User Localisation using Visual Sensing and RF Signal Strength

Ciarán Ó Conaire[1], Keith Fogarty[2], Conor Brennan[2], Noel E. O'Connor[1]
[1]CLARITY: Centre for Sensor Technologies, [2]RF Modelling and Simulation Group
Dublin City University, Ireland
oconaire@eeng.dcu.ie

Abstract

In this paper, we simulate a museum scenario where we are concerned with identifying the exhibit that a person is observing. Many different technologies exist for determining the location of a user, including GPS, GSM and RF-id tags. Due to their low-power and passive nature, along with their relatively good accuracy, we examine the use of image-based localisation, alongside RF-based localisation for this task. In image-based localisation, we investigate if it is possible to determine a user's location given an image captured at their current location. In RF-based localisation, we can attempt localisation using a set of signal strength readings from detected wireless networks. As the two sources of data are complementary, we investigate different fusion strategies and measure the resulting increase in performance from using both systems.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*sensor fusion, tracking*

Keywords

Localisation, SURF, RF, wireless network, signal strength, image matching.

1 Introduction

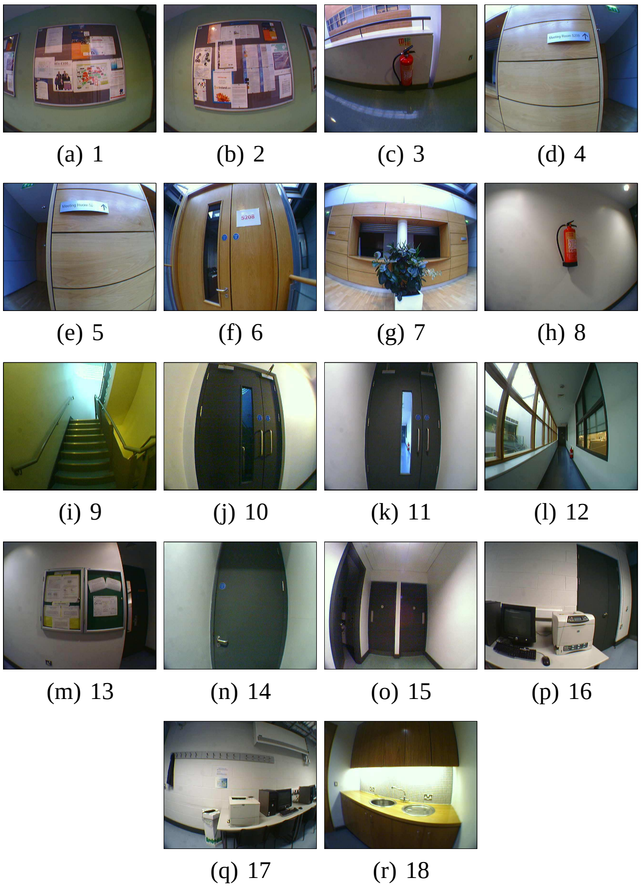
In this paper we report on our initial work on building combined image-based and RF-based localisation systems. The work is motivated by the goal of providing low-cost localisation technology for our preferred end-user application scenario, corresponding to a location aware museum guide. To this end, we construct a simulated museum with multiple exhibits, and simulate a user device which captures images and wireless signal strength readings. We show how this sensed data can be used to accurately locate the user to one of the hypothesized exhibits.

In section 2 we motivate our work by outlining our application scenario. Then in 3, we describe our setup and how the data was captured. Section 4 describes the two localisation algorithms and how they evaluate the likelihood of the person being at each exhibit. Evaluation results, shown in section 5, indicate that both methods perform well individually but that significant performance improvement can be achieved by fusing the two methods. Finally, we give some brief conclusions in section 6.

