BUILDING PAIR-WISE VISUAL WORD TREE FOR EFFICENT IMAGE RE-RANKING

Shiliang Zhang[1], Qingming Huang[2], Yijuan Lu[3], Wen Gao[1], Qi Tian[4]

[1]Key Lab of Intelli. Info. Process., Inst. of Comput. Tech., CAS, Beijing 100080, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Computer Science, Texas State University, San Marcos, TX, 78666
[4]Department of Computer Science, University of Texas at San Antonio, TX, 78249
{slzhang, qmhuang, wgao}@jdl.ac.cn, yl12@txstate.edu, qitian@cs.utsa.edu

ABSTRACT

Bag-of-visual Words (BoW) image representation is getting popular in computer vision and multimedia communities. However, experiments show that the traditional BoW representation is not as effective as it is desired. One of the most important reasons for its ineffectiveness is that, the traditional BoW representation lost the spatial information in images. To overcome this problem, we propose the pair-wise visual word tree, within which each visual word keeps both the appearance and spatial information between two interest points in image. Thus, the corresponding novel BoW representation preserves the spatial structure in image. Based on the pair-wise visual word tree, we propose an efficient topic word selection algorithm, which utilizes the Latent Semantic Analysis to discover the most expressive visual words for different image categories. An efficient strategy is then utilized to combine the selected topic words for image re-ranking. Massive experiments show that the novel BoW representation shows promising performance. Meanwhile, the proposed image re-ranking strategy shows the state-of-the-art precision and promising efficiency.

Index Terms—Image Analysis, Image Processing

1. INTRODUCTION

Bag-of-visual Words (BoW) representation is popular in multimedia and vision tasks, including video event detection [1], object recognition [2, 3], large-scale image retrieval [4-6], *etc*. Traditionally, a visual vocabulary is trained by clustering a large number of local feature descriptors. The exemplar descriptor of each cluster is called a visual word. In previous works, various numbers of visual words are generated for BoW representation. There are two observations [2-5]: 1) more visual words generally results in better performance; 2) the performance will be saturated when the number of visual words reaches certain levels. Intuitively, larger number of visual words indicates more fine-grained partitioning of the descriptor space. Hence the visual words become more distinctive. However, the second observation strongly implies that the descriptive ability of a visual word is limited, no matter how fine-grained it is.

Fig. 1. Traditional BoW representation lost the spatial clues

One shortcoming exists in the traditional visual words, which might be an important reason for their limited descriptive power. The traditional visual words only keep the appearance information in the small local image patches around the interest points [7]. Thus, the corresponding BoW representation lost most of the spatial information in images. A toy example showing this shortcoming is illustrated in Fig.1. In the figure, the two images show different semantics. Unfortunately, since the visual word lost the spatial cues, their BoW representations (*i.e.*, the visual word histograms) are nearly identical.

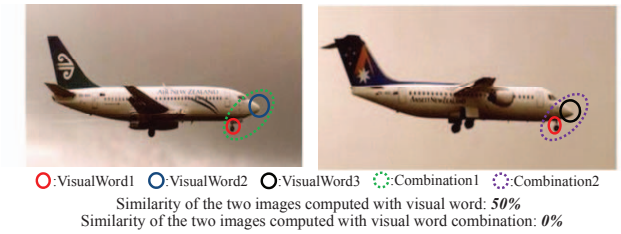
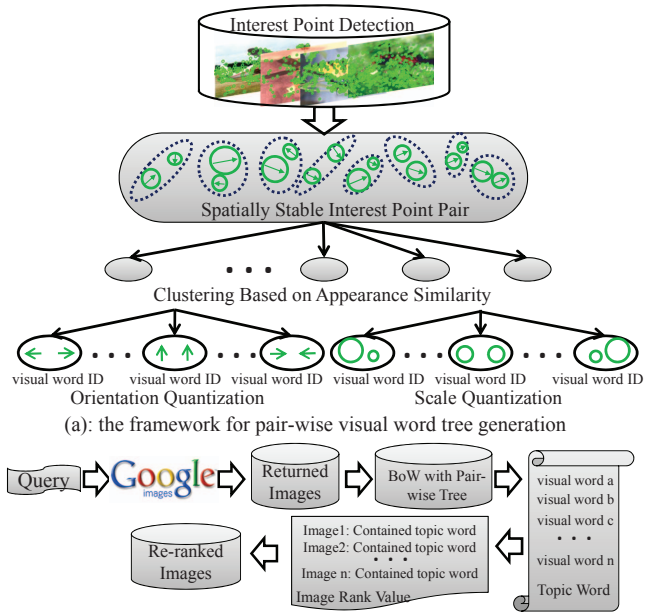


