

# Associating Locations Between Indoor Journeys from Wearable Cameras

Jose Rivera-Rubio<sup>(✉)</sup>, Ioannis Alexiou, and Anil A. Bharath

Imperial College London, London, UK  
jose.rivera@imperial.ac.uk

**Abstract.** The main question we address is whether it is possible to crowdsource navigational data in the form of video sequences captured from wearable cameras. Without using geometric inference techniques (such as SLAM), we test video data for its location-discrimination content. Tracking algorithms do not form part of this assessment, because our goal is to compare different visual descriptors for the purpose of location inference in highly ambiguous indoor environments. The testing of these descriptors, and different encoding methods, is performed by measuring the positional error inferred during one journey with respect to other journeys along the same approximate path.

There are three main contributions described in this paper. First, we compare different techniques for visual feature extraction with the aim of associating locations between different journeys along roughly the same physical route. Secondly, we suggest measuring the quality of position inference relative to multiple passes through the same route by introducing a positional estimate of ground truth that is determined with modified surveying instrumentation. Finally, we contribute a database of nearly 100,000 frames with this positional ground-truth. More than 3 km worth of indoor journeys with a hand-held device (Nexus 4) and a wearable device (Google Glass) are included in this dataset.

## 1 Introduction

There is increasing interest in technologies that perform the indoor localisation of a user with respect to his or her surroundings. Many of the applications of such a technology are in commerce, allowing mobile devices, such as smartphones, to be more context-aware. However, there are many assistive contexts in which accurate user localisation could have a strong role to play. These include the ability of a user to request assistance in a public space, allowing him or her to be found, and guidance or assistance directed towards them. A more general and wide-ranging possibility is the use of computer vision to contribute to the guidance of an individual. With the emergence of wearable cameras, the potential contributions of computer vision to the navigational context, particularly for visually-impaired users, is enormous. This work explores a complementary approach to visual localisation than using geometric and Simultaneous Localization and Mapping (SLAM)-based techniques. Location is inferred through answering

visual queries that are submitted against the paths of other users, rather than by explicit map-building or geometric inference. This mimics current hypotheses about at least one component of localisation in mammalian vision, where different localisation mechanisms are thought to co-exist; see, for example, the review article by Hartley and others [7]. We test the ability to localise from visual content – not self-motion – in a new dataset of *visual paths* [18], containing more than 3 km of video sequences in which ground-truth is acquired using modified surveying equipment. The dataset can be used to assess localization accuracy using any number of techniques that involve vision, including SLAM. The results suggest that, even without tracking, good localization of a user, even in ambiguous indoor settings, can be captured. The application to wearable camera technology – whereby image cues are harvested from volunteered journeys, then used to help other users of the same space – is the eventual goal of this work, a natural extension to recently reported approaches based on harvesting environmental signals [28].

