J. Vis. Commun. Image R. 24 (2013) 786–793

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Laplacian affine sparse coding with tilt and orientation consistency for image classification

Chunjie Zhang [a,*], Shuhui Wang [b], Qingming Huang [a,b], Chao Liang [c], Jing Liu [d], Qi Tian [e]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049 Beijing, China
[b] Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[c] National Engineering Research Center for Multimedia Software, Wuhan University, 430072 Wuhan, China
[d] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[e] Department of Computer Sciences, University of Texas at San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 26 October 2012
Accepted 6 May 2013
Available online 17 May 2013

Keywords:
Image classification
Affine transformation
Sparse coding
Laplacian matrix
Tilt and orientation
Smooth constraints
Object categorization
Bag-of-visual words model

ABSTRACT

Recently, sparse coding has become popular for image classification. However, images are often captured under different conditions such as varied poses, scales and different camera parameters. This means local features may not be discriminative enough to cope with these variations. To solve this problem, affine transformation along with sparse coding is proposed. Although proven effective, the affine sparse coding has no constraints on the tilt and orientations as well as the encoding parameter consistency of the transformed local features. To solve these problems, we propose a Laplacian affine sparse coding algorithm which combines the tilt and orientations of affine local features as well as the dependency among local features. We add tilt and orientation smooth constraints into the objective function of sparse coding. Besides, a Laplacian regularization term is also used to characterize the encoding parameter similarity. Experimental results on several public datasets demonstrate the effectiveness of the proposed method.

1. Introduction

As a fundamental problem in computer vision, image classification has attracted many researchers' attention. A lot of image classification models have been proposed to solve this problem, for example, bag-of-visual words model (BoW) [1], part-based model [2]. The BoW model is widely used both for its simplicity and good performance in real world applications. The histogram based representation makes the BoW model robust to scale and rotation variances. Typically, the BoW model can be divided into three steps: (i) Local region selection and description; (ii) Visual codebook construction and local feature encoding; (iii) Histogram based image representation and classifier training.

Among the three steps, the codebook construction and local feature encoding process plays a vital role for efficient image classification. Traditional image classification methods [1] use the *k*-means clustering method or its variants to construct codebook and view the cluster centers as the visual words. Each local feature is quantized to the nearest visual word using the Euclidean distance. However, this hard assignment method can cause severe information loss [3], especially when the local features are on the boundary. To alleviate the information loss, many works [3–6] have been done by softly encoding local features. This helps to improve the final image classification performance. However, the *k*-means based codebook construction method should be used with non-linear kernels for efficient image classification which has a time complexity of $O(n^3)$. This problem is more sever when large image datasets are used. To reduce the computational cost, Yang et al. [5] proposed to use sparse coding along with max pooling for image representation and use linear SVM classifier instead. This achieved the state-of-the-art performance for image classification on several public datasets.

Besides, the BoW model has no information of local features' spatial information which is also very important for robust image classification. To incorporate the spatial information of local features, spatial pyramid matching (SPM) was proposed by Lazebnik et al. [7]. The SPM tried to combine the spatial information by dividing each image into increasingly finer sub-regions. This simple but efficient algorithm is widely used by researchers since its introduction. Inspired by the SPM, many other methods which consider the spatial information were also proposed [8,9].

The above mentioned methods actually assume that the extracted local features are discriminative enough for robust and

* Corresponding author.
E-mail addresses: cjzhang@jdl.ac.cn (C. Zhang), shwang@jdl.ac.cn (S. Wang), qmhuang@jdl.ac.cn (Q. Huang), liangchao827@gmail.com (C. Liang), jliu@nlpr.ia.ac.cn (J. Liu), qitian@cs.utsa.edu (Q. Tian).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.05.004

efficient image classification. However, images are often captured under diverse conditions, such as different poses, scales, illuminations or camera parameters. If we can generate more efficient features that can cope with these variations, we would be able to make the image classification algorithm more robust and effective. Inspired by this, a lot of work have been made by proposing more discriminative features [10–15], exploring the relationship among local features [16–20] or making various transformations to images or local features [21–23]. The invention of new discriminative features has been a hot research topic in computer vision and needs careful design and experiments. Before we can get more discriminative features, it would be more effective and efficient if we can make good use of the existing local features. Inspired by this, Kulkarni and Li [21] proposed to use affine sparse coding to cope with the affine transformation of images for better image classification. However, the affine sparse coding technique has no constraints on the tilt and orientations of the transformation which are also very useful for image classification. During the local feature encoding process, besides minimizing the reconstruction error, the tilt and orientation of local features should also be combined. Moreover, in affine sparse coding, the mutual dependence among local features are also not considered. This means similar local features may be encoded with dissimilar parameters. It is more natural that visually similar local features should be encoded with similar coding parameters. If we can take the tilt and orientation of local features as well as the encoding parameter similarities into consideration, we would be able to further improve the image classification performance over affine sparse coding [21].