2 Application Scenario

Audio-tours are frequently used in museums and galleries to guide visitors around the exhibition space and provide them with information on specific exhibits. These devices are usually not location aware but require some form of user interaction, such as punching in an exhibit ID number in order to localise the wearer and deliver relevant audio content. In fact, sometimes wearers are required to follow a pre-defined route so that the audio matches the exhibit being viewed. Making the device aware of the wearer's location would allow visitors to wander the museum freely, instead of following a fixed path. The tour experience could be personalised by providing more audio content on the exhibits deemed to be of particular interest, based on the amount of time the users spends at the exhibit, for instance. A location aware guide would also cut down on user interaction, leading to a more immersive museum experience overall. Furthermore, on returning the location-aware guide, the visitor's path through the museum, and time spent at each exhibit, could be analysed providing for a variety of personalised services that could be delivered offline. This could include a personalised web-accessible multimedia summary of the visit that provides extra information on the exhibits of particular interest – similar to the application presented in [1]. Information on how visitors navigate the museum could also be aggregated over time and used by museum planners, helping them determine the exhibits that are most frequently visited and whether exhibits are badly-positioned based on recognising that they are rarely visited.

A key barrier to adoption of location-aware guides is the potentially expensive and intrusive technical infrastructure required by localisation technologies, such as RF identification as used in previous works [2]. GPS provides accurate localisation without any technical infrastructure in the museum but does not work indoors. GSM has shown potential for providing good localisation [3] but does not reliably provide the accuracy required for this application. In this paper we consider achieving localisation via a combination of imaging and RF signal strength that could feasibly be implemented on a low-cost device that could be used as a platform for next generation museum guides. The benefit of the approach is that it requires no infrastructure beyond the wireless network. Imaging may initially seem a strange choice since traditionally photography has been banned in museums. However, in our scenario, the capture device is


(a) 1 (b) 2 (c) 3 (d) 4

(e) 5 (f) 6 (g) 7 (h) 8

(i) 9 (j) 10 (k) 11 (l) 12

(m) 13 (n) 14 (o) 15 (p) 16

(q) 17 (r) 18

Figure 1. Examples of images of the 18 **exhibits used to simulate a museum scenario.**

provided by the museum to a user who returns it before leaving. The end-user would not be able to download or otherwise have access to the image data – it would remain the sole property of the museum, thereby ensuring the protection of museum and lender copyright. Also of course, we do not use potentially damaging flash photography, another traditional concern related to allowing photography in museum spaces.

3 Simulated Museum

3.1 Mock-up

Figure 1 shows example images of the 18 *exhibits* that were used in our museum scenario. All exhibits were located on the 2nd floor of the Engineering Building in DCU. Their exact locations can be seen in figure 2. Exhibits were chosen so as to closely resemble typical museum exhibits, such that: (i) They are distinct from their immediate surroundings, (ii) Some exhibits appear visually very similar to each other (exhibits 10 and 11, and exhibits 4 and 5, for example) and (iii) Some exhibits are physically very close to each other, therefore appearing very similar in *RF-space*. Additionally, we assigned exhibit 12 to be an *exhibit space*, as exhibits are not always so cleanly localised. Our test-bed is therefore challenging for both image-based and RF-based localisation, and simulates the difficulties in a real-world museum setting.

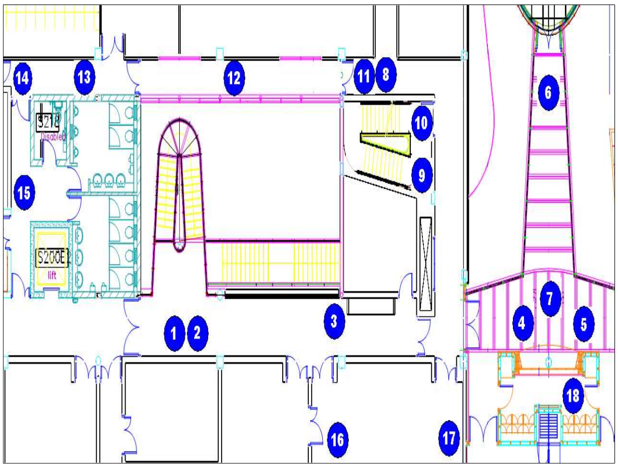


Figure 2. Map of exhibit placement.

