LETTER Action Recognition Using Weighted Locality-Constrained Linear Coding

Jiangfeng YANG^{†a)}, Member and Zheng MA^{†b)}, Nonmember

SUMMARY Recently, locality-constrained linear coding (LLC) as a coding strategy has attracted much attention, due to its better reconstruction than sparse coding and vector quantization. However, LLC ignores the weight information of codewords during the coding stage, and assumes that every selected base has same credibility, even if their weights are different. To further improve the discriminative power of LLC code, we propose a weighted LLC algorithm that considers the codeword weight information. Experiments on the KTH and UCF datasets show that the recognition system based on WLLC achieves better performance than that based on the classical LLC and VQ, and outperforms the recent classical systems.

key words: action recognition, action representation, weighted localityconstrained linear coding

1. Introduction

In recent work on object recognition and action recognition, the bag-of-features (BoF) is one of the most popular models for feature design. There are three stages in BoF-based human action recognition: extracting local features from videos, obtaining a video representation vector via these local features, and classifying action videos with a classifier based upon the video representation vector. To obtain the video representation vector, several feature coding and pooling schemes have been developed. Many authors used k-means and vector quantization (VQ) for feature coding, as well as average-pooling to group these feature codes to generate the video representation vector. To reduce the quantization error caused by k-means and VQ that assign one codeword for a feature, soft vector quantization (SVQ)[3] and sparse coding (SC) [21] are used to encode local features for action recognition tasks.

However, the local features usually reside on nonlinear manifolds [20]. Neither SVQ nor SC can preserve the nonlinear manifold structure. The manifold is nonlinear and not Euclidean in its own space, but is linear and Euclidean in a local region [23]. For this reason, Yu *et al.* [7] provided a Local Coordinate Coding (LCC) to encode feature with locality-constrained, Wang *et al.* [8] introduced an improved version of LCC named Locality-constrained Linear Coding (LLC) to reduce computation cost (see Fig. 2).



Fig. 1 The proposed weighted locality-constrained linear coding (WLLC). (a) sample data **x** and 4 selected bases. (b) selected bases {**b**_{*i*}} and their weight information { w_i }. (c) coding result by classical LLC without considering weight information. z_i denotes coding coefficient related to **b**_{*i*}. (d) weight normalization. (e) corrected coding result \overline{z}_i using normalized weight.



Fig. 2 Comparison between VQ, SC and LLC. The selected bases for representation are highlighted in black.

Due the properties of LLC, such as better reconstruction, local smooth sparsity, and analytical solution, it has attracted much attention.

However, LLC ignores the weight information present in the codebook during coding. A codebook usually contains three types information: sample information, spatiotemporal (ST) position and weight information. Sample information is represented as codeword vector; ST position relationship between codewords can be measured by their Euclidean distance; and weight value of a codeword is decided by the percent of total training samples assigned to it, for instance, if 3 out of total 10 training samples is assigned to a clustering center (or a codeword), the codeword weight is 0.3. Original LLC concentrates on the first two types information only, and ignores the last one. To improve the discrimination performance of LLC code, we propose a weighted locality-constrained linear coding (WLLC) method that makes use of the weight information (see Fig. 1).

Manuscript received July 3, 2014.

Manuscript revised September 2, 2014.

Manuscript publicized October 31, 2014.

[†]The authors are with School of Communication and Information Engineering, University of Electronic Science and Technology of China, P. R. China.

a) E-mail: 369322023@qq.com

b) E-mail: wallsonyang@163.com

DOI: 10.1587/transinf.2014EDL8134

Our paper has two contributions:

- using the weight of codewords, we propose a WLLC algorithm for improving the discriminative power of LLC code.
- the weight information of codewords can be beneficial to boost the performance of recognition system based on WLLC.

The rest of this paper is organized as follows. Related work is presented in Sect. 2. Then, WLLC is proposed in Sect. 3. Experimental results and analysis are shown in Sect. 4. Next, the conclusions are drawn in Sect. 5. Finally, acknowledgements is provided.

2. Related Work

Let us consider a codebook denoted by $\mathbf{B} = {\mathbf{b}_i \in \mathbb{R}^d, i = 1, \dots, K}$. The codebook is constructed on a subset of local descriptors ${\mathbf{x}_i \in \mathbb{R}^d, i = 1, \dots, N}$ extracted from the training dataset.