In this paper, we propose a Laplacian affine sparse coding with tilt and orientation consistency algorithm which combines the tilt and orientations of affine local features as well as the dependence among affine local features. To combine the tilt and orientation of affine local features, we propose to add tilt and orientation smooth constraints into the objective function of affine sparse coding. Besides, a Laplacian regularization term with histogram intersection similarity measurement is also used to measure the similarity of affine local features. This new formulation can generate more discriminative coding parameters which can then be used for image representation. Besides, the consideration of local features' similarity during the encoding process also helps to reduce the encoding error and preserve as much information as possible. Max pooling is then used to extract the final image representation and multi-class linear SVM classifiers are trained to predict the category of images. Experiments on several public dataset demonstrate the effectiveness of the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm (LASC-TOC). We give the flowchart of the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm in Fig. 1.

The rest of this paper is organized as follows. In Section 2, we give the related work. The detail of the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm is given in Section 3. We give the experimental results in Section 4 and conclude in Section 5.

2. Related work

To classify an image based on its semantic content, the BoW model [1] has been widely used. However, traditional *k*-means clustering based codebook generation and nearest neighbor assignment based local feature quantization method may cause sever information loss [3], especially when the local features are on the boundary of different visual words. To reduce the information loss, many works [4–6] have been done by softly encoding local features. Gemert et al. [4] used kernel codebook by soft assignment of local features at heavy computational cost. To speed up the training process, Yang et al. [5] proposed to use sparse coding along with max pooling for image representation. Although proven effective, the sparse coding has no constraints on the coding parameters. This may cause information loss because max pooling is then used to extract information for image representation. Negative coding parameters have no influences on the final image representation. To solve this problem, Zhang et al. [6] proposed to use non-negative sparse coding with max pooling instead.

Moreover, the BoW model has no information about the spatial layout of local features while spatial information plays a vital role for efficient image classification. Lazebnik et al. [7] proposed spatial pyramid matching algorithm to combine weak spatial information. This is achieved by dividing images into increasingly finer sub-regions. Inspired by this, Bosch et al. [8] represented shape with a spatial pyramid kernel and improved the final performance. Component based image representation [9] is also proposed to combine the spatial information at heavy computational cost.

SIFT feature is often used in the BoW model, since the discriminative power of SIFT is limited, Many other features [10–15] are also used. Belongie et al. [10] proposed to use shape context while Serre et al. [11] tried to categorize objects with features inspired by visual cortex. Viola and Jones [12] proposed a rapid object detection method using a boosted cascade of simple Harr-like features. Sande et al. [13] proposed to transform images into different color spaces and use color SIFT for visual applications. To avoid the scale and orientation calculation of SIFT feature, Dalal and Triggs [14] proposed to use histograms of oriented gradients (HoG) for fast computation in object detection. Xie et al. [15] used bin-ratio information for category and scene classification. The explore of local feature's relationship for better image classification is also studied by researchers [16–20]. Grauman and Darrell [16] proposed pyramid match kernel which tried to explore the relationship of local features in the kernel space. Wu et al. [17] bundled features together for large scale partial-duplicate web image search. Wang et al. [18] used feature context for image classification and object detection which achieved good performance. Yao et al. [19] tried to classify actions and measure action similarity by modeling the mutual context of objects and human poses while Lee and Grauman [20] proposed to use object-graphs for context-aware category discovery. Before we can get more discriminative features, it would be more effective and efficient if we can make good use of the existing local features. Inspired by this, Kulkarni and Li [21] proposed to transform images with affine transformations and then extract SIFT features to cope with the affine transformation of images for better image classification. Zhang et al. [22] made Harr-like transformation of local features and improved the classification performance. Zhang et al. [23] proposed a simple but effective method by resizing images to generate descriptive visual words for visual applications.

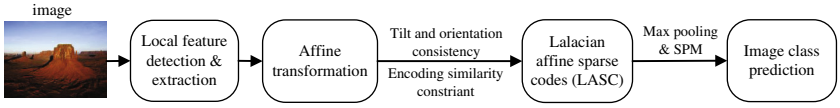

Fig. 1. Flowchart of the proposed image classification method using Laplacian affine sparse coding with tilt and orientation consistency algorithm.

3. Laplacian affine sparse coding with tilt and orientation consistency for image classification

In this section, we give the details of the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm. First, affine transformation is used to cope with the tilt and rotation changes caused during image capturing process. Local features (typically, SIFT feature is used) are then extracted on these transformed images and encoded for image representation. We add tilt and orientation constraints to the objective function of sparse coding. Besides, the similarities of local features are also considered during the sparse coding process. Finally, max pooling with spatial pyramid matching is used to represent images and multi-class linear SVM classifiers are trained for image category prediction.

