Feature set reduction for image matching in large scale environments

Nabeel khan
Computer Science Department
University of Otago
nabeel@cs.otago.ac.nz

Brendan McCane
Computer Science Department
University of Otago
mccane@cs.otago.ac.nz

Steven Mills
Computer Science Department
University of Otago
steven@cs.otago.ac.nz

ABSTRACT

Image matching in large scale environments is challenging due to the large number of features used in typical representations. In this paper we investigate methods for reducing the number of SIFT (Scale invariant feature transform) features in an image based localization application. We find that reductions of up to 59% in the number of features can result in improved performance of a naive matching algorithm for highly redundant data sets. However, those improvements do not carry over to visual bag of words, where a more moderate feature reduction (up to 16%) is often needed to maintain performance similar to the non-reduced set. Our reduced features have performed better than other robust feature descriptors namely HoG, GIST and ORB on all data sets with naive matching. The main contribution of this paper is the compact feature representation of a large scale environment for robust 2D image matching.

1. INTRODUCTION

Image based matching is an ongoing and an active research problem in the computer vision community. It has been used for several different applications, with the most prominent being navigation or localization [2, 4, 14, 19]. Scaling up image databases to a very large scale (billions of images) is of great importance for image matching techniques to become maximally useful.

The standard procedure in image matching systems is to convert all trained images into a set of scale, rotation and view-point invariant features and then search for query image features in the database to find the closest match. The number of extracted features tends to scale linearly with the number of database images with a relatively high linear coefficient (often more than 1000 features per image).

In this paper we investigate the feasibility of reducing the number of features so that it scales linearly per scene independent of the number of images per scene. Our approach makes use of the observation that many images of the same scene share common features, and we propose four methods for reducing the total number of features.

1.1 Related Work

The problem of scene recognition has been addressed by several authors in the past. In [8], the idea of reduced features is used for precise matching in an indoor environment having 296 images. The SIFT features are extracted after partitioning the environment into locations, followed by query feature probabilistic integration thus yielding an estimate of the most likely location. They kept only 10% of the originally detected features and have reported good location recognition rates. Although they report very good results, it is not clear how well the technique will scale to very large databases as features are kept based on their discrimination ability.

In [6], Visual BoW based on homographies between SIFT features along with a voting scheme is used for scene matching in an indoor environment. The focus of the work is good localization in office buildings where scene matching becomes challenging due to high similarity among the images. No attempt was made to reduce the number of features required for matching.

Recent advances in structure-from-motion research [18] have made it possible to construct 3D models for environments on a large scale thus making it possible to perform 2D to 3D matching for scene localization. This idea was used in [4] which matched query features with 3D points (F2P). The focus of the work was precise matching of city scale images in large urban environments. They proposed a compact way to represent the 3D features and used a vocabulary tree for robust matching. In contrast, 3D points are matched with query features (P2F) in [10]. The off-line 3D model is compressed via a suitable selection of 3D points on the basis of visibility. The main contribution of this work is the efficient and precise matching of 3D to 2D points.

Priority based searching for efficient 2D to 3D matching is proposed in [16]. They have used different ways to reduce the 3D points and reported superior results over [10]. Direct 2D-3D matching techniques (F2P) are slower compared to direct 3D-2D techniques (P2F). The main contribution of this work is the considerable efficiency gain with the direct matching technique. All the mentioned works use SIFT features [11] for construction of 3D models.

Our work is inspired by [16, 10] who reduced the 3D points from the point cloud by averaging corresponding descriptors related to the 3D points. We propose and evaluate a 2D track based approach for feature reduction. The paper is structured as follows. Section 2 discusses our proposed methodology for feature reduction. Section 3 describes the dataset and performance metrics for our experiments. The experimental results are presented in Section 4 and the article is concluded in Section 5.

2. TRACK BASED FEATURE REDUCTION

In several feature matching schemes (e.g. visual bag-of-words), each image is represented as a set of features. The central idea of our approach is to represent each location (e.g. the kitchen) as a bag of features, where each location would typically have several images associated with it. The reduction in the number of features for a bag comes from identifying similar features in multiple images of the same location and storing that feature only once. *Tracks* represent the similar features that are matched across multiple images of the same location. The essential idea is very simple, but there are multiple ways in which tracks can be generated and Section 2.1 describes four variants. A 96 dimenionsal version of SIFT features is used in the current work [7], however any robust descriptor similar to SIFT can be used.