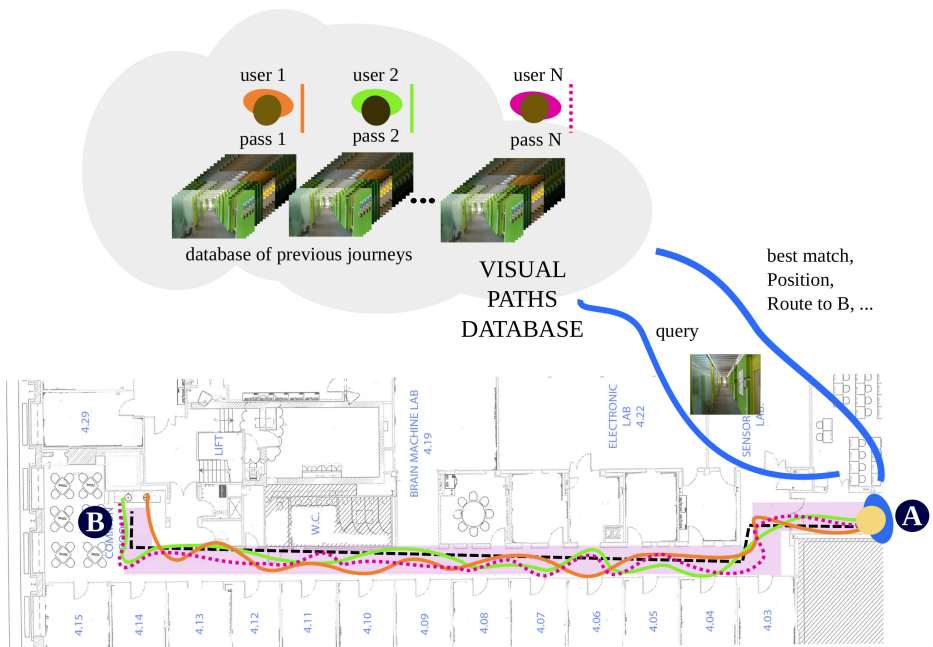
## 2 Related Work

### 2.1 Early Findings in Robotics

Early work by Matsumoto *et al.* [14] suggested the concept of a “view-sequenced route representation” in which a robot performed simple navigation tasks by correlating current views to those held in a form of training database. Similar ideas can be seen on the work by [15], using the difference between frames of detected vertical lines to estimate changes in position and orientation. Their results were constrained to controlled robot movement, and therefore arguably of limited applicability to images obtained from human self-motion. Tang *et al.* also used vertical lines as features [24], but from from omni-directional cameras; their technique relied on estimating positional differences between playback and training sequences to achieve robot navigation. Tang introduced odometers as well, therefore fusing vision with self-motion sensing. This is, in fact, what one might expect a working system to do. However, fusing sensor data makes it difficult to really assess and tune the contribution of individual sensing cues, particularly one as complex as vision, where several visual processing strategies could be applied: optic flow, feature detection and tracking, stereo, etc.

### 2.2 Emerging Methods

The mapping of outdoor navigational routes has progressed rapidly in the past 2 decades, with satellite-based positioning and radio-strength indicators providing high-quality navigation capability over scales of around 10 m or less. In an *indoor* context, localization technology is in its infancy [17, 21, 28]. For indoor localization, there has been remarkable work from Google and crowdsourced sensor information and maps [9]. The potential to use retrieval-based visual localization systems, such as the proposed by the NAVVIS team, are relatively



**Fig. 1.** A sample path (Corridor 1, C1) illustrating the multiple passes through the same space. Each of these passes represents a sequence that is either stored in a database, or represents the queries that are submitted against previous journeys. In the assistive context, the user at point A could be a blind or partially sighted user, and he or she would benefit from solutions to the association problem of a query journey relative to previous “journey experiences” along roughly the same path, crowdsourced by  $N$  users that may be sighted.

computationally intensive, but provide a source of data that is often neglected in human navigation systems. Nevertheless, the NAVVIS team demonstrated that estimating the position of a robot was possible, and provided a dataset acquired from a camera-equipped robot with ground truth [8]. They also expanded early work on visual localization based on SIFT descriptors [16] to one that uses a Bag-of-Features. This is an important step, as it allows scalable operation in larger datasets, or a subset of data to be cached on a smartphone or wearable device for low-latency operation during active navigation [19, 20].

### 2.3 Biological Motivation

Over the past 40 years, research into mammalian vision has uncovered remarkable details about the way in which neurons in the brain respond to the environment of an animal. One of the areas known to be strongly associated with memory is also implicated in localization: the hippocampus. Evidence suggests that there are at least three sources of explicit localization encoding

in hippocampal cells. For example in rodents, cells have been found to display elevated firing rates when the animal is in specific locations within an environment, but the responses fall into different “features” of the location of the animal. Some cells appear to participate in a joint encoding, with individual cells responding to more than one location (grid cells). Other cells appear to use various cues to localise themselves relative to boundaries, as evidenced by firing rates that encode “distance to boundary”. From detailed experiments in insect vision, we know that optical flow is one of the contributing sources of such information, and quite similar mechanisms are found in higher animals [11]. The third type of hippocampal localization cell motivates this work: hippocampal place cells [7]. These cells display elevated firing when an animal is in a specific location, and they can also be found in humans [3]. To be clear, each cell that is characterised as a place cell has the property that it displays significantly elevated firing rates *only* when an animal is in a particular spatial location. The nature of these experiments cannot rule out the possibility that such cells participate in a joint encoding, but the “simplistic” view of place cells is “one-cell, one-location”.

## 2.4 This Work

Our usage context is related to aspects of previous work, but is motivated by the idea that there are significant opportunities to use computer vision in assistive contexts. Whilst often considered power and compute intensive compared to other sources of sensor data, visual data is almost singularly rich in the navigation context. There are only a few examples of its use in assistive technology, where techniques such as ultrasound, intelligent canes and standard localization technologies are dominant. However, due to the emergence of wearable cameras and highly connected devices that can process video data efficiently (e.g. general purpose graphics processors, embedded on phones), the opportunity to harness visual data for navigation is very attractive.

The dominant technique for localization and mapping in computer vision is SLAM. However, we consider that the convergence of crowdsourcing approaches to “map out” physical spaces is not supported by this technique. In other words, the approaches we can use with crowdsensing of *signal* data to learn navigational routes has not been applied to *visual* data. Of course, in using visual information, one would certainly seek to support it with other forms of sensor such as Received Signal Strength Indication (RSSI) data, magnetometers, and tracking algorithms [17, 19, 20]. However, in *assessing* and evaluating its performance, it is hard to isolate factors that affect the quality of visual information when it is included as part of a sensor fusion approach. Thus, we focus in this work on purely visual methods, with the purpose of teasing out aspects of the algorithms that represent, in a location-specific way, the location of a person with a camera.