3.1. Affine sparse coding

SIFT feature has been widely used since its introduction for visual applications. However, images may undergo varied pose, illumination or scale changes which means using the SIFT feature alone may not be able to cope with these variations. To alleviate this problem, affine SIFT feature (ASIFT) is proposed [24]. The ASIFT first makes affine transformation of images and then extract dense SIFT features on these transformed images. Formally, the affine transformation map is given by

$$A = \lambda \begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} cos\alpha & -sin\alpha \\ sin\alpha & cos\alpha \end{bmatrix} \quad (1)$$

where $\lambda > 0$ and t controls the tilt, ψ is camera spin and $\alpha \in [0, \pi)$. λt is the determinant of A. The affine distortion which is caused by changes in the optical axis orientation can be described by the latitude and longitude camera parameters α and θ ($t = 1/cos\theta$). The longitude parameter α can be simulated by horizontally rotating an image from the frontal position while the latitude parameter can be simulated by directional t-subsampling [24,25]. Morel and Yu [24] experimentally found that setting 6 different tilts on a finite number of rotation angles α is enough for most real world applications. Without loss of generality, we set ψ to 0 in this paper. Since images are hardly rotated more than 90°, we use a maximum of six tilts and corresponding rotations [21]. This is achieved by first obtain the tilt factor $t = 2^{i/2}$, where $i = 1, 2, \ldots, 6$. Then we obtain α for each tilt factor t as $k/t * 72$ with $k/t * 72 < 180°$, $k = 1, 2, 3, \ldots$. Finally, we calculate the affine transformation of the input image for all t and α. Dense SIFT features are extracted from these affine transformed images which are then used for image representation. In this way, since we can make use of the original image as well as the transformed images, we can get more discriminative local features than traditional methods which helps to improve final image classification performance. If we set $t = 1$ and $\alpha = 0$, the affine SIFT method will degenerate to traditional SIFT based image classification method.

After obtaining the affine SIFT features, we can use them to construct the codebook and encoding these features. Recently, the sparse coding becomes a popular choice because of its good performance for image classification. Formally, let $X = [x_1, x_2, \ldots, x_N]$ be N affine SIFT features, where $x_i \in \mathbb{R}^{d\times 1}$, $i = 1, 2, \ldots, N$. The corresponding codebook is $B = [b_1, b_2, \ldots, b_K] \in \mathbb{R}^{d\times K}$. Affine sparse coding tried to minimize the reconstruction error with sparsity regularization as:

$$min_{B,C} \sum_{i=1}^{N} \|x_i - Bc_i\|^2 + \lambda_1 \|c_i\|_1 \quad (2)$$

where $C = [c_1, c_2, \ldots, c_N]$ are the sparse coding parameters for the N local features. λ_1 is the parameter which controls the sparsity of C. For each local region, the sparse reconstruction errors of these corresponding affine SIFT features are then calculated. The one with minimum reconstruction error is chosen to represent this local region.

3.2. Laplacian affine sparse coding with tilt and orientation consistency

We can use the affine SIFT features for codebook construction and local feature encoding directly. However, this strategy does not consider the correlations among affine SIFT features. For example, the tilt and orientation information. Fig. 2 shows a toy example of this problem. The local features 'A' and 'B' in the original image (the left one) are the same as local features 'a' and 'b' in the rotated image (the right one), hence the extracted SIFT features should be very similar. Unfortunately, this is often violated in real applications, especially when dense sampling of SIFT features are used which is a favorite choice of researchers. Although the affine sparse coding algorithm [21] tried to choose the optimal local feature by minimizing the reconstruction error, there is no guarantee that the similarity shown in Fig. 2 holds. This means if we can combine the tilt and orientation information of affine local features, we will be able to reduce the computational cost by choosing a fraction of the extracted affine SIFT features instead of using all of them. Besides, it can also make the final classifier more robust since we can get ride of some noisy local features. To make use of the tilt and orientation information of local features, we propose to add tilt factor and rotation angle constraints to the objective function of affine sparse coding as:

Fig. 2. A toy example showing the necessity of considering the tilt and orientation of affine local features. The local features 'A' and 'B' in the original image are the same as local features 'a' and 'b' in the rotated image, hence the extracted SIFT features should be very similar. Unfortunately, this is often violated in real applications, especially when dense local feature extraction is used.

$$min_{B,C}\sum_{i=1}^{N}\|x_i - Bc_i\|^2 + \lambda_1\|c_i\|_1 \tag{3}$$

$$s.t. \quad \sum_{i=1}^{N}\sum_{j=1}^{N}\|t_i - t_j\|^2 + \|\alpha_i - \alpha_j\|^2 < \lambda_2$$

where $t_i,\ \alpha_i,\ i = 1, 2, \ldots, N$ are the tilt factor and rotation angle of the ith SIFT feature respectively. λ_2 is the scale and orientation constraint parameter.

Moreover, similar affine SIFT features may be encoded by quite different parameters due to the sensitiveness of sparse coding. This means the dependence information of affine SIFT features is lost. To alleviate this problem, we propose to use a Laplacian constraint with affine sparse coding. Specifically, we add a regularization term which considers local feature's similarity into the optimization problem of (3) to ensure that similar affine SIFT features are encoded with similar parameters. In this way, we can not only increase the effectiveness of affine sparse coding but also reduce the information loss during local feature encoding process [26]. This is achieved by solving the following optimization as:

$$\begin{aligned} &min_{B,C}\sum_{i=1}^{N}\|x_i - Bc_i\|^2 + \lambda_1\|c_i\|_1 + \lambda_3/2\sum_{ij}\|c_i - c_j\|^2 W_{ij} \\ &= min_{B,C}\|X - BC\|_F^2 + \lambda_1\|c_i\|_1 + \lambda_3 tr(CLC^T) \end{aligned} \tag{4}$$