Figure 1: An example of a track showing one similar feature traced across three images of an indoor scene.

2.1 Track Generation

A greedy method is used in our work to identify feature matches for each pair of images in a scene to generate the tracks [17, 9]. Two features are said to match if the Euclidean distance in SIFT feature space is less than 170. The distance threshold of 170 has been determined empirically for SIFT features in [7].

The initial tracks only contain matched features from two images. However these tracks are expanded to include matched features from other images in two ways:

1. **Strict:** if a newly found feature matches all existing features of a track, then add the new feature to the existing track.

2. **Less Strict:** if a newly found feature matches with the original feature, then add the new feature to the existing track.

Non-matching features of a scene are stored in singleton tracks.

While identifying feature correspondences between images, we record the proportion of features for each image matched with the remaining images of the same scene. The process is done for every image of the scene and information is stored in an image correlation matrix of size $n \times n$, where n refers to the total number of images for a particular scene. This matrix indicates the similarity between scene images on the basis of overlapping features and is used for image grouping as discussed in Section 2.3.

2.2 Reduced Feature Set

Each track is represented as the average of all features contained within it. A scene is represented as a set of tracks. We have evaluated four different scene description variants:

ST All strict tracks are included.

LT All less strict tracks are included.

STF Strict tracks excluding singletons are included.

LTF Less strict tracks excluding singletons are included.

The argument for removing singleton tracks is that since they have not matched between training images of the same scene, it is less likely they will match a query image. Eliminating singleton tracks also produces a large reduction in the number of features needed to represent a scene.

2.3 Scene Representation

Scenes can be represented by one of the four proposed reduced feature sets. Every scene can have a single reduced feature set and this makes sense for image matching in simple scenes where all images are quite similar. However in complex scenes, some images within a scene are totally different e.g. images of two corners of a room. Therefore multiple sets per scene may lead to better classification performance. However the number of features remains the same regardless of the number of feature sets per scene.

We use the image correlation matrix to generate multiple feature sets per scene. Our splitting method starts with the first available scene image and finds corresponding similar images on the basis of a similarity threshold *(S_Threshold)*. The resulting initial group may be expanded, as some images in the initial group may have other similar images. An incremental approach is used and the group is considered final once no further expansion is possible. This is a simple image clustering method and we report on the results of image clustering in the trained data set in Section 4.4. We have used different thresholds in our experiments to generate different number of sets per scene for performance analysis.

3. DATASETS AND METRICS

We have used four data sets in the course of the experiments for our work.

1. **David Nister (DN):** The data set contains 4 images each of 2500 objects [12]. We have used the first 500 objects from the database with 1500 images for training and 500 for testing (the first image of each object is the test image).

2. **Pasadena Buildings (PB):** The data set contains 6 photos of the facades of 103 houses in the Pasadena area and 22 buildings from the Caltech campus taken at different times with varying viewpoints [1]. The first image of every building is used for testing while the remaining images are used for training.

3. **Owheo (OW):** The dataset contains 1534 indoor images of an office building [5]. The trained images comprise of corridors, central halls, labs etc and a total

of 25 scenes are identified. A *HTC Wildfire S* smart phone is used to capture 750 test images while navigating inside the building. The test and trained images are taken from different cameras.

4. **Commerce (CM):** The data set contains 864 indoor images of an office building [5]. The trained images comprise of corridors, central halls, stairs etc and a total of 14 scenes are identified from the data set. 234 test images are captured from the building via a *HTC Wildfire S* smart phone.

A sample image from each data set is shown in Figure 2. The following definitions are used to define the performance metrics for our system:

M_T Total number of query images correctly matched.

Q_T Total number of query images passed to the system.

C_S Total number of images of a scene grouped on the basis of similarity threshold ($S\_Threshold$).

Q_S Total number of images of a scene.

W_S Total number of images of a scene wrongly grouped.