Fig. 2. The illustration of the magnified quantization error

Aiming at this problem, lots of works have been conducted to capture the spatial configuration between visual words [2, 3, 4, 6, 8, 9]. In general, this is achieved by identifying visual word combinations sharing stable spatial relationships. *E.g.*, in [2] the authors select the most discriminative visual word pairs for object recognition; visual word correlogram and correlation are utilized for object recognition in [3]. However, since quantization error exists in visual words, such error could be magnified when combining several visual words together. This results in the low repeatability of the generated combinations. As shown

in Fig.2, because of the quantization error, the similar image patches can be assigned with different visual words. In the visual word combinations, such errors are magnified, making these combinations invalid in representing the similarity between the two images. Several works are reported trying to suppress the quantization error and improve the repeatability of visual word combinations. *E.g.*, Visual Phrase is proposed in [8] by grouping semantically similar visual word pairs together; Visual Synset is proposed as semantically similar visual phrases in [9]. These methods, although show promising performance in specific tasks, are expensive to compute and are not scalable.



(a): the framework for pair-wise visual word tree generation

(b): the framework for the proposed image re-ranking

Fig. 3. The framework of the proposed algorithms

To generate visual words preserving both appearance and spatial information, pair-wise visual word tree is proposed in this paper. As illustrated in Fig. 3(a), pair-wise visual word tree is built by clustering a large number of Spatially Stable Interest Point Pairs (SSIPPs), which are defined as the co-occurred interest points within a constrained spatial distance. Since each interest point pair contains the appearance (*i.e.*, the feature descriptors of the interest points) and spatial clues (*i.e.*, the scale and orientation relationships between the two interest points), the generated visual word is designed to be more descriptive. Topic words are defined as the visual words most related with the semantics of certain image category. As illustrated in Fig. 3(b), with the pair-wise visual word tree, the images returned from the image search engines are represented as BoW representations. Then, the words most related with the query (*i.e.*, the topic words) are identified. Finally, the rank of each image is computed with their contained topic words.

The contributions of our work can be summarized as: 1) the pair-wise visual word tree is proposed. The BoW image representation preserving both the appearance and spatial information can be efficiently achieved with it. 2) The proposed topic word based image re-ranking shows the state-of-the-art precision with promising efficiency.

The rest of the paper is organized as follows. Section 2 introduces the pair-wise visual word tree generation. Topic word based image re-ranking is presented in Section 3. Section 4 discuses our experiments. The paper is finally concluded in Section 5.

2. PAIR-WISE VISUAL WORD TREE GENERATION

As illustrated in Fig. 3(a), before the pair-wise visual word tree generation, we use the DoG (Difference of Gaussian) [7] to detect interest points. From each interest point, we extract the information shown in Fig. 4. Each interest point is denoted as $P(S, D, O)$. More details about the interest point detection, scale and orientation can be found in [7]. Based on the interest points, we proceed to introduce the SSIPP detection and pair-wise visual word tree generation.

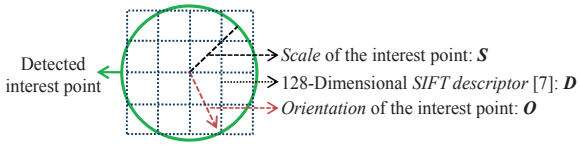


Fig. 4. The extracted information form interest point [7]

2.1. Spatially Stable Interest Point Pair Detection

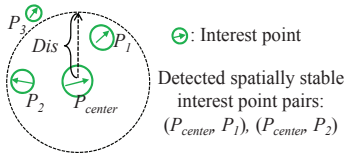


Fig. 5. The spatially stable interest point pair detector

We utilize the detector illustrated in Fig. 5 to identify the SSIPPs. In the figure, a circle with radius *Dis* is centered at an interest point. Each interest point within the circle composes a SSIPP with the centered interest point. The detected SSIPP is denoted as: (P_a, P_b), where P_a and P_b stand for the corresponding two interest points. The *Dis* is computed with Eq. 1 to achieve scale-invariance.

$$Dis = S_{center} \bullet \lambda \quad (1)$$

where S_{center} is the scale of the centered interest point, λ is a parameter controlling the constraint of co-occurrence. Larger λ is necessary for identifying stable spatial relations and overcoming the sparseness of the interest point pairs. However, large λ also increases the computational cost and the occurrence of noise. We experimentally set λ as 6, a good trade-off between efficiency and performance.

