SELECTING REPRESENTATIVE AND DISTINCTIVE DESCRIPTORS FOR EFFICIENT LANDMARK RECOGNITION

Sheng Gao, Joo-Hwee Lim

SELECTING REPRESENTATIVE AND DISTINCTIVE DESCRIPTORS FOR EFFICIENT LANDMARK RECOGNITION

Sheng Gao and Joo-Hwee Lim
Institute for Infocomm Research (I^{2}R), A-Star, Singapore, 119613
{gaosheng, joohwee}@i2r.a-star.edu.sg

ABSTRACT

To have a robust and informative image content representation for image categorization, we often need to extract as many as possible visual features at various locations, scales and orientations. Thus it is not surprised that an image has a few hundreds or even thousands of visual descriptors. This raises huge cost of computation and memory. To eliminate the problem, we can only select the most representative and distinctive descriptors and discard the other non-informative features when training the image category models. This paper will present a Markov chain based algorithm to learn a measure of the descriptor importance in order to weigh the degree of representativeness and distinctiveness. From the measures the descriptor selection algorithm is derived. The presented approach starts from constructing a graph with each node being a descriptor to characterize the pair-wise descriptor similarity and then the PageRank algorithm is exploited to estimate the stationary distribution of the graph whose values are the indicator of the descriptor importance. We evaluate the proposed approach on the STOIC-101 landmark dataset. Our experiments demonstrate the Markov chain based descriptor selection can select the most informative descriptors to distinguish the landmarks. Even with the large reduction of the size of descriptors, the classification accuracy is still competitive or overcomes compared with the system without any descriptor selection.

Index Terms – scene recognition, Markov chain, classification accuracy, PageRank, sample selection

1. INTRODUCTION

In image categorization such as generic object recognition, scene recognition, etc., we often need to extract sufficient visual descriptors for characterizing the image category. To retain as much as possible information and to have a robust representation for scales, rotation, transformation, clutter, etc., the descriptors are always extracted from the various locations, scales and orientations. It is typical to find an image is characterized by a few hundreds or even thousands of visual features. For example, in the paper the average size of SIFT descriptors [1] is ~655 for a 320x240 image. Therefore, for thousands of images in the training set, there will be millions of descriptors. The large size of features not only raises huge cost of computation and memory but also may cause the learned image category models bias due to the dependency among the descriptors.

In most of learning algorithms, it is often assumed that the samples (image or descriptor) are drawn independently and uniformly from the sample space. For instance, in the bag-of-words based image representation, when we build a visual codebook using the k-means clustering on the pooled descriptors, we assume each descriptor is independent and it is uniformly sampled, i.e. each descriptor having equal contribution in the objective function of k-means clustering. However, the assumption has the defects in practice. For example, in [8] the authors discuss that the cluster centers of k-means are drawn irresistibly towards denser regions of the sample distribution and thus the clusters tend to be tightly clustered near the dense regions and sparsely spread in the sparse. The method in [8] addresses the defect using an undersampling framework combined with on-line clustering and mean-shift. Besides of the undersampling method, the sample bias can be tackled using the unlabeled data (e.g. [11]) to correct the sample distribution of the training set. In the approach, the unlabeled data is used to weigh the samples in the training set so that the sample distribution of the training set and the test set will match. The approach breaks the uniform assumption about the training samples.

In the paper, we present a Markov chain based algorithm to learn a measure of the descriptor importance in order to weigh the degree of representativeness and distinctiveness and then apply the measure to choose the most representative and distinctive descriptors for landmark recognition. Firstly, a graph is constructed based on the pair-wise similarity analysis among the descriptors. The graph node is a descriptor while the node links to the other are sought using the k-nearest neighbor method. After the graph is ready, secondly, we exploit the PageRank algorithm to estimate a stationary distribution of the graph [9] and the PageRank scores are the indicator of the descriptor importance in the dataset. Based on the descriptor importance distribution, we can derive the informative measure for weighing the degree of representativeness and distinctiveness of the descriptors for characterizing the landmarks. Two schemes are developed in the paper for descriptor selection (See Section 2): one is to measure the degree only using the category-specific descriptor importance and another is to measure the degree combing the category-specific importance, which is estimated from the descriptors from one category, and category-free importance, which is estimated from the descriptors from all categories. The selected descriptors can be used directly for classification, for instance, using the k-NN classifier, or be used as a pre-stage for some complex training algorithms (See Section 3).

The proposed method is a novel application of Markov chain theory to weigh the importance of descriptors based on the pair-wise dependency analysis of descriptors and the link structure analysis of the graph. The estimated distribution of the degree of representativeness and distinctiveness of descriptors has many applications. For example, we can do the sample selection based on the measure. But this selection is one time rather than multiple iterative steps like the undersampling in [8]. It is thus cheap. The learned distribution of descriptors can be integrated into training the classifier like in [11]. But it is superior to [11] in that the weights are estimated only from the training set without the need of the extra unlabelled dataset, which is often the test set.

