Dual-sensor fusion for indoor user localisation

Milan Redžić, Conor Brennan and Noel E O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland
{milan.redzic,brennanc,oconnorn}@eeng.dcu.ie

ABSTRACT

In this paper we address the automatic identification of indoor locations using a combination of WLAN and image sensing. Our motivation is the increasing prevalence of wearable cameras, some of which can also capture WLAN data. We propose to use image-based and WLAN-based localisation individually and then fuse the results to obtain better performance overall. We demonstrate the effectiveness of our fusion algorithm for localisation to within a $8.9m^2$ room on very challenging data both for WLAN and image-based algorithms. We envisage the potential usefulness of our approach in a range of ambient assisted living applications.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless communication; H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms

Algorithms, Measurement, Experimentation

Keywords

indoor localisation, WLAN, SURF vocabulary tree, fusion

1. INTRODUCTION

Wearable camera technology has evolved to the point whereby small unobtrusive cameras are now readily available, e.g. the Vicon Revue[1]. This has allowed research effort to focus on analysis and interpretation of the data that such devices provide [1]. Even in the absence of bespoke platforms such as the Vicon Revue, any smart phone can be turned into a wearable camera. The Campaignr[2] configurable micropublishing platform has shown the capability of mobile platforms to act as WLAN (and more general sensor) data gathering hubs. Our group is developing a device using an Android-based smart phone worn on a lanyard around the neck that in addition to image capture also senses a variety of other modalities e.g. motion, GPS, Bluetooth, WLAN. The idea is to use this platform in a variety of ambient assisted living applications as well as assistive technology for the memory and visually impaired. Whilst outdoor localisation is taken care of on this platform via GPS, indoor localisation is still an unsolved issue.

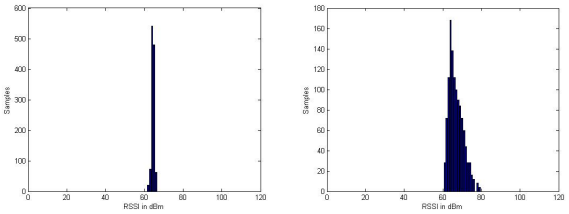
Although GPS has become synonymous with user localisation, indoors its signals are weak or non-existent. Using WLAN as a solution has given promising results, but its performance is subject to change due to multipath propagation and changes in the environment, such as number of persons present in a given location (see fig.1(a) which illustrates the effect of the presence of humans on WLAN received signal strength (RSS) data), variable orientation, temporary changes to building layout, etc) [2]. Performance also depends on the material the building is made from, size of spaces where measurements take place, antenna orientation, directionality, etc. [2]. Whilst image-based localisation techniques have provided some promising results, as in [3, 4], the limitations continue to be due to occlusion, changes in lighting, noise and blur [5]. In this paper, we propose an approach that combines image and WLAN data to leverage the best of both of these complementary modalities. We present the results for locating a user to within a specific office in a building or to a location within that office.

Most existing localisation methods are based on a single modality. In fact, to our best knowledge, even in other application domains there are only a few techniques based on fusion of RF and image sensing methods. In [6] the authors consider an active tracking system, consisting of a camera mounted on pan-tilt unit and a 360° RFID detection system to efficiently track humans in crowds. First they applied a particle filtering method to fuse heterogeneous data and then controlled the system motion to follow the person of interest. Other work describes an object tracking scheme using a different particle filtering model [7]. It consists of a camera observation method based on color features of the target and a WLAN-based localisation system.

In our previous work [8], we showed a proof of concept of how RSSs and image matching data could be fused to do coarse localisation for a small number of locations. Here we introduce a more precise WLAN-based algorithm, vocabulary tree concept for image-based localisation and a novel more complex and effective function in the fusion process.

*Area chair: Alexander Hauptmann
[1]http://www.viconrevue.com
[2]http://www.campaignr.com

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.



(a) Effect of users' presence on WLAN RSSI histogram in an office space: no users present (left), users present and moving (right)

(b) Sample images collected of office spaces

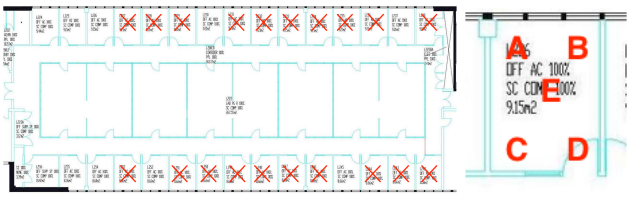
Figure 1: Examples of WLAN-based and visual sensing.

The approach is verified on a much larger and more challenging dataset. Our paper is structured as follows: Section 2 describes the experimental setup including data capture while section 3 discusses WLAN and image-based algorithms respectively. Section 4 introduces a novel fusion function while section 5 presents results for each modality individually and demonstrates that fusion outperforms any one modality leading to very accurate results overall.