The first step in doing this is, therefore, to a) collect data that allows us to determine how plausible it is to infer the location of one user relative to others that have made the same journey using visual data *alone*; b) apply matching techniques between data sets, treating some video data as a “journey” database, and other

data as one or more queries. The general principle of the data acquisition takes the form of experiments in which ground-truth is measured using modifications to fairly standard surveying equipment. We now describe this more fully.

### 3 The Dataset

In order to allow different approaches to be compared, and as a community resource to develop this technique, the *RSM dataset* is made publicly available at <http://rsm.bicv.org> [18].

#### 3.1 Existing Datasets

Datasets for evaluating visual localization methods have historically been tied to specific publications and their function was often limited to demonstrate the performance of particular metrics. This has led to a number of datasets that were difficult to adapt to new work, or simply impossible to use because they were never released to the community.

*Historical Datasets.* Early work described in Section 2.1 used custom-planned datasets for their specific evaluation objectives. This led to datasets [14, 15, 24] containing very short sequences, of few meters of length, that could not be used to assess localization performance at human scale.

*SLAM datasets and the NAVVIS Dataset.* SLAM datasets, found in the robotics community, have a variety of scopes and recorded distances: large indoor spaces [23], outdoor itineraries [1], and up to the scale of a few km car ride [22]. They are also heterogeneous in terms of the precision and nature of the ground truth: some use GPS, others the Microsoft Kinect to capture depth [23], while others use the Vicon motion capture system. While the ground truth is often precise (up to the level of GPS, Kinect or Vicon precision), these have usually targeted outdoor comparisons; indoor comparisons focused at geometric reconstruction or pose estimation rather than localisation.

To the best of our knowledge, with the exception of NAVVIS, SLAM datasets have had rather restricted distances, not addressing real-world navigation on the scale of buildings. The NAVVIS project described in Section 2.2 first introduced a more generalistic dataset that could evaluate visual localization and navigation at human scale for robotic applications. Our proposed dataset takes the evaluation and the principle closer to the assistive context than the robot-centric approach of the NAVVIS team: our data and evaluation context introduces the particularities of human motion, both from hand-held and a wearable camera.

#### 3.2 Visual Paths

We define a “visual path” as the video sequence captured by a moving person in executing a journey along a particular physical path. For the construction of

our dataset, the *RSM dataset of visual paths*, a total of 60 videos were acquired from 6 corridors of a large building. In total, 3.05 km of data is contained in this dataset at a natural indoor walking speed. For each corridor, ten passes (i.e. 10 separate visual paths) are obtained; five of these are acquired with two different devices with 30 videos each. One device was a LG Google Nexus 4 phone running Android 4.4.2. The video data was acquired at approximately 24-30 fps at two different resolutions,  $1280 \times 720$  and  $1920 \times 1080$  pixels. The second device was a Google Glass (Explorer edition) acquiring at a resolution of  $1280 \times 720$ , and at a frame rate of 30 fps. Table 1 summarizes the acquisition. As can be seen, the length of the sequences varies within some corridors, due to a combination of different walking speeds and/or different frame rates. Lighting also varied, due to a combination of daylight/night-time acquisitions, and occasional prominent windows that represent strong lighting sources in certain parts of some corridors. Changes were also observable in some videos from one pass to another, due to the presence of changes and occasional appearance from people. In total, more than 90,000 frames of video are labelled with positional ground-truth in a path relative manner. The dataset is publicly available for download at <http://rsm.bicv.org> [18].

### 3.3 Ground Truth Acquisition

A surveyor’s wheel (Silverline) with a precision of 10 cm and error of  $\pm 5\%$  was used to record distance, but was modified by wiring its encoder to a Raspberry Pi running a number of measurement processes. The Pi was synchronised to network time enabling synchronisation with timestamps in the video sequence. Because of the variable-frame rate of acquisition, timestamp data from the video was used to align ground-truth measurements with frames. This data was used to access the accuracy of associating positions along journeys through frame indexing and comparison.







## 4 Retrieval Methods for Visual Localisation

We include results from unmodified, widely-used frame and sequence-based descriptor implementations reported in the image and video categorization and retrieval literature. We also implemented our own methods for greater control of parameter tuning and a more consistent comparison of the possible choices of spatial derivatives, temporal derivative/smoothing and spatial pooling. We describe the two classes of methods as “standard” and “projective”; the latter refers to the fact that our implementations are all performed by linear projections onto spatial weighting functions, and are created by a cascade of convolution operations, followed by spatial sub-sampling.