It is well-known that the Markov chain is applied for measuring the popularity of web documents, e.g. PageRank in Google. We find that there are many research works on extending PageRank to other applications such as feature selection [10], word sense disambiguation [12], and visual image search [13], etc. In these applications, the link structures are not natural available, which is quite different from the web documents where the links are ready. So the first step is to generate the link structure by analyzing the similarity among the elements. After that the similar operation is done as scoring and ranking the web pages.

In the paper, we only evaluate the descriptor selection for landmark recognition on the STOIC-101 dataset, which consists of 101 attractive landmarks in Singapore and 3,539 images totally.

The paper is organized as follows. In Section 2, we will present the proposed Markov chain based descriptor selection algorithm. In Section 3, we report the experimental results on multiclass landmark recognition. Finally, we conclude our findings in Section 4.

2. MARKOVE CHAIN BASED VISUAL DESCRIPTOR SELECTION

In millions of descriptors extracted from the training images, it is expected that high dependency among the descriptors should exist. It is not effective and efficient to use all of them to train the image category models. In fact, it is not necessary. We can only select the most representative and distinctive visual descriptors in training the image category models. The question is how to measure the quantity of representativeness and distinctiveness. In the section, we introduce the Markov chain based algorithm for measuring the importance of descriptors in characterizing the image category content. Then the descriptor selection method can be derived from the measures.

2.1. Weighing Importance of Visual Descriptors

Similar to PageRank [9], where the random surfer visiting the webpage document is modeled by Markov chain, here we use the Markov chain to model the visual descriptor selection. In the chain, each state is a visual descriptor and its links are sought by comparing the pair-wise descriptor similarity and being determined based on the *K*-nearest neighbor method. If we assume the set has *N* descriptors and adjacency matrix, $A_{N\mathrm{x}N}$, to characterize the structure of descriptor relations. The jump from one state to another is controlled by the transition probability matrix, $P_{N\mathrm{x}N}$, where the $(\mathrm{i,j})$-th element measuring the probability jumping from the *i-th* state to the *j-th* state. The stationary distribution of the Markov chain is the probability distribution, $Q=\left[q(x_1),q(x_2),\cdots,q(x_N)\right]$. $q(\cdot)$ is a probability of a descriptor being visited by the sampler. The distribution is the eigenvector of the following equation as,

$$Q = P^T Q \tag{1}$$

In practice, a residual probability, *d*, is often added into Eq.(1). So the stationary distribution is the solution of modified Eq. (1) as,

$$Q = \frac{1-d}{N} \cdot I + \left(1-d\right) \cdot P^T Q \qquad (2),$$

where *I* is *N*-dimensional vector filling with 1 in all elements. Eq. (2) is a penalized version of Eq. (1), i.e. a uniform prior probability is added in each node. Its solution is found using the iterative algorithm used in PageRank [9].

The stationary distribution is a good measure of descriptor's importance in the database. The descriptor's importance measures how many descriptors are pointed to it and are similar to it. If the descriptor has a lot of links, it implies that the descriptor is representative and it encodes richer information than the other descriptors with the lower importance. Thus, rather than using all descriptors, we can summarize the content of the dataset by only retaining the most informative descriptors whose importance values are higher than the threshold.

2.2. Measuring Representativeness and Distinctiveness

As the estimation of descriptor importance in a dataset is solved, we can exploit the measure in order to derive a quality to weigh the degree of representativeness and distinctiveness of a visual descriptor for charactering the content of image category. Based on the measure, we can tell which descriptor is representative for a category and which one is distinctive for discriminating one category from another.

We derive the idea from the weighing of the word terms in the text document using the tf-idf method, i.e. *term frequency - inverse term frequency*, for information retrieval [15]. Term frequency is the probability of a term occurred in a document while inverse term frequency is a measure of the importance of the term across the documents, which is calculated from the number of documents containing the term and the total document number in the set. The former is a measure of representativeness of the term and the latter is a measure of distinctiveness or discrimination of the term. The higher the term frequency is in a document, the more representative the term is for the document. Similarly, the term which has the higher inverse term frequency is more distinctive and holds much more capability of discrimination to distinguish one document from another.