In the original BoW method [1], coding local descriptors is performed with hard assignment. Each local descriptor is assigned to the nearest visual word, i.e.,

$$z_{i,j} = \begin{cases} 1, & \text{if } j = \arg\min_{j=1,\cdots,K} ||\mathbf{x}_i - \mathbf{b}_j||_2^2, \\ 0, & \text{otherwise}, \end{cases}$$
(1)

with \mathbf{z}_i the code of size K associated to the descriptor \mathbf{x}_i .

As reported in [2], [3], [11], such coding scheme has several limitations, mainly the sensitivity to distortion errors of the codebook, and the significant reconstruction error caused by it. Using sparse coding [6] (SC) as an alternative has significantly improved its robustness to these problems. Therefore, coding is performed by solving the ℓ_1 -norm regularized approximate problem:

$$\mathbf{z}_{i} = \arg\min_{\mathbf{z} \in \mathfrak{R}^{K}} ||\mathbf{x}_{i} - \mathbf{B}\mathbf{z}||_{2}^{2} + \lambda ||\mathbf{z}||_{1}, \lambda \in \mathfrak{R}$$
(2)

where **B** denotes a over-complete codebook with *K* atoms, **z** denotes the reconstruction coefficient associated to sample \mathbf{x}_i . Nevertheless, this optimization problem is computationally expensive and leads to non-consistent encoding of similar descriptors [8]. Indeed, it might select different bases for similar descriptors due to the over-completeness of the codebook, which results in large deviations in representing similar local features.

As suggested in [5], locality is more essential than sparsity, as locality must lead to sparsity but not necessary vice versa. Therefore, authors of [8], [9] proposed more efficient and consistent coding methods relying on the locality property introduced by [7]. Their hypothesis is that descriptors approximately reside on a lower dimensional manifold in an ambient descriptor space. Then, using Euclidean distances for assigning descriptors to visual words is only meaningful within a local region. Hence, local bases are selected to perform the coding. The formulation of original LLC [8] is the following:

$$\mathbf{z}_{i} = \underset{\mathbf{z} \in \mathfrak{R}^{K}}{\arg \min} \|\mathbf{x}_{i} - \mathbf{B}\mathbf{z}\|_{2}^{2} + \lambda \|\mathbf{d}_{i} \odot \mathbf{z}\|_{2}^{2},$$

s.t. $\mathbf{1}^{T} \mathbf{z}_{i} = 1$ (3)

where $\mathbf{d}_i = \exp(\frac{\operatorname{dist}(\mathbf{x}_i, \mathbf{B})}{\sigma})$, and distance vector $\operatorname{dist}(\mathbf{x}_i, \mathbf{B}) = [\operatorname{dist}(\mathbf{x}_i, \mathbf{b}_1), \cdots, \operatorname{dist}(\mathbf{x}_i, \mathbf{b}_K)]^T$ the Euclidean distances between \mathbf{x}_i and the basis vectors; and parameter σ controls the weight decay speed for the locality; \odot denotes the element-wise multiplication. We usually further normalize \mathbf{d}_i to be between (0, 1] by subtracting max(dist(x_i, \mathbf{B})) from dist(x_i, \mathbf{B}). The constraint $\mathbf{1}^T \mathbf{c}_i = 1$ makes sure the shift-invariant requirements of the LLC code. Note that the LLC code in (3) is not sparse in the sense of ℓ_0 norm, but is sparse in the sense that the solution only has few significant values. In practice, we simply threshold those small coefficients to be zero.

3. Weighted Locality-Constrained Linear Coding

In [24], Leibe *et al.* learned an Implicit Shape Model (ISM) based upon local appearance features of images/videos for object detection tasks. An ISM consists of two components: a class-specific alphabet (the codebook) of local appearances that are prototypical for the object category, and a spatial probability distribution which specifics where each codebook entry may be found on the object. In [25], Gall *et al.* employed Hough forest to partition the input sample space, and the set leaves in the Hough forest are regarded as an implicit appearance codebook, each leaf node contains its center position and the spatial distribution of its training sample. Inspired by the success of the work in [24], [25] that incorporated the spatial probability distribution around a codeword/leaf node, we propose an improved version of LLC named Weighted LLC (WLLC).

To achieve good classification performance, the coding scheme should generate similar codes for similar descriptors. Following this requirement, the locality regularization term $\|\mathbf{d}_i \odot \mathbf{z}\|_2^2$ in Eq. (3) presents two attractive properties, which ensure the model LLC+BoF outperform either VQ+BoF or SC+BoF:

- Better reconstruction. In VQ, each descriptor is represented by a single basis in the codebook, as illustrated in Fig. 2 (a). Due to the large quantization errors, the VQ code for similar descriptors might be very different