Local Visual Similarity Descriptor for Describing Local Region

Xianglin Huang, Ye Xu, Lifang Yang
Faculty of Science and Technology, Communication University of China

ABSTRACT

Many works have devoted to exploring local region information including both the information of the local features in local region and their spatial relationships, but none of these can provide a compact representation of the information. To achieve this, we propose a new approach named Local Visual Similarity (LVS). LVS first calculates the similarities among the local features in a local region and then forms these similarities as a single vector named LVS descriptor. In our experiments, we show that LVS descriptor can preserve local region information with low dimensionality. Besides, experimental results on two public datasets also demonstrate the effectiveness of LVS descriptor.

Keywords: Bag-of-features, spatial pyramid, spatial information, local region.

1. INTRODUCTION

The significant performance improvement induced by preserving spatial information of local features has attracted much attention in recent years. One of the most predominant works is spatial pyramid matching (SPM) [10], which has emerged as a popular framework to serve various works [2-8, 13]. Its basic idea is to partition an image into multiple increasingly finer blocks and then concatenates all the pooling vectors for each block to form a final image representation. Nevertheless, this approach only preserves the global spatial information of the local features within an image. To solve this problem, many works have devoted to exploring local region information, which includes both the information of the local features in local region and their spatial relationships. These works can be roughly divided into three classes. The first class [1, 2, 3] groups the spatially close visual words into visual phrases and then represents an image as a histogram of these phrases. The major problems of this class are combinatorial explosion and inadequate description on geometric information. These problems are addressed in the second class [4, 5, 6] to some degree. Different from obtaining visual phrases after feature quantization, the second one first concatenates the neighboring local features into joint features and then learns codebook over these features by K-means or sparse coding. In [5], local features are concatenated in different directions to preserve more geometric information. One shortcoming of this class is the high dimensionality of the joint feature. For example, a joint feature of size $d \times 9$ is required to represent a local region including 9 local features, where d is the dimensionality (e.g., 128 for SIFT [9]) of local feature. The third class [7, 8] employs graph to accurately describe the spatial relationships among visual words and then performs graph based clustering algorithm to learn visual graph codebook. In [8], a fast sub-graph detection algorithm is adopted for higher classification performance, but it still incurs high computational cost.

All the above approaches preserve local region information more or less, but none of these can provide compact representation vector for local region. To achieve this, we propose a new approach named Local Visual Similarity (LVS), which does not belong to any one of the above classes. LVS first calculates the similarities among the local features in a local region and then forms these similarities as a single vector named LVS descriptor. The only weak assumption followed in this paper is that local features approximately reside on a low-dimensional and irregular manifold. This assumption is also the reason why LVS descriptor is capable of preserving local region information. Our experiments show that LVS descriptor becomes more precise as more pairs of local features involve in the similarity calculation. Following the common classification pipeline [13], we also evaluate the classification performance of LVS descriptor on 15 Scenes [10] and Caltech 101 [11] dataset, yielding the promising classification results.

2. LOCAL VISUAL SIMILARITY DESCRIPTOR

In this paper, we aim to produce compact representation vector named LVS descriptor for local region, which can preserve not only the information of the local features in a local region but also their spatial relationships. To obtain a LVS descriptor, we only need to perform two steps sequentially: (1) calculating the similarities among the local features

Ninth International Conference on Machine Vision (ICMV 2016), edited by Antanas Verikas, Petia Radeva,
Dmitry P. Nikolaev, Wei Zhang, Jianhong Zhou, Proc. of SPIE Vol. 10341, 103410S
© 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2268689

in a local region; (2) concatenating these similarities to form a LVS descriptor. The two steps are shown in Fig. 1. Since the similarities among local features are recorded in a sequence, their spatial relationships are preserved accurately.