The following evaluation metrics are used:

T_P The true positive rate refers to the correct matching performance of the system.

$$T_P = \frac{M_T}{Q_T}. \tag{1}$$

S_C The scene clustering rate refers to the grouping (in %) of similar images for all scenes. Where n refers to the total number of scenes. This metric is used in conjunction with E_S.

$$S_C = \sum_{i=1}^{n} \frac{C_S}{Q_S} \tag{2}$$

E_S The scene clustering error rate refers to the grouping of non-similar images for all scenes.

$$E_S = \sum_{i=1}^{n} \frac{W_S}{C_S} \tag{3}$$

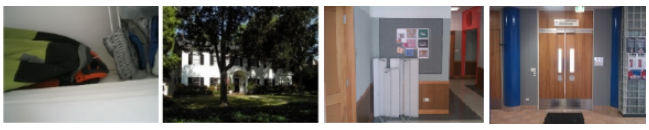


Figure 2: Sample images from David, Pasadena, Owheo and Commerce data sets.

4. RESULTS

4.1 Feature Reduction Statistics

Some algorithms produce fewer features than SIFT and are considered to be more robust. Therefore the following feature descriptors are included for comparison:

1. **Histogram of gradients (HoG):** These descriptors characterize the image via the distribution of local intensity gradients [3]. In our experiments, one HoG feature is generated per image and each feature is a vector of 81 values. The trained image giving the minimum Euclidean distance against the query is selected as the best match.

2. **GIST:** These descriptors use a set of perceptual dimensions such as naturalness, openness, ruggedness etc to represent the dominant spatial structure of a scene [13]. In our experiments, one GIST feature with a vector of 512 values is generated per image. The trained image giving the minimum Euclidean distance against the query is selected as the best match.

3. **ORB:** These are efficient binary descriptors based on orientations and offer robust matching [15]. In ORB, the number of features is fixed per image. We have used ORB with 200 features and hamming distance with a threshold of 50 is used to identify the query and trained matches. The trained image with a maximum number of matches is selected as the best match.

The feature reduction statistics of our track based approach and the other feature types on the four standard datasets are shown in Table 1. The table shows that feature reduction is significantly higher for the indoor environment due to a higher similarity between the images. A higher similarity threshold ($S\_Threshold$) leads to fewer groupings of images per scene and results in a large number of feature sets for every scene. On the other hand, zero similarity threshold means maximum grouping between the scene images and therefore results in a single feature set per scene. Table 1 also shows that HoG, GIST and ORB generate fewer features compared to normal SIFT.

As can be seen from Table 1 the sets which exclude singleton tracks are much more agressive at reducing the number of features per location. It should also be noted that in image collections with relatively few images per location (Nister and Pasadena sets), there is a very large reduction when singletons are excluded - predominantly because there is very little overlap between the images.

4.2 Naive Matching

In this section, we have evaluated the performance of the reduced feature sets via naive matching [7]. In naive matching, the features are extracted from the trained data and are stored. In image classification, all nearest neighbors of the query image features are found in the image collection. If the nearest neighbor is within a distance threshold in feature space, then a feature correspondence is recorded. The matched image is selected as that image with the most feature correspondences from the collection.

Table 1: Feature reduction via track based approach on all data sets. Legend: S_THRESH, Similarity threshold.

Datasets	Scenes	S_THRESH	Feature sets	Feature Type	Total Features	Feature Reduction
Owheo	-	-	**1534**	**HoG**	**1534**	-
	-	-	**1534**	**GIST**	**1534**	-
	-	-	**1534**	**ORB**	**306800**	-
	-	-	**1534**	**Normal SIFT**	**399238**	-
	25	0%	25	ST (SIFT)	335141	16.05%
		8%	65	ST-8 (SIFT)		
		0%	25	LT (SIFT)	239970	40%
		20%	292	LT-20 (SIFT)		
		0%	25	STF (SIFT)	198334	50.32%
		0%	25	LTF (SIFT)	88736	77.77%
David Nister	-	-	**1500**	**HoG**	**1500**	-
	-	-	**1500**	**GIST**	**1500**	-
	-	-	**1500**	**ORB**	**300000**	-
	-	-	**1500**	**Normal SIFT**	**482519**	-
	500	0%	500	ST (SIFT)	422797	12.38%
		0%	500	LT (SIFT)	418354	13.30%
		0%	500	STF (SIFT)	84652	82.45%
		0%	500	LTF (SIFT)	80209	83.37%
Pasadena	-	-	**500**	**HoG**	**500**	-
	-	-	**500**	**GIST**	**500**	-
	-	-	**500**	**ORB**	**100000**	-
	-	-	**500**	**Normal SIFT**	**248535**	-
	125	0%	125	ST (SIFT)	238344	4.1%
		0%	125	LT (SIFT)	236945	4.66%
		0%	125	STF (SIFT)	17338	93.02%
		0%	125	LTF (SIFT)	15992	93.56%
Commerce	-	-	**864**	**HoG**	**864**	-
	-	-	**864**	**GIST**	**864**	-
	-	-	**864**	**ORB**	**172800**	-
	-	-	**864**	**Normal SIFT**	**315556**	-
	864	0%	125	ST (SIFT)	260916	17.31%
		0%	864	LT (SIFT)	178290	43.49%
		0%	864	STF (SIFT)	164974	58.67%
		0%	864	LTF (SIFT)	84528	73.21%