2. EXPERIMENTAL SETUP

For our experimental test bed we use 20 offices on the second floor of a building (see fig. 2), where the average size of an office is $8.9m^2$. Within each office we use 5 calibrations points (CP), A, B, C, D & E. Each orientation of a CP (North, South, West and East) is represented with 8 (640×480 pixels) images taken with a camera (see fig. 1(b) for examples), and 300 RSSI (received signal strength indication) observations taken with a laptop. Every CP is represented using data from all four orientations together. In total we had $5,000$ images, of which $3,200$ were used for training and $1,800$ for testing, and $125,000$ signal strengths observations of which $120,000$ were used for training and $5,000$ for testing. One observation consists of received signal strengths from all active access points (in the best case the total number of access points: 14 in our case). Offices are chosen to be next to each other and moreover, look very similar inside, thus resulting in very challenging data for both WLAN and image-based localisation methods.

In a test we used one image and one signal strength observation per CP and tested how precisely we could localise to a given CP. Clearly, if we can localise to a CP, we can localise to within the office that contains that CP. We present results for localizing to a given office whereby the office selected is based on the 1^{st} ranked results corresponding to one of the CPs for that office, even if the top ranked CP is incorrect. However, we wanted to understand how many



(a) (b)

Figure 2: (a) Map of office locations – red crosses indicate offices used; (b) Calibration points $ABCDE$ within an office

(and which) CPs are necessary as this has an impact on the manual data collection effort required to perform accurate localisation. We examined localisation precision for 5 different combinations of 1, 2 and 3 CPs per office (giving 5 different sets of 20, 40 and 60 locations respectively in total). Precision (P) and average precision (AVP) were used as performance measures.

3. LOCALISATION METHODS

3.1 WLAN-based localisation

Probabilistic WLAN-based localisation techniques presume a priori knowledge of the probability distribution of the user's location [9, 2]. We decided to employ a Naive Bayes method [9] which takes into account the access points' (APs) signal strength values (RSSI) and also the frequency of the appearance of these APs. A signature for each CP is defined as a set of W distributions of signal strengths of W APs and a distribution representing the number of appearances of W APs received at this CP. We denote by $C \in \{1, 2, ..., K\}$ the CP random variable where K is the number of CPs, $X_m \in \{1, 2, ..., W\}$ represents the m^{th} AP random variable, $Y_m \in \{1, ..., V\}$ is the signal strength that corresponds to m^{th} AP where W is number of APs, M is number of APs of an observation and V is number of discrete values of signal strength. From a set of N training observations $D = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_n\}$ where $\mathbf{o}_n = (c^{(n)}, x_1^{(n)}, y_1^{(n)}, ..., x_M^{(n)}, y_M^{(n)})$, $n = 1, .., N$ we can then estimate the signature parameters.

The joint distribution $P(C, X_1, Y_1, ..., X_M, Y_M)$ is given by

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

where the distribution $P(C = c)$ could be assigned as uniform (without any loss of confidence).

If the identity function is $\mathbb{I}(a, b) = 1$ if $a = b$ else $= 0$, the sufficient statistics are:

$$n_c = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}(c^{(n)}, c) \tag{2}$$

$$n_c^{(x)} = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}(c^{(n)}, c)\mathbb{I}(x_m^{(n)}, x) \tag{3}$$

$$n_{c,x}^{(y)} = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{I}(c^{(n)}, c)\mathbb{I}(x_m^{(n)}, x)\mathbb{I}(y_m^{(n)}, y) \quad (4)$$

The parameters of $P(X_m|C)$ are estimated as

$$\hat{\pi}_c^{(x)} = \frac{n_c^{(x)} + 1}{n_c + W} \quad (5)$$

and the parameters of $P(Y_m|C, X_m)$ as

$$\hat{\gamma}_{c,x}^{(y)} = \frac{n_{c,x}^{(y)} + 1}{n_c^{(x)} + V} \quad (6)$$

Eventually at the prediction step we have:

$$l_i = P(c) \prod_{m=1}^{M} P(x_m|c)P(y_m|c, x_m) \quad (7)$$

The algorithm chooses the location which maximises l_i as being the user location. We rescaled these probabilities to sum to one and denoted their new values as the CP confidences, p_i.