By scanning each interest point with the detector, a collection of SSIPPs can be generated. Because of the limited interest point number in each image and the properly selected λ , this process can be finished efficiently.

2.2. Pair-wise Visual Word Tree Generation

With the collected SSIPPs, the pair-wise visual word tree is generated with two steps: appearance based clustering and pair-wise spatial relationship quantization.

2.2.1. Clustering based on appearance similarity

To cluster the SSIPPs with their appearance clues, we first define the appearance similarity between two SSIPPs as

$$\mathrm{Sim}(I,J)=\mathrm{Min}\left(\underbrace{\left(\mathrm{M}(D_I^a,D_J^a)+\mathrm{M}(D_I^b,D_J^b)\right)}_{\mathrm{O1}_{I,J}},\underbrace{\left(\mathrm{M}(D_I^a,D_J^b)+\mathrm{M}(D_I^b,D_J^a)\right)}_{\mathrm{O2}_{I,J}}\right) \quad (2)$$

where Sim(*I*, *J*) is the similarity between SSIPP *I* and *J*. D_I^a is the SIFT of the interest point P_a. M(,) denotes the distance metric. $\mathrm{O1}_{I,J}$ and $\mathrm{O2}_{I,J}$ stand for the two possible match orders between *I* and *J*. Eq. 2 selects the best match order to compute the largest similarity between two SSIPPs.

We utilize the cosine distance as the M(,). Hierarchical *K*-means is employed to implement the clustering task for its high efficiency. Note that, during the clustering, two match orders between SSIPP *I* and the cluster center *C* can be produced, *i.e.*, $\mathrm{O1}_{I,C}$ and $\mathrm{O2}_{I,C}$. According to the Eq. 2, the new cluster centers ($\hat{D}_C^a$, $\hat{D}_C^b$) are updated with Eq. 3:

$$\hat{D}_C^a=\mathrm{Mean}\left(\sum_I^{\{\mathrm{O1}_{I,C}\}}D_I^a+\sum_I^{\{\mathrm{O2}_{I,C}\}}D_I^b\right)\quad \hat{D}_C^b=\mathrm{Mean}\left(\sum_I^{\{\mathrm{O1}_{I,C}\}}D_I^b+\sum_I^{\{\mathrm{O2}_{I,C}\}}D_I^a\right) \quad (3)$$

where *I* stands for a SSIPP in *C*. $\{\mathrm{O1}_{I,C}\}$, $\{\mathrm{O2}_{I,C}\}$ are two SSIPP sets whose best match orders with the old cluster center of *C* are $\mathrm{O1}_{I,C}$ and $\mathrm{O2}_{I,C}$, respectively.

The result of hierarchical *K*-means clustering is the appearance based tree, where each node preserves the appearance information. Spatial information will then be assigned to each leaf node in the next section.

2.2.2. Pair-wise spatial relationship quantization

The spatial relationship in each SSIPP is defined as: the scale ratio (*i.e.*, scale relationship) and the included angel (*i.e.*, orientation relationship) between the contained two interest points. We quantize the spatial information hierarchically. Suppose the scale ratio and the included angel are quantized into α and β scales, respectively. Then, each leaf node in the appearance based tree is divided into α new nodes, representing different scale relationships. Similarly, each new node is divided into β visual words. For an appearance based tree with *K* leaf nodes, the final pair-wise visual word tree contains $K\bullet\alpha\bullet\beta$ visual words.

Suppose *I* is a SSIPP falling in leaf node *L* by searching its nearest nodes in the tree hierarchically. Its scale ratio and included angel R_I^S, R_I^O are computed in Eq. 4.

$$R_I^S=\begin{cases}\log(S_I^a/S_I^b) & if:\mathrm{O1}_{I,L}\\ \log(S_I^b/S_I^a) & if:\mathrm{O2}_{I,L}\end{cases}\quad R_I^O=\begin{cases}\mathrm{N}(O_I^a-O_I^b) & if:\mathrm{O1}_{I,L}\\ \mathrm{N}(O_I^b-O_I^a) & if:\mathrm{O2}_{I,L}\end{cases} \quad (4)$$

where S_I^a and O_I^a are the scale and orientation of interest point P_a in *I*. $\mathrm{O1}_{I,L}$, $\mathrm{O2}_{I,L}$ stand for the match orders between *I* and *L*. N(θ) normalizes the angle θ into [$-\pi$, π]. Based on R_I^S and R_I^O, *I* can be assigned with corresponding visual word, which keeps its appearance and spatial information.