### 4.1 Standard Methods

*Keypoint based SIFT (KP-SIFT)*. The original implementation of Lowe’s SIFT descriptor follows the identification of interesting points, each with assigned

**Table 1.** A summary of the dataset with thumbnails

	Photo	Length (m)			No. of frames		
		Avg	Min	Max	Avg	Min	Max
C1		57.9	57.7	58.7	2157	1860	2338
C2		31.0	30.6	31.5	909	687	1168
C3		52.7	51.4	53.3	1427	1070	1777
C4		49.3	46.4	56.2	1583	1090	2154
C5		54.3	49.3	58.4	1782	1326	1900
C6		55.9	55.4	56.4	1471	1180	1817
Total		3.042 km			90,302 frames		

intrinsic scales and orientations within the image that are likely to be stable, known as the ‘‘SIFT keypoints’’ [13]. This is widely used across many computer vision applications from object recognition to motion detection and SLAM. We used the standard implementation from VLFEAT to compute  $\nabla f(x, y; \sigma)$  where  $f(x, y; \sigma)$  represents the scale-space embedding of image  $f(x, y)$  within a Gaussian scale-space at scale  $\sigma$ . We also filtered out small local maxima in scale-space. The resulting descriptors are sparsely spread through each video frame.

*Dense SIFT (DSIFT).* The Dense-SIFT [12, 25] descriptor is a popular and fast alternative to keypoint based SIFT. This DSIFT descriptor was calculated by dense sampling of the smoothed estimate of  $\nabla f(x, y; \sigma)$ . We used dense SIFT from VLFEAT toolbox using  $\sigma = 1.2$ , with a stride length of 3 pixels. This process yielded around 2,000 descriptors per frame, each describing a patch of roughly  $10 \times 10$  pixels in the frame. Spatial scale is fixed with this approach, though the descriptor structure is otherwise the same as for the sparse keypoints.

*HOG3D.* The **HOG 3D** descriptor (HOG3D) [10] was introduced with the aim of extending the very successful two-dimensional histogram of oriented gradients technique [5], to space-time fields, in the form of video sequences. HOG 3D seeks computational efficiencies by smoothing using box filters, rather than Gaussian spatial or space-time kernels. This allows three-dimensional gradient estimation across multiple scales using *integral video* representations, a direct extension of the integral image idea [27]. The gradients from this operation are usually performed across multiple scales. We used the dense HOG 3D option from the implementation of the authors, and the settings yielded approximately 2,000, 192-dimensional descriptors per frame of video.

## 4.2 Projective Descriptors

This grouping of descriptors is based on distinct implementations of spatial and/or temporal filtering. In this sense, there are exact or minor variations on

the gradient-based methods considered in the previous section. However, what is common to all of the methods below is that the initial filtering is converted into descriptors using projections against spatial weighting functions, one for each descriptor element. This approach is similar to a soft-histogram approach, but allows greater flexibility in tuning the bin weightings.

*Single Frame Gabor descriptors (SF\_GABOR).* This is an odd-symmetric Gabor-based descriptor. For this, we used 8-directional spatial Gabor filters previously tuned on PASCAL VOC data [6] in order to encode the image gradient field. Each filtering operator produces a filtered image plane, denoted  $\mathbf{G}_{k,\sigma}$ . Spatial pooling of these image planes was performed by the spatial convolution  $\mathbf{G}_{k,\sigma} * \Phi_{m,n}$ .  $\Phi_{m,n}$  represent *spatial pooling functions* that are generated by spatial sampling of the function:

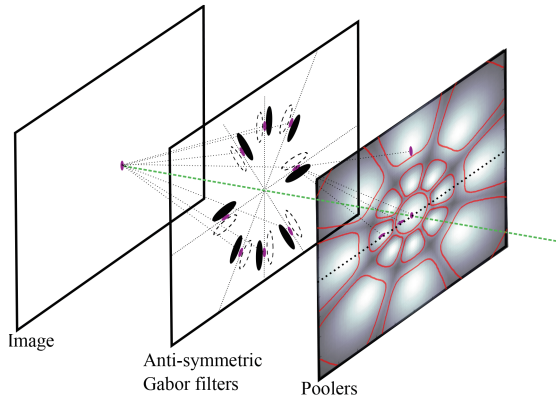
$$\Phi(x, y; m, n) = e^{-\alpha \left[ \log_e \left( \frac{x^2 + y^2}{d_n^2} \right) \right]^2 - \beta |\theta - \theta_m|} \quad (1)$$

We used  $\alpha = 4$  and  $\beta = 0.4$  in our implementation. The values of  $m$  and  $n$  were selected to “collect” filtered image data over 8 angular regions and with the weighting roughly peaking around distances  $d_1 = 0.45$  and  $d_2 = 0.6$  away from the centre of each pooling region, for a total of 17 pooling regions across each of the eight filtering channels. In the ( $m = 0$ ) central region, there is no angular variation. The resulting fields – one field for each pooling region for each directional channel – are sub-sampled to produce dense 136-dimensional descriptors, each representing a  $10 \times 10$  image region, yielding approximately 2,000 descriptors per image frame when the result of the convolution is sub-sampled. The pooling regions are illustrated in Fig. 2.