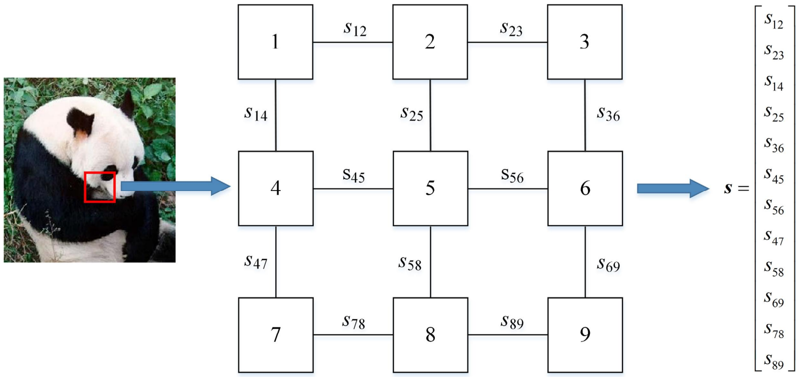

1
2
3
4
5
6
7
8
9
s_{12}
s_{23}
s_{14}
s_{25}
s_{36}
s_{45}
s_{56}
s_{47}
s_{58}
s_{69}
s_{78}
s_{89}
$\boldsymbol{s} = [s_{12}, s_{23}, s_{14}, s_{25}, s_{36}, s_{45}, s_{56}, s_{47}, s_{58}, s_{69}, s_{78}, s_{89}]$

Figure 1. Illustration on the production process of LVS descriptor. The local region in the red box includes 9 local features with their indices $i = 1, 2, \ldots, 9$. s_{ij} denotes the similarity between the ith local feature and the jth one, and $\boldsymbol{s}$ is the LVS descriptor for the region.

The production process of LVS descriptor can also be well illustrated from the viewpoint of feature space. As mentioned in Section 1, local features approximately reside on a low-dimensional and irregular manifold, and the features in a local region are the data points on the manifold. After performing similarity calculation (e.g., l_2 distance), a graph is built over these data points and the distances between the nodes are recorded in a LVS descriptor. Given a local region and its LVS descriptor, if the similar LVS descriptors to the given one are only obtained around the local features in the region, then the region is well-described. In general, a local region is described more precisely with more complicated graph. As shown in Fig. 2, the graph in Fig. 2b is much better than the one in Fig. 2a in terms of preserving local region information.

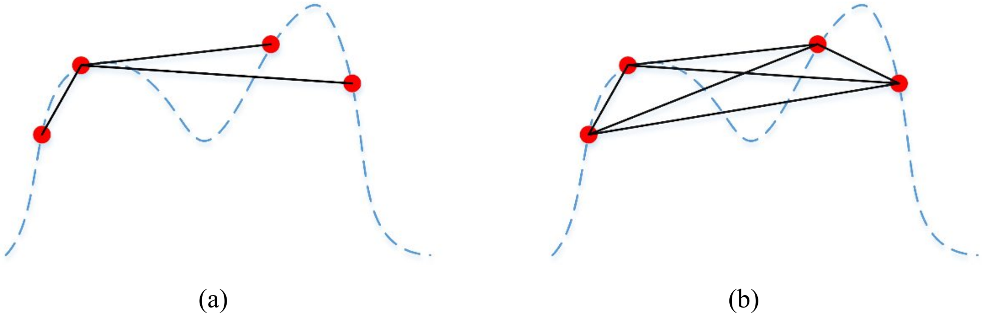


Figure 2. Graphs built over 4 local features within a 2 × 2 local region. The red points denote the local features located at a low-dimensional manifold, and the black lines indicate which pairs of nodes involve in similarity calculation.

We further ameliorate LVS descriptor by taking the following two practices. One is that we can calculate a similarity vector instead of a scalar by utilizing the structure characteristic of local feature, as shown in Fig. 3b. This practice improves the preciseness of LVS descriptor especially when there are a few local features in a local region (verified in section 3.1). The other is that we can freely rule the pairs of local features involving in similarity calculation. In our experiments, this practice leads to comparable classification performance with more compact LVS descriptors.

