An Alternative to IDF: Effective Scoring for Accurate Image Retrieval with Non-Parametric Density Ratio Estimation

Yusuke Uchida, Koichi Takagi, Shigeyuki Sakazawa
KDDI R&D Laboratories, Inc.
{*ys-uchida, ko-takagi, sakazawa*}*@kddilabs.jp*

Abstract

In this paper, we propose a new scoring method for local feature-based image retrieval. The proposed score is based on the ratio of the probability density function of an object model to that of background model, which is efficiently calculated via nearest neighbor density estimation. The proposed method has the following desirable properties: (1) a sound theoretical basis, (2) effectiveness than IDF scoring, (3) applicability not only to quantized descriptors but also to raw descriptors, and (4) ease and efficiency of calculation and updating. We show the effectiveness of the proposed method empirically by applying it to a bag-of-visual words-based framework and a k-nearest neighbor voting framework.

1. Introduction

With the advances in both stable interest region detectors [6] and robust and distinctive descriptors [5], local feature-based image or object retrieval has attracted significant attention. It has also become applicable to large-scale databases owing to the bag-of-visual words (BoVW) framework [12]. In the BoVW framework, local feature points or regions are first detected in an image, then feature descriptors are extracted from them. These feature vectors are quantized into visual words (VWs) using a visual codebook, resulting in a histogram representation of VWs. In many cases, image similarity is measured by the L_2 distance between the normalized histograms. As the histograms are generally sparse, an inverted index data structure and a voting function enables an efficient similarity search. The equivalency between L_2 distances and scores obtained with the voting function is described in [3] in detail. In order to emphasize distinctive VWs, the inverse document frequency (IDF) scoring [12] has been widely used, and shown to be effective.

Though the BoVW framework realizes efficient retrieval, some degradation of accuracy is caused by quantization [2]. Two major approaches are proposed to alleviate quantization error: post-filtering approaches [3, 4, 14] and multiple assignment approaches [10, 3, 7]. In the post-filtering approaches, after the VW-based matching, unreliable matches are filtered out according to the estimated distances between query and reference descriptors. In the multiple assignment approaches, a query descriptor can be matched not only with reference descriptors in the nearest VW, but also with reference descriptors in the several nearest VWs. Although quantization error is alleviated, all the above methods still depend on IDF scoring in voting, which is designed for words or quantized descriptors. In other words, the score is still *quantized*.

In this paper, we propose a new scoring method applicable to both quantized and unquantized descriptors. The proposed score is based on the ratio of the probability density function of an object model to a background model, which is efficiently calculated in an on-the-fly manner via nearest neighbor density estimation. In experiments, we show the effectiveness of the proposed scoring method by applying it to a BoVW framework and a k-nearest neighbor voting framework.

2. Proposed Approach

In this section, we first present the formulation of the proposed scoring method, starting with a classification problem. Then, in order to make it applicable to large-scale image retrieval, an approximation is introduced. Finally, the score is calculated via non-parametric density ratio estimation.

2.1. Probabilistic formulation

Given a query image Q, the objective is to find a similar image $R_{\hat{j}}$ from a large number of reference images $R_1, \cdots, R_{n_C}$. Considering it as a classification

problem, we start with maximum-a-posteriori estimation: $\hat{j} = \arg\max_j p(R_j|Q)$. Assuming $p(R_j)$ is uniform, the maximum-a-posteriori estimation reduces to a maximum likelihood estimation:

$$\hat{j} = \arg\max_j p(R_j|Q) = \arg\max_j p(Q|R_j). \quad (1)$$

Letting $\mathcal{Q} = \{q_1, \cdots, q_n\}$ denote the descriptors of the query image Q, with the naive Bayes assumption, we get:

$$p(Q|R_j) = p(q_1, \cdots, q_n|R_j) = \prod_{i=1}^{n} p(q_i|R_j). \quad (2)$$

As pointed out in [2], if we assume all query descriptors are derived from only the object model of R_j, $p(Q|R_j)$ tends to be too small even if Q and R_j share the same object. In [2], the problem is alleviated by estimating $p(q_i|R_j)$ using a few dozen images representing the same class. As this is not practical for large-scale image or object retrieval, we model $p(q_i|R_j)$ by a mixture of the object model of R_j and a background model distinct from R_j:

$$p(q_i|R_j) = \lambda p(q_i|\mathcal{R}_j) + (1-\lambda)p(q_i), \quad (3)$$

where $\mathcal{R}_i$ denotes a set of descriptors in the reference image R_j. If we consider the descriptors $\mathcal{Q}$ and $\mathcal{R}_j$ as words, this is identical to LM (language modeling)-RSV [11] in the area of information retrieval (IR). Combining Eqs. (1)–(3), we obtain:

$$\begin{aligned}\hat{j} &= \arg\max_j \prod_{i=1}^{n} p(q_i|R_j) = \arg\max_j \sum_{i=1}^{n} \log p(q_i|R_j) \\ &= \arg\max_j \sum_{i=1}^{n} \log(\lambda p(q_i|\mathcal{R}_j) + (1-\lambda)p(q_i)) \\ &= \arg\max_j \sum_{i=1}^{n} \log(\frac{\lambda}{1-\lambda}\frac{p(q_i|\mathcal{R}_j)}{p(q_i)} + 1). \quad (4)\end{aligned}$$

