingerprint and observation are used interchangeably in this work.

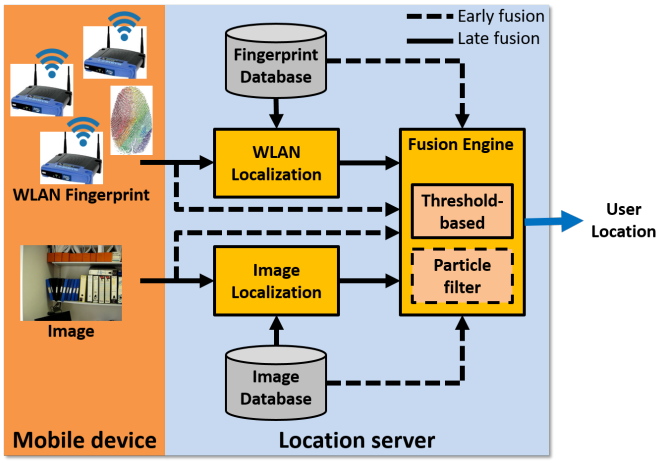


Fig. 1. Localization system architecture.

lecting and fingerprinting signal strength measurements with wireless mobile devices from the ubiquitous WLAN infrastructure inside buildings; see [18] and references therein for survey of recent advances. Many localization solutions complement WLAN with inertial sensors (i.e., accelerometer, gyroscope, magnetometer, barometer) and floorplan maps, like the *Anyplace* indoor navigation service [19], or further augment it with ambient light and sound signals as in *SurroundSense* [20]. This has been stirred with availability of sensor-rich mobile devices.

UnLoc is an unsupervised indoor localization system that leverages smartphone sensors to compute the displacement and direction of users to avoid the need for war-driving for populating the fingerprint database [21]. *LiFS* utilizes the spatial relation of RSS fingerprints, so that the collected fingerprints are distributed according to collectors' mutual distances in real world [22]. *WILL* is another system that combines WLAN fingerprints with user movements to infer user location without site survey or knowledge of AP locations [23].

Authors in [24] present a WLAN-based system that employs principal component analysis in an efficient mechanism for replacing sets and subsets of available APs. In [25], by reducing both the volume of collected data and the number of data collection points, the radio map can be successfully rebuilt using an interpolation approach. Along the same line, *SEAMLOC* uses a novel interpolation algorithm, based on the specification of robust, range and angle-dependent likelihood functions [4]. Authors in [26] discuss the reduction of severe fluctuations of RSS and propose a scheme that efficiently extracts the signal for user localization.

In this work, for the WLAN-based localization module we build upon and extend the Naive Bayes approach [17]. We explicitly modelled signal strength distributions coming from available APs together with distributions of frequency of appearance of these APs, and eventually used them in our WLAN-based indoor localization framework, as described in Section 3.

2.2 Vision-based localization

Vision-based localization has drawn attention due to the rich information contained in image measurements, due to its passive nature, and the fact that vision provides the most of the human sensory information. Few methods employ the visual vocabulary tree using Scale Invariant Feature Transform (SIFT) features [27], [28], while some others such as [29], [30] use landmarks to implement indoor localization. The landmarks represent features or group of features detected from the images. During the searching period, features which are detected from the query image are matched to the landmarks.

Based on a series of images or video sequences one is able to construct a topological map, and then to refine it by employing learning vector quantization [31]. In the online phase, similar regions in the query image are detected using a nearest neighbor rule.

Localization based on stereo-imaging has also been studied as stereo images can provide depth insights for 3D reconstruction [29], [32]. An indoor localization algorithm based on an efficient database search using robust matching algorithms is presented in [33].

What differentiates this approach from other image-based localization solutions is in employing a verification step mechanism, based on bidirectional image feature matching of a vocabulary tree framework, which refines the location predictions and thus improves the final user location. Moreover we employ a heuristic to fix the cluster centers of the vocabulary tree.

2.3 Hybrid and fusion-based localization

Even though some early vision-based localization systems used only image processing and matching techniques, several recent solutions rely on the combination of location estimates derived with camera and other technologies. For instance, authors in [13] propose a particle filter for fusing positioning information from cellular base stations and images.

Alternatively, imagery and other sensory data can be directly fused to determine user location. For example, *RAVEL* (Radio And Vision Enhanced Localization) fuses visual information coming from cameras and WLAN readings [34], while [35] proposes a camera-assisted region-based magnetic field fingerprinting technique. Going further by fusing more sensors, *Travi-Navi* is a vision-guided navigation system that employs magnetic field distortions and WLAN signals to achieve robust and effective user indoor tracking [36]. Similarly, the system proposed in [37] combines opportunistic WLAN signals and magnetic field readings together with camera-based positioning in areas with fewer magnetic disturbances to assist magnetic field positioning. In our previous work [2], we introduced a WLAN-based algorithm and an image matching framework to support image-based localization coupled with a simple hybrid process.

Bayesian filtering is a powerful tool for processing and fusing location-dependent data from diverse sources. For

instance, Kalman filter is used for target tracking in collaborative camera sensor networks [38], while an error-state Kalman filter is proposed in [39] that combines measurements from moving vision sensors and radio ranging equipment to estimate user position over time.

Particle filter is a sequential Monte Carlo method based on Bayesian inference that enables fusion of heterogeneous measurements, non-linear relationships between measurements and the target state and estimates non-Gaussian posterior distributions [40]. In the context of indoor localization, apart from [1], [21], and [36], authors in [41] employ particle filter for fusing inertial sensory data on android phones.

Fusion of data from a network of security cameras and RSS fingerprint observations is presented in [14] to enable the simultaneous tracking of multiple individuals inside indoor environments. Another work addresses object tracking with a solution that consists of a camera recording method based on color features of the target and a WLAN-based localization algorithm [5].

Authors in [15] describe an object tracking scheme that employs a sensor fusion approach composed of visual and location information estimated from WLAN signal strength values. Switching between fusion-based (i.e., image and WLAN) and purely WLAN-based location is decided as follows: in areas where an image can be taken the system gives priority to fusion, whereas in areas that images cannot cover priority is given to WLAN.

In [12] the authors discuss an approach that combines WLAN-based localization and static camera tracking. The purpose of fusing WLAN and video data is to reduce localization error in the rooms where there is a camera, in contrast to using only WLAN that still offers room level accuracy when no cameras are present.

Our work is closer to the systems discussed in [15] and [12]; however, the proposed solution employs a novel image-based localization algorithm and fuses image and WLAN signals by means of a threshold-based or a particle filter algorithm to trade off positioning error and computational time/energy consumption depending on the application scenario. As compared to previously mentioned approaches these two fusion methods are complementary in terms of how we integrate sensing modalities and interpret results, and also we propose a hybrid fusion as a viable way when trading-off efficiency and accuracy of such a localization system.

3 WLAN-BASED LOCALIZATION

Probabilistic WLAN localization techniques based on fingerprinting start with the acquisition of training observations consisting of RSS information at Calibration Points (CP) distributed along a dense grid throughout the building [17], [42]. To calculate the probability of a user being at a particular CP given the RSS values that he/she observes, we employ a Naive Bayes method, which represents an extension of the Bayes and Naive Bayes classifiers. This algorithm takes into account the RSS values of WLAN APs and also the frequency of the appearance of these APs.

A signature for each CP is defined as a set of W distributions of RSS values from W APs and a distribution representing the number of appearances of W APs received at this CP. $C \in \{1, 2, ..., K\}$ denotes the CP random variable where K is the number of CPs, $X_m \in \{1, 2, ..., W\}$ represents the m^{th} AP random variable, $Y_m \in \{s_1, ..., s_V\}$ is the RSS value received from the m^{th} AP, where W is the number of APs, M is the number of APs of an observation and V is the number of discrete RSS values. $D = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_N\}$ is a set of N training observations where the n^{th} training observation is defined as $\mathbf{o}_n = (c^{(n)}, x_1^{(n)}, y_1^{(n)}, ..., x_M^{(n)}, y_M^{(n)})$, for $n = 1, .., N$, where $x_m \in X_m$ and $y_m \in Y_m$ for $m = 1, .., M$. It is not necessary that each AP produces receivable signals at each CP, and indeed whether an AP signal can be obtained at a CP can vary with time depending on the state of the radio channel. The joint distribution $P(C, X_1, Y_1, ..., X_M, Y_M)$ is given by

$$P(C) \prod_{m=1}^{M} P(X_m|C)P(Y_m|C, X_m). \tag{1}$$

Using the Naive Bayes approach and one testing observation $\mathbf{o}$ the likelihood that the user is at location c can be written as

$$P(c|\mathbf{o}) \propto P(c) \prod_{m=1}^{M} P(x_m|c)P(y_m|c, x_m). \tag{2}$$

Based on (2), we can obtain a ranking of the CPs according to $P(c|\mathbf{o})$, i.e., the first, the second, the third and so on CP where the user is most likely located.

