# Binary Feature-based Image Retrieval with Effective Indexing and Scoring

Yusuke Uchida
KDDI R&D Laboratories, Inc.
Saitama, Japan

Shigeyuki Sakazawa
KDDI R&D Laboratories, Inc.
Saitama, Japan

Shin'ichi Satoh
National Institute of Informatics
Tokyo, Japan

***Abstract*—In this paper, we propose a stand-alone mobile visual search system based on binary features and bag of visual words framework. The contribution of this paper is two-fold: (1) a visual word-dependent substring extraction method is proposed; (2) a modified version of the local NBNN scoring method is proposed in the context of image retrieval. The proposed system improves retrieval accuracy by 11% compared with a conventional method without increasing the database size.**

## I. Introduction

Local feature-based image or object retrieval has become a popular research topic. In particular, binary features such as ORB, FREAK, and BRISK have attracted much attention due to their efficiency [1]. With the increasingly wide-spread use of mobile devices such as Android phones or iPhones, mobile visual search (MVS) has become one of the major applications of image retrieval and recognition technology. While some research focuses on server-client systems in the context of MVS, the purpose of our research is to achieve fast and accurate recognition with lower memory requirements on mobile devices; in this paper, we call the latter type of MVS "local MVS". Local MVS does not require any server and it works without a network, realizing faster recognition. Thus, it is suitable for recognizing medium sized databases: i.e., recognizing catalogs, paintings in a museum, or cards in a collectible card game. The difficulty in local MVS lies in indexing of local features because it is necessary to fit the database to memory on mobile devices while maintaining retrieval accuracy. In other words, managing the trade-off between the memory size of the database and the accuracy of image retrieval is very important.

Indexing features is an essential component of efficient retrieval or recognition. The most widely adopted framework is the bag of visual words (BoVW) framework [2]. In the BoVW framework, local features of an image are quantized into visual words (VWs), resulting in a histogram representation of VWs. As the histograms are generally sparse, an inverted index data structure and a voting function enables an efficient similarity search [2], [3]. Though the BoVW framework allows efficient retrieval, some degradation of accuracy is caused by quantization because, in the BoVW framework, two features are matched if and only if they are assigned to the same VW [3]. Therefore, quite different features are sometimes matched in the BoVW framework. One of the frameworks commonly used to overcome this degradation is Hamming embedding [3]. In this framework, a short binary substring is stored for each feature and, after VW-based matching, unreliable matches are filtered out according to the distances between the substrings.

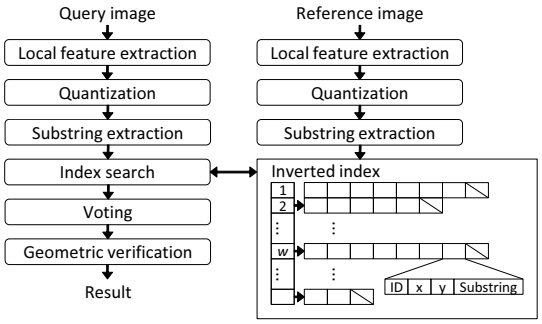


Fig. 1. The framework of the proposed method.

Although many methods have been proposed to improve the BoVW framework for continuous features such as SIFT, there are not many studies on indexing recent binary features for the purpose of image retrieval. In [4], the BoVW framework has been adopted to index recent binary features, referred to as bag of binary words (BoBW). In [5], a variant of the Hamming embedding method is proposed for binarized features, where the first $a$ bits are used to form VWs, and the next $b$-bit substring is stored in an inverted index. However, this approach results in non-optimal performance because some bits are not informative (mean value is significantly different from 0.5) and other bits are highly correlated, and these statistics are differ markedly from one VW to another. In this paper, to solve this problem, we propose a VW-dependent substring extraction method, which adaptively extracts effective substrings. As the scoring method in [5] has not been considered in depth, we also propose the use of a modified version of a local NBNN-based scoring method for image retrieval, which was first proposed for image classification. It provides a theoretical basis for scoring in voting and the proposed modification improves performance by using adaptive density estimation.

## II. Proposed Approach

Figure 1 shows the framework of the proposed method, which is almost the same as the Hamming embedding method [3] except for the substring generation and scoring methods. In the indexing step (offline), binary reference features are extracted from reference images and quantized into VWs. The substrings of reference features are generated and stored in an inverted index for efficient search. In the search step (online), each query feature of a query image votes scores to reference images according to the distances between the query feature substring and reference feature substrings.

TABLE I. COMPARISON OF THE PROPOSED METHOD WITH CONVENTIONAL METHODS.

| | book covers | business cards | cd covers | dvd covers | landmarks | museum paintings | print | video frames | average |
|---|---|---|---|---|---|---|---|---|---|
| BoBW [4] | 0.610 | 0.173 | 0.427 | 0.465 | 0.080 | 0.486 | 0.125 | 0.584 | 0.369 |
| [5] | 0.874 | 0.463 | 0.752 | 0.811 | 0.197 | 0.671 | 0.423 | 0.824 | 0.627 |
| PROP | **0.916** | **0.535** | **0.807** | **0.897** | **0.253** | **0.718** | **0.542** | **0.853** | **0.690** |
| PROP+GW | 0.943 | 0.602 | 0.849 | 0.930 | 0.278 | 0.740 | 0.568 | 0.900 | 0.726 |
| PROP+LN(a) | 0.927 | 0.515 | 0.830 | 0.924 | 0.282 | 0.758 | 0.501 | 0.909 | 0.706 |
| PROP+LN(b) | **0.955** | **0.609** | **0.873** | **0.944** | **0.289** | **0.773** | **0.570** | **0.914** | **0.741** |