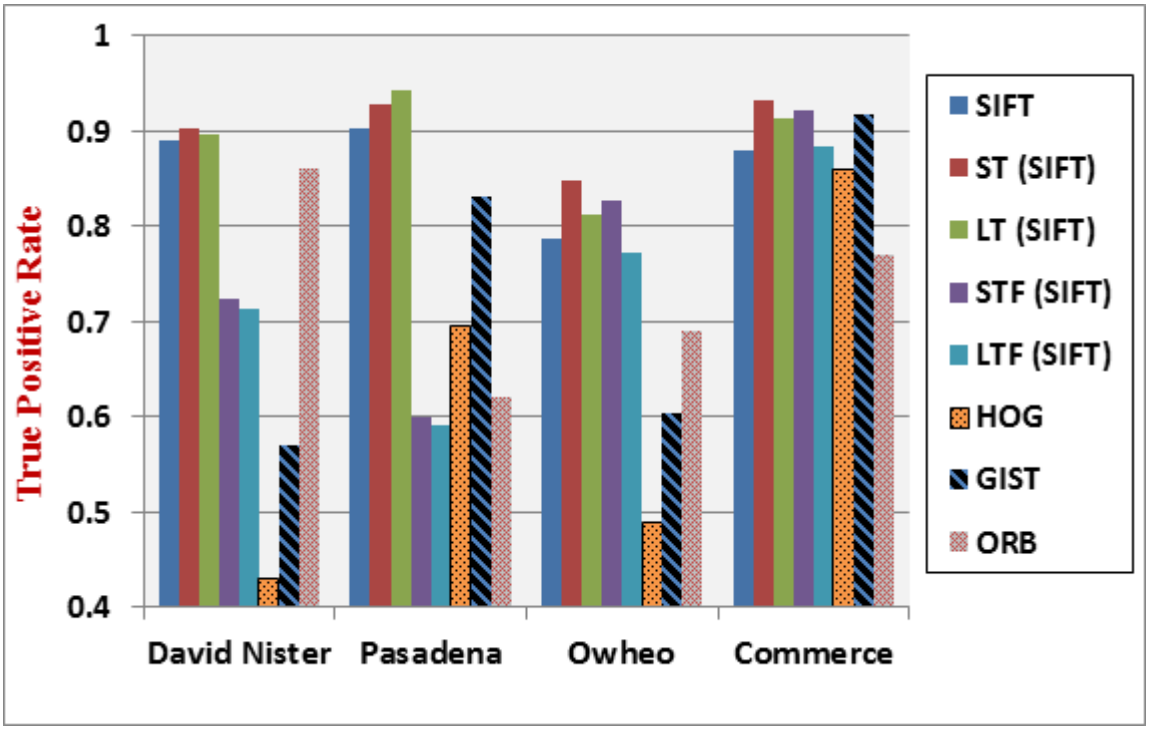


Figure 3: True positive rate (T_P) for normal unreduced and compact features on the all data sets.

In naive matching, either single or multiple feature sets per scene can be used as we normally look for 1-1 correspondence between the query and the scene features. We have therefore used a single reduced feature set per scene and the true positive rate (T_P) for all features on the four datasets via naive matching is shown in Figure 3. The results show that the reduced feature sets perform well compared to other unreduced features. Normal SIFT features give a stable performance over the other feature descriptors.