$$s.t. \quad \sum_{i=1}^{N}\sum_{j=1}^{N}\|t_i - t_j\|^2 + \|\alpha_i - \alpha_j\|^2 < \lambda_2$$

where λ_3 is the regularization parameter which control the relative importance of affine sparse coding parameter smoothness. Let $T = [t_1, t_2, \ldots, t_N], \vec{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$. W is the similarity matrix. Its Laplacian matrix is $L = D - W$ where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. To generate the similarity matrix, we can use the Euclidean similarity, the L_1 similarity or histogram intersection similarity. We choose to use histogram intersection because its effectiveness has been proven by many researchers [26,27]. Besides, there is no parameter needed to be tuned for the histogram intersection similarity. The histogram intersection similarity is defined as follows:

$$W(c_i, c_j) = \sum_{k=1}^{K} min(c_{ik}, c_{jk}) \tag{5}$$

To save computational cost, we use approximate method to construct the Laplacian matrix W by first finding the k nearest neighbor of c_i and calculate the corresponding histogram intersection similarity, the other W_{ij} are set to 0. We set $k = 5$ in this paper, as [26] did.

After the encoding parameters are obtained, we can make categorization of images. We follow [5] and use max pooling to extract information from the coding parameters as this strategy has been proven very effective. To combine the spatial information of local features, spatial pyramid matching (SPM) with three pyramids ($L = 0, 1, 2$) is also used [7]. Multi-class linear SVM classifiers are then trained to predict images' classes.

3.3. Implementation

Fixing the tilt and orientation constraints, the optimization problem of (4) is not convex for B and C simultaneously, but is convex for B when C is fixed and vice versa. Hence, we try to optimize $B,\ C,\ T$ and $\vec{\alpha}$ iteratively while keeping the other three fixed. If we set T and $\vec{\alpha}$ to 0 and do not make affine transformation to images, the proposed Laplacian affine sparse coding algorithm will degenerate to [26]. Hence the Laplacian sparse coding algorithm proposed in [26] can be viewed as a special case of the Laplacian affine sparse coding algorithm.

Algorithm 1. The proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm

Input:
The local features $X,\ \lambda_1,\ \lambda_2,\ \lambda_3$, threshold parameter γ and max iteration number *maxiter*;
Output:
The learned codebook B and coding parameters C;
1: construct the Laplacian matrix W
2: for $iter = 1, 2, \ldots, maxiter$
3: Find the optimal codebook B with $C,\ T$ and $\vec{\alpha}$ fixed by solving problem (6);
4: Find the optimal coding parameter C with $B,\ T$ and $\vec{\alpha}$ fixed by solving problem (7);
5: Check the tilt and orientation constraints in problem (4).
If satisfied
go to step 6
Else
find the local feature with the largest tilt deviation and delete it;
Find the local feature with the largest orientation deviation and delete it;
Updata the Laplacian matrix W.
6: Check the change of objective value of (4), stop if it is below γ; else go to step 2;
7: **return** $B,\ C$;

During the codebook construction process, we first randomly choose some local features to construct the Laplacian matrix. When optimizing over B while keeping $C,\ T$ and $\vec{\alpha}$ fixed, problem (4) can be simplified as:

$$min_B\|X - BC\|_F^2 \tag{6}$$

The column of B is normalized to avoid scaling problem. When we try to optimize over C while keeping the others fixed. Problem (4) can be solved as:

$$min_C\|X - BC\|_F^2 + \lambda_1\|c_i\|_1 + \lambda_3 tr(CLC^T) \tag{7}$$

This can be solved by optimizing over each local feature separately by feature sign search [28]. To find the optimal T and $\vec{\alpha}$, we iteratively remove the local feature that causes the largest tilt and orientation deviation until the constraints in (4) are satisfied. The Laplacian matrix is also updated at the same time, this can be achieved by deleting the corresponding row and column of W. The optimization of $\vec{\alpha}$ can be done in a similar way as T. This process is iterated either a max iteration number is achieved or the changes of objective function falls below a pre-defined threshold. Algorithm 1 gives the procedure of the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm. After the codebook is learned, we can encode the extracted affine SIFT features by solving problem (4) while keeping B fixed.

4. Experiments

We evaluate the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm (LASC-TOC) for image classification on several public datasets: the Caltech 256 dataset [29], the UIUC-Sport dataset [30] and the Scene 15 dataset [7].

4.1. Parameter setting

We use the SIFT feature as the local region description, as [4–7] did. To be consistent with previous work and for fair comparison, we densely extract SIFT features with overlap. The overlap pixel is set to 6 and the smallest patch size is 16×16 pixels as this setting is found to be more effective than the original settings in [7]. These extracted SIFT features are then normalized with ℓ_2 norm. The codebook size is fixed to 1000 for all the datasets. To construct the codebook, we randomly choose 10^5 features for each dataset. As to the spatial pyramid matching, we follow [7] and use the first three layer ($1 \times 1, 2 \times 2, 4 \times 4$) to incorporate the spatial information of local features. We randomly choose the training images and repeat this process for five times. Instead of re-implementing other algorithms, we directly compare with the results reported by researchers for fair comparison. For multi-class SVM classifier training, we use the code provided by Yang et al. [5].