Taking into account the above two practices, we formulate LVS descriptor in the following. Given a local region r, we denote by $U = \{(\boldsymbol{f}_p^i, \boldsymbol{f}_q^i), i = 1, 2, \ldots, N\}$ the set of the pairs of local features involving in similarity calculation, in which $\boldsymbol{f}_q \in \mathbb{R}^d$ and $\boldsymbol{f}_q \in \mathbb{R}^d$ indicate the pth and qth local feature in the region r, respectively. Let g: $\mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^m$ be the similarity operation, LVS descriptor can be defined as:

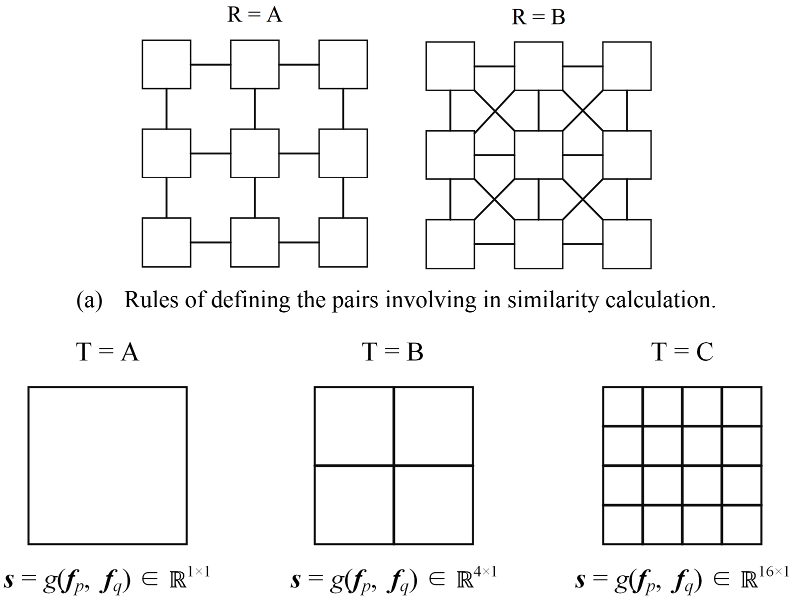
$$\boldsymbol{s} = \left[\boldsymbol{s}_1^{\mathrm{T}}, \boldsymbol{s}_2^{\mathrm{T}}, \mathrm{K}, \boldsymbol{s}_N^{\mathrm{T}}\right]^{\mathrm{T}} \in ¡^{mN\times 1}$$

$$\boldsymbol{s}_i = g(\boldsymbol{f}_p^i, \boldsymbol{f}_q^i) \in ¡^{m\times 1}.$$

Here, the similarity operation g could be l_p distance, or histogram distance [12], or other sophisticated distances.

3. EXPERIMENTS AND RESULTS

In this section, we first investigate LVS descriptor in Section 3.1 to 3.3 from three aspects, then compare LVS descriptor with the joint SIFT feature (named *macrofeature* in [6]) in Section 3.4, and finally report the classification performance of the combination of LVS and SIFT descriptor in Section 3.5. Following [6], a joint SIFT feature is the concatenation of the local features in a local region, for example, an 1152-dimensional joint SIFT feature for a 3 × 3 local region including 9 SIFT features.


R = A
R = B
T = A
T = B
T = C
$s = g(f_p, f_q) \in \mathbb{R}^{1\times1}$
$s = g(f_p, f_q) \in \mathbb{R}^{4\times1}$
$s = g(f_p, f_q) \in \mathbb{R}^{16\times1}$

(a) Rules of defining the pairs involving in similarity calculation.

(b) Types of similarity vector.

Figure 3. Illustration on the rule R of defining the pairs involving in similarity calculation and the type T of similarity vector.

We conduct these experiments on Caltech 101 and 15 Scenes dataset. Caltech 101 dataset is a challenging object recognition benchmark. It has 102 classes and each class has 31 to 800 images. 15 Scenes dataset consists of 4492 images covering 15 scene classes. The number of images per class varies from 260 to 440. For Caltech 101 dataset, we use 30 training images per class and 50 testing images per class. For 15 Scenes dataset, 100 images per class are used for training and the rest for testing. Dense SIFT is extracted from each image on a regular grid of 16 × 16 pixels. The step-size is fixed to 8 pixels for 15 Scenes dataset and 6 pixels for Caltech 101 dataset. After obtaining the SIFT features within an image, we extract the LVS descriptors for the image over the neighborhoods of all the SIFT features. K-means is employed to learn codebook over the training LVS descriptors, and the codebook size is set to 1024 for 15 Scenes dataset and 2048 for Caltech 101 dataset. We adopt Localized Soft-assignment Coding (LSC) [13] for feature coding and perform max pooling to obtain the final image representation. The spatial pyramid with the levels of 1 × 1, 2 × 2 and 4 × 4 is used to preserve the spatial information of LVS descriptors and linear SVMs are used for classification. In particular, we use Caltech 101 dataset to investigate LVS descriptor in Section 3.1 to 3.3, and l_2 distance for similarity calculation throughout all the experiments unless otherwise indicated.

