Local Feature Reliability Measure Using Multiview Synthetic Images for Mobile Visual Search

Kohei Matsuzaki*, Yusuke Uchida*†, Shigeyuki Sakazawa*, Shin'ichi Satoh‡
*KDDI R&D Laboratories, Inc., †University of Tokyo, ‡National Institute of Informatics
*Saitama, Japan, †Tokyo, Japan, ‡Tokyo, Japan
*{ko-matsuzaki, ys-uchida, sakazawa}@kddilabs.jp, ‡satoh@nii.ac.jp

Abstract

In this paper, we propose a new database (DB) construction method for the mobile visual search (MVS) system based on the local feature and bag-of-visual-words framework. In MVS, quantization error is unavoidable and causes performance degradation. Typical approaches for visual search extract features from a single view of reference images, though such features are insufficient to manage the quantization error. In this paper, we generate multiview synthetic images and extract local features. These features are resampled according to our novel reliability measure in order to reduce the DB size. Experiments on the three datasets show that the proposed method successfully constructs a robust DB with same size. The proposed method improved the mean average precision compared with a conventional method without changing the searching procedure.

1. Introduction

Image retrieval on mobile devices (MVS: mobile visual search) has been studied [2, 1, 12, 18]. Many of MVS systems are based on the Bag-of-Visual-Words (BoVW) framework [14]. In the BoVW framework, local features such as SIFT [7] are extracted from a query image. Those local features are quantized to representative vectors called Visual Words (VWs), and a similarity search is performed based on histogram of the VWs. In this paper, we focus on MVS system based on the BoVW framework. While some researches focus on server-client systems [12, 18], the purpose of our research is to achieve more accurate retrieval with lower memory requirements on stand-alone mobile devices. Since this system stores a database (DB) on a mobile device, smaller DB size is preferred due to small amount of available memory. However, it is challenging to achieve both small DB size and high accuracy. In MVS, the change of view is caused by capturing the query image on mobile devices. It induces quantization errors, decreasing the retrieval accuracy. Many of existing methods that alleviate quantization errors induce an increase of the DB size.

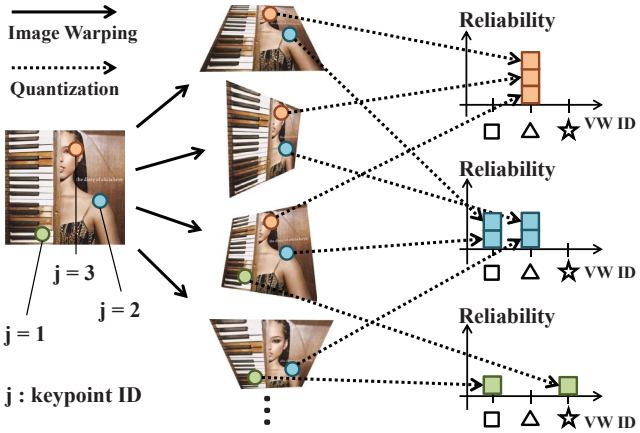

Figure 1. Illustration of the proposed method. Circles represent keypoints. Squares represent quantization results.

As a countermeasure to the quantization error, multiple assignment (MA) [11] is often used to assign a single feature to multiple VWs that represent the k-nearest neighbor. Mikulík et al. achieved a more accurate MA than assigning to the k-nearest neighbors by assigning to k VWs that learned from features extracted from different viewpoint images [8]. However, these countermeasures increase the DB size k times. In the meantime, feature selection is an effective method for reducing the DB size. Wang et al. reduced the DB size by selecting the informative feature [17]. However, this approach is not realistic in practical use since it requires several images taken by the same object from different viewpoints. In order to solve this problem, we propose using synthetic images simulating viewpoint changes from a single image like ASIFT [9], then performing the feature selection. ASIFT realizes a robust image matching result. However, if we apply its strategy in the context of image retrieval, we have to register all synthetic images to the DB, which increases the DB size massively.

In this paper, we propose a DB construction method to improve the search accuracy without increasing the DB size. We aim to achieve higher accuracy retrieval with fewer features of DB by selecting robust features across various

image warping. Figure 1 illustrates the proposed method. When there is a limitation on the DB size, it is desirable to register the feature according to certain reliability measure. In the proposed method, we define the feature reliability by measuring feature robustness in keypoint detection and quantization error. As shown in Figure 1, we generate various synthetic images from a reference image and extract features from them individually. We measure how frequently features detected from the same location of an object are quantized to the same VW, and use it as the feature reliability, e.g., an orange feature is the most reliable because it obtained same quantization result three times. Finally, we construct the DB while preventing an increase of its size by selecting features based on the feature reliability.