*Space-time Gabor (ST\_GABOR)* Functions have been used in activity recognition, structure from motion and other applications [2]. We performed convolution between the video sequence and three one-dimensional Gabor functions along each spatial dimension i.e.  $x$  or  $y$ , or along  $t$ . The one-dimensional convolution is crude, but appropriate if the videos have been spatially smoothed. The spatial extent of the Gabor was set to provide one cycle of weight oscillation over roughly a 5 pixel distance, both for the  $x$  and  $y$  spatial dimensions. The filter for the temporal dimension used a wavelength of around 9 frames. We also explored symmetric Gabor functions, but found them less favourable.

After performing three separate filtering operations, each pixel of each frame is assigned a triplet of values corresponding to the result of each filtering operation. The three values are treated as being components of a 3D vector. Over a spatial extent of around  $16 \times 16$  pixels taken at the central frame of the 9-frame support region, these vectors contribute weighted votes into descriptor bins according to their azimuth and elevations, with the weighting being given by the length of the vector. This is similar, but not identical, to the initial stages of the HOG3D filter. Pooling is then performed using the spatial lobe pattern illustrated in Fig. 2. Each frame had approximately 2,000, 221 dimensional ST\_GABOR descriptors.





**Fig. 2.** This illustrates the nature of the spatial pooling used in the projective descriptors. The regions are produced from Eq. 1, generating non-negative spatial filters that collect (pool) filtered data over a  $10 \times 10$  pixel region. Because of the spatial symmetry, the masks can be applied to the Gabor filtered video frame outputs by spatial convolution. These regions were obtained as a result of optimisation of parameters of Eq. 1 using a metric similar to mean absolute precision (mAP).

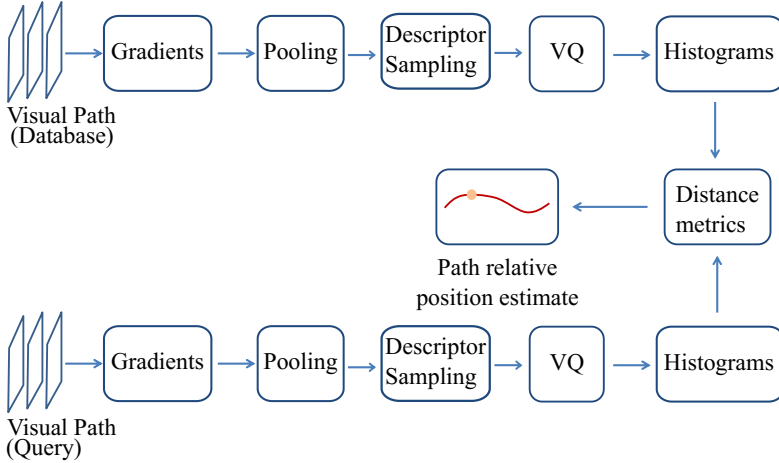
*Space-Time Gaussian.* This descriptor consisted of spatial derivatives in space, combined with smoothing over time (**ST\_GAUSS**). In contrast to the strictly one-dimensional filtering operation used for the **ST\_GABOR** descriptor, we used two  $5 \times 5$  gradient masks for the  $x$  and  $y$  directions based on derivatives of Gaussian functions, and an 11-point Gaussian smoothing filter in the temporal direction with a standard deviation of 2. 8-directional quantization was applied to the angles of the gradient field, and weighted voting with the gradient magnitude was used to populate the bins of a 136-dimensional descriptor. Like the **ST\_GABOR** descriptor, the pooling regions were as shown in Fig. 2. The number of descriptors produced was equivalent to the other methods described for patch-based indexing.

## 5 Evaluation Framework

### 5.1 BOVW Pipeline

In order to test the ability to localise position based on the visual structure of either a short sequence of frames or individual frame information, we adopted a retrieval structure for efficient mapping of the visual descriptors, sparsely or densely populating an image, into a single frame or vignette-level representation. The approach is based on fairly standard retrieval architectures used for image categorization – the Bag-of-Visual Words (BOVW)– and is illustrated in Figure 3.

For the vector quantization, hard assignment (HA) was used to encode each descriptor vector by assignment to a dictionary entry. The data set was partitioned by selecting  $M - 1$  of the  $M$  video sequences of passes through each



**Fig. 3.** Video sequences from wearable and hand-held cameras are processed using a customized BOVW pipeline. Variants of the gradient operators, pooling operators, quantization and distance metrics are described in Section 4.

possible path. This ensured that queries were *never* used to build the vocabulary used for testing the localization accuracy. The dictionary was created by applying the  $k$ -means algorithm on samples from the video database. We fixed the dictionary size to 4,000 (clusters, words); this allows comparison with the work of others in related fields, such as [4].