3.2 Data capture

We captured image data using a SenseCam [4], which is a wearable camera with a wide-angle lens. To create an exhibit image database, 153 images were captured of the 18 exhibits, giving an average of 8 or 9 database images per exhibit. Wireless network signal strength readings were captured using a laptop running the NetStumbler program [1]. For each reading, the network BSSID and signal strength were stored. For the testing set, data was gathered in a similar way and the two sources were aligned by synchronising both clocks before capturing a dataset. Our datasets contain a total of 3132 images and approximately 230,000 wireless signal strength samples. To realistically simulate our museum scenario, we included images where the exhibit was fully or partially occluded.

4 Localisation Algorithms

4.1 Image-based localisation

Kosecka and Yang [5] perform image-based localisation using SIFT features [6] which are known to be robust to illumination and viewpoint changes. Their approach is to create an image database of known locations and perform a search over the entire database to find a match for a query image. Similar to SIFT and other interest point descriptors (many of which are evaluated in [7]), the recently proposed SURF method [8] locates interest points and extracts an invariant descriptor for each point. However, SURF achieves greater computational efficiency by using integral images. Examples of SURF matches are shown in figure 3. In order to efficiently locate relevant images in a large database, Nistér and Stewnius [9] propose the use of a *vocabulary tree* of SIFT descriptors.

We use a variation on the Nistér method to efficiently match query images to our image database[10]. All SURF features are extracted from all 153 images in the database, giving approximately 170,000 feature descriptors. Each feature is associated with the image it came from. A SURF descriptor is a vector of size 64 and these descriptors can be compared using their Euclidian distance. The features

[1] http://www.netstumbler.com/

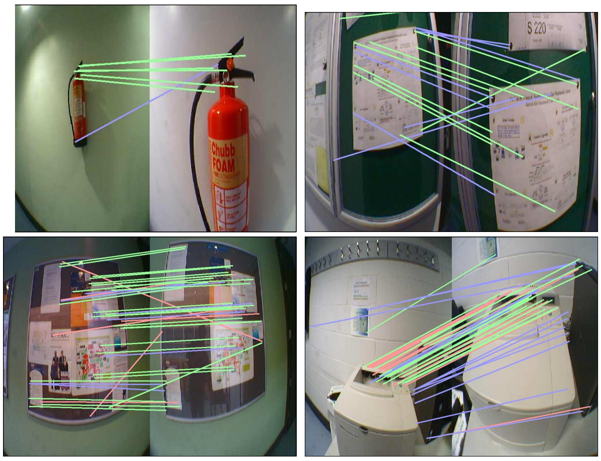


Figure 3. Examples of SURF point matches between different images of the same exhibit. Red lines are matches from the left image to the right image, and blue lines are matches from the right image to the left image. Green lines are bi-directional matches.

were split into two groups based on the sign of the Laplacian of the descriptor, allowing us to halve our search time. For each group, a hierarchical tree is created by clustering the descriptors using the K-means algorithm recursively. We used $K = 16$ which initially created 16 clusters, then within each cluster, 16 more clusters, and so on. This tree representation allows the features to be matched efficiently.

Given a query image, its SURF descriptors are extracted. A match for each descriptor is found using the hierarchical tree, and since this match is labelled with the database image from which it was extracted, it therefore casts one vote for the exhibit that this image belongs to. After each descriptor has voted for an exhibit, we then have a ranked list of exhibits, from the most likely exhibit to the least likely. It is possible to then have a *verification stage* using bi-directional matching [10] where the exhibit images matches are confirmed. This was found to increase performance, but was not used in this work, as it adds computational complexity.

4.2 RF-based localisation

In [11], Bahl and Padmanabhan present an RF-based system for location estimation that uses signal-strength information from wireless network base-stations at known locations. Using 3 base-stations they take signal strength samples at multiple locations and orientations, and use a nearest-neighbour approach to localise the user. Our approach is different in that all wireless base-stations are not always detected. Also, we are trying to identify exhibits, rather than exact locations.