2. Baseline of Visual Search System

Figure 2 shows the framework of the visual search system as a baseline. Our proposed method is based on the binary local feature and BoVW framework, and its extension as shown in Figure 2. In the index step, reference features are extracted from reference images and quantized to VWs. They are stored in the inverted index with additional data useful in the search called *substring*. In the search step, query features are extracted from the query image in the same way as the index step. Then, scores are voted for reference images to calculate image similarities. Finally, geometric verification is performed to reference images with top similarities.

In the following, we briefly describe the extension methods of this framework used in this paper.

Multiple Assignment. Feature vectors extracted from the reference image are assigned to the k-nearest neighbor VWs in the feature space in order to alleviate the quantization error. Multiple assignment (MA) is performed on the reference side [11].

Weighting based on Substring as an alternative to Hamming embedding. Hamming embedding (HE) converts the feature vector into compact codes, and stores them in the inverted index [4]. HE improves the search accuracy by using the Hamming distance between query feature codes and reference feature codes to weighting in the voting process. In this paper, we use substring (SS) for weighting as an effective alternative to HE for the binary local feature as done in [16]. SS generates a short binary string by extracting bits of specific positions of the binary feature vector. As with HE, SS measures the Hamming distance between each feature string. In order to boost the performance, SS learns the positions of distinctive bits for the w-th VW from training images. For example, if the training result is $Sw = \{4, 25, 70, 87\}$, the 4th, 25th, 70th, and 87th bit of each binary feature assigned to w-th VW are extracted.

Weak Geometric Consistency. Weak geometric consistency (WGC) improves the search accuracy by filtering local features that are not consistent in orientation and scale [4]. Since WGC based on the scale does not contribute to accuracy as shown in [15], we only use orientation.

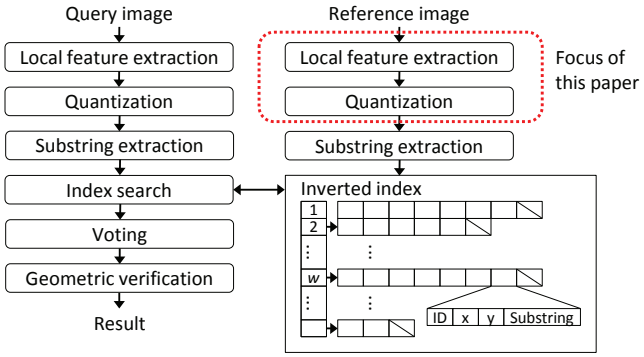

Figure 2. Framework of the visual search system.

Figure 3. Example of synthetic images (same scaling factor).

3. Proposed Approach

3.1. Overview

The purpose of the proposed method is to alleviate quantization errors without increasing the DB size. The proposed method extracts features from each synthetic image, and estimates the reliability of features based on statistics. This process is performed in the local feature extraction process and quantization process in Figure 2. We construct a DB where the specified number of features is registered based on the reliability. In this DB construction process, we reduce the burstiness of the VW directly and DB size simultaneously. In order to improve SS and WGC performance, we average features extracted from the synthetic images.

3.2. Synthetic Image Generation

We generate synthetic images from a virtual viewpoint P_i by warping the reference image in order to simulate various query images captured by mobile devices. We perform a uniform sampling of the virtual viewpoints position as done in [3]. In this paper, we empirically use 26 viewpoints where elevation exceeds 45 degrees in 71 viewpoints [6] because a finer sampling interval results in a larger amount of calculation. For each viewpoint, we also generate multi-scale images to simulate scale changes by using the scaling factor of 1, $1/\sqrt{2}$, and $1/2$. We calculate homography matrix H_i corresponding to P_i as done in [3], and generate synthetic images as shown in Figure 3.

3.3. Keypoint Tracking

We track keypoints detected from the same point of the object in order to collect keypoints that are robust to disturbance. We detect keypoints from the reference image and synthetic images. Let K_j and k_j denote the j-th keypoint and its position detected from the reference image, and let Q_l and q_l denote the l-th keypoint and its position detected from the synthetic image. We match K_j with Q_l using the homography matrix H_i, which generated i-th synthetic image (i.e., ground truth) : we find $q_{l'}$ corresponding to K_j using the following equations:

$$l' = \arg\min_{l} \|H_i k_j - q_l\|^2 \quad (1)$$

$$\|H_i k_j - q_{l'}\|^2 \leq t \quad (2)$$

where t is a reprojection error threshold. In this paper, $t = 3$ in imitation of a typical value used in calculating the inlier by using the RANSAC algorithm (e.g., OpenCV). We repeat the above process for all synthetic images corresponding to P_i. We then obtain the tracking result of K_j as *track* $T_j = \{K_j, Q_{l'1}, Q_{l'2}, \cdots\}$. The keypoints in the track T_j are expected to be extracted from the same point as K_j. In the following chapter, we do not use Q_l that failed in this tracking.