The three parameters λ_1, λ_2, λ_3 are the most important parameters in this paper. Yang et al. [5] found that the performance is best when λ_1 is set to 0.3–0.4. Besides, as found by Gao et al. [26], the performance is good when λ_3 is set to 0.1 (with $\lambda_1 = 0.3$) or 0.2 (with $\lambda_1 = 0.4$). For image sets with large inter class variation, λ_2 should be set to a larger value compared to images set with small inter class variation. Hence, we set λ_2 according to the type of image sets. Specifically, for the Caltech 256 and UIUC-Sport dataset, we set $\lambda_1 = 0.1$, $\lambda_2 = 0.1$ N and $\lambda_3 = 0.3$. For the Scene 15 dataset, we set $\lambda_1 = 0.2$, $\lambda_2 = 0.05$ N and $\lambda_3 = 0.4$. The max iteration number in Algorithm 1 is set to 50. We also use the LLC technique [31] to speed up the sparse coding process and improve the performance.

4.2. Caltech 256 dataset

The Caltech 256 dataset has 256 image categories with a total of 29,780 images. This dataset is introduced as an extension of the Caltech 101 dataset. Images of the Caltech 256 dataset have larger intra class and inter class variations compared with the Caltech 101 dataset. There are at least 80 images in each class of the Caltech-256 dataset. Fig. 3 gives some image classes with the classification accuracy of the Caltech 256 dataset. We evaluate the proposed Laplacian affine sparse coding with tilt and orientation consistency algorithm using 15, 30, 45 and 60 training images respectively. For each class, we randomly pick the training images and use the rest images for testing.

Table 1
Performance comparison on the Caltech-256 dataset.

Methods	15 Images	30 Images	45 Images	60 Images
KCSPM [4]	–	27.17 ± 0.46	–	–
SPM [29]	–	34.10	–	–
SPM [5]	23.34 ± 0.42	29.51 ± 0.52	–	–
ScSPM [5]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LLC [31]	34.36	41.19	45.31	47.68
LScSPM [26]	30.00 ± 0.14	35.74 ± 0.10	38.54 ± 0.36	40.43 ± 0.38
ASC [21]	37.67	43.10	46.90	49.84
LASC	38.36 ± 0.64	43.85 ± 0.59	47.15 ± 0.46	49.90 ± 0.53
LASC-TOC	**38.83 ± 0.56**	**44.20 ± 0.72**	**47.37 ± 0.48**	**49.95 ± 0.52**

Table 2
Performance comparison on the UIUC Sport dataset.

Methods	Classification rate
Li [30]	73.40
ScSPM [5]	82.74 ± 1.46
HIK + OCSVM [27]	83.54 ± 1.13
LScSPM [26]	85.31 ± 0.51
LASC	87.85 ± 0.64
LASC-TOC	**88.53 ± 0.68**

Table 1 gives the performance comparison of the proposed LASC-TOC with other methods. To show the effect of imposing tilt and orientation consistency, we also give the performance of using laplacian affine sparse coding without tilt and orientation consistency (Abbrev. LASC). This can be achieved by setting λ_2 to a large enough number in problem (4). We can see from Table 1 that the proposed LASC-TOC achieved the state-of-the-art performance. Compared with LASC, the consideration of tilt and orientation consistency can preserve the structure information to some extent, hence helps to improve the image classification performance over LASC. Compared with soft assignment methods, such as kernel codebook [4] or sparse coding [5], LASC-TOC can alleviate information loss by making affine transformation of images and consider the smoothness of sparse coding. The LScSPM algorithm also considers the similarity of local features during the encoding process, however, the proposed LASC-TOC algorithm goes one step further by using affine transformation for feature extraction and tilt and orientation consistency. The affine transformation enables us to

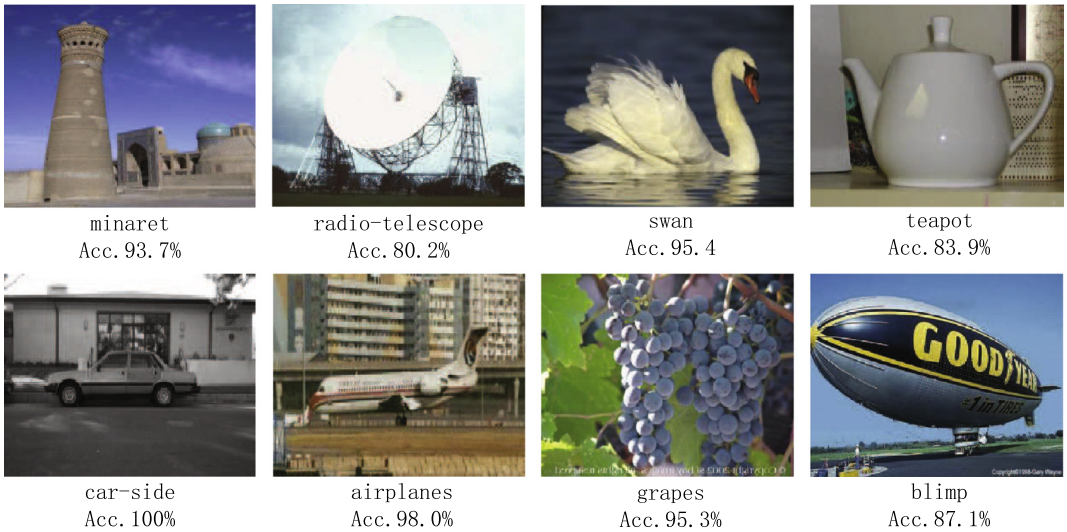

Fig. 3. Performance of LASC-TOC on the Caltech 256 dataset showing some classes with classification accuracies.