In our system, the signal strength of a particular wireless network at each exhibit is modelled as a probability distribution. This model is created by gathering signal samples and computing a histogram of the signal strengths. We used a fine-scale histogram with 200 bins representing signal strengths of $-200dBm$ to $-1dBm$. To improve the models robustness to noise, we smoothed the histograms with a Gaussian filter of size 11 and $\sigma = 1.5$. All histograms are then converted to probability distributions by normalising them to sum to one. In total, 9 different wireless networks were detected during the RF model building stage. Each exhibit is therefore modelled as 9 histograms, one per network. If a network was never detected at an exhibit, the histogram is set to a uniform distribution. Histograms for exhibit 1 are shown in figure 4.

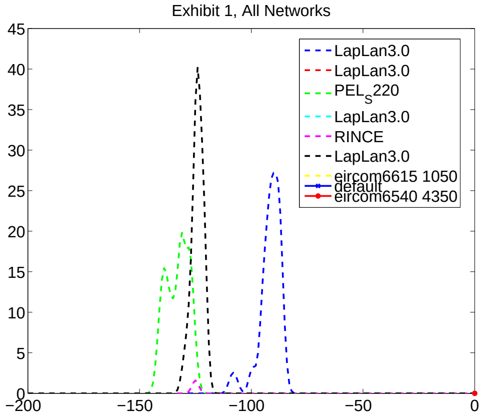


Figure 4. Signal strength histograms for exhibit 1. Signal strength in dBm on the x-axis. Non-unique network names (SSIDs) shown. Each network has a unique BSSID (MAC address), not shown.

To determine the current exhibit, we gather signal strength samples for a window period of 7 seconds. Longer sampling times will result in greater accuracy, but we chose 7 seconds as a compromise between accuracy and the length of time a person would wait at a particular exhibit. Given these samples, smoothed histograms are computed, as in the training phase. For two normalised histograms, h_x and h_y, the histogram intersection is computed as:

$$I(h_x, h_y) = \sum_{i=1}^{200} \min\{h_x(i), h_y(j)\} \tag{1}$$

The similarity between a query sample window and an exhibit is computed as:

$$S(h, e) = \prod_{n=1}^{9} I(h_n, H_n^e) \tag{2}$$

where h_n is the query histogram of network n and H_n^e is histogram of network n from exhibit e. If a network n was not detected in the window, we set $I(h_n, H_n^e) = 1$ for all exhibits.

5 Results

5.1 Individual algorithms

Table 1 shows some interesting results for each approach. If we evaluate the methods based on their *most likely exhibit*, then the image-base method does best, getting it right 74.90% of the time, compared to 59.83% for the RF-based method. When we examine the average distance error, the RF-based method performs better, indicating that when it guesses incorrectly, its guess is usually physically close by.

Dataset	Exhibit Correct		Distance Error	
ID	Image	RF	Image	RF
3	80.13	62.33	2.44	1.74
4	82.82	64.98	2.16	2.61
5	65.71	55.91	4.42	1.51
6	69.44	56.21	4.50	1.61
7	73.10	52.18	3.37	3.62
8	60.00	54.98	5.83	1.69
9	72.22	63.29	3.62	2.56
10	78.95	54.71	2.57	1.87
11	76.43	57.21	3.00	1.74
12	82.58	69.11	1.98	1.53
13	82.53	67.23	1.76	1.51
MEAN	74.90%	59.83%	3.24m	2.00m

Table 1. Percentage correct guesses on identifying the current exhibit and the distance error (in metres) for both methods. Datasets 1 **and** 2 **were used to learn the** *RF* **exhibit model.**

We further illustrate the performance of both systems in figure 5, where the position of the correct exhibit in the rank list is examined. Examining just the top ranked guess, the image-based method does best. However, if we take the top-N guesses (with $N > 1$), then the RF-based method outperforms the image-based method. It seems that when the image-based method does not get the correct exhibit as its best guess, the information it provides is not very useful. On the other hand, while the RF method does not always select the correct exhibit, its guess will usually be physically very close to the correct exhibit, and it will have the correct exhibit ranked highly. We should bear in mind, however, that the image database was created 6 months before the test datasets were collected, whereas the *RF* models were created immediately before the datasets were collected. Also, the verification stage mentioned in section 4.1 was not used, which would be expected to improve image-based performance.