3.4. Feature Reliability Measure

Although keypoints are detected from the same location of an object, their feature vectors could be changed by disturbance and then quantized to different VWs. We rely on VWs that are observed frequently at the same location of an object in various synthetic images. We measure how frequently features are quantized to the same VW and use it as a feature reliability measure. Then, we construct the DB according to the reliability measure.

We quantize all features in the track T_j to VWs, and we obtain the reliability score $s(j, v)$ that represents the frequency of VW v that appeared in the track T_j. Figure 4 shows the overview of our feature selection method based on the reliability score. Figure 4 corresponds to the Figure 1 regarding the reliability score, keypoint ID, and VW ID. That is, each keypoint is connected with the frequency where features are quantized to the same VW. For example, in Figure 1, an orange keypoint (keypoint ID is 3) is successful in tracking for the three synthetic images, and all of the three features are quantized to VW ID "△". Therefore, its reliability score becomes $s(3, \triangle) = 3$ in Figure 4. We select features in descending order of the score $s(j, v)$, and discard the rest when they reach the specified number. In Figure 4, the features corresponding to the three highest scores are selected.

The following effects are expected when we select a specified number of features according to the score $s(j, v)$:

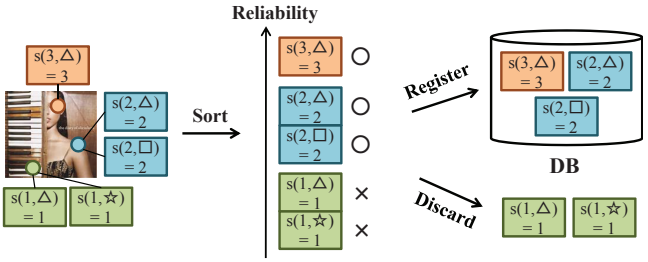

Figure 4. Feature selection based on the feature reliability measure

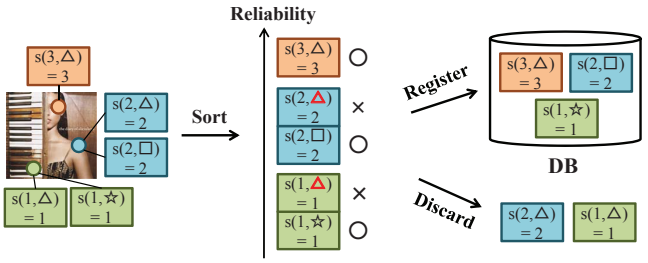

Figure 5. The overview of non-bursty selection

Useful Feature Selection. In the proposed method, features with high repeatability in detection and less prone to quantization error are preferentially selected. It is expected that these features tend to be matched with the correct query features in the search step and significantly improve the final result.

Adaptive Multiple Assignment. As a result of the proposed method, multiple (non-fixed n) VWs could be assigned to the single keypoint because we select features as a pair of the keypoint and the VW. For the keypoint that has high repeatability of both detection and quantization, the n becomes small. For the keypoint that has high repeatability of detection and low repeatability of quantization, the n becomes large. Therefore, we can realize a more efficient MA with respect to the memory storage.

3.5. Non-bursty Selection

Jégou et al. argued that burstiness of the visual element corrupts the visual similarity measure [5]. They propose alleviating this burstiness phenomenon by scoring, but it is impossible to reduce the DB size by their approach. Therefore, we propose a feature selection method to reduce the burstiness of the VW directly and DB size simultaneously. In our selection method, if multiple features are quantized to the same VW, only a feature with the highest reliability score is registered and the others are discarded. Figure 5 shows the overview of our Non-bursty selection method. In Figure 5, the feature with $s(3, \triangle)$ is selected firstly. According to the score $s(j, v)$, the feature with $s(2, \triangle)$ should be selected secondly. However, unlike Figure 4, the feature with $s(2, \triangle)$ is discarded because the feature with the same VW "△" is already registered. By selecting a speci-

fied number of features using this method, we can suppress the burstiness of the VW directory and reduce the DB size simultaneously. If the number of registered features did not reach the specified number, we repeat the same process to select new features from the discarded features in the previous process.