Finally, we get the voting score s_{ij}:

$$s_{ij} = \log(\frac{\lambda}{1-\lambda}\frac{p(q_i|\mathcal{R}_j)}{p(q_i)} + 1). \quad (5)$$

For each q_i, the voting score s_{ij} is assigned to each R_j. The resulting $\sum_i s_{ij}$ corresponds to the similarity measure between Q and R_j.

2.2. Approximation with nearest neighbors

In the above formulation, it is required to calculate s_{ij} for all R_j. Similarly, $\min_{r\in\mathcal{R}_j} ||q_i - r||^2$ should be calculated for all R_j in [2]. Letting n_C denote the number of classes and n_D denote the average number of descriptors in an image, the the calculation of s_{ij} for all R_j has a time cost of $O(n_C \cdot \log(n_D))$ with efficient (approximate) nearest neighbor search algorithms [8, 1, 4]. This does not become a fatal flaw in classification problems where $n_D \gg n_C$. However, it is intractable in large-scale image retrieval problem where $n_C \gg n_D$ because n_C corresponds to the number of images or objects in a database. In order to make it tractable, the following simple approximation is adopted. We assume the nearest neighbor descriptors $\mathcal{D}(q_i)$ of q_i (e.g., k-nearest neighbors of q_i) were obtained against all reference descriptors. Then, $p(q_i|\mathcal{R}_j)$ is calculated only for R_j at least one of whose descriptors appears in $\mathcal{D}(q_i)$, and otherwise we assume $p(q_i|\mathcal{R}_j) = 0$. Because the voting score s_{ij} becomes 0 if $p(q_i|\mathcal{R}_j) = 0$, the voting is performed efficiently. With this approximation, the computational cost is reduced from $O(n_C \cdot \log(n_D))$ to $O(\log(n_C \cdot n_D))$.

2.3. Non-parametric density ratio estimation

Finally, the voting score s_{ij} is calculated using $\mathcal{D}(q_i)$. We assume $\mathcal{D}(q_i)$ can be decomposed into m disjoint sets $\mathcal{D}_1(q_i), \cdots, \mathcal{D}_m(q_i)$:

$$\mathcal{D}(q_i) = \bigcup_{t=1}^{m} \mathcal{D}_t(q_i),\ \mathcal{D}_t(q_i) \bigcap \mathcal{D}_{s\neq t}(q_i) = \emptyset. \quad (6)$$

We also assume that these disjoint sets are ordered:

$$\mathcal{D}_1(q_i) > \mathcal{D}_2(q_i) > \cdots > \mathcal{D}_m(q_i), \quad (7)$$

so that they satisfy

$$t < s \ \Leftrightarrow\ p(q_i|r \in \mathcal{D}_t(q_i)) > p(q_i|r \in \mathcal{D}_s(q_i)). \quad (8)$$

For each $\mathcal{D}_t(q_i)$ ($1 \le t \le m$), and for each $\mathcal{R}_j$ one of whose descriptors appears in $\mathcal{D}_t(q_i)$, the densities $p(q_i|\mathcal{R}_j)$ and $p(q_i)$ in Eq. (5) are estimated via k-nearest neighbor density estimation:

$$p(q_i|\mathcal{R}_j) = \frac{n_{tj}}{|\mathcal{R}_j| \cdot V_t},\ p(q_i) = \frac{\sum_{s=1}^{t} |\mathcal{D}_s(q_i)|}{|\mathcal{R}_{all}| \cdot V_t}, \quad (9)$$

where n_{tj} is the number of descriptors of R_j that appear in $\mathcal{D}_t(q_i)$, $\mathcal{R}_{all}$ is all reference descriptors $\bigcup_j \mathcal{R}_j$, V_t is the volume of a hypersphere with radius $\sqrt{||q_i - \hat{r}_t||^2}$, and $\hat{r}_t \in \mathcal{D}_t(q_i)$ is the farthest descriptor from q_i. Combining Eqs. (5) and (9), we obtain:

$$s_{ij} = \log(\frac{\lambda}{1-\lambda}\frac{n_{tj} \cdot |\mathcal{R}_{all}|}{\sum_{s=1}^{t} |\mathcal{D}_s(q_i)| \cdot |\mathcal{R}_j|} + 1). \quad (10)$$

More concrete examples of the formulation are shown in the following section. One advantage of this scoring method is that the up-to-date score is efficiently calculated in an on-the-fly manner using $\mathcal{D}(q_i)$, even if the database is modified. The only requirement is to store the number of descriptors $|\mathcal{R}_j|$ in each reference image.