5.2 Fusion

To evaluate fusion performance, we use *average precision* as performance measure. Using the error distance or the percentage of correct guesses only evaluates the top guess. Average precision takes the entire exhibit ranking into account and is therefore more informative and will potentially be a better judge of how the systems will perform in tracking, when extra contextual information will be available.

If we have R images (or signal strength sample windows) to evaluate, then average precision is computed as:

$$AP = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{P_i} \tag{3}$$

where P_i is the position of the correct match in the i^{th} test. A perfect score of $AP = 1$ is achieved only when $P_i = 1$ for $i \in \{1,2,\ldots,R\}$.

As a baseline fusion strategy, we allow each method to assign a confidence to each exhibit, and then we add the confidences of both methods. For the image-based method, the confidence for each exhibit is computed by dividing the number of votes it received by the total number of votes. For the RF-based method, the confidence for an exhibit is computed by dividing its similarity score $S(h,e)$ by the sum of the similarity scores for all exhibits, $\sum_e S(h,e)$. The results shown in table 2 show that this fusion strategy outperforms the best of either method. On average, it improves the average precision by 0.054.

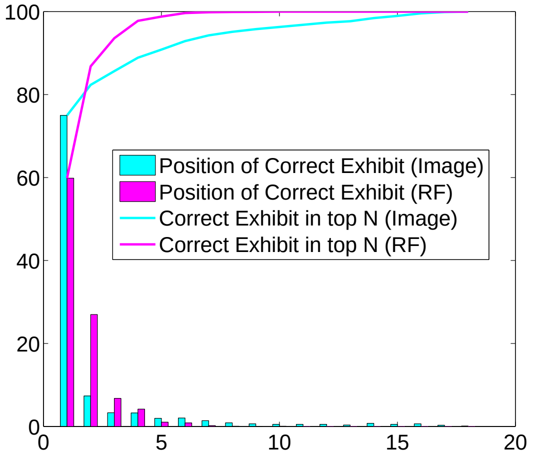


Figure 5. Position of correct exhibit in ranked list provided localisation methods. While the image-based method provides a more accurate "best guess", the RF-based method more often has the correct exhibit in its top matches. In the top 4 **guesses, the RF method has the correct exhibit** 97.76% **of the time, whereas the image-based method manages** 88.88%. **The average precision values for the methods are** 0.820 **(Image) and** 0.770 **(RF).**

5.3 Tracking

Instead of simply examining a single image or window of RF samples, the images (or windows) before and after give contextual information that can be useful in improving the classification accuracy. For example, with reference to figure 2, it is almost impossible for a user to be at exhibit 14 and then to be at exhibit 18 within a few seconds. We represent all possible paths for the user as a directed graph, where the nodes are labelled (e,i), where e is the exhibit ID and i is the image number. Each node (x,i) is connected to node $(y,i+1)$, for $x,y \in \{1,2,\ldots,18\}$. A likelihood, $L(e,i)$, is associated with each node depending on its image/RF data. A transition likelihood, $C(a,b)$, is also associated with each directed edge, representing the likelihood of travelling from exhibit a to exhibit b in a given time.

For single-modality tracking, we set $L(e,i)$ equal to the confidence of exhibit e (see section 5.2). To perform fusion, numerous strategies were investigated, including simply summing confidences, as in section 5.2. The best performing fusion approach was to set $L(e,i)$ equal to the confidence from the image data if the RF confidence for that exhibit was above a threshold T, and to set it to zero otherwise. T was learned from training data. Essentially, this fusion method uses the signal strength data as a filter, removing lo-

Dataset	Average Precision			Improvement
ID	Image	RF	Fusion	Over Best
3	0.853	0.762	0.893	0.040
4	0.869	0.802	0.876	0.007
5	0.756	0.802	0.952	0.150
6	0.776	0.764	0.878	0.102
7	0.821	0.723	0.847	0.026
8	0.684	0.732	0.790	0.059
9	0.803	0.789	0.897	0.094
10	0.855	0.749	0.862	0.007
11	0.830	0.742	0.863	0.033
12	0.889	0.836	0.928	0.040
13	0.884	0.809	0.920	0.036
MEAN	0.820	0.774	0.882	0.054

Table 2. Fusion of both methods: On average, the fusion method chooses the correct exhibit 79.21% of the time and has an average distance error of $1.256m$.