The results show that ST and LT feature sets have performed well across all datasets and in some cases can produce significantly smaller feature sets (maximum of 44% reduction). STF and LTF reduced sets under-perform on David Nister and Pasadena data sets due to a significant reduction of features i.e. more than 80%. STF and LTF cannot represent the scenes effectively and this leads to a poor matching performance. In our indoor data sets, each location has a large number of images; about 50 on average. This results in a feature reduction of up to 59% for STF which is sufficient to represent every indoor scene effectively. However LTF again under-performs due to a higher feature reduction (i.e. more than 70%). This highlights that at best, the number of features can be effectively reduced by about half. It is likely that our scheme is only useful for highly redundant image collections.

4.3 Visual BoW

We have analyzed the performance of reduced features via naive matching first as it has a smaller error rate compared to visual BoW. As discussed before, a single scene set (one bag per scene) often leads to reduced performance in image matching. Therefore we have used multiple sets per scene for our visual BoW experiments.

The final set of experiments is conducted using visual BoW with normalized term frequency scheme (ntf) in the same way as is done in [6]. An inverted index is used to retrieve 100 image clusters most similar to the query image. The retrieved image clusters are then ranked by BoW and the top ranked image cluster is considered the best match for the query image. The following terms are used in these results:-

1. ***LT, ST, STF and LTF*** refer to the cases when every scene is represented by a single feature set.
2. ***LT-8, LT-10, ST-20 etc.*** refer to the cases when a specific similarity threshold (8%, 10%, 20%) is used in C_S and the scene is represented by multiple feature sets.
3. ***SIFT*** refers to the case when normal unreduced features are used.

The results are displayed in Figures 4a to 4d and show that reduced features do not perform consistently as well as normal SIFT features across all data sets. However, for highly redundant collections such as Owheo and Commerce, the number of features can be reduced by 50% or more with the STF method with only a small reduction in performance in the case of Owheo, and slightly improved performance in the case of Commerce. For less redundant collections however, only a modest reduction in the number of features appears feasible.

4.4 Scene Clustering (S_C)

As discussed earlier, higher grouping of scene images (S_C) may result in a higher grouping error rate (E_S). The higher error E_S may lead to poor performance in classification tasks. In this section, we have analyzed the performance of our image clustering algorithm. An earlier version of our indoor Owheo data set is used for this analysis as it contains more scenes, although the number of images per scene is lower (15 on average). The error rate E_S is identified from the generated image groups for every scene against different similarity thresholds as shown in Table 2. The results show the image clustering algorithm is not the cause of the classification errors previously reported.

Table 2: Scene clustering and error rates for Owheo and David Nister data sets

	$\mathbf{S_C}$	$\mathbf{E_S}$		$\mathbf{S_C}$	$\mathbf{E_S}$
OW (ST-8)	76%	4%	**DN (ST-8)**	76%	0%
OW (ST-20)	42 %	1.2%	**DN (ST-20)**	47%	0%
OW (ST-40)	17 %	0%	**DN (ST-40)**	15%	0%

5. CONCLUSION

In this paper, we have investigated the effectiveness of reducing the number of features in image localisation tasks. We had hoped that a significant reduction in the number of features could be obtained with only a small reduction in performance. This would have allowed the visual BoW method to be applied on compute and memory limited devices such as mobile phones. Unsurprisingly, we found that the size of the reduction depended heavily on the collection used. For image collections with many redundant images, the number of features can be reduced by more than 50% with a small improvement in performance for naive matching. Curiously, an improvement in naive matching does not necessarily carry over to improvements in visual BoW, indicating that the ranking function used is not optimal for reduced features and alternative ranking functions should be investigated.