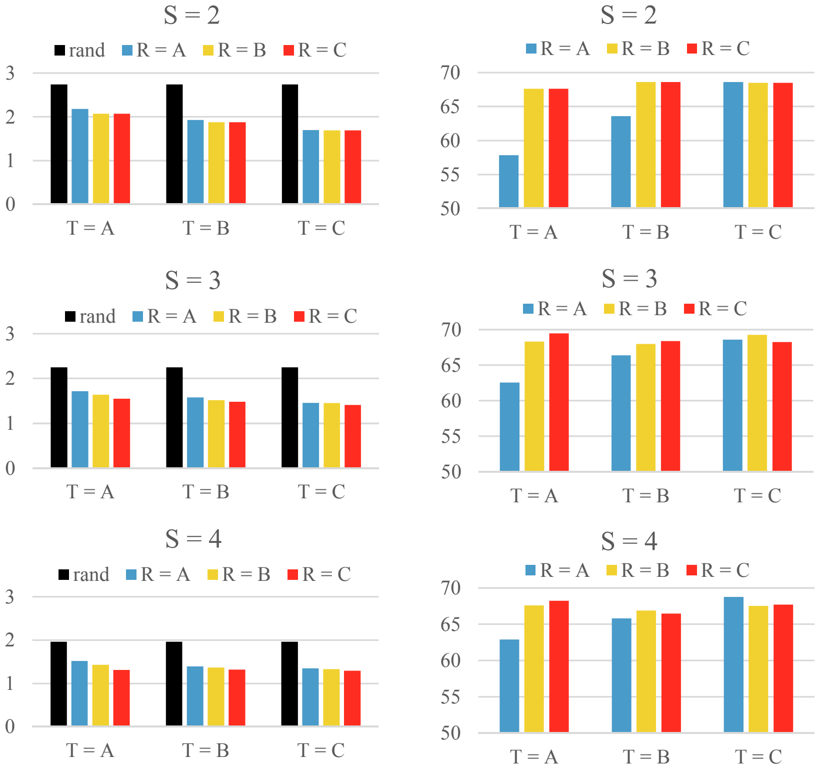
For comprehensive analysis, we produce various LVS descriptors by combining the three factors: the side length S of local region, the rule R (Fig. 3a) of defining the pairs involving in similarity calculation and the type T (Fig. 3b) of similarity vector. In addition to the rules defined in Fig. 3a, we also test the rule (R = C) which asks each local feature in a local region to perform similarity calculation with all other features in the region. For simplicity, we assume that local region is square in this paper.

3.1 LVS descriptor preserving local region information

To investigate whether LVS descriptor can preserve local region information and what LVS descriptor is better, the most immediate method is to restore a LVS descriptor to all the possible local regions and then measure the average difference between the real region and these possible regions. However, this method is too expensive. Instead, we take a simple method by considering the following two factors: (1) similar local regions should have similar LVS descriptors;

(2) the local regions restored from similar LVS descriptors should be similar. Firstly, a LVS descriptor is randomly selected from the set of LVS descriptors obtained from 1000 images. Afterwards, the distance d between the joint SIFT feature corresponding to the selected LVS descriptor and its nearest one is measured in the joint SIFT feature space. Next, we find the nearest LVS descriptor to the selected one in the LVS descriptor space. At the end, we measure the distance d' between the joint SIFT feature corresponding to the nearest LVS descriptor and the one corresponding to the selected LVS descriptor. The above process is performed 1000 times separately, and the average ratio over 1000 ratios $\tau = d' / d$ is reported as the final evaluation result. Apparently, the closer the average ratio to 1, the more precise LVS descriptor is.