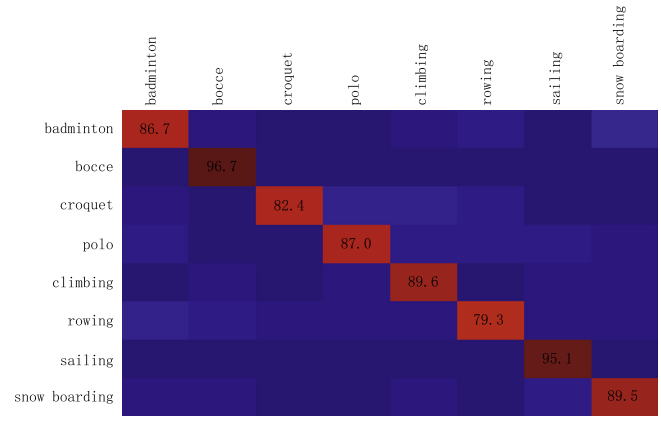

Fig. 4. Confusion matrix on the UIUC Sport dataset. The accuracy decreases from red to green. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Table 3
Performance comparison on the Scene 15 dataset.

Methods	Classification rate
SPM [7]	81.40 ± 0.50
KCSPM [4]	76.67 ± 0.39
SPM [5]	76.73 ± 0.65
ScSPM [5]	80.28 ± 0.93
HIK + OCSVM [27]	84.00 ± 0.46
LScSPM [26]	89.75 ± 0.50
LASC	90.17 ± 0.58
LASC-TOC	**90.36 ± 0.63**

extract more discriminative local features which will eventually help to improve the final image classification performance. Besides, LASC and LASC-TOC consistently outperform LScSPM by around eight and nine percent respectively. This demonstrates the effectiveness of affine transformation as well as tilt and orientation consistency for boosting the image classification performance. Moreover, the relative improvement of LASC-TOC over ASC [21] decreases with the increasing number of training images. We believe this is because the effect of smoothness constraints is less important as the training images increases, the learned classifier are more discriminative to make correct categorization of images with more training samples.

Table 4
Mean and standard derivation of the tilt and orientation as well as the average reconstruction error of 1000 randomly sampled local features on the Scene 15 dataset. ASC: affine sparse coding, LASC-TOC: Laplacian affine sparse coding with tilt and orientation consistency.

Algorithm	Tilt	Orientation	Reconstruction error
ASC	0.31 ± 0.15	0.27 ± 0.12	**0.28 ± 0.09**
LASC-TOC	**0.19 ± 0.07**	**0.13 ± 0.08**	0.36 ± 0.17

4.3. UIUC Sport dataset

The UIUC Sport dataset contains eight categories (*badminton, bocce, croquet, polo, rock climbing, rowing, sailing* and *snow boarding*) of 1792 images. Each class has 137 to 250 images. We follow the same experimental setup as [5,26,27] for fair comparison. 70 images per class are randomly selected and we use the rest of images for testing.

Table 2 shows the performance results on the UIUC Sport dataset. We also give the confusion matrix in Fig. 4. We can have similar conclusions as on the Caltech 256 dataset. The proposed LASC-TOC outperforms the LScSPM method by about 3.2%, this again demonstrates the effectiveness of doing affine transformation with tilt and orientation consistency. Besides, the tilt and orientation consistency helps to improve the performance of LASC by 0.7%.

4.4. Scene 15 dataset

The Scene 15 dataset has 15 categories (*bedroom, suburb, industrial, kitchen, livingroom, coast, forest, highway, insidecity, mountain, opencountry, street, tallbuilding, office* and *store*) of 4485 images with 200 to 400 images per category. The images are diverse and range from indoor to outdoor categories. To compare with previous work, we randomly choose 100 images per class as the training data and use the rest images for testing.

Table 3 gives the performance comparison on the Scene 15 dataset. Again the proposed LASC-TOC algorithm achieved good performance. The improvement of LASC-TOC over LScSPM is not so obvious compared with the results on the Caltech 256 dataset and the UIUC Sport dataset. We believe this is because images of the Scene 15 dataset are relatively easy to separate and the improvement of affine transformation is not so obvious. To analysis the details of the results, we also give the confusion matrix in Fig. 5. We can see from Fig. 5 that the indoor classes are harder to classify than outdoor classes, as [5,7,26] found. This also shows