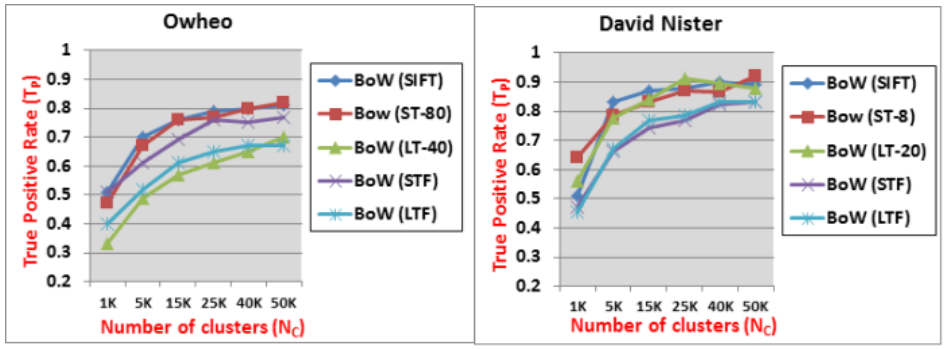
6. REFERENCES

[1] M. Aly, P. Welinder, M. Munich, and P. Perona. Scaling object recognition: Benchmark of current state of the art techniques. In *Proc. of Workshop on Emergent Issues in Large Amounts of Visual Data*, 2009.

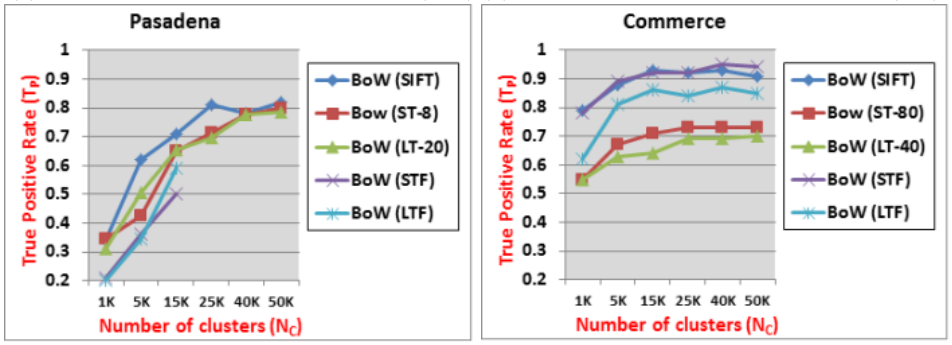
[2] R. Castle, G. Klein, and D. W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *Proc. of IEEE International Symposium on Wearable Computers*, pages 15–22, 2008.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[4] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. of Conference of*



(a) True positive rate for Owheo dataset (OW) (b) True positive rate for David data set (DN)



(c) True positive rate for Pasadena (PB) (d) True positive rate for Commerce (CM)

Figure 4: True positive rate for unreduced normal SIFT and reduced features on all data sets. Similarity thresholds (S_THRESH) of 8%, 20%, 40% , 80% are used for images clustering in our experiments and the best results are reported.

Computer Vision and Pattern Recognition, pages 2599–2606, 2009.

[5] N. Khan. Indoor environment images of owheo building, dunedin, nz, 2012.

[6] N. Khan, B. McCane, and G. Wyvill. Homography based visual bag of word model for scene matching in indoor environments. In *Proc. of Image and Vision Computing New Zealand*, 2011.

[7] N. Khan, B. McCane, and G. Wyvill. Sift and surf performance evaluation against various image deformations on benchmark dataset. In *Proc. of Conference on Diigital Image Computing Techniques and Applications*, pages 501–506, 2011.

[8] F. Li and J. Kosecka. Probabilistic location recognition using reduced feature set. In *Proc. International Conference on Robotics and automation*, pages 3405–3410, 2006.

[9] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proc. of European Conference on Computer Vision*, pages 427–440, 2008.

[10] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *Proc. of European conference on Computer vision*, 2011.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Interational Journal of Computer Vision*, 60(2):91–110, 2004.

[12] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of Conference of Computer Vision and Pattern Recognition*, pages 2161–2168, 2006.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[14] D. Robertson and R. Cipolla. An image-based system for urban navigation. In *Proc. of British Machine Vision Conference*, pages 819–828, 2004.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. Orb: An efficient alternative to sift or surf. In *Proc. of International Conference on Computer Vision*, pages 2564–2571, 2011.

[16] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. pages 667–674, 2011.

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *Proc. of SIGGRAPH*, pages 835–846, 2006.

[18] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80:189–210, 2008.

[19] W. Zhang and J. Kosecka. Image based localization in urban environments. In *Proc. International Symposium on 3D Data Processing, Visualization and Transmission*, 2006.