3. TOPIC WORD BASED IMAGE RE-RANKING

Image re-ranking is a research topic catching more and more attentions [10, 11] in recent years. The goal is to resort the images returned by text-based search engines according to their visual appearances to make the top-ranked images more relevant to the query. In our algorithm, the returned images are first represented as BoW representations with the pair-wise visual word tree. Then, the visual words most related with the query are identified for image re-ranking.

3.1. Topic Word Selection

For the images retrieved with a query *Q*, we utilize the Latent Semantic Analysis (LSA) [12] to compute the importance of each visual word to *Q*. The most important ones are identified as topic words. Proposed in natural language processing, LSA analyzes the relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents. Similarly, we treat each image as a document and each visual word as a term. Then, for the image set, we build a $m\times n$ sized term-document matrix *M*, where *n* is the number of documents, and each *m*-dimensional vector is a visual word histogram. According to LSA, *M* can be decomposed with Singular Value Decomposition in Eq. 5.

$$M=U\sum V^T \quad (5)$$

where *U* and *V* are orthonormal matrices and $\sum$ is a $k\times k$ sized diagonal matrix. Each diagonal element in $\sum$ represents a latent topic found in *M*. We keep the largest *t* elements and set the rest to zero, resulting in a new matrix $\sum^*$. Intuitively, since many returned images are related with the query *Q* and show similar appearances (*i.e.*, similar visual topics), it is reasonable to keep the most dominant latent topics and filter the noisy ones. In the paper, *t* is experimentally set as $0.1\bullet k$. By replacing $\sum$ with $\sum^*$ in Eq. 5, we get the new matrix M^*, with which, the importance of each visual word to query *Q* *i.e.*, the w_i is computed in Eq. 6. Then, the visual words with high importance can be selected as the topic words for *Q*.

$$w_i=\sum_{j=1}^{n}M_{i,j}^* \quad (6)$$

3.2. Topic Word based Image Re-ranking

Based on topic words, we utilize the strategy illustrated in Eq. 7 to compute the rank value of each image.

$$Rank^{(i)}=\sum_j^T tfidf_j^{(i)}\bullet w_j \quad (7)$$

where $Rank^{(i)}$ denotes the rank value of image *i*. *T* is the total number of the topic words, which is experimentally set as 200. $tfidf_j^{(i)}$ stands for the TF-IDF (*i.e.*, Term Frequency $\bullet$ Inverse Document Frequency [5]) of the topic word *j* in image *i*. With Eq. 7, the image re-ranking task can be finished by sorting the images according to their rank values.

4. EXPERIMENTS

4.1. Dataset Collection

To implement convincing experiments, we collect a dataset with ground truth. We first download images form Google Image with keywords of location such as “Great Wall”,

"Eiffel Tower", *etc.* From the downloaded images, we selected 40 categories, within which we keep 250 relevant images and 100 irrelevant ones. Finally, we build a dataset containing 14000 images, all of which are annotated with positive or negative tags. In addition, with the downloaded image dataset, we extracted about 5 million SSIPPs for pair-wise visual word tree generation in the experiments.

4.2. Comparisons and Evaluations

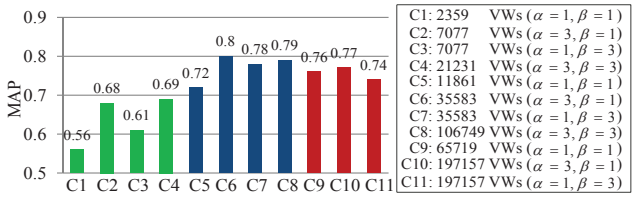


Fig. 6. The MAP with different parameters

We first run image re-ranking with different visual word numbers and different spatial quantizations. MAP (Mean Average Precision) computed in Eq. 8 is adopted to measure the performance of image re-ranking. Fig. 6 presents the experimental results.

$$MAP = \text{Mean}\left(\left(\sum_{i=1}^{250} correct_i^{(Q)}/i\right)\Big/250\right), Q = 1,2,...,40 \qquad (8)$$

where, $correct_i^{(Q)}$ is the number of positive images in the top i re-ranked images for query Q.