3. Experimental evaluation

In this section, we show the effectiveness of the proposed scoring method by applying it to the BoVW and

k-nearest neighbor (k-NN) voting frameworks.

3.1 Experimental setup

Experiments were performed on the University of Kentucky recognition benchmark dataset[1] provided by the authors of [9]. It includes 2,550 different objects or scenes. Each of these objects is represented by four images taken from four different angles, making 10,200 images in all. These images are used as both reference and query images. Mean average precision (MAP) [9, 3] is used as an indicator of retrieval performance. A visual codebook with size 20,000 is created using a dataset distinct from the images introduced above.

3.2. BoVW framework

The proposed scoring method will now be applied to the BoVW framework. In this case, $\mathcal{D}(q_i)$ is defined as a set of reference descriptors that are quantized into the same VW as q_i. This consists of only a single disjoint set ($m = 1$):

$$\mathcal{D}(q_i) = \mathcal{D}_1(q_i) = \{r \in \mathcal{R}_{all} \mid \mathrm{q}(r) = \mathrm{q}(q_i)\}, \quad (11)$$

where $\mathrm{q}(r)$ and $\mathrm{q}(q_i)$ denote the identifiers of the corresponding VWs of r and q_i after quantization. Then, s_{ij} is calculated using frequencies of VWs:

$$s_{ij} = \log\left(\frac{\lambda}{1-\lambda}\frac{tf_j^{\mathrm{q}(q_i)} \cdot |\mathcal{R}_{all}|}{tf_{all}^{\mathrm{q}(q_i)} \cdot |\mathcal{R}_j|} + 1\right), \quad (12)$$

where tf_j^w represents the frequency of the w-th VW in R_j, and tf_{all}^w the frequency of the w-th VW in all reference images.

Figure 1 shows the MAP scores obtained with different scoring methods as a function of λ. The voting score used for each method is as follows: tf for Baseline, $tf \cdot idf$ for IDF, and s_{ij} for density ratio estimation (DRE), respectively. It is clear that DRE achieves better performance than Baseline and IDF, and the accuracy is not so sensitive to the choice of λ. It can be said that the DRE scoring method is effective even if applied to the pure BoVW framework. The best MAP score of 0.811 is achieved with relatively low λ ($\lambda = 0.06$), which implies that there are a small number of features useful for object recognition [13].

3.3. Exact k-NN voting framework

The proposed scoring method is applied to the k-NN voting framework, where (approximate) k-nearest neighbor descriptors of q_i are extracted from all reference descriptors, and the corresponding reference images obtain the corresponding scores [4]. In this case, $\mathcal{D}(q_i)$ is defined as a set of the k-nearest neighbors of q_i, and $\mathcal{D}_t(q_i)$ contains only the t-th nearest neighbor: $m = k$, $n_{tj} = 1$, and $\sum_{s=1}^{t} |\mathcal{D}_s(q_i)| = t$. Finally, s_{ij} becomes very simple:

$$s_{ij} = \log\left(\frac{\lambda}{1-\lambda}\frac{|\mathcal{R}_{all}|}{t \cdot |\mathcal{R}_j|} + 1\right). \quad (13)$$

Although *exact* k-NN search requires impractical retrieval time, it gives us the upper bound performance of the proposed method, which is useful in the evaluation of the *approximate* k-NN-based system. Therefore, we first evaluate the exact k-NN version. Figure 2 shows the comparison of scoring methods in the exact k-NN voting framework. The fixed voting score 1.0 is used for Baseline because the IDF scoring is not applicable, and s_{ij} with $\lambda = 0.1$ is used for DRE. The best MAP scores of 0.912 and 0.902 are achieved by DRE with $k = 20$ and by Baseline with $k = 10$.

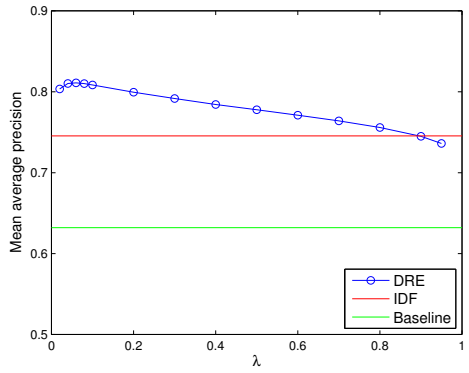

Figure 1: Comparison of scoring methods in the BoVW framework.

