Associating locations from wearable cameras

Jose Rivera-Rubio
http://www.bicv.org
Ioannis Alexiou
i.alexiou09@imperial.ac.uk
Anil Bharath
a.bharath@imperial.ac.uk
Riccardo Secoli
r.secoli@imperial.ac.uk
Luke Dickens
luke.dickens@imperial.ac.uk
Emil Lupu
e.c.lupu@imperial.ac.uk

Imperial College London
South Kensington Campus, UK

1 Motivation and contributions

In this paper, we address a specific use-case of wearable or hand-held camera technology: indoor navigation. We explore the possibility of crowdsourcing navigational data in the form of video sequences that are captured from wearable or hand-held cameras. Without using geometric inference techniques (such as SLAM), we test video data for navigational content, and algorithms for extracting that content. We do not include tracking in this evaluation: our purpose is to explore the hypothesis that visual content, on its own, contains cues that can be mined to infer a person's location. We test this hypothesis through estimating positional error distributions inferred during one journey with respect to other journeys along the same approximate path.

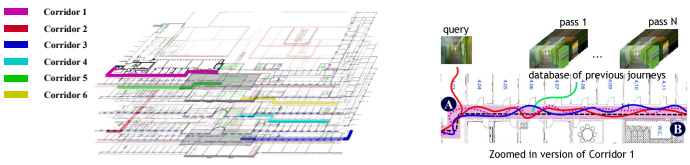


Figure 1: Maps of the recording locations (left). A sample path (Corridor 1, C1) with the multiple passes overlaid (right). Each of these passes represents a database sequence.

The contributions of this work are threefold. First, we propose alternative methods for video feature extraction that identify candidate matches between query sequences and a database of sequences from journeys made at different times. Secondly, we suggest an evaluation methodology that estimates the error distributions in position inference with respect to a ground truth. We assess and compare standard approaches in the retrieval context, such as SIFT [2] and HOG3D [1], to establish positional estimates. The final contribution is a publicly available database comprising over 90,000 frames of video-sequences with positional ground-truth. The data was acquired along more than 3 km worth of indoor journeys with a hand-held device (Nexus 4) and a wearable device (Google Glass).

2 The RSM dataset

The dataset contains 3.05 km of journey data. For each corridor, ten passes (i.e. 10 separate visual paths) were obtained. Five of these videos were acquired with the hand-held Nexus, and the remainder with Glass. The dataset is publicly available at [3].

	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		

Table 1: A summary of the dataset with thumbnails.

3 Methods

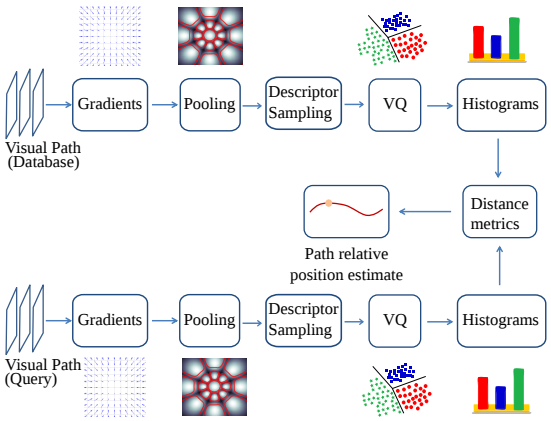
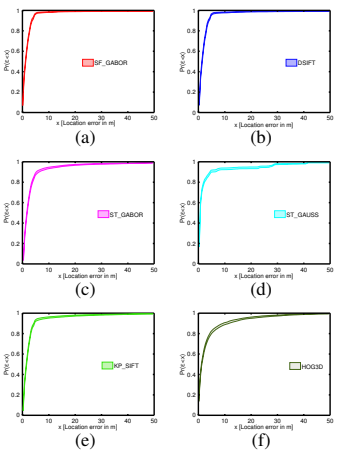
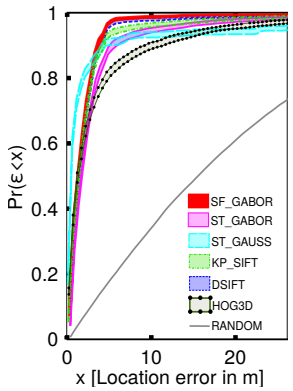


Figure 2: The stages in processing image sequences from database and query visual paths are illustrated above.

4 Evaluation





(g) Comparison between the error distribution obtained with the different methods. The results for a random test (RANDOM) were introduced as a "sanity check".

Figure 3: Cumulative Distribution Function of the methods under study.

[1] A Kläser, M Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, pages 995–1004, 2008.

[2] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[3] Jose Rivera-Rubio, Ioannis Alexiou, and Anil A. Bharath. RSM dataset, 2014. URL `http://rsm.bicv.org`.