3.6. Feature Averaging

In order to improve the performance of SS and WGC described in Section 2, we use the average of features in the same track T_j and quantized to the same VW v — the number of features averaged in this step is equal to $s(j, v)$. Usually, the features extracted from the reference image are used in SS and WGC. However, in the proposed method, multiple features extracted from synthetic images are assigned to the pair of j and v. Therefore, we average feature vectors and orientations extracted from synthetic images and use them in SS and WGC instead of the features of the K_j. This averaging corresponds to the maximum likelihood estimation of the feature in terms of a pair of j and v. That is, we improve the SS and WGC by using the maximum likelihood feature vectors and orientations under the constraint as the same keypoint location and the same quantization result. Regarding the feature vector, we average each dimension, and binarize the average value. Regarding the orientation, we average unit vectors with the angles of features and use the angle of the averaged vector in WGC. For orientation, we also tried a median value, but we obtained a slightly worse result than averaging.

4. Experimental Evaluation

We conducted two experiments. The first is an evaluation of the contribution of non-bursty selection and feature averaging. The second is a comparison of the proposed method and the conventional MA. Let "Prop" denote a method without both non-bursty selection and feature averaging. Let "Prop (NBS)" denote Prop with non-bursty selection. Let "Prop (FA)" and "Prop (NBS + FA)" denote Prop and Prop (NBS) with feature averaging respectively.

4.1. Datasets and Parameters

In the first experiment, we use the Stanford mobile visual search dataset[1] (SMVS). SMVS consists of 8 classes such as book, CD, and so forth. There are 1,193 clean reference images and 3,269 query images taken with mobile devices. In the second experiment, we use the three publicly available datasets in order to evaluate performance on the datasets with different types and sizes, namely SMVS, INRIA Holidays dataset[2] (Holidays), and University of Kentucky Benchmark[3] (UKB). Holidays contains 1491 images consisting of 500 subsets. UKB consists of 10,200 images taking 2,550 objects from four different viewpoints. For all datasets, we reconfigure both the query images and the reference images VGA size. As an indicator of retrieval performance, we use the mean average precision (MAP) as in [10]. Due to the efficiency, we adopt the ORB feature [13], where 900 features are extracted from 4 scales on average.

[1] http://web.cs.wpi.edu/ claypool/mmsys-dataset/2011/stanford/
[2] https://lear.inrialpes.fr/ jegou/data.php#holidays/
[3] http://vis.uky.edu/ stewe/ukbench

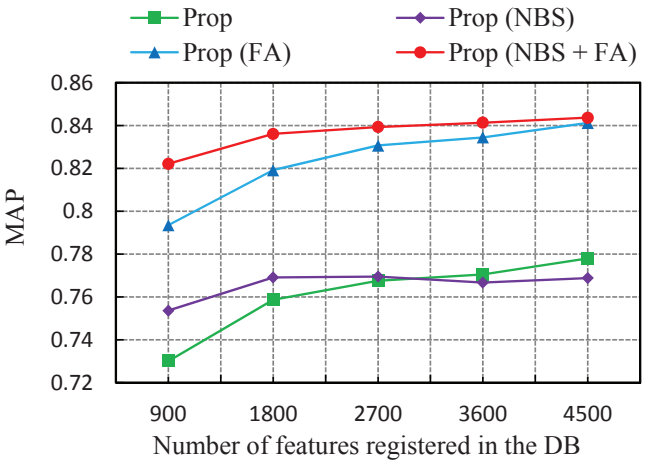

Figure 6. Contribution of each element of the proposed method. Prop vs. Prop (NBS) or Prop (FA) vs. Prop (NBS + FA) represents the effect of the non-bursty selection. Prop and Prop (NBS) vs. Prop (FA) and Prop (NBS + FA) represents the effect of the feature averaging.

4.2. Impact of Each Element of the Proposed Method

We evaluate non-bursty selection and feature averaging. The number of features registered to the DB is set to 900, 1800, 2700, 3600, and 4500. Figure 6 shows the average MAP of the 8 classes of each method as a function of the DB size. We can see a tendency that the accuracy decreases as the DB size is shrinks. It is shown that the non-bursty selection contributes to the maintaining of accuracy when the DB size is small, by comparing with or without NBS. That is, the result suggests that it is better to impose a non-burstiness constraint in comparison to registering features in reliable order without that constraint. However, if DB size is too large, the result is adversely affected by the constraint as shown by the comparison of Prop and Prop (NBS). This is because the DB includes more unreliable features as the number of registered features increases. Prop (NBS) tends to register unreliable features compared to the Prop because of its constraints. However, it is shown that the feature averaging constantly contributes regardless of the DB size, by comparing with or without FA. Thus the average of features extracted from synthetic images is more reliable than the feature extracted from a single reference image.