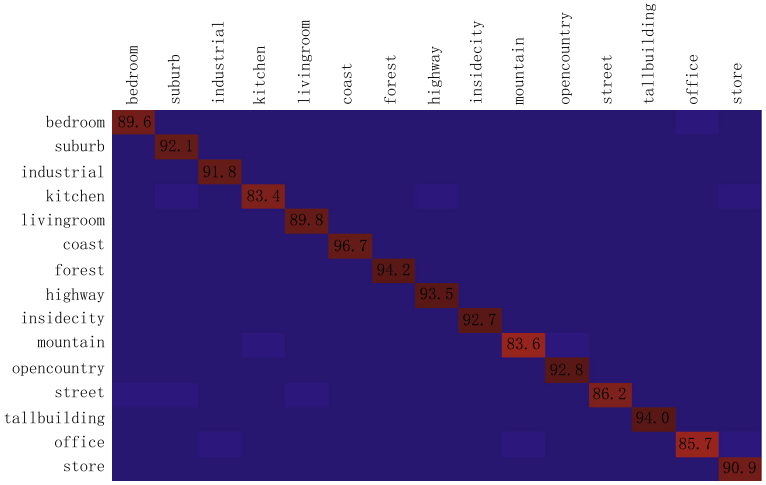

Fig. 5. Confusion matrix on the Scene 15 dataset. The accuracy decreases from red to green. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

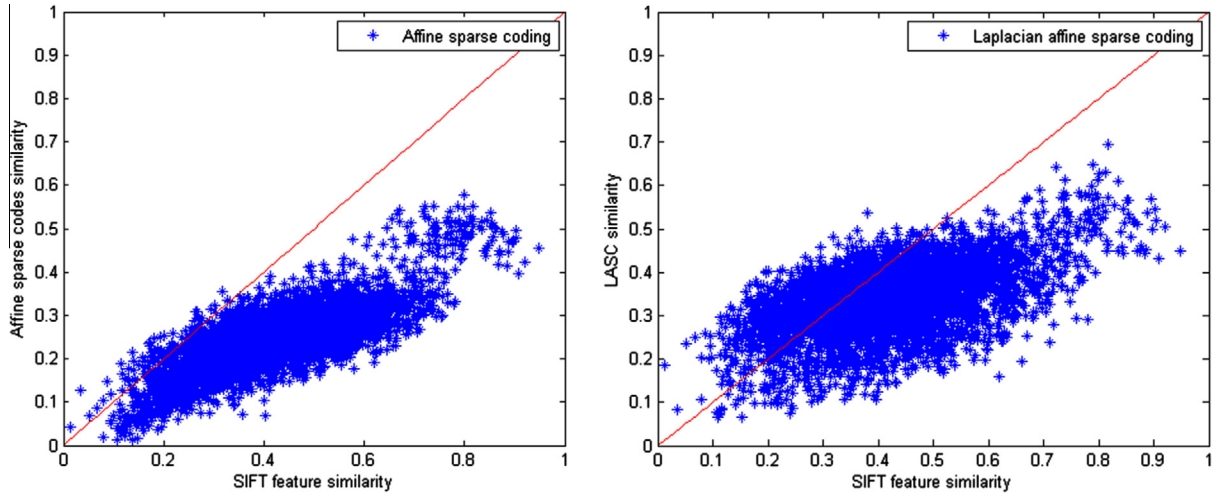

Fig. 6. Similarity correspondence of Laplacian affine sparse codes and affine sparse codes on the Scene 15 dataset. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the effectiveness of the proposed method for robust image classification.

4.5. Analysis of Laplacian affine sparse coding with tilt and orientation consistency

The main contribution of our method lies in two aspects. On one hand, the usage of tilt and orientation constraints help to choose the most discriminative affine SIFT features jointly instead of choosing them by merely minimizing the reconstruction error. On the other hand, the Laplacian regularization term ensures smoothness of the coding parameters. To show the effectiveness of the usage of tilt and orientation constraints, we give their mean and standard derivation as well as the mean reconstruction error on the Scene 15 dataset in Table 4. This is achieved by randomly sampling 1000 local features from the dataset. The tilt and orientation are pre-normalized for easy comparison.

We can see from Table 4 that the mean and standard derivation of the tilt and orientation of the proposed LASC-TOC is smaller than affine sparse coding. Besides, the LASC-TOC also performs better than affine sparse coding for image classification in Table 3. These results demonstrate the effectiveness of imposing tilt and orientation constraints into the objective function for image classification. Moreover, we can see from Table 4 that LASC-TOC performs not as good as affine sparse coding when minimizing the reconstruction error. We believe this is because the objective of sparse coding and image classification are inherently different, the sparse coding tries to minimize the reconstruction error while image classification aims to separate images of different classes correctly. A good local feature encoding scheme should be able to encode local features efficiently for classification instead of merely minimizing the reconstruction error. In fact, the reconstruction error of affine sparse coding is the lower bound of LASC-TOC. If we impose no tilt and orientation as well as smoothness constraints, the reconstruction error of LASC-TOC will equal to affine sparse coding.

To illustrate the influence of Laplacian regularization for similarity preservation, we plot the similarity correspondence of Laplacian affine sparse codes and affine sparse codes on the Scene 15 dataset in Fig. 6. This is achieved by calculating the similarities of about 4000 randomly sampled features. We can see from Fig. 6 that the LASC-TOC preserves more similarity information of local features than affine sparse coding. This demonstrates the effectiveness of Laplacian regularization for preserving the similarities.