Exhibit Correct (%)			Distance Error (m)		
Image	RF	Fused	Image	RF	Fused
89.21	65.16	94.11	0.96	1.22	0.26

Table 3. Tracking results using either data source and the combination of both sources.

cations that are deemed unlikely, since RF readings cannot easily distinguish nearby exhibits.

A connected path through the graph, $\{E_1, E_2, ..., E_N\}$, represents a user's path through the museum and its score is determined by summing the likelihoods for each node and edge in the path, shown in equation 4:

$$L(E_1, 1) + \sum_{j=2}^{N} C(E_{j-1}, E_j) + L(E_j, j) \tag{4}$$

with $C(a,b) = -(D/(t\sigma))^2$, where D is the geodesic (walking) distance between exhibits a and b, t is the time difference between the captured images. A value for σ was learned from training data for each of the three systems (image, rf, fused). Datasets 3-7 were used for training and parameters were chosen to optimise the classification score. We determine the optimum (most likely) path $\{E_1, E_2, ..., E_N\}$ using the Viterbi algorithm. Table 3 shows the results comparing tracking performance for the three systems. Not only does the combination of both sources increase the classification performance by almost 5%, the distance error is dramatically reduced to almost one quarter the error of either single-modality system.

6 Conclusions

In this work, we showed that both image-based and RF-based systems can demonstrate robust performance in localisation, even with challenging data. As complementary sources of data, the fusion of the two approaches improves performance, leading to significant increases in both positional and classification accuracy.

The best fusion strategy we found was to use the RF data as a possibility filter, and remove locations that are deemed unlikely, since by its nature, signal strength readings cannot easily discriminate nearby locations. The visual data is then used to decide between the remaining candidates, or if visual data fails, the user's motion priors will constrain the location estimation. Future work will examine the integration of both systems into one device, such as the N95 phone, which has imaging and wireless capability and also investigate how the proposed methods scales to a larger number of exhibits.

Acknowledgement

This work is supported by Science Foundation Ireland under grant 07/CE/I1147. Part of this work was supported by a fellowship funded by the School of Electronic Engineering, Dublin City University.

7 References

[1] L. Aroyo, Y. Wang, R. Brussee, P. Gorgels, L. Rutledge, and N. Stash, "Personalized museum experience: The rijksmuseum use case," in *In Proceedings of Museums and the Web*, San Francisco, USA, April 2007.

[2] T.Y. Liu, T.H. Tan, and Y.L. Chu, "The ubiquitous museum learning environment: Concept, design, implementation, and a case study," in *Sixth International Conference on Advanced Learning Technologies*, Kerkrade, The Netherlands, July 2006, pp. 989–991.

[3] A. Varshavsky and et al., "Are gsm phones the solution for localization?," in *7th IEEE Workshop on Mobile Computing Systems and Applications*, 2006.

[4] C. Ó Conaire, N. E. O'Connor, A. Smeaton, and G. J. F. Jones, "Organising a daily visual diary using multi-feature clustering," in *Proc. of 19th annual Symposium on Electronic Imaging*, 2007.

[5] J. Kosecka and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," in *17th International Conference on Pattern Recognition*, Aug 2004, vol. 4, pp. 319–322.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *9th European Conference on Computer Vision (ECCV'06)*, May 2006.

[9] D. Nistér and H. Stewnius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2006, pp. 2161–2168.

[10] C. Ó Conaire, M. Blighe, and N. E. O'Connor, "Sensecam image localisation using hierarchical surf trees," in *International MultiMedia Modeling Conference*, 2009.

[11] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proceedings of the IEEE Infocom*, March 2000.