In the context of image categorization, the document may be an image or an image category which have a lot of visual descriptors to characterize their content. In the paper, we would like to measure the descriptors in the category, i.e. a category as a document. The *term* refers to the visual descriptor. We first run PageRank on the pooled visual descriptors from one category, saying the *i-th* category, and obtains the category-specific importance vector, $Q_i = \left[q_i\left(x_1^i\right), q_i\left(x_2^i\right), \cdots, q_i\left(x_{N_i}^i\right) \right]$, where x_j^i is the *j-th* descriptor in the *i-th* category and N_i is the size of descriptors. $q_i\left(x_j^i\right)$ is the popularity of the term x_j^i in the *i-th* category and will be treated as the term frequency. After the term frequency is obtained, we then run PageRank on the pooled descriptors from all categories to get the category-free importance vector,

$$Q = \left[q\left(x_1^1\right), \cdots, q\left(x_{N_1}^1\right), q\left(x_1^2\right), \cdots, q\left(x_{N_2}^2\right), \cdots, q\left(x_1^C\right), \cdots, q\left(x_{Nc}^C\right) \right].$$

Each element is the term frequency of the descriptor in the whole dataset. Its inverse simulates the inverse term frequency.

Similar to tf-idf, we derive a measure from the category-specific importance and the category-free importance as in Eq. (3)

$$w_j^i = q_i\left(x_j^i\right) \big/ q\left(x_j^i\right) \qquad (3)$$

The value of w_j^i weighs the degree of representativeness and distinctiveness for the descriptor, x_j^i .

2.3. Visual Descriptor Selection

As the measure of representativeness and distinctiveness of descriptors is ready, the descriptor selection is cheap. For example, we can simply choose top-K descriptors, which have highest values in Eq. (3), for each category to summarize the category content. Due to the descriptor selection, the size of visual descriptors is decreased and huge computation cost is reduced. It speeds up the model training and it allows us to evaluate more complex and precise models for image categorization classification. More interestingly and importantly, sample selection will not worsen the classification accuracy, even with the half reduction of the training set. In contrast, we observe the accuracy is improved. It will be shown in Section 3.

3. EXPERIMENTS

In the section, we will evaluate the effect on the performance of the proposed visual descriptor selection algorithm in the image landmark recognition. The landmark dataset, *STOIC-101*, includes 101 landmarks in Singapore and has totally 3,539 images with the size of 240x320. The landmark images are captured in large variations of scale, clutter, and lighting condition. Some details can refer to [4, 7]. The dataset is further split into two parts: the sub-set A, i.e. *SA*, has 3,090 images and the sub-set B, i.e. *SB*, has 449 images.

128-dimensional SIFT descriptors are extracted using the detector developed by Low [1]. Then the eigenspace is estimated using the principle component analysis (PCA) on the *SA* set, which has 2,023,507 visual descriptors, and 56 eigenvectors are retained. So the 128-dimensional SIFT descriptor is mapped into the 56-dimensional vector.

3.1. Effect of Size of Selected Descriptors

We first study the effect of the size of selected descriptors on the classification accuracy. We use *SB* as the training set and *SA* as the test set. Rather than recognizing 101 landmarks, we only select 5 landmarks due to the memory and computation cost. Thus the small training set (*SB_T*) and test set (*SA_E*) are constructed using corresponding images in the 5 landmarks. It results in the *SB_T* having 23 images with 15,120 descriptors and the *SA_E* having 152 images with 99,248 descriptors. We investigate how the classification accuracy will be affected when the descriptors in the training set is decreased. The k-NN classifier is adopted for simplicity and we make the decision based on the voting.

Two sample selection schemes are discussed. The first is to select informative descriptors for a landmark only based the category-specific importance, and the second is to select descriptors based on the combination of the category - specific importance (See Eq. (2)) and the category-free importance (See Eq. (3)). Then the selected category-specific descriptors are used as the template exemplars in the k-NN classification. The baseline is the system without descriptor selection. The number of selected descriptors per landmark is 500 (2,500), 1,000 (5,000), and 2,000 (9,315), respectively. The number in the parentheses is the total size of selected descriptors for 5 landmarks (Note: some categories may have descriptors less than the threshold). We draw their classification accuracies in Fig.1 (Blue bar: first scheme; Red bar: second scheme, Yellow bar: baseline).

From the figure, we can conclude that: 1) combing the category-free importance with the category-specific importance obviously improves classification accuracy; The improvement is significant when the sample size is smaller, e.g. 34.7% relative accuracy improvement achieved for 500 descriptors per landmark and 10.4% for 1,000 descriptors per landmark; 2) As the number of selected descriptors is increasing, the classification accuracy is improving; 3) both schemes works better than the baseline system without sample selection; even with 500 descriptors per landmark, we can achieve 63.82%, which is much better than the baseline system having only 59.21% ; 4) the recognition speed is significantly increased due to the large reduction of sample size, e.g. ~1s for 500 descriptors vs. ~6s for the baseline.