5. Conclusions

This paper proposed a novel image classification method by Laplacian affine sparse coding with tilt and orientation consistency. Besides minimizing the reconstruction error of traditional sparse coding, we also impose a regularization constraint to ensure the smoothness of tilt and orientation of affine local features. Moreover, a Laplacian regularization term with histogram intersection similarity is also used to characterize the similarity of affine transformed local features during the sparse coding process. This helps to reduce the information loss and improve the image classification performance. Experimental results on several public datasets show the effectiveness and efficiency of the proposed Laplacian affine sparse coding with tilt and orientation consistency method.

Our future work will concentrate on how to speed up the encoding process of local features. Besides, the usage of other local feature transformations will also be studied.

Acknowledgments

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR): 201204268, China Postdoctoral Science Foundation: 2012M520434, National Basic Research Program of China (973 Program): 2012CB316400, National Natural Science Foundation of China: 61025011, 61272329, 61202325.

References

[1] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of International Conference on Computer Vision, Nice, France 14–17, October 2003, pp. 1470–1477.
[2] R. Fergus, P. Perona, A. Zisserman, A sparse object category model for efficient learning and exhaustive recognition, in: Proceedings of Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 380–387.
[3] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proceedings of Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23, June 2007.
[4] J.C. Gemert, J.M. Geusebroek, C.J. Veenman, A. Smeulders, Kernal codebooks for scene categorization. in: Proceedings of European Conference on Computer Vision, October 2008, pp. 696–709.
[5] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, FL, USA, June 2009, pp. 1794–1801.
[6] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition, in: Proceedings of Computer Vision and, Pattern Recognition, 2011, pp. 1673–1680.

[7] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of Computer Vision and Pattern Recognition, New York, USA, 17–22, June 2006, pp. 2169–2178.
[8] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the International Conference on Image and Video Retrieval, 2007, pp. 401–408.
[9] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, S. Ma, A boosting sparsity-constrained bilinear model for object recognition, IEEE Multimedia 19 (2) (2012) 58–68.
[10] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.
[11] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, in: Proceedings of Computer Vision and, Pattern Recognition, 2005, pp. 994–1000.
[12] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of Computer Vision and, Pattern Recognition, 2001, pp. 511–518.
[13] K. Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1582–1596.
[14] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of Computer Vision and Pattern Recognition, Montbonnot, France, June 2005, pp. 886–893.
[15] N. Xie, H. Ling, W. Hu, X. Zhang, Use bin-ratio information for category and scene classification, in: Proceedings of Computer Vision and, Pattern Recognition, June 2010, pp. 2313–2319.
[16] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: Proceedings of International Conference on Computer Vision, Oct. 2005, pp. 1458–1465.
[17] Z. Wu, Q. Ke, J. Sun, Bundling features for large scale partial-duplicate web image search, in: Proceedings of Computer Vision and, Pattern Recognition, 2009, pp. 25–32.
[18] X. Wang, X. Bai, W. Liu, L. Latecki, Feature context for image classification and object detection, in: Proceedings of Computer Vision and Pattern Recognition, June 2011, pp. 961–968.
[19] B. Yao, A. Khosla, L. Fei-Fei, Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses, in: Proceedings of International Conference on Machine Learning, June 2011.
[20] Y. Lee, K. Grauman, Object-graphs for context-aware category discovery, in: Proceedings of Computer Vision and Pattern Recognition, June 2010, pp. 1–8.
[21] N. Kulkarni, B. Li, Discriminative affine sparse codes for image classification, in: Proceedings of Computer Vision and Pattern Recognition, 20–25, June 2011, pp. 1609–1616.
[22] C. Zhang, Q. Huang, J. Liu, Q. Tian, C. Liang, Image classification using Harr-like 634 transformation of local features with coding residuals, Signal Processing 93 (8) (2012) 2111–2118.
[23] S. Zhang, Q. Tian, G. Hua, Q. Huang, W. Gao, Generating descriptive visual words and visual phrases for large-scale image applications, IEEE Transactions on Image Processing 20 (9) (2011) 2664–2677.
[24] J. Morel, G. Yu, ASIFT: a new framework for fully affine invariant image comparison, SIAM Journal on Imaging Sciences 2 (2) (2009) 438–469.
[25] G. Yu, J. Morel, A fully affine invariant image comparison method, Acoustics, Speech and Signal Processing (2009) 1597–1600.
[26] S. Gao, I. Tsang, L. Chia, P. Zhao, Local features are not lonely-Laplacian sparse coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, CA, USA, June 2010, pp. 3555–3561.
[27] J. Wu, J. Rehg, Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: Proceedings of International Conference on Computer Vision, Kyoto, Japan, Sep. 2009, pp. 630–637.
[28] H. Lee, A. Battle, R. Raina, A. Ng, Efficient sparse coding algorithms, in: Proceedings of Neural Information Processing Systems, British Columbia, Canada, December 2006, pp. 801–808.
[29] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Technical Report, CalTech, 2007.
[30] L. Li, L. Fei-Fei, What, where and who? Classifying events by scene and object recognition, in: Proceedings of International Conference on Computer Vision, Rio, Brazil, October 2007.
[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, CA, USA, June 2010, pp. 3360–3367.