The resulting dictionaries were then used to encode the descriptors, both those in the database and those from queries. The frequency of occurrence of atoms was used to create a histogram of visual words “centered” around each frame of the video sequence (visual path) in a database, and the same process was used to encode each possible query frame from the remaining path. Histograms were all  $L_2$ -normalized.

## 5.2 Localization Using “kernelized” Histogram Distances

Once histograms had been produced, a kernelized-version in [26] of a distance measure in 4,000-dimensional space was used to compare the similarity of histograms in a query frame with the database entries. A variety of kernel functions exist, such as the popular Hellinger kernel, but we found the  $\chi^2$  best for this problem. For a random subset of the  $M - 1$  videos captured over *each* path in the dictionary, the query is generated from the remaining journey. Each query frame,  $H_q$ , results in  $M - 1$  separate comparison vectors, each containing the distance of each frame to the query. We identified the best matching frame,  $\hat{m}$  from pass  $\hat{p}$  across all of the  $M - 1$  vectors. This is done using:

$$L(\hat{p}, \hat{f}) = \arg \max_{p,f} \{K_D(H_q, H_{p,f})\} \quad (2)$$

$H_{p,f}$  denotes the series of normalized histogram encodings, indexed by  $p$  drawn from the  $M - 1$  database passes, and  $\hat{f}$  denotes the frame number within that pass.  $K_D$  denotes so-called “kernelized” distance measure [26]. The estimated “position” of a query,  $L$ , was that corresponding to the best match given by Eq. 2; this position is always relative to that of another journey along roughly the same route; the accuracy and repeatability of this in associating location between passes was evaluated using distributions of location error distributions and area-under-curve criteria derived from these distributions.

### 5.3 Measurements of Performance

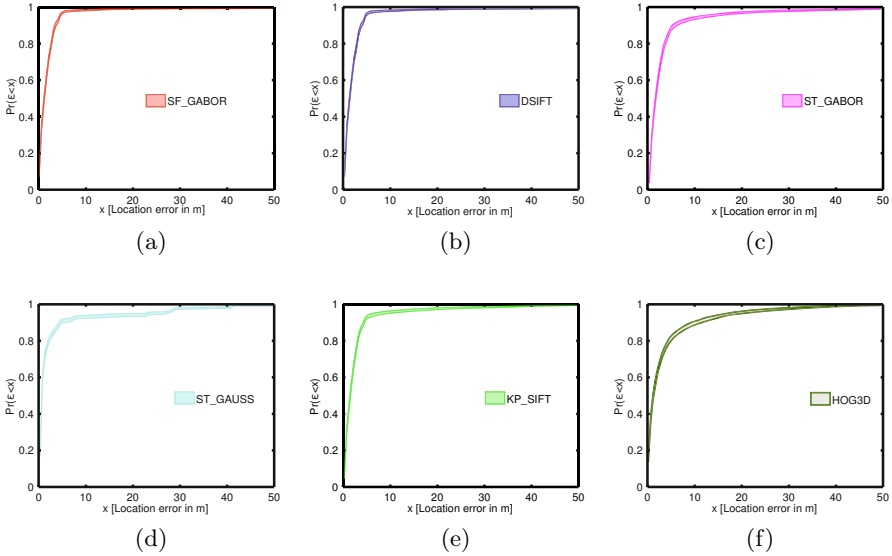
We quantify the accuracy of being able to associate *locations* along physical paths in corridors within the dataset described in Section 3. By permuting the paths that are held in the database and randomly selecting queries from the remaining path, we were able to obtain the error in localization. Repeated runs with random selections of groups of frames allowed variability in these estimates to be obtained, including that due to different paths being within the database. To estimate these distributions, we measured the absolute error in localization as a distance,  $\epsilon$ , relative to route ground truth, summarizing this as estimates of  $P(\epsilon < x)$ . For this, we used the ground-truth information acquired as described in Section 3.

### 5.4 Cumulative Distribution Functions

In Fig. 5, we compare the error distributions of all techniques. In Figs. 4(a) to 4(b), we provide separate assessments of the *variability* in error distribution when 1 million permuted queries are performed by cycling through 1,000 permutations of 1,000 randomly selected queries. This Monte-Carlo approach to testing accuracy allows the stability of approaches to be assessed. The graphs here suggest high reproducibility of retrieval performance (small shaded areas between lower and upper traces of each graph). All the results were generated with videos resized down to  $208 \times 117$  pixels; these are also supplied with the dataset.

### 5.5 Area-Under-Curve Comparisons

We calculated the average absolute positional error (in m) and the standard deviation of the absolute positional errors (Table 2). All queries were again performed by adopting the leave-one-out strategy, but because of the high repeatability of results (as seen in Fig. 4), we did not apply random frame-level sampling. Standard deviations of the absolute error distribution are also provided. Table 2 also provides the area-under-curve (AUC) values obtained from the CDFs of Fig. 4.