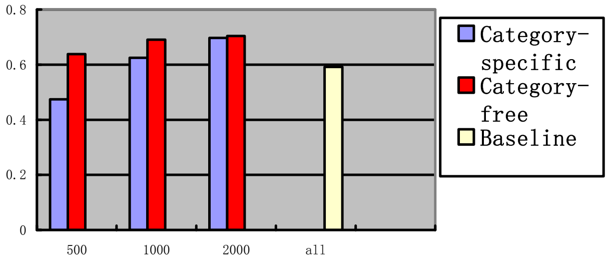

Fig.1 Classification accuracy for 2 selection schemes in various size of selected descriptors and the baseline

3.2. Large-scale Evaluation

Now we will report the evaluation results of the effect on the performance of the selected descriptor size in the whole 101 landmarks. The training set is the *SA* having 3,090 images and the test set is the *SB* having 449 images. Totally there are 2,023,507 descriptors in the training set, ~20,034 per landmark. The bag-of-words method is used to derive one high-dimensional vector per image. Then we apply the MC MFoM learning algorithm to train a linear classifier for maximizing macro-averaging F1 measure [6]. We estimate the landmark-specific visual codebook having 32 keywords per landmark and then the 101 landmark-specific codebooks are pooled into a universal codebook having 3,232 visual keywords. The universal codebook is used to quantize the descriptors which results in a normalized 3,232-dimensional histogram. The MC MFoM based classifier is trained on the bag-of-words image representation.

Considering about 2 millions of descriptors in the training set, calculating the category-free importance of descriptors will consume huge memory and computation cost. In the paper, we only report the results on the first scheme (See sub-section 3.1) of descriptor selection. The size of selected descriptors is 1,000, 5,000 and 10,000 per landmark based on the category-specific descriptor importance. Then the visual keywords are learned from the selected descriptors using k-means clustering. The baseline system is trained without any descriptor selection. Table 1 summarizes their classification accuracy.

Table 1 Classification accuracy on 101 landmarks

	1,000	5,000	10,000	Baseline
Acc(%)	62.14	74.83	76.84	75.95

Similar findings are observed from Table 1 in the large-scale evaluation. In case of only keeping about half size of descriptors, i.e. 10,000, the classification accuracy still overcomes the baseline. The accuracy is improving with the increasing size of selected samples.

4. CONCLUSION

In the paper we presented a Markov chain based algorithm to learn the importance of descriptors by analyzing the pair-wise similarity and building a graph to characterize the dependency among the descriptors. The descriptor importance is estimated using the PageRank algorithm. We study the application of category-specific importance and category-free importance in measuring the degree of representativeness and distinctiveness of descriptor and the application in descriptor selection for landmark recognition. Our experiments on the small and large scale sets demonstrate that the proposed approach not only reduces the size of descriptors in the training set, but also improves the classification accuracy due to informative features being selected while the noisy features being compressed.

5. REFERENCES

[1] D.G. Lowe, “Object recognition from local scale-invariant features”, Proc. of ICCV’99.

[2] G. Csurka, et al., “Visual categorization with bags of keypoints”, Proc. of Workshop on Statistical Learning in Computer Vision, ECCV’04.

[3] J. Farquhar, et al., “Improving bag-of-keypoints image categorization”, Technical report, University of Southampton, 2005.

[4] J.-H. Lim, et al., “Scene recognition with camera phones for tourist information access”, Proc. of ICME’07.

[5] R. O. Duda, P. E. Hart & D. G. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[6] S. Gao, et al., “A MFoM learning approach to robust multiclass multi-label text categorization”, Proc. of ICML’04.

[7] Y. Q. Li & J.-H. Lim, “Outdoor place recognition using compact local descriptors and multiple queries with user verification”, Proc. of ACM Multimedia’07 (short paper).

[8] F. Jurie & B. Triggs, “Creating efficient codebooks for visual recognition”, Proc. of ICCV’05.

[9] S. Brin & L. Page, “The anatomy of a large-scale hypertextual web search engine”, Computer Networks and ISDN Systems, 30, (1998).

[10] D. Ienco, et al., “Using PageRank in feature selection”, Proc. of Italian Symposium on Advanced Database Systems, 2008.

[11] J.Y. Huang, et al., “Correcting sample selection bias by unlabeled data”, Proc. of NIPS’07.

[12] R. Mihalcea, “Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling”, Proc. of HLT/EMNLP’05.

[13] Y. Jing & S. Baluja, “VisualRank: applying PageRank to large-scale image search”, IEEE Trans. on PAMI, pp.1877-1890, Vol.30, No.11, 2008.

[14] Ricardo B. Y. & Berthier R.-N., Modern Information Retrieval, Addison Wesley, 1999.