Fig. 4a shows the average ratios for various LVS descriptors. The results indicated by the black bins are the baselines, which are obtained by randomly selecting 1000 joint SIFT features as the fake ones corresponding to 1000 nearest LVS descriptors. Clearly, these black bins are higher than others, which means that LVS descriptor is capable of preserving local region information. Besides, LVS descriptor becomes more precise with finer similarity vector and the same tendency can also be found as more pairs of local features involve in similarity calculation. Fig. 4b reports the classification accuracies for these LVS descriptors. Overall, the higher accuracies are acquired with the LVS descriptors with the lower average ratios. In addition, it is worth noting that the LVS descriptors for the larger local region (e.g., 4 × 4) result in slightly inferior accuracies.



(a) (b)

Figure 4. Evaluation results for various LVS descriptors. (a) average ratios; (b) classification accuracies on Caltech 101 dataset.

3.2 Normalization on LVS descriptor

We check out the influence of common normalization operations on classification performance. The LVS descriptor obtained when S = 3, R = C and T = A is adopted for evaluation. Fig. 5b reports the classification accuracies for several normalization operations: l_2, l_1 and $l_{1.5}$. As can be seen, the performance drops significantly after applying normalization operation on LVS descriptor. The reason is that normalization operation reduces the graph information recorded in LVS descriptor, making more dissimilar local regions have similar LVS descriptors. This explanation is demonstrated by the increase of average ratio $\tau = d' / d$ in Fig. 5a.

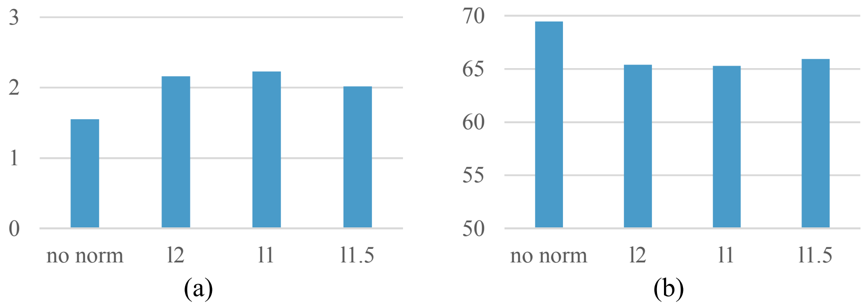


Figure 5. Evaluation results for normalization on LVS descriptor. (a) average ratios; (b) classification accuracies on Caltech 101 dataset.

3.3 Impact of similarity operation on classification performance

In the definition of LVS descriptor, the similarity between two local features could be measured by l_p distance, or histogram distance, or other sophisticated distances. Here, several common distance metrics, l_2, l_1 and histogram distance, are chosen for evaluation. We can see in Fig. 6b that l_1 and histogram distance lead to inferior accuracies to l_2 distance. The reason can be found in Fig. 6a. As shown, LVS descriptor obtained by l_2 distance performs significantly better in terms of preserving local region information.

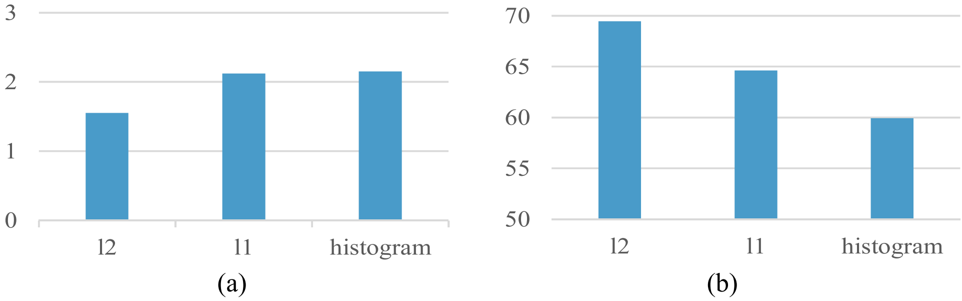


Figure 6. Evaluation results for various similarity operations. (a) average ratios; (b) classification accuracies on Caltech 101 dataset.