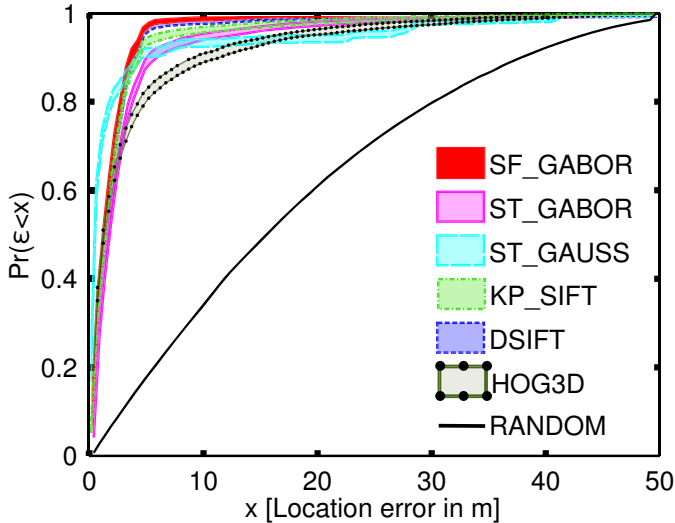
**Fig. 4.** Comparison between the error distributions obtained with the different methods. Note the high reproducibility of the performance results. The origin of the variability within each curve is explained in Section 5.4.

## 6 Results

One of the clear distinctions that we found, whether we used standard methods or the projective version of descriptors, is that single frame methods worked better than multiple-frame methods. This can be seen by comparing the top and bottom rows of Table 2. The results show that localization is achieved with

**Table 2.** Summaries of average absolute positional error and standard deviation of positional errors for different descriptor types.  $\mu_\epsilon$  is the average absolute error, and  $\sigma_\epsilon$  is the standard deviation of the error, both in metres. Top: single frame methods. Bottom: spatio-temporal methods.

Method	Error summary (m)		AUC (%)	
	$\mu_\epsilon$	$\sigma_\epsilon$	Min	Max
SF_GABOR	<b>1.59</b>	0.11	96.11	<b>96.39</b>
DSIFT	1.62	0.11	95.96	96.31
KP_SIFT	2.14	0.17	94.58	95.19
ST_GAUSS	<b>2.11</b>	0.24	94.82	<b>95.57</b>
ST_GABOR	2.54	0.19	93.90	94.44
HOG3D	4.20	1.33	90.89	91.83



**Fig. 5.** Comparison between the error distributions obtained with the different methods. The results for a random frame test (RANDOM) were introduced as a “sanity check”.

high accuracy in terms of CDF and AUC without a large difference between the applied methods, despite the big diversity in their complexity. Absolute errors show significant differences between methods, with average absolute errors in the range of 1.5 m to 4.20 m. Single frame methods (SF\_GABOR, KP\_SIFT and DSIFT) perform slightly better than spatio-temporal ones. This is not surprising, as the spatio-temporal methods might be too affected by the self motion over fine temporal scales.

In spite of using image retrieval methods in isolation, this performance is in the range of methods reviewed in Section 2 that include tracking, other sensors or estimate motion. We emphasise that no tracking was used in estimating position: this was deliberate, in order that we could assess performance in inferring location from the visual data fairly. Introducing tracking will, of course, improve localization performance, and could reduce query complexity. Yet, tracking often relies on some form of motion model, and for pedestrians carrying or wearing cameras, motion can be quite unpredictable.

## 7 Conclusion

We have presented several contributions in the topic of indoor localization using visual path matching from wearable and hand-held cameras. We provide an evaluation of six local descriptor methods: three custom designed and three standard image (KP\_SIFT and DSIFT) and video (HOG3D) matching methods

as baseline. These local descriptions follow a standard bag-of-words and kernel encoding pipeline before they are evaluated with the ground truth. The code for the local descriptors and the evaluation pipeline is available on the web page [18]. We also make available a large dataset with ground truth of indoor journeys to complete the evaluation framework.

The results show that there is significant localization information in the visual data even without using tracking, and that errors as small as 1.5 m can be achieved. We have split the results in two: a) Absolute positional errors that help to discern between image description methods and assess their localization capabilities; and b) error distributions that can be used to build a model for inclusion in a Kalman or particle filtering approach that is appropriate for human ambulatory motion.

We plan to introduce tracking as part of our future work and make use of the error distributions to build human motion models. There are, of course, numerous other enhancements that one could make for a system that uses visual data; integration of data from other sensors springs to mind, such as inertial sensing, magnetometers and RSSI. Although the fusing of independent and informative data sources leads to improvements in performance, we would argue that the methods applied to infer location from each information source should be rigorously tested, both in isolation and as part of an integrated system. This will ensure that real-world systems perform well in standard use, but are also somewhat robust to the sensor failure. With their very good standalone performance, we anticipate that using vision and associating the journeys of several users through their visual paths could play an important role in localization.