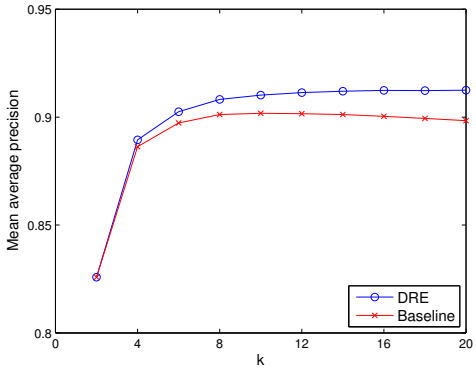

Figure 2: Comparison of scoring methods in the exact k-NN voting framework.

[1] http://www.vis.uky.edu/~stewe/ukbench/

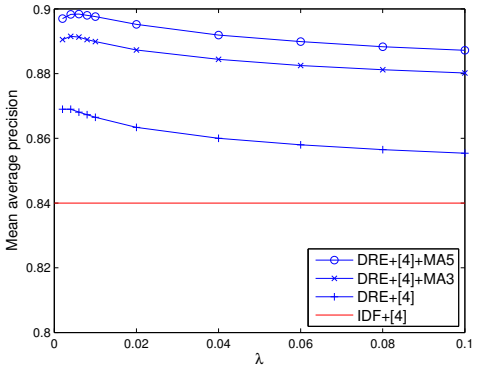

Figure 3: Comparison of scoring methods in the approximate k-NN voting framework.

3.4. Approximate k-NN voting framework

Finally, we evaluate the proposed scoring method in combination with the state-of-the-art product quantization-based approximate k-NN search method [4]. The parameters recommended in [4] are used[2]. The voting score of the proposed method is calculated by Eq. (13). The IDF scoring is also applicable because q_i is quantized in the process of the approximate k-NN search[3]. Figure 3 shows the comparison of scoring methods in the approximate k-NN voting framework. It is shown that the accuracy is significantly improved by using the proposed scoring method instead of the IDF scoring. In the experiment, we chose the best k ($k = 6$) for the IDF scoring. For the proposed scoring, we used an adaptive k depending on the frequency $tf_{all}^{\mathrm{q}(q_i)}$ of the assigned VW $\mathrm{q}(q_i)$ as $k = \sqrt{tf_{all}^{\mathrm{q}(q_i)}}$. This choice was inspired by the rule of thumb in k-NN density estimation: use the $\sqrt{n}$ nearest samples out of n samples in estimation. This slightly improved the best accuracy of DRE+[4] from 0.866 ($k = 14$) to 0.869, and even better, it frees us from a difficulty in choosing k.

The DRE method is also evaluated in combination with multiple assignment (MA) [10] for different numbers of assignments (3 and 5). It is shown that DRE with 5 MA achieves a MAP score of 0.898 with $\lambda = 0.006$, which is a satisfactory result compared with the MAP score of 0.912 obtained by the exact k-NN search.

4. Conclusion

In this paper, we have proposed a new scoring method for local feature-based image retrieval, which is based on the ratio of the probability density function of an object model to that of a background model. The effectiveness of the proposed method was confirmed by applying it to the bag-of-visual words-based framework and the k-NN voting framework. The proposed method can also be applicable to hierarchical vocabulary [9] or learned vocabulary [7], where the matched descriptors can be ordered. In the future, kernel density estimation can be used for more accurate density estimation.

[2]We set the size of the codebooks for product quantization $k^* = 256$ and the number of vector decomposition $m = 8$.

[3]We adopted the non-exhaustive version of [4] called IVFADC.

References

[1] A. Andoni. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proc. of FOCS*, pages 459–468, 2006.

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proc. of CVPR*, pages 1–8, 2008.

[3] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, 2010.

[4] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. on PAMI*, 33(1):117–128, 2011.

[5] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on PAMI*, 27(10):1615–1630, Oct. 2005.

[6] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *IJCV*, 60(1-2):43–72, Nov. 2005.

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

[8] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *Proc. of VISAPP*, pages 331–340, 2009.

[9] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of CVPR*, pages 2161–2168, 2006.

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

[11] T. Roelleke and J. Wang. Tf-idf uncovered: A study of theories and probabilities. In *Proc. of SIGIR*, pages 435–442, 2008.

[12] J. Sivic and A. Zissermane. Video google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, pages 1470–1477, 2003.

[13] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features. In *Proc. of WS-LAVD*, 2009.

[14] Y. Uchida, M. Agrawal, and S. Sakazawa. Accurate content-based video copy detection with efficient feature indexing. In *Proc. of ICMR*, 2011.