From C1, C5, and C9 in the Fig. 6, we can conclude that the number of leaf nodes in appearance based tree is important for the final performance. The results of different spatial quantizations indicate that the spatial clue is helpful for improving precision. Especially from C6, C7 and C10, C11, scale is more important than the orientation. This might be because the detected scale clue is more robust than the orientation. Moreover, we observe that too finer spatial quantization decreases the performance. In Fig. 6, C6 with 11861 leaf nodes, α=3 and β=1 shows the best performance.

In order to illustrate the advantage of the pair-wise visual word tree as well as to compare our topic word based image re-ranking with the state-of-the-art algorithm. We implement the following algorithms in the next experiment:

- *A1*: traditional visual word tree with 31973 visual words.
- *A2*: pair-wise visual word tree with 35583 visual words (11861 leaf nodes in appearance based tree, α=3 and β=1).
- *A3*: the state-of-the-art VisualRank [10] algorithm.

From the experimental result illustrated in Fig. 7(a), it is clear that our pair-wise visual word tree outperforms the traditional visual word tree and shows similar *MAP* with the state-of-the-art VisualRank. Thus, the effectiveness of our pair-wise visual word tree and the proposed re-ranking algorithm can be clearly illustrated. In addition, it is necessary to point out that our algorithm is still very efficient. Obviously in Fig. 7(b), our algorithm shows similar efficiency with the traditional visual word tree and is faster for about 17 times than the VisualRank. Thus, we could conclude that our pair-wise visual word tree and topic word based image re-ranking is effective and efficient.

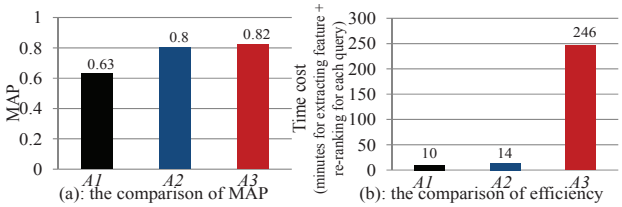


Fig. 7. The comparison between different algorithms

5. CONCLUSIONS

In this paper, we propose the pair-wise visual word tree and topic word based image re-ranking. Generated by clustering a large number of SSIPPs, the visual word in the pair-wise visual word tree preserves both the appearance and spatial information. Based on the LSA, we propose an efficient topic word selection algorithm. With the pair-wise visual word tree and the selected topic words, an efficient image re-ranking algorithm is proposed. Experiments illustrate that the novel pair-wise visual word shows better performance than the traditional visual word tree. In addition, the proposed image re-ranking algorithm presents promising efficiency and the state-of-the-art precision.

6. ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China: 60833006, in part by National Basic Research Program of China (973 Program): 2009CB320906, in part by Beijing Natural Science Foundation: 4092042, in part by the start-up funding from Texas State University, and in part by DHS Grant N0014-07-1-0151.

7. REFERENCES

[1] D. Xu and S. F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *T-PAMI*, 30(11): 1985-1997, Nov. 2008.

[2] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," *CVPR*, pp. 1-8, 2008.

[3] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearanceand shape by correlations," *CVPR*, 2006.

[4] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," *ACM Multimedia*, pp. 75-84, 2009.

[5] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *CVPR*, pp. 2161-2168, 2006.

[6] Z. Wu, Q. Ke, and J. Sun, "Bundling features for large-scale partial-duplicate web image search," *CVPR*, 2009.

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

[8] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," *CVPR*, pp.1-8, 2007.

[9] Y. Zheng, M. Zhao, S. Y. Neo, T. Chua, and Q. Tian. "Visual synset: a higher-level visual representation," *CVPR*, pp. 1-8, 2008

[10] Y. Jing and S. Baluja, "VisualRank: applying PageRank to large-scale image search," *T-PAMI*, 30(11): 1877-1890, Nov. 2008.

[11] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. Hua, "Bayesian video search reranking," *ACM Multimedia*, 2008.

[12] S. Deerwester, S. Dumais, and R. Harshman, "Indexing by latent semantic analysis," *JASIST*, 41(6): 391-407, 1990.