3.4 Performance comparison between LVS descriptor and the joint SIFT feature

To further evaluate LVS descriptor, we compare LVS descriptor to the joint SIFT feature on 15 Scenes and Caltech 101 dataset. Here, the evaluated LVS descriptor is obtained when S = 3, R = C and T = A, and the joint SIFT feature is a vector of size 128 × 9 accordingly. Table 1 reports the classification accuracies and the numbers in parentheses denote the dimensionality. As can be seen, while LVS descriptor results in inferior classification accuracies to the joint SIFT feature, its dimensionality is much lower than that of the joint SIFT feature, incurring low computational cost and storage complexity.

Table 1. Results for LVS descriptor and the joint SIFT feature on Caltech101 and 15 Scenes dataset.

	LVS descriptor (**36**)	Joint SIFT feature (**1152**)
Caltech 101	69.5 ± 0.47	72.9 ± 0.73
15 Scenes	74.5 ± 0.56	78.6 ± 0.61

3.5 Performance of the combination of LVS and SIFT descriptor

In this section, we evaluate the performance of the combination of SIFT descriptor and the LVS descriptor obtained when S = 3, R = C and T = A. Here, the final image representation is the concatenation of the one obtained by the LVS descriptor and the one by SIFT descriptor. From Table 2, we find that the combination of LVS and SIFT descriptor gives better classification performance than both LVS and SIFT descriptor used separately.

Table 2. Results for the combination of LVS and SIFT descriptor on Caltech101 and 15 Scenes dataset.

	LVS descriptor	SIFT descriptor	LVS + SIFT
Caltech 101	69.5 ± 0.47	74.1 ± 0.95	**75.0 ± 0.67**
15 Scenes	74.5 ± 0.56	82.6 ± 0.14	**83.2 ± 0.58**

4. CONCLUSION

In this paper, we proposed a new approach for preserving local region information. Different from the existing works, our approach can produce compact representation vector named LVS descriptor for local region. Our experiments show that LVS descriptor becomes more precise with more complicated graph. Besides, the classification results on Caltech 101 and 15 Scenes dataset also demonstrates the effectiveness of LVS descriptor. The current works we are pursuing are to improve the preciseness of LVS descriptor, such as incorporating the angle information of graph into LVS descriptor and weighting the elements in LVS descriptor to highlight the discriminative information in local region.

REFERENCES

[1] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," IEEE Transaction on Image Processing. **20**(9), 2664-2677 (2011).

[2] M. M. Farhangi, M. Soryani, M. Fathy, "Informative visual words construction to improve bag of words image representation," IET Image Processing. **8**(5), 310-318 (2014).

[3] T. Li, T. Mei, I. Kweon, and X. Hua, "Contextual bag-of-words for visual categorization," IEEE Transaction on Circuits and Systems for Video Technology. **21**(4), 381-392 (2011).

[4] N. Morioka, S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in Europe Conference on Computer Vision (ECCV), 692-705 (2010).

[5] N. Morioka, S. Satoh, "Learning directional local pairwise bases with sparse coding," in British Machine Vision Conference (BMVC), 1-11 (2010).

[6] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in Computer Vision and Pattern Recognition (CVPR), 2559-2566 (2010).

[7] M. Dammak, M. Mejdoub, C.B. Amar, "Histogram of dense subgraphs for image representation," IET Image Processing. **9**(3), 184-191 (2015).

[8] F. B. Silva, S. Goldenstein, S. Tabbone, and R. S. Torres, "Image classification based on bag of visual graphs," in International Conference on Image Processing, 4312-4316 (2013).

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision. **60**(2), 91-110 (2004).

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in Computer Vision and Pattern Recognition (CVPR), 2169-2178 (2006).

[11] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories," in Computer Vision and Pattern Recogition (CVPR), 1-8 (2004).

[12] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in IEEE International Conference on Computer Vision (ICCV), 1458-1465 (2005).

[13] L. Liu, L. Wang, X. Liu, "In Defense of Soft-assignment Coding," in IEEE International Conference on Computer Vision (ICCV), 2486-2493 (2011).