	Scoring side		Database side		SMVS	Holidays	UKB
	WGC	SS	NBS	FA			
MA					0.427	0.375	0.497
MA	x				0.525	0.470	0.681
MA		x			0.730	0.577	0.727
MA	x	x			0.760	0.578	0.718
Prop					0.574	0.403	0.530
Prop	x				0.612	0.506	0.710
Prop		x			0.733	0.574	0.732
Prop	x	x			0.768	0.582	0.722
Prop (NBS)	x	x	x		0.770	0.580	0.699
Prop (FA)	x	x		x	0.831	0.630	**0.759**
Prop (NBS+FA)	x	x	x	x	**0.839**	**0.631**	0.756

Table 1. Compare the proposed method with multiple assignment on the three dataset in terms of mean average precision. MA = multiple assignment, WGC = weak geometric consistency, SS = weighting based on substring: see Section 2. NBS = non-bursty selection, FA = feature averaging: see Section 3.

4.3. Compare the Proposed Method with Multiple Assignment

We compare the proposed method with the reference side MA [11] as a conventional method. We compare the search accuracy using the DB constructed by each method while keeping the searching procedure and the DB size the same. The number of features registered to the DB is set to 2700. This corresponds to the same DB size when the number of assignments is three in MA. Though WGC and SS are not used in [11], we combine MA with them to compare with the proposed method. Table 1 shows all the experimental results. As shown, the proposed method achieves the best accuracy in all datasets. Comparing Prop and MA, Prop generally outperforms MA if we use the same searching procedure (i.e., Scoring side in Table 1). This is because Prop effectively alleviates the quantization error. Prop (NBS) is sometimes worse than MA. This could be related to the fact that non-bursty selection can select unreliable features when the DB size is large. In this experiment, we use the large size DB to compare with MA but it is in fact desirable to use smaller size DB. Going back to Figure 6, we can see that non-bursty selection works more effectively with the smaller size DB. Prop with FA always significantly outperforms MA. This is because FA provides the optimal orientations and feature vectors for WGC and SS respectively.

5. Conclusion

In this paper, we proposed a new database construction method for mobile visual search. The proposed method achieves the image retrieval alleviating quantization error without changing the searching procedure. Experiments on the three datasets show that the proposed method is superior to multiple assignment when the database size is the same.

References

[1] D. Chen and B. Girod. Memory-efficient image databases for mobile visual search. *IEEE MultiMedia Magazine*, 21(1):14–23, 2014. 1

[2] B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham. Mobile visual search. *IEEE Signal Processing Magazine*, 28(4):61–76, 2011. 1

[3] S. Hinterstoisser, V. Lepetit, S. Benhimane, P. Fua, and N. Navab. Learning real-time perspective patch rectification. *IJCV*, 91(1):107–130, 2011. 2

[4] H. Jégou, D. Matthijs, and S. Cordelia. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008. 2

[5] H. Jégou, D. Matthijs, and S. Cordelia. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009. 3

[6] D. Kurz, O. Thomas, and B. Selim. Representative feature descriptor sets for robust handheld camera localization. In *ISMAR*, pages 65–70, 2012. 2

[7] D. G. Lowe. Distinctive image feature from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[8] A. Mikulík, M. Perdoch, O. Chum, and J. Matas. Learning a fine vocabulary. In *ECCV*, pages 1–14, 2010. 1

[9] J. M. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sciences*, 2(2):438–469, 2009. 1

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007. 4

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8, 2008. 1, 2, 5

[12] H. Qi, M. Stojmenovic, K. Li, Z. Li, and W. Qu. A low transmission overhead framework of mobile visual search based on vocabulary decomposition. *IEEE MultiMedia Magazine*, 16(7):1963–1972, 2014. 1

[13] E. Rubee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *ICCV*, pages 2564–2571, 2011. 4

[14] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003. 1

[15] S. S. Tsai, D. M. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod. Fast geometric re-ranking for image based retrieval. In *ICIP*, pages 1029–1032, 2010. 2

[16] Y. Uchida, S. Sakazawa, and S. Satoh. Binary feature-based image retrieval with effective indexing and scoring. In *GCCE*, pages 319–320, 2014. 2

[17] Z. Wang, Q. Zhao, D. Chu, F. Zhao, and L. J. Guibas. Select informative features for recognition. In *ICIP*, pages 2477–2480, 2011. 1

[18] Y. Wu, S. Lu, T. Mei, J. Zhang, and S. Li. Local visual words coding for low bit rate mobile visual search. In *ACM MM*, pages 989–992, 2012. 1