**Indexing step.** Before indexing, two training procedures are required. First, in order to create VWs, the k-means algorithm is performed on training binary features as done in [4]. The other training procedure is for substring generation. While the method proposed in [5] uses fixed positions of bits for the substring, we adaptively change the positions for each VW. For this purpose, we construct substring dictionary $\mathcal{S}_w$, which defines the positions of useful bits for the $w$-th VW. For example, in the case where $\mathcal{S}_w = \{4, 25, 70, 87\}$, the 4th, 25th, 70th, and 87th bit of each binary feature assigned to $w$-th VW are selected, resulting in a 4-bit substring. The substring dictionary is constructed by using the algorithm used in ORB [6], where informative (mean value is close to 0.5) and non-correlated bits are selected. In the indexing step, the reference binary feature is first quantized by using VWs, and the substring is extracted using $\mathcal{S}_w$. In this paper, we set $|\mathcal{S}_w| = 64$; 64-bit substrings are used. Then, the following information is stored to the $w$-th list of the inverted index as shown in Figure 1: image identifier (2 bytes), the position $(x, y)$ (2+2 bytes), and the substring (8 bytes).

**Search step.** In the search step, each binary query feature of a query image votes scores to reference images by the following procedure. First, the binary query feature is assigned to the nearest neighbor VW. The substring of the binary query feature is generated in the same manner as in the indexing step. Then, the distances between the query substring and reference substrings in the corresponding list of the inverted index are calculated. Finally, scores are voted to the $K$-nearest neighbor reference features. It is known that weighting scores according to their distances improves performance. The most common way of doing this weighting is to use the Gaussian function $\exp(-d^2/\sigma^2)$ [7], where $d$ is the Euclidean or Hamming distance between the query feature and reference feature and $\sigma$ is an adjustable parameter. However, this approach has little theoretical basis and is not optimal. In this paper, we propose the use of a modified version of the local NBNN (LN) method [8], which has a theoretical background in the derivation of its score. Although LN was originally proposed for image classification, we show that this method also works well in image retrieval. In LN, for each query $q$, a score of (a) $d_K^2 - d_k^2$ is voted to the corresponding image of the $k$-th nearest neighbor feature, where $d_x^2$ represents the distance between $q$ and its $x$-th nearest neighbor feature. In this paper, we modify this original formulation to (b) $(d_K/d_k)^2 - 1$. This modification has the effect of adaptively changing the kernel radius in kernel density estimation, resulting in more appropriate scoring.

## III. EXPERIMENTAL EVALUATION

In the experiments, the Stanford mobile visual search dataset[1] is used. It contains eight classes of images and each class consists of 100 reference images and 400 query images. As an indicator of retrieval performance, mean average precision (MAP; higher is better) [3] is used. We adopt the ORB feature [6], where 900 features are extracted from 4 scales on average. The number of VWs is fixed at 1024 in all methods and experiments.

Table I summarizes the experimental results; the MAP scores of eight classes are shown for each method. First, the effectiveness of the proposed substring generation method is evaluated. For fair evaluation, traditional tf-idf scoring [9] is used for all methods. Comparing simple BoBW [4] with the method proposed in [5][2], we can see that the use of substring improves accuracy dramatically. However, the proposed method can further improve accuracy by adaptively generating substring. Second, we evaluate the proposed scoring method. The proposed substring method with Gaussian weighting (GW) is used as a conventional method. From Table I, it is shown that, while the original LN scoring method (LN(a)) is inferior to GW, the proposed LN scoring method (LN(b)) outperforms GW by 1.5% and the method in [5] by 11% in MAP. As the overhead of the proposed method is negligible, the proposed system can improve retrieval accuracy with the same memory requirement and almost the same computational cost as conventional methods.

## IV. CONCLUSION

In this paper, we proposed a stand-alone mobile visual search system based on binary features. In our system, a VW-dependent substring extraction method and a new scoring method are used. It is shown that the proposed system can improve retrieval accuracy with the same memory requirement as conventional methods.

## REFERENCES


[1] J. Heinly, E. Dunn, and J.-M. Frahm, "Comparative evaluation of binary features," in *Proc. of ECCV*, 2012, pp. 759–773.

[2] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, 2003, pp. 1470–1477.

[3] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IJCV*, vol. 87, no. 3, pp. 316–336, 2010.

[4] D. Gálvez-López and J. D. Tardós, "Real-time loop detection with bags of binary words," in *Proc. of IROS*, 2011, pp. 51–58.

[5] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," in *Proc. of MM*, 2012.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proc. of ICCV*, 2011, pp. 2564–2571.

[7] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. of CVPR*, 2009, pp. 1169–1176.

[8] S. McCann and D. G. Lowe, "Local naive bayes nearest neighbor for image classification," in *Proc. of CVPR*, 2012.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.


[1] http://web.stanford.edu/~dmchen/mvs.html

[2] We used first 10 bits to define 1024 ($= 2^{10}$) VWs, and the next 64 bits as substring.