3.2 Image-based localisation

Some efficient image retrieval and image-based localisation methods have been proposed so far [3]. Here we used a variation of the method shown in [10] to efficiently match query images that belong to a specific CP to the image training dataset of all CPs. Instead of SIFT, 64-dimensional SURF descriptors are used [4]. In our approach we make use of the sign of Laplacian (the trace of the Hessian matrix) which was already computed during the detection stage. Using the sign of the Laplacian we only compare those features that have the same type of contrast (dark on light background or light on dark background). All SURF features were extracted from all $3,200$ images in the database, giving $794,146$ feature descriptors. Each feature was associated with the image and the CP to which it belonged. To compare them we used the Euclidean distance. The features were split into two groups based on the sign of the Laplacian which halves our search time. For each group, we created a hierarchical tree clustering the descriptors using the K-means algorithm repeatedly. In the case of a query image, a match for each of its descriptor was found using the $+1$ or -1 hierarchical tree. Since each was labelled with the image and CP from which it was extracted, it therefore casts one vote for that CP. After each descriptor had voted for a CP, we then had a ranked list of CPs, from the most to the least likely. Similarly, we assigned a confidence for each CP (q_i) as the ratio of the number of votes associated with that CP and the total number of votes.

4. DATA FUSION

To perform fusion, we take confidences p_i and q_i from both sensing modalities P and Q into account, where in our case P and Q were WLAN and image sensing methods respectively. Here, i refers to a given CP. If we sort these confidences we can denote the first ranked, the second ranked, the third ranked, etc. (sorted) confidence by p_{max1}, p_{max2}, p_{max3}, etc. respectively (or by q_{max1},q_{max2}, etc. for the Q modality). We decided to use a large training dataset of confidences of different CPs. This would help in building a robust fusion function which would be reliably used on (unknown) testing data. First, let us define $P_{ij} = p_{maxi} - p_{maxj}$ and similarly $Q_{ij} = q_{maxi} - q_{maxj}$. Observing P_{12} and Q_{12} in many training confidence pairs we concluded that for values P_{12} and/or Q_{12} beyond some reliably large threshold, we were sure that the correct CP (location) was the 1^{st} ranked one, based either on P or Q (or both). We established the thresholds T_1 and T_2 for P and Q and moreover we deduced that introducing multiplication and/or addition functions under some conditions can improve precision even more (or at least average precision). It is important to note that even if the 1^{st} ranked confidence belongs to the correct location the algorithm would discard it if P_{12} (or Q_{12} or both) is below this(these) threshold(s). Also we found that the ranking of the correct location did not fall below some positions in both sets of rankings. In general, these are the m^{th} position for P and the n^{th} position for Q modality. The fusion function is thus as follows (eq. 8), where f_i represents fusion confidence and k_i confidence of the method to which $min(n, m)$ corresponds. The location output by the algorithm is the one with the maximum value of the fusion confidence:

$$f_i = \begin{cases} p_i, & P_{12} \geq Q_{12} \wedge P_{12} \geq T_1 \wedge Q_{12} \geq T_2 \\ q_i, & Q_{12} \geq P_{12} \wedge P_{12} \geq T_1 \wedge Q_{12} \geq T_2 \\ p_i, & P_{12} \geq T_1 \wedge Q_{12} < T_2 \\ q_i, & Q_{12} \geq T_2 \wedge P_{12} < T_1 \\ p_i q_i, & T_3 \leq P_{12} \leq T_4 \wedge T_5 \leq Q_{24} \leq T_6 \\ p_i + q_i, & T_7 \leq P_{12} \leq T_8 \wedge T_9 \leq Q_{24} \leq T_{10} \\ k_i, & \text{else} \end{cases} \quad (8)$$

5. RESULTS

An example of the benefits of fusion, when 2 CPs are observed as individual locations (BE), is shown in figure 3. It shows the behaviour of precision considering the top N ranked results, thus illustrating how often each modality returned the correct location as the top ranked result, 2^{nd} ranked results, and so on (bars in the graph) and also how precision increases if the top N ranked results are considered (lines in the graph). In the top N, for $N = 1 \ldots 5$, the fusion approach outperforms both WLAN and image-based methods reaching precision of 91.82%. Also it can be seen that correct location rank doesn't drop below 8^{th} for WLAN, and 12^{th} for the image-based method. In this example we have $AVP_W = 75.18\%$, $AVP_I = 68.14\%$ and $AVP_F = 80.94\%$ for the WLAN, image-based and the fusion approach respectively.

The left hand side of table 1 shows results when using 1, 2 and 3 CPs per office (every CP represents a different location), using WLAN data only (P_W), image data only (P_I) and the fusion of both modalities (P_F). For 2 and 3 CPs we show a selection of results, corresponding to the best performing ones, rather than all possible combinations. The right hand side of the table shows results when we take into account the 1^{st} ranked result that is not the correct one but that belongs to a CP within that particular office. This gives the precision to a particular office, denoted by P_{WO}, P_{IO} and P_{FO} obtained using WLAN-based, image-based and the fusion method respectively. From the table it is clear that fusion of WLAN and images significantly improves the performance of using either approach on its own. Moreover, on average, P_W, P_I and P_F decrease while P_{WO}, P_{IO} and P_{FO}