## References

1. Bosse, M.: Simultaneous Localization and Map Building in Large-Scale Cyclic Environments Using the Atlas Framework. *The International Journal of Robotics Research* **23**(12), 1113–1139 (2004)
2. Bregonzio, M.: Recognising action as clouds of space-time interest points. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 1948–1955 (June 2009). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5206779>
3. Burgess, N., Maguire, E.A., O’Keefe, J.: The human hippocampus and spatial and episodic memory. *Neuron* **35**(4), 625–641 (2002)
4. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: *Proceedings of the British Machine Vision Conference 2011* (1), 76.1–76.12 (2011). <http://www.bmva.org/bmvc/2011/proceedings/paper76/index.html>
5. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)* 1, 886–893 (2005). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1467360>
6. Everingham, M., Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* **88**(2), 303–338 (2009). <http://www.springerlink.com/index/10.1007/s11263-009-0275-4>

7. Hartley, T., Lever, C., Burgess, N., O'Keefe, J.: Space in the brain: how the hippocampal formation supports spatial cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**(1635), 20120510 (2014)
8. Huitl, R., Schroth, G.: TUMindoor: An extensive image and point cloud dataset for visual indoor localization and mapping. In: *International Conference on Image Processing* (2012). [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6467224](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6467224)
9. Kadous, W., Peterson, S.: Indoor maps: the next frontier. In: *Google IO* (2013)
10. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: *British Machine Vision Conference*. pp. 995–1004 (2008). <http://eprints.pascal-network.org/archive/00005039/>
11. Layton, O.W., Browning, N.A.: A Unified Model of Heading and Path Perception in Primate MSTd. *PLoS Computational Biology* **10**(2), e1003476, February 2014. <http://dx.plos.org/10.1371/journal.pcbi.1003476>
12. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *Computer Vision and Pattern Recognition*. vol. 2, pp. 2169–2178. IEEE (2006)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **60**, 91–110 (2004). <http://www.springerlink.com/index/H4L02691327PX768.pdf>
14. Matsumoto, Y., Inaba, M., Inoue, H.: Visual navigation using view-sequenced route representation. In: *International Conference on Robotics and Automation*, pp. 83–88. No., April 1996. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=503577](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=503577)
15. Ohno, T., Ohya, A., Yuta, S.: Autonomous Navigation for Mobile Robots Referring Pre-recorded Image Sequence. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS 1996*. vol. 2, pp. 672–679. IEEE (1996). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=571034>
16. Park, S., Jung, S., Song, Y., Kim, H.: Mobile robot localization in indoor environment using scale-invariant visual landmarks. In: *18th IAPR International Conference in Pattern Recognition*, pp. 159–163 (2008). <http://www.eurasip.org/Proceedings/Ext/CIP2008/papers/1569094833.pdf>
17. Quigley, M., Stavens, D.: Sub-meter indoor localization in unmodified environments with inexpensive sensors. In: *Intelligent Robots and Systems*, pp. 2039–2046. IEEE, October 2010
18. Rivera-Rubio, J., Alexiou, I., Bharath, A.A.: RSM dataset (2014). <http://rsm.bicv.org>
19. Schroth, G., Huitl, R.: Mobile visual location recognition. *IEEE Signal Processing Magazine*, 77–89, July 2011. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=5888650](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5888650)
20. Schroth, G., Huitl, R.: Exploiting prior knowledge in mobile visual location recognition. In: *IEEE ICASSP*, pp. 4–7 (2012). [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6288388](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6288388)
21. Shen, G., Chen, Z., Zhang, P., Moscibroda, T., Zhang, Y.: Walkie-Markie: Indoor Pathway Mapping Made Easy. In: *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13) USENIX*, pp. 85–98 (2013). [http://research.microsoft.com/en-us/um/people/moscitho/Publications/NSDI\\_2013.pdf](http://research.microsoft.com/en-us/um/people/moscitho/Publications/NSDI_2013.pdf)
22. Simpson, R., Cullip, J., Revell, J.: The Cheddar Gorge Data Set (2011)
23. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, October 2012

24. Tang, L., Yuta, S.: Vision based navigation for mobile robots in indoor environment by teaching and playing-back scheme. In: International Conference on Robotics and Automation, pp. 3072–3077 (2001). <http://ieeexplore.ieee.org/xpls/abs.all.jsp?arnumber=933089>
25. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008). <http://www.vlfeat.org/>
26. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010)
27. Viola, P., Jones, M.: Robust real-time object detection. *International Journal of Computer Vision*, pp. 1–25 (2001). <http://www.starocceans.net/documents/CRL-2001-1.pdf>
28. Wang, H., Sen, S., Elgohary, A., Farid, M., Youssef, M.: Unsupervised Indoor Localization. In: *MobiSys*. ACM (2012). <http://synrg.ee.duke.edu/papers/unloc.pdf>