In the absence of any other information the *a priori* probability distribution of the user location, $P(C = c)$, is presumed to be uniform. The distribution of AP x given a location c, $P(X_m = x|C = c)$ is multinomial, the probability of signal strength y given location c and AP x, $P(Y_m = y, C = c, X_m = x)$, is normalized histogram. Using the identity function

$$\mathbb{I}(e_1, e_2) = \begin{cases} 1 & \text{for} \quad e_1 = e_2 \\ 0 & \text{for} \quad e_1 \neq e_2 \end{cases}, \tag{3}$$

in a maximal likelihood estimation framework the sufficient statistics are

$$n_c = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}(c^{(n)}, c), \tag{4}$$

$$n_c^{(x)} = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}(c^{(n)}, c)\mathbb{I}(x_m^{(n)}, x), \tag{5}$$

in which we observe the frequency of appearance of the APs while in

$$n_{c,x}^{(y)} = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}(c^{(n)}, c)\mathbb{I}(x_m^{(n)}, x)\mathbb{I}(y_m^{(n)}, y), \tag{6}$$

we take into account its corresponding RSS values which will be eventually used to calculate conditional probabilities of APs and signal strengths. We evaluate these probabilities as follows. The probability of AP x given location c, $P(X_m = x|C = c)$ is given by

$$P(X_m = x|C = c) = \frac{n_c^{(x)} + 1}{n_c + W}, \tag{7}$$

while the probability of signal strength y given location c and AP x, $P(Y_m = y|C = c, X_m = x)$ is given by

$$P(Y_m = y|C = c, X_m = x) = \frac{n_{c,x}^{(y)} + 1}{n_c^{(x)} + V}. \tag{8}$$

These are estimates of the signature parameters, for every AP and also for every RSS value that can be observed from that AP.

We rescaled the corresponding probabilities of the candidate CPs in (2) to sum to one and denoted their new values as the CP confidences, p_i. To calculate the final user location we used the Minimum Mean Square Error (MMSE) estimation algorithm given by

$$\mathbf{r}_W = \frac{\sum_{i=1}^{K} p_{maxi}\mathbf{CP}_{maxi}}{\sum_{i=1}^{K} p_{maxi}}, \tag{9}$$

where the first, the second, ... , k^{th} ranked CP positions, corresponding confidence values, and the user location output are denoted by $\mathbf{CP}_{max1}$, $\mathbf{CP}_{max2}$, ... , $\mathbf{CP}_{maxK}$, p_{max1}, p_{max2}, ... , p_{maxK}, and $\mathbf{r}_W$, respectively.

4 Novel Image-Based Localization

For the image-based localization, we use a feature point based approach, that employs a variation of a vocabulary tree supported by bidirectional matching, to obtain the user location; see Figure 2. Beside extending vocabulary tree concept, three novel contributions are proposed:

- Use of quantized features and a two-brench vocabulary tree to speed up the setup and the localization process.
- Re-estimation procedure for fixing cluster centers of the hierarchical vocabulary tree of the SURF features as a part of an extended κ-means algorithm.
- A bidirectional matching approach used to reorder locations previously ranked by the vocabulary-tree based method.

Speeded Up Robust Features (SURF) is an image detector and descriptor[3], robust to lighting, viewpoint changes, and changes in scale [43]. It uses a Haar wavelet approximation of the determinant of Hessian blob detector

$$\mathcal{H}(\mathbf{l}, \sigma) = \begin{bmatrix} L_{\xi\xi}(\mathbf{l}, \sigma) & L_{\xi\zeta}(\mathbf{l}, \sigma) \\ L_{\xi\zeta}(\mathbf{l}, \sigma) & L_{\zeta\zeta}(\mathbf{l}, \sigma) \end{bmatrix}, \tag{10}$$

where $L_{\xi\xi}(\mathbf{l}, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial\xi^2} g(\sigma)$ with the image I in point $\mathbf{l}$, and

3. SURF feature and descriptor vector are used interchangeably.

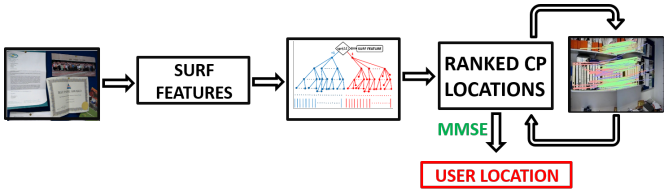


Fig. 2. A block diagram of the image-based localization. After extracting the SURF features from an input image, we propagate the features through the vocabulary tree to obtain the ranked CP locations. The ranking list is refined using the bidirectional matching. By employing the MMSE algorithm on the refined ranking list and corresponding confidence values, we obtain the user location.

similarly for $L_{\xi\zeta}(\mathbf{l}, \sigma)$ and $L_{\zeta\zeta}(\mathbf{l}, \sigma)$. ξ and ζ denote orthogonal coordinate axes of a two dimensional Cartesian coordinate system associated with the image.

A SURF interest point must be selected at distinct location in image (T-junctions, corners, blobs) and its neighborhood is represented by a descriptor vector. Haar wavelet responses in ξ and ζ direction within circle of radius $6s$ around that interest point (s is scale at which the interest point was detected) were calculated. The horizontal and vertical responses within the window are summed and yield a local orientation vector. The longest such vector among all windows gives the orientation of the interest point. Then, a square region of size $20s$ around interest point is split into 16 small sub-squares (4×4 within one square). Then, four Haar wavelet responses at 5×5 regularly spaced sample points are computed respectively: $\sum d_\xi$, $\sum d_\zeta$, $\sum |d_\xi|$ and $\sum |d_\zeta|$. This gives a SURF descriptor vector of length 64 for that interest point.

Every interest point in the first image can be compared to every point in the second image by calculating the Euclidean distance between their descriptor vectors. A pair (match) is detected, if distance of the nearest is less than T times the distance of the second nearest neighbor. Since this measure is asymmetrical (matching from the second to the first image) those that appear in both directions are called *bidirectional matches* (see Figure 3).

The SURF features from all database images were associated with the image and their CP of origin. The features were split into two groups (denoted ± 1 respectively) based on the sign of the Laplacian, which halves the search time. For each group, we created a hierarchical tree clustering the descriptor vectors using the extended κ-means algorithm repeatedly. This partitioning of U features into κ disjoint subsets S_j each containing U_j features, minimizes the sum-of-squares criterion

$$\mathbf{J} = \sum_{j=1}^{\kappa} \sum_{u \in S_j} \left\| \mathbf{l}_u - \boldsymbol{\mu}_j \right\|^2, \tag{11}$$

where $\mathbf{l}_u$ is a vector representing the u-th data point and $\boldsymbol{\mu}_j$ is the geometric centroid of all data points in S_j.

The algorithm consists of the re-estimation procedure as follows. Initially, the features are assigned at random

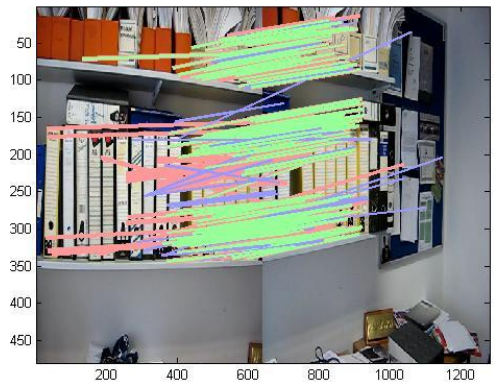


Fig. 3. SURF matching between a testing and a training image (separated vertically at pixel 640 on the horizontal axis) during the testing stage. Unidirectional matches in right image that correspond to left image are represented with red lines (vice versa for blue lines) and bidirectional matches with the green lines.

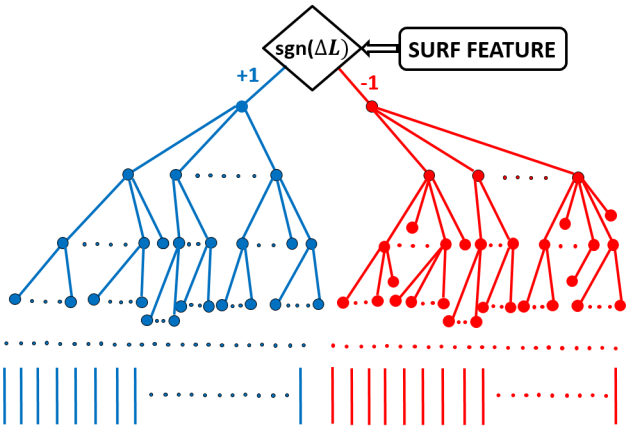


Fig. 4. Proposed vocabulary tree used in image-based localization. The red and blue circles denote the vocabulary tree clusters' centers. The red and blue vertical lines denote the tree's last level SURF descriptor vectors.

to the sets. For step 1, the centroid is computed for each set. In step 2, every feature is assigned to the cluster whose centroid is closest to that feature. These two steps are alternated until a stopping criterion is met. For the first two or even three levels of the hierarchical tree κ cluster centers were found by calculating the mean value of several previously calculated cluster centers. In other words, for ι iterations there are ι cluster centers vectors, each of length κ. Then the mean value for each dimension was calculated resulting to a vector of length ι representing κ cluster centers. For the higher tree levels this process is not necessary as cluster centers are already properly *fixed*. This approach only requires linear memory, $O(\kappa+U)$, in the number of cluster centers κ and feature points U.

For a query image, its SURF descriptor vectors and the (corresponding) signs of the Laplacian were extracted and a match for each descriptor vector was found using $+1$ or -1 hierarchical tree (see Fig. 4). Since the match was labeled with the image and location from which it was extracted it, therefore, casted one vote for its associated location. After each descriptor vector had voted for a location locations were ranked from the most likely to least likely. A *verification stage* was employed by using the bidirectional matching to reorder the top 5 previously ranked locations. Firstly, ranking obtained by the descriptor vectors was weighted by the normalized bidirectional matching location scores and again normalized, thus associating normalized votes with each CP. A confidence is assigned for each CP, denoted by q_i, and defined as the ratio of normalized votes associated with that CP and total number of the normalized votes. Similar to Section 3, we calculate user location, denoted by $\mathbf{r}_I$, using the MMSE estimation algorithm given by

$$\mathbf{r}_I = \frac{\sum_{i=1}^{K} q_{maxi}\mathbf{CP}_{maxi}}{\sum_{i=1}^{K} q_{maxi}}. \quad (12)$$

5 Threshold-Based Fusion Method

To perform threshold-based fusion, we take the confidences p_i and q_i from both sensing modalities P (WLAN) and Q (image) into account. Here, i refers to a given CP. The first ranked, the second ranked, the third ranked, etc. sorted confidences are denoted by p_{max1}, p_{max2}, p_{max3}, etc. respectively (and similarly for Q).

Let us define $P_{ij} = p_{maxi} - p_{maxj}$ and similarly $Q_{ij} = q_{maxi} - q_{maxj}$. We used a separate training and validation dataset to derive the fusion function and to define the threshold values. Observing P_{12} and Q_{12} in many confidence pairs, which were derived using the validation dataset, we concluded that for values P_{12} and/or Q_{12} beyond some reliably large thresholds, we were sure that the nearest CP (location) was the 1^{st} ranked one, based either on P or Q (or both). These reliably large thresholds, denoted by T_P and T_Q for P and Q modality respectively, are equal to

$$T_P = \min_{VAL_P} \{P_{VAL_P 12}\} \quad (13)$$

$$T_Q = \min_{VAL_Q} \{Q_{VAL_Q 12}\} \quad (14)$$

and are derived based on the validation datasets (denoted by VAL_P and VAL_Q for P and Q modality, respectively). Moreover, we deduced that introducing multiplication ($p_i q_i$) and/or addition ($p_i + q_i$) functions under some conditions, i.e. using more thresholds, can decrease the positioning error even more. But to avoid over-fitting we have not used the additional multiplication/addition eventually.

We found that the ranking of the correct location did not fall below some positions in both sets of rankings (the α_P^{th} position for P and the α_Q^{th} position for Q modality). If none of the conditions is satisfied we decided to take the ranking of the modality to which $min(\alpha_P, \alpha_Q)$ corresponds. The steps in the fusion process are

$$f_i = \begin{cases} p_i, & P_{12} \geq Q_{12} \wedge P_{12} > T_P \wedge Q_{12} > T_Q \\ q_i, & Q_{12} > P_{12} \wedge P_{12} > T_P \wedge Q_{12} > T_Q \\ p_i, & P_{12} > T_P \wedge Q_{12} < T_Q \\ q_i, & Q_{12} > T_Q \wedge P_{12} < T_P \\ \beta_i, & \text{else} \end{cases}. \quad (15)$$

Here f_i represents the fusion confidence, while β_i is the confidence of the method to which $min(\alpha_P, \alpha_Q)$ corresponds. Similar to Sections 3 and 4, we calculate the user location, denoted by $\mathbf{r}_{Ft}$, using the MMSE estimation algorithm given by

$$\mathbf{r}_{Ft} = \frac{\sum_{i=1}^{K} f_{maxi} \mathbf{CP}_{maxi}}{\sum_{i=1}^{K} f_{maxi}}. \quad (16)$$

This threshold-based fusion approach is evaluated in Section 8 using a test dataset that is different from the training and the validation datasets.

6 Fusion based on particle filter

In this approach, the sequentially arriving RSS and image measurements are fused with location predictions from the user kinematic model using a Sampling Importance Resampling (SIR) particle filter [44]. The particle filter method projects the state of the user to be tracked (particles) one step ahead. This is followed by RSS and image measurement acquisition to assign weights to particles and generate a probability distribution. Next, the motion and measurement models are introduced together with the description of the particle filter.

6.1 Motion Model and Kinematic Prior Propagation

The position and velocity of a user at time step $k = 1, \ldots, K$ within a building interior is described by vector $\boldsymbol{\mathcal{X}}_k = [\chi_k \ \dot{\chi}_k \ \psi_k \ \dot{\psi}_k]^T$, where (χ_k, ψ_k) are the positions in the χ and ψ dimension, and $(\dot{\chi}_k, \dot{\psi}_k)$ are the corresponding velocities. The user motion is described as

$$\boldsymbol{\mathcal{X}}_k = \mathbf{F}\boldsymbol{\mathcal{X}}_{k-1} + \mathbf{Q}\mathbf{v}_{k-1}, \quad (17)$$

$$\mathbf{F} = \begin{bmatrix} 1 & \delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \delta t \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and δt is the time difference between state transitions. $\mathbf{Q}$ is a diagonal process noise covariance matrix, and $\mathbf{v}_k$ denotes a zero-mean, unit variance Gaussian process that models velocity errors. The model in (17) is associated with the kinematic prior distribution $P(\boldsymbol{\mathcal{X}}_k | \boldsymbol{\mathcal{X}}_{k-1})$.

Particles are then defined which represent realizations of possible user states $\boldsymbol{\mathcal{X}}_{h,k} = [\chi_{h,k} \ \dot{\chi}_{h,k} \ \psi_{h,k} \ \dot{\psi}_{h,k}]^T$ where $h = 1, \ldots, H$ and H is the total number of particles used. Each particle $h = 1, \ldots, H$ is propagated one step ahead using the kinematic model in (17) as

$$\boldsymbol{\mathcal{X}}_{h,k} = \mathbf{F}\boldsymbol{\mathcal{X}}_{h,k-1} + \mathbf{Q}\mathbf{v}_k \quad (18)$$

which is equivalent to sampling from the kinematic prior distribution $P(\boldsymbol{\mathcal{X}}_k | \boldsymbol{\mathcal{X}}_{h,k-1})$.

6.2 Measurement Models

The training measurement set is used as information on the correspondence of measurements to user locations, which we have incorporated in the WLAN and image measurement models detailed in the following.

6.2.1 WLAN Measurements Model

The signal strength measurements $y_m^{(n)}$, $m = 1, \ldots, M$, $n = 1, \ldots, N$ of the training set collected at CP $i = 1, \ldots, K$ at location $(\chi_{CP,i}, \psi_{CP,i})$ are assumed to be Gaussian distributed with mean and variance that are estimated using the training set measurements as

$$\hat{\mu}_{i,m} = \frac{1}{N} \sum_{n=1}^{N} y_m^{(n)} \quad (19)$$

$$\hat{\sigma}_{i,m}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (y_m^{(n)} - \hat{\mu}_{i,m})^2. \quad (20)$$

The received signal measurement likelihood $P_{rss,i}$ for each CP and from M APs using (7) is given by

$$P_{rss,i} = \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\hat{\sigma}_{i,m}^2}} e^{-\frac{(y_{rss,m} - \hat{\mu}_{i,m})^2}{2\hat{\sigma}_{i,m}^2}} P(x_m | i). \quad (21)$$

6.2.2 Image Measurements Model

For the image measurements the training set measurements is used to provide normalized votes q_i on the CP from which it is likely to have obtained a given image y_{img} in the current time step as described in Section 4. Therefore, the normalized votes can be interpreted as how likely it is that a test image was taken at a certain location. The image measurement likelihood is then approximated to be equal to the normalized votes as

$$P_{img,i} = q_i. \quad (22)$$

This is then used as probability distribution that indicates the probability that an image was taken at each of the locations $(\chi_{CP,i}, \psi_{CP,i})$, $i = 1, \ldots, K$.

6.3 Particle Weighting with Measurements

During localization, the particles are assigned weights based on RSS and image measurements generated based on a true user state $\boldsymbol{\mathcal{X}}_k = [\chi_k \ \dot{\chi}_k \ \psi_k \ \dot{\psi}_k]^T$ at location (χ_k, ψ_k), which is unknown to the tracker.

6.3.1 Likelihood based on RSS measurements

A RSS measurement for the current time step that is due to the true user position (χ_k, ψ_k) is taken from the test dataset. The location of a CP $(\chi_{CP,\bar{i}}, \psi_{CP,\bar{i}})$ is identified that is nearest to (χ_k, ψ_k) where testing set received signal measurements exist which is indexed by $\bar{i} = \underset{i}{\text{argmin}} \ ||[\chi_k, \psi_k] - [\chi_{CP,i}, \psi_{CP,i}]||_2^2$. Then, a measurement from the test dataset, denoted as $y_{rss,m}$, is selected

for each AP corresponding to CP $\bar{i}$. In addition, for each particle proposed location from $(\chi_{h,k}, \psi_{h,k})$ a location for which training measurement data exist is identified as $(\chi_{CP,i_h}, \psi_{CP,i_h})$ where $i_h = \underset{i}{\text{argmin}} \, ||[\chi_{h,k}, \psi_{h,k}] - [\chi_{CP,i}, \psi_{CP,i}]||_2^2$. Index i_h then defines likelihood distribution given in (21) for particle h with mean and variance $\mu_{h,m} = \hat{\mu}_{i_h,m}$ and $\sigma_{h,m}^2 = \hat{\sigma}_{i_h,m}^2$ in (19) and (20) respectively for each AP $m = 1, \ldots, M$. The likelihood when using RSS measurements is now given by

$$P_{rss,h,i} = \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma_{h,m}^2}} e^{-\frac{(y_{rss,m} - \mu_{h,m})^2}{2\sigma_{h,m}^2}} P(x_m|i). \quad (23)$$

for each particle h and each AP m.

6.3.2 Likelihood based on image measurements

Image measurements that arise due to the true user state are taken from the image test dataset. The index of the CP nearest to the true user state where images were collected is identified as $\tilde{i} = \underset{i}{\text{argmin}} \, ||[\chi_k, \psi_k] - [\chi_{CP,i}, \psi_{CP,i}]||_2^2$. Then, an image is drawn uniformly at random from the test dataset of CP i denoted as y_{img}. Then, for each particle proposed state $\chi_{h,k}$ the location for which training set image measurements exist is identified as $(\chi_{CP,\tilde{i}_h}, \psi_{CP,\tilde{i}_h})$ where $\tilde{i}_h = \underset{i}{\text{argmin}} \, ||[\chi_{h,k}, \psi_{h,k}] - [\chi_{CP,i}, \psi_{CP,i}]||_2^2$ and the likelihood for each particle n based on the image measurements is taken as the normalized votes in (22) as $P_{img,h} = q_{img,\tilde{i}_h}$.

6.3.3 Particle weighting

Considering both RSS and image measurements and assuming that they are mutually independent, the weight of each particle using the RSS and image likelihoods is

$$\bar{w}_{h,k} = P_{img,h} \prod_{m=1}^{M} P_{rss,h,m} P(x_m|i_h). \quad (24)$$

Weights are then normalized as $w_{h,k} = \frac{\bar{w}_{h,k}}{\sum_{h=1}^{H} \bar{w}_{h,k}}$ and the estimates of the posterior distribution and the estimate of the user state are respectively given by

$$\hat{P}(\boldsymbol{\mathcal{X}}_k | \{y_{rss,m}\}_{m=1}^{M}, y_{img}) = \sum_{h=1}^{H} w_{h,k} \delta(\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}_{h,k}) \quad (25)$$

$$\hat{\boldsymbol{\mathcal{X}}}_k = \sum_{h=1}^{H} w_{h,k} \boldsymbol{\mathcal{X}}_{h,k} \quad (26)$$

which is followed by particle resampling [44]. The particle filter fusion algorithm is outlined in Table 1. We used 600 particles, when using more the accuracy did not improve. In Section 8 we evaluate the performance of this algorithm and the positioning error is given as the Euclidean distance between the estimated position in (26) and the true user position.

TABLE 1
The RSS-Image Particle Filter Fusion Tracking Algorithm

- For each particle $h = 1, \ldots, H$
  - Propose a new user state at time k as (18) $\boldsymbol{\mathcal{X}}_{h,k} = \mathbf{F}\boldsymbol{\mathcal{X}}_{h,k-1} + \mathbf{Q}\mathbf{v}_k$
  - Collect RSS $\{y_{rss,m}\}_{m=1}^{M}$ and image y_{img} measurements
  - Calculate likelihoods and weights $w_{h,k}$ (24)
- Calculate estimate of posterior distribution (25) $\hat{P}(\boldsymbol{\mathcal{X}}_k | \{y_{rss,m}\}_{m=1}^{M}, y_{img})$
- Calculate estimate $\hat{\boldsymbol{\mathcal{X}}}_k$ of posterior (26)
- Particle resampling

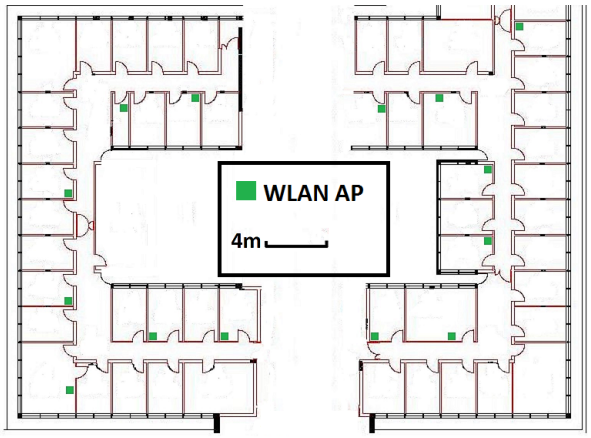


Fig. 5. Map of offices used in the experiments where green squares denote the 14 APs. The office room average size is equal to $12.9\,\mathrm{m}^2$.

7 Experimental Setup

For our experiments we use 36 offices (see Figure 5 where most offices were employed) in the School of Electronic Engineering, Dublin City University, Ireland. Within each office we use 5 CPs, denoted A, B, C, D, E, which are placed at each corner of the office and at its center, as shown in see Figure 6). Each orientation of a CP (N, S, W and E) has 8 (640×480 pixel) images and 150 associated RSS observations taken with Canon PowerShot A560 camera and Dell Inspiron laptop with Intel Core 2 Duo Processor T5250 (2.0 GHz, 2 MB L2 cache, 667 MHz FSB), memory of 2×2048 MB, 667 MHz Dual Channel DDR2 SDRAM, SATA Hard Drive with 450 GB, and Intel PRO/Wireless 3945ABG card, respectively. However, the acquisition and processing methods used to develop the proposed localization approach and run the pilot tests do not restrict the mass adoption of the approach on the wide variety of commercial devices, including tablets, smartphones, and mobile robots, that are equipped with WLAN and image modules.

In this work, we use the RSS Indicator (RSSI) value reported by the WLAN adapter, which is defined as the absolute RSS value. RSS data were captured using InSSIDer[4] software. An observation consists of RSS readings from up to 14 WLAN APs. Note that these

4. http://www.metageek.net/products/inssider/

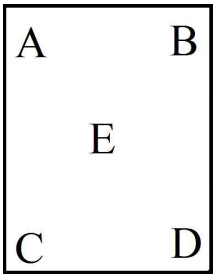


Fig. 6. Calibration points $ABCDE$ within an office.

Fig. 7. Some of the images used in the experiments.

APs are part of the university infrastructure for the provision of wireless connectivity and we did not install any additional APs.

We gathered $120,000$ images, of which $83,000$ were used for training and $17,000$ for testing, and $210,000$ signal strengths observations of which $160,000$ were used for training and $20,000$ for testing. To derive threshold values and the fusion function in the threshold-based fusion method, we have used an independent set of $20,000$ images and 30,000 signal strength observations as the validation dataset. During image and RSS data collection, the user was standing still at the CPs. During the training stage image and WLAN data was taken at the CPs, while during the validation and testing stage it was taken at arbitrary points.

Offices are next to each other and look very similar inside thus resulting in very challenging data for both WLAN and image-based localization methods (see examples in Figure 7). Each CP is associated with several datasets using data taken at different times of the day and different days to demonstrate the robustness of the localization approaches. During localization, the user collected one image and one RSS fingerprint from the test dataset and investigated the distance between the estimated and the true user location.

8 Performance Evaluation

We assess the performance of our system in terms of the trajectory matching accuracy (in $\%$) defined as the Euclidean distance between the true and estimated user locations. Specifically, we report the mean positioning error $\mathcal{E}_p$ together with the 95% confidence interval given by $\mathcal{E}_p \pm 1.96\sigma/\sqrt{n}$, where σ denotes the standard deviation of the positioning error and n is the number of test samples. Essentially, the 95% confidence interval indicates that $\mathcal{E}_p$ falls within the interval with a high degree of certainty.

Our assessment is performed under different conditions by considering parameters such as varying number of CPs per office, number of WLAN APs, number of training images, and number of training RSS fingerprints. To avoid introducing any bias in the results by considering a specific subset of CPs per office or subset of WLAN APs, etc., the reported results are further averaged over all possible combinations of each parameter.

Later, our system is compared against other solutions with respect to positioning error and computation time.

8.1 Effect of number of CPs per office

In this experiment, we vary the number of CPs per office while we fix the number of WLAN APs to 14 and the number of training RSS fingerprints per CP is equal to 600. The statistics of the positioning error for different methods are depicted in Figure 8. In particular, the height of each bar and the whiskers indicate the mean value $\mathcal{E}_p$ and the 95% confidence interval, respectively.

It is clear that the fusion of WLAN and images improves accuracy compared to using either WLAN or image as standalone localization methods. In each case $\mathcal{E}_p$ decreases when the number of CPs per office increases. However, this comes at the expense of higher data collection effort and time for populating the database with training RSS fingerprints and images, thus Figure 8 provides a guideline for this trade-off. Even though it pays off in terms of positioning error to survey more CPs per room (e.g., $\mathcal{E}_p$ drops from around $4\,\mathrm{m}$ to $2\,\mathrm{m}$ when 5 CPs, instead of 1 CP, are considered), some applications might tolerate higher error but have strict setup time constraints (e.g., data collection completed in a few hours, rather than few days).

8.2 Effect of number of APs

In this case, 5 CPs per office were used and we vary the number of WLAN APs while the number of training RSS fingerprints and the number of images per CP is equal to 600 and 32, respectively. Figure 9 shows the trend of $\mathcal{E}_p$ for increasing number of APs. The Particle filter and Threshold-based fusion methods are considerably better than the standalone WLAN localization method, reaching around $2\,\mathrm{m}$ with 8 APs. For reference, our Image method achieves $\mathcal{E}_p = 2.7\,\mathrm{m}$.

As expected, $\mathcal{E}_p$ does not decrease significantly when more than 8 APs are considered. This is in line with what has been reported in the related literature in the past; a few APs (i.e., low dimension of the RSS fingerprints) are not enough to sufficiently distinguish between locations, while using more APs beyond a certain point (i.e., high dimension of the RSS fingerprints) does not improve the

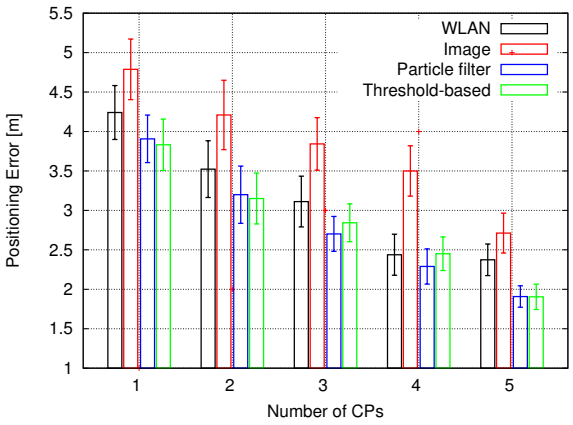


Fig. 8. Positioning error of the WLAN, Image, Particle filter fusion and Threshold-based fusion methods for variable number of CPs per office.

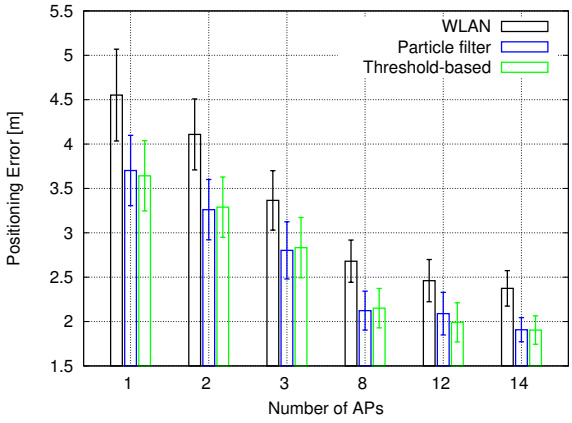


Fig. 9. Positioning error of the WLAN, Particle filter fusion and Threshold-based fusion methods as the number of APs increases.

discriminative capability of the fingerprints. Moreover, factors such as the inherent uncertainty in RSS data, due to noise and measurement errors, and especially the modeling errors that result from the finite number of calibration points, all place a limit to the possible localization accuracy that cannot increase by increasing the number of APs. This result suggests that reasonably accurate localization can be achieved in new WLAN deployments with lower budget and quicker installations.

8.3 Effect of number of training images

In this experiment, we consider 14 WLAN APs, 5 CPs per office, and the number of training RSS fingerprints per CP is equal to 600. The bar chart in Figure 10 illustrates the improvement in $\mathcal{E}_p$ as the number of training images per CP increases. Increasing the number of images per CP is achieved by equally using more images per orientation in every CP.

As expected, the performance of all methods improves when more training images are considered and the fusion methods consistently attain lower $\mathcal{E}_p$ by around $1\,\mathrm{m}$ compared to the standalone Image localization method.

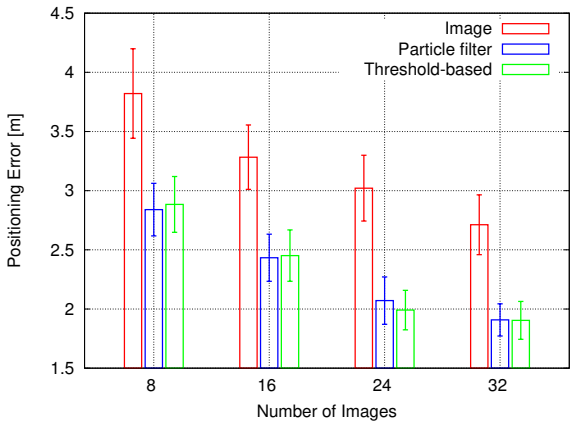


Fig. 10. Positioning error of the Image, Particle filter fusion and Threshold-based fusion methods as the number of training images per CP increases.

The WLAN method does not depend on the number of training images and delivers $\mathcal{E}_p = 2.4\,\mathrm{m}$ in all cases. This accuracy level is reached by the fusion methods using 16 training images per CP, while doubling the number of images further reduces $\mathcal{E}_p$ by $0.5\,\mathrm{m}$. Therefore, there is again a trade-off between the positioning error and the setup time of the system that increases significantly when more training images are captured.

8.4 Effect of number of training RSS data

In this experiment, we consider 14 WLAN APs, 5 CPs per office, and 32 training images per CP. Figure 11 shows how $\mathcal{E}_p$ decreases when the amount of training RSS fingerprints increases. Clearly, the two fusion methods utilizing both RSS and image data outperform the WLAN method in all cases. The Image localization method achieves $\mathcal{E}_p = 2.7\,\mathrm{m}$.

Similarly to the number of training images discussed previously, there is a trade-off between the positioning error and the setup time of the system that increases significantly when more training RSS fingerprints are collected in every CP. For instance, collecting 600 RSS fingerprints per CP, instead of 400, only improves $\mathcal{E}_p$ by around $0.4\,\mathrm{m}$ for the fusion methods.

8.5 Particle filter for a dynamic scenario

The particle filtering method performs data fusion from multiple heterogeneous sources producing measurements that have a non-linear relationship to the target state. In addition, the particle filtering method is capable of incorporating target kinematic information into the estimation process. Moreover, the particle filter is able to handle measurements that arrive asynchronously or at irregular intervals by continuing to propagate the belief on the target state via regularly updating the state using the kinematic prior and updating the predictions on the target state with knowledge from new measurements when new measurements do become available.

In the results presented so far, where the target kinematic properties did not include a high uncertainty

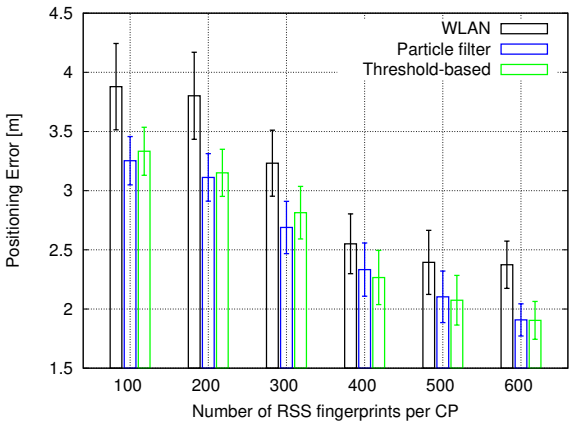


Fig. 11. Positioning error of the WLAN, Particle filter fusion and Threshold-based fusion methods as the number of training RSS fingerprints per CP increases.

due to the semi-stationary pattern of motion, the particle filter method did not achieve significant accuracy improvement as compared to the threshold-based approach. To demonstrate the effectiveness of the particle filter in a dynamic scenario, we rearranged the order of the collected data appropriately to emulate a user walking along a path that passes from one office to the next assuming a typical user walking speed. We produced 13 different trajectories of the same path by considering test measurements from nearby randomly selected locations within each office. We fixed algorithm-specific parameters to their optimal values, i.e., 5 CPs, 14 APs, 32 images, 600 RSS fingerprints per CP, and 600 particles as in the above experiments. In this case, the average of the mean positioning errors pertaining to these 13 trajectories is $1.566\,\mathrm{m}$ with a confidence interval of $0.32\,\mathrm{m}$ compared to $1.908\,\mathrm{m}$ in static localization, as shown in Figures 8-11.

8.6 Hybrid fusion

As images are one of the most energy-consuming data sources we would like to avoid using them continuously, but rather employ them only when necessary, e.g., in case the WLAN method is not expected to have good accuracy. Therefore, we investigate the effect on positioning error in case we use images only when the number of sensed APs in the observed RSS fingerprint is less than 3. The intuition is that according to the previous analysis in Section 8.2 the positioning error of the WLAN method degrades significantly below that value. Thus, the hybrid approach provides a practical way for indoor localization where users move around freely enjoying reliable WLAN RSS-based location information and stop to take a picture of the surroundings when the system detects that the user is at an unknown location or in a region with few detected APs that may result in poor WLAN RSS-based localization.

We modified the *original* Particle filter and Threshold-based fusion methods, which use the captured image in every localization test, to create their *hybrid* variants that use images sporadically based on the number of sensed APs. In other words, the *hybrid* methods compute user location using only WLAN RSS data most of the time and fuse it with image data only if one or two APs are sensed in the RSS fingerprint.

Images used in every test	Particle filter	Threshold-based
Yes (*original* fusion)	1.908	1.904
No (*hybrid* fusion)	2.337	2.400

TABLE 2
Positioning error of the *hybrid* fusion methods.

Table 2 reports $\mathcal{E}_p$ for the *hybrid* fusion methods. The *hybrid* methods deliver higher positioning error by around $0.5\,\mathrm{m}$ compared to the *original* fusion methods. However, this is compensated by time efficiency due to the less frequent sampling and processing of images during localization, as analyzed in the following Section 9.

9 Computation Time

A series of experiments was conducted with a goal to compare computation time of the five localization solutions, namely the WLAN, the image, the threshold-based fusion, the particle filter fusion, and the hybrid fusion. Note that we report results only for the particle filter hybrid fusion solution for brevity. We used the same laptop with the specifications described in Section 7, while all algorithms were implemented in Java.

Computation times are calculated using a number of tests and we report the average value. We assume that each test is equally time consuming. We denote $\bar{t}_W$ and $\bar{t}_I$ the average computation times for obtaining the user's location with WLAN and image modalities given by

$$\bar{t}_W = \bar{t}^a_W + \bar{t}^l_W, \tag{27}$$

$$\bar{t}_I = \bar{t}^a_I + \bar{t}^l_I, \tag{28}$$

where $\bar{t}^a$ and $\bar{t}^l$ denote the times required for the acquisition of the corresponding modality and the localization process, respectively.

Acquisition of a WLAN RSS takes in our case, $\bar{t}^a_W = 1\,\mathrm{s}$, while determining the user location with the WLAN localization algorithm takes $\bar{t}^l_W = 0.058\,\mathrm{s}$. Thus, the whole process takes in total $\bar{t}_W = 1.058\,\mathrm{s}$. On the other hand, the acquisition of a 640×480 pixel image takes $\bar{t}^a_I = 0.21\,\mathrm{s}$, while the localization process takes $\bar{t}^l_I = 3.962\,\mathrm{s}$. In total, this gives $\bar{t}_I = 4.172\,\mathrm{s}$.

For the threshold-based fusion one has to take one RSS observation and one image for each test and subsequently use the image-based and the WLAN-based location results in the fusion process. In the case of particle filter fusion only acquisitions of the WLAN and the image data are taken into account in addition to the particle filter localization process. For the hybrid fusion, which is based on the particle filter, an image is acquired only when the WLAN-based location is considered unreliable (i.e., when less than 3 APs are sensed in the measured fingerprint) in addition to the localization time

	W	I	Ft	Fp	H
Time (s)	1.058	4.17	5.68	8.87	5.96

TABLE 3
Average computation time of the WLAN (W), Image (I), Threshold-based fusion (Ft), Particle filter fusion (Fp), and Hybrid fusion (H) methods.

of the hybrid fusion. Therefore, the average computation times for the fusion methods are given by

$$\bar{t}_{Ft} = \bar{t}_W + \bar{t}_I + \bar{t}^l_{Ft}, \tag{29}$$

$$\bar{t}_{Fp} = \bar{t}^a_W + \bar{t}^a_I + \bar{t}^l_{Fp}, \tag{30}$$

$$\bar{t}_H = \bar{t}^a_W + \rho\bar{t}^a_I + \bar{t}^l_H, \tag{31}$$

where ρ is a parameter pertaining to the hybrid fusion that denotes the percentage of tests where an image was acquired. In our tests we observed that $\rho = 21.5\%$.

Table 3 summarizes the computation time for each localization method. Among the fusion methods we observe that the Hybrid method is the most power efficient; however, this comes at the expense of around $0.4\,\mathrm{m}$ degradation in the positioning error compared to the Particle filter fusion method, as shown before in Table 2. The latter method proves to be the most demanding in terms of run time. Therefore, the Threshold-based fusion method is a good compromise between time-efficiency and positioning error.

We note that using a more powerful computer or server, instead of a laptop, and also applying techniques to parallelize the computations for the particles (now they are performed sequentially) could easily reduce the time to compute the user's location from a few seconds to less than one second. Thus, the proposed system would be applicable to practical real-life applications for localizing individuals that move at normal walking speed inside a building. In this case, the latency of the system (i.e., the time required to compute the user location) depends on the time to acquire a WLAN measurement (i.e., $1\,\mathrm{s}$).

Moreover, energy consumption of a particular method running on a device follows the method's computation time in a monotonically increasing fashion, i.e., as the method's computation time increases, the energy consumption of the method running on the device increases as well. Due to lack of space we omit the corresponding results related to energy consumption of the aforementioned methods running on the laptop.

10 Comparison with other methods

The proposed fusion methods are compared against two state-of-the-art methods presented in [15] and [12].

The system in [15] uses WLAN-based localization, which follows a Naive Bayes approach, together with image-based localization. In particular, the system estimates the user's location from the scanned WLAN RSS values using a modified version of the centroid algorithm; see [15] for more details. In the image-based localization, a fusion algorithm is employed based on images extracted from video using the *FFmpeg* application. The target is modeled as a simple three dimensional cylindrical object but using a single camera with multiview perspective. Images captured from cameras are degenerated to two dimensional planar images.

	Threshold-based	Particle filter	[15]	[12]
Time (s)	5.68	8.87	9.61	8.67

TABLE 4
Average computation time of the Threshold-based fusion, Particle filter fusion, and the methods in [15], [12].

In [12], WLAN-based localization is achieved by comparing RSS fingerprints in the database (collected offline) and the RSS fingerprint taken by a client (observed online). Similarly to our approach, if an AP is present in the observed fingerprint at the location of the device, but not present in a database fingerprint, then the matching between them should be low and a penalty is applied to handle this. This was also the case if an AP was missing in the observed fingerprint, but was present in the fingerprint in the database. In the image/video-based part, foreground segmentation is employed followed by how human shapes are extracted and mapped to floor plan as it is explained in detail in [12]. When both WLAN and camera data are available, then the two measurements are combined with a naive Bayesian approach. In the following, we compare the Particle filter and Threshold-based fusion methods with the methods in [15] and [12] in terms of mean positioning error $\mathcal{E}_p$.

In the first experiment, we consider 5 CPs per office, while the number of training images and RSS fingerprints per CP is 32 and 600, respectively. Figure 12 shows the performance comparison as we vary the number of APs. The proposed fusion methods outperform the methods in [15] and [12]. In particular, $\mathcal{E}_p$ is lower by around $1.5\,\mathrm{m}$ and $1\,\mathrm{m}$ when only one AP or two APs are considered, respectively. Even though the positioning error does not drop considerably when more than 8 APs are considered, fusion methods are still more accurate in the mean by about $0.5\,\mathrm{m}$ when 8 or more APs are used.

In the second experiment, we fix the number of APs to 14 and vary the number of training RSS fingerprints per CP. The results are depicted in Figure 13. We observe similar behavior for all methods and the proposed fusion methods deliver positioning error below $3\,\mathrm{m}$ when 300 RSS training fingerprints are used.

Finally, we present results related to computation time in Table 4. We observe that our Threshold-based fusion method is very efficient and achieves significant savings in terms of computation time compared to competing methods in [15] (i.e., 41% reduction in computation time) and [12] (i.e., 35% reduction). On the other hand, the Particle filter fusion method outperforms the method in [15], but is slightly worse than the method in [12]. In this case, the *hybrid* version of the Particle filter fusion method performs better than [12] in terms of time, while still providing lower positioning error.

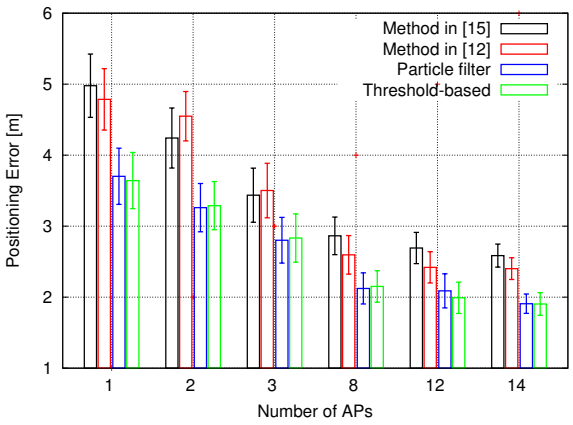


Fig. 12. Positioning error of the methods in [15], [12], Particle filter fusion and Threshold-based fusion methods as the number of APs increases.

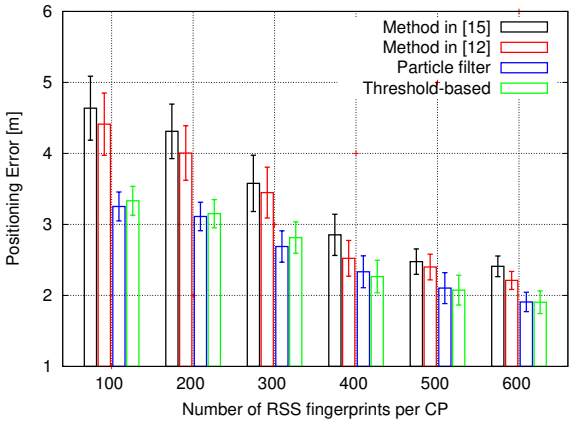


Fig. 13. Positioning error of the methods in [15], [12], Particle filter fusion and Threshold-based fusion methods as the number of training RSS fingerprints per CP increases.

11 Conclusion

In this work we investigate the combination of two complementary data sources for indoor localization and propose a novel image-based localization algorithm, as well as two strategies for fusing either the locations induced by WLAN and image localization algorithms or the raw WLAN and image measurements directly. The results demonstrate that the fusion methods achieve lower positioning error than any individual modality, while outperforming competing fusion approaches.

Both our fusion methods deliver similar accuracy; however, they have different features that make each one of them the preferred solution depending on the application scenario. For instance, the Threshold-based fusion method is more light-weight, i.e., it has lower computational complexity, resulting in lower run time and energy consumption. On the other hand, it requires the collection of a separate validation dataset and subsequent fine-tuning for selecting algorithm-specific thresholds, thus increasing the system setup time. In this case, the more flexible Particle filter fusion method can be used instead. The flexibility of the particle filter algorithm is demonstrated when used as a hybrid fusion approach able to trade off positioning error with reduction in computational time. Finally, it can fuse measurements from additional heterogeneous sources (additionally to image and RSS data) if available.

Future work will investigate the use of dynamic confidence-based weighting between the WLAN and image modalities in both fusion approaches. Such adaptive fusion scheme is expected to further improve the positioning error at no additional time-energy cost. In addition, the use of different WLAN bands at $2.4\,\mathrm{GHz}$ and $5\,\mathrm{GHz}$, and Bluetooth beacons and a fusion of WLAN, IMUs, and image data can be used to improve performance and improve the versatility of our localization system. A possible research direction would be to leverage WLAN Channel State Information (CSI) information instead of RSS as in the Dynamic-MUSIC algorithm [6] to further improve WLAN-based localization accuracy due to the higher resolution of the CSI measurements.

References

[1] A. Rai, K.K. Chintalapudi, N.V. Padmanabhan and R. Sen, *Zee: zero-effort crowdsourcing for indoor localization*, In Proceedings of the MobiCom, pp. 293-304, 2012.

[2] M. Redzic, C. O'Conaire, C. Brennan, and N.E. O'Connor, *A hybrid method for indoor user localization*, In 4th European Conference on Smart Sensing and Context (EuroSSC), 2009.

[3] B. R. Chang, H. F. Tsai, C. P. Young, *Intelligent data fusion system for predicting vehicle collision warning using vision/GPS sensing*, Expert Systems with Applications, vol. 37, pp. 2439-2450, 2010. Elsevier.

[4] M. D. Redžić, C. Brennan, and N. E. O'Connor, *SEAMLOC: Seamless Indoor localization Based on Reduced Number of Calibration Points*, IEEE Transactions on Mobile Computing, vol. 13, no. 6, pp. 1326–1337, June 2014.

[5] S. Mazuelas, A. Bahillo, R.M. Lorenzo, P. Fernandez, F.A. Lago, E. Garcia, J. Blas and E.J. Abril, *Robust Indoor Positioning Provided by Real-Time RSSI Values in Unmodified WLAN Networks*, IEEE Journal of Selected Topics in Signal Processing, vol.3, no.5, pp.821-831, October 2009.

[6] X. Li, S. Li, D. Zhang, J. Xiong, Y. Wang and H. Mei, *Dynamic-MUSIC: Accurate device-free indoor localization*, International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), 2016.

[7] L. Ledwich and S. Williams, *Reduced SIFT features for image retrieval and indoor localization*, In Australian Conference on Robotics and Automation, 2004.

[8] S. Zirari, P. Canalda and F. Spies, *WiFi GPS based combined positioning algorithm*, In IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS), 2010.

[9] R. Hansen, R. Wind, C.S. Jensen and B. Thomsen, *Seamless Indoor/Outdoor Positioning Handover for Location-Based Services in Streamspin*, In Tenth International Conference on Mobile Data Management: Systems, Services and Middleware (MDM), 2009.

[10] M. Kourogi, N. Sakata, T. Okuma and T. Kurata, *Indoor/Outdoor Pedestrian Navigation with an Embedded GPS/RFID/Self-contained Sensor System*, In 16th International conference on Artificial Reality and Telexistence (ICAT), 2006.

[11] N. Viandier, F.D. Nahimana, J. Marais and E. Duflos, *MRERA (Minimum Range Error Algorithm): RFID - GPS Integration for vehicle navigation in urban canyons*, In IEEE Position, Location and Navigation Symposium, 2008.

[12] S. Van den Berghe, M. Weyn, V. Spruyt and A. Ledda, *Combining wireless and visual tracking for an indoor environment*, In Proceedings of the second IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN), 21-23. September 2011.

[13] J. Jiao, Z. Deng, L. Xu and F. Li, *A Hybrid of Smartphone Camera and Basestation Wide-area Indoor Positioning Method*, KSII Transactions on Internet & Information Systems. Feb. 2016, Vol. 10, Issue 2, pp. 723-743.

[14] C. Nielsen, J. Nielsen, V. Dehghanian, *Fusion of Security Camera and RSS Fingerprinting for Indoor Multi-Person Tracking*, International Conference on Indoor Positioning and Indoor Navigation, Alcal de Henares, 2016.

[15] T. Miyaki, T. Yamasaki and K. Aizawa, *Multi-Sensor Fusion Tracking Using Visual Information and WiFi Location Estimation*, In Proceedings of the First ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2007.

[16] A. Joki, A.J. Burke and D. Estrin, *Campaignr: A Framework for Participatory Data Collection on Mobile Phones*, Papers, Center for Embedded Network Sensing, UC Los Angeles, 2007.

[17] L. Jiang, H. Zhang, and Z. Cai, *A Novel Bayes Model: Hidden Naive Bayes*, IEEE Transactions on Knowledge and Data Engineering, 21(10):1361-1371, October 2009.

[18] S. He and S. H. G. Chan, *Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons*, IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 466–490, 2016.

[19] D. Zeinalipour-Yazti, C. Laoudias, K. Georgiou, and G. Chatzimilioudis, *Internet-based indoor navigation services*, IEEE Internet Computing, vol. PP, no. 99, pp. 101–114, 2016.

[20] M. Azizyan, I. Constandache, and R. R. Choudhury. *SurroundSense: mobile phone localization via ambience fingerprinting*, In Proceedings of the 15th annual international conference on Mobile computing and networking (MobiCom '09).

[21] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury. *No need to war-drive: unsupervised indoor localization*, In Proceedings of the 10th international conference on Mobile systems, applications, and services (MobiSys '12).

[22] Z. Yang, C. Wu, and Y. Liu. *Locating in fingerprint space: wireless indoor localization with little human intervention*, In Proceedings of the 18th annual international conference on Mobile computing and networking (Mobicom), 2012.

[23] C. Wu, Z. Yang, Y. Liu, and W. Xi, *WILL: Wireless Indoor Localization without Site Survey*, IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 4, pp. 839-848, April 2013. doi: 10.1109/TPDS.2012.179

[24] S.-H. Fang and T. Lin, *Principal component localization in indoor WLAN environments*, IEEE Transactions on Mobile Computing, 11(1):100-110, January 2012.

[25] X. Chai and Q. Yang, *Reducing the Calibration Effort for Probabilistic Indoor Location Estimation*, IEEE Transactions on Mobile Computing, vol. 6, no. 6, pp. 649-662, June 2007.

[26] S.-H. Fang, T. Lin, and K.C. Lee, *A novel algorithm for multipath fingerprinting in indoor WLAN environments*, IEEE Transactions on Wireless Communications, 7(9):3579-3588, September 2008.

[27] T. Sattler, B. Leibe, and L. Kobbelt, *Fast image-based localization using direct 2D-to-3D matching*, In IEEE International Conference on Computer Vision (ICCV), 2011.

[28] Y. Li, N. Snavely, and D.P. Huttenlocher, *Location recognition using prioritized feature matching*, In European Conference on Computer Vision (ECCV), 2010.

[29] H. Lategahn and C. Stiller, *City gps using stereo vision*, In IEEE International Conference on Vehicular Electronics and Safety (ICVES), 2012.

[30] D. Sinha, M.T. Ahmed, and M. Greenspan, *Image Retrieval using Landmark Indexing for Indoor Navigation*, In Canadian Conference on Computer and Robot Vision (CRV), 2014.

[31] R. Mautz, *Indoor positioning technologies*, Doctoral dissertation, Habilitationsschrift, ETH Zurich, 2012.

[32] H. Lategahn and C. Stiller, *Vision-only localization*, IEEE Transactions on Intelligent Transportation Systems, 15(3): 1246-1257, 2014.

[33] Y. Huang, H. Wang, K. Zhan, J. Zhao, P. Gui, and T. Feng, *Image-based localization for indoor environment using mobile phone*, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-4/W5, 2015.

[34] S. Papaioannou, H. Wen, A. Markham, and N. Trigoni, *Fusion of Radio and Camera Sensor Data for Accurate Indoor Positioning*, In Proceedings of the 11th IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS '14).

[35] Y. Du, T. Arslan, A. Juri, *Camera-aided Region-based Magnetic Field Indoor Positioning*, International Conference on Indoor Positioning and Indoor Navigation, Alcal de Henares, 2016.

[36] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao, *TraviNavi: self-deployable indoor navigation system*, In Proceedings of the 20th annual international conference on Mobile computing and networking (MobiCom '14), 2014.

[37] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao, *Magicol: Indoor Localization Using Pervasive Magnetic Field and Opportunistic WiFi Sensing*, IEEE Journal on Selected Areas in Communications, vol. 33, pp. 1443-1457, 2015.

[38] C. Laoudias, P. Tsangaridis, M. Polycarpou, C. Panayiotou, C. Kyrkou, and T. G. Theocharides, *Cooperative fault-tolerant target tracking in camera sensor networks*, In IEEE International Conference on Communications (ICC), 2015, pp. 6634–6639.

[39] T. Oskiper, C. Han-Pang, Z. Zhu, S. Samarasekera and R. Kumar, *Multi-modal sensor fusion algorithm for ubiquitous infrastructure-free localization in vision-impaired environments*, In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2010.

[40] I. Kyriakides, *Target tracking using adaptive compressive sensing and processing*, Signal Processing, vol. 127, pp. 44–55, Oct. 2016.

[41] F. Li, C. Zhao, G. Ding, J. Gong, C. Liu, F. Zhao, *A reliable and accurate indoor localization method using phone inertial sensors*, In Proceedings of the ACM Conference of Ubiquitous Computing (UbiComp), 2012, pp. 421-430.

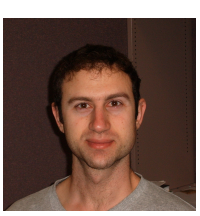
[42] A. Haeberlen, E. Flannery, A.M. Ladd, A. Rudys, D.S. Wallach, and L.E. Kavraki, *Practical robust localization over large-scale 802.11 wireless networks*, In: Proceedings of MobiCom, pages 70-84, 2004.

[43] H. Bay, T. Tuytelaars and L.V. Gool, *SURF: Speeded Up Robust Features*, Computer Vision and Image Understanding (CVIU), 110(3):346-359, August 2008.

[44] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, *A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking*, IEEE Transactions on Signal Processing, vol. 50, no. 2, pp. 174–188, Feb. 2002.

Milan D. Redžić received M.Sc. degree from Faculty of Electrical Engineering, University of Belgrade, Serbia in 2006. and Ph.D. degree from CLARITY: Centre for Sensor Web Technologies, DCU, Ireland in 2012. both in electronic engineering. He is currently a Principal Video Intelligence Consultant in Huawei Ireland Research Center in Dublin, Ireland, working on deep learning for visual recognition related projects. Before that he was with IBM Connections Lab and SAP Predictive Analytics (both in Dublin, Ireland) where he was involved in machine-learning and big data related topics. During his Ph.D. studies, and after as a post-doctoral researcher, he was working on different location-sensing projects both in indoor and outdoor scenarios. His research interests include indoor/outdoor localization, deep learning, computer vision, multi-sensor fusion and similarity measures.

Christos Laoudias is currently a senior research fellow at KIOS Research Center, University of Cyprus contributing to various projects related to localization, tracking, and navigation in telecommunication and smart camera networks. Before that he was leading the geolocation technology research in Huawei Ireland Research Center, Dublin. He holds an Engineering Diploma in Computer Engineering and Informatics (2003) and an M.Sc. in Integrated Hardware and Software Systems (2005) from the University of Patras, Greece, and a Ph.D. in Computer Engineering from the University of Cyprus (2014). During his doctoral studies he was involved in several award-winning indoor localization prototype systems and received the Alpha Bank Cyprus Award for "Creative Research and Innovation". His research interests include positioning and tracking, fault-tolerant algorithms, mobile and pervasive computing, and location-based services.



Ioannis Kyriakides is currently an Associate Professor at the Engineering Department at the University of Nicosia. He received his B.S. degree in Electrical Engineering in 2003 from Texas A&M University. He received his M.S. and Ph.D. degrees in 2005 and 2008 respectively from Arizona State University. His research interests include Bayesian target tracking, sequential Monte Carlo methods, and adaptive compressive sensing and processing.