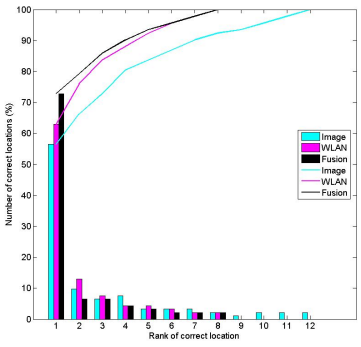


Figure 3: Number of correct locations (in %) found on the N^{th} rank (bars); Number of correct locations (in %) found in the top N ranks (lines)

CP	P_W	P_I	P_F	P_{WO}	P_{IO}	P_{FO}
A	65.22	50.00	69.57	65.22	50.00	69.57
E	69.57	60.87	73.91	69.57	60.87	73.91
C	69.57	56.52	78.26	69.57	56.52	78.26
B	73.91	58.70	82.61	73.91	58.70	82.61
D	71.74	63.04	76.09	71.74	63.04	76.09
AB	61.96	47.83	69.57	65.22	60.87	71.74
BE	63.04	56.52	72.83	71.74	70.65	76.09
ED	66.30	54.35	73.91	73.91	60.87	78.26
AC	64.13	46.74	71.74	75.00	53.26	78.26
BC	68.48	53.26	73.91	77.17	58.70	80.43
ABE	55.07	46.38	63.77	69.57	62.32	75.36
AEC	58.70	45.65	66.67	72.46	65.22	79.71
EBD	58.70	49.28	70.29	76.81	63.77	84.06
ABD	61.59	47.83	69.57	79.71	69.57	83.33
ACD	63.04	44.20	71.74	81.16	65.94	84.78

Table 1: Localisation results: P_W, P_I, P_F are precision results for considering each CP as a separate location; P_{WO}, P_{IO}, P_{FO} are precision results for localising to a specific office

increase when the number of CPs per office increases. This is expected since the data within an office are very similar, thus making the algorithms choose the nearby CP instead of the correct one. For images we have even more complex situation as locations that are not physically close by look similar as well. When we consider localisation to an office many incorrect 1^{st} guesses become correct especially when the number of CPs in an office increases. Thus, in the case of 3 CPs, one can notice a large increase in precision, where on average it increased by 15.52%, 18.72% and 13.04% for WLAN-based, image-based and the fusion method respectively.

The performance variation for the localisation to within an office obtained by using a variable number of CPs also gives an interesting conclusion. Whilst the best results are naturally always obtained by using all 5 CPs for each office, we can see that using only one CP produces reasonably good performance: 82.61% precision for the best result (calibr. point B), 76.09% on average. This is important as it means that the manual data collection stage for model creation outlined in section 2 is viable as it only needs to be performed once (i.e. at one CP) per office in order to obtain reasonably accurate performance.

6. CONCLUSION

In this work, we presented results combining two complementary data sources for indoor localisation. By fusing WLAN signal strengths and image data, we achieve better performance than any individual modality. Future work will investigate the possibility of seamlessly tracking a user indoors, using dynamic and adaptive confidence-based weighting between these modalities, and more sophisticated classifiers such as neural networks.

7. ACKNOWLEDGMENTS

This work is supported by Science Foundation Ireland under grant 07/CE/I1147.

8. REFERENCES

[1] S. Karaman, J. Benois-Pineau, R. Meandgret, V. Dovgalecs, J.-F. Dartigues, and Y. Gaeandstel. Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *ICPR*, pages 4113–4116, 2010.

[2] A.S. Paul and E.A. Wan. RSSI-based Indoor Localization and Tracking Using Sigma-Point Kalman Smoothers. *IEEE Journal of Selected Topics in Signal Processing*, 3(5):860–873, 2009.

[3] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization by combining an image-retrieval system with Monte Carlo localization. *IEEE Transactions on Robotics*, 21(2):208 – 216, 2005.

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

[5] M. Redzic, C. O'Conaire, C. Brennan, and N. O'Connor. Multimodal Identification of Journeys. In *International Conference on Database and Expert Systems Applications (DEXA)*, pages 283–287, 2010.

[6] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat. Vision and RFID data fusion for tracking people in crowds by a mobile robot. *CVIU*, 114(6):641–651, June 2010.

[7] T. Miyaki, T. Yamasaki, and K. Aizawa. Tracking persons using particle filter fusing visual and Wi-Fi localizations for widely distributed camera. In *ICIP*, volume 3, pages 225–228, 2007.

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

[9] K. Tran, D. Phung, B. Adams, and S. Venkatesh. Indoor Location Prediction Using Multiple Wireless Received Signal Strengths. In *Proceedings of Australasian Data Mining Conference (AusDM)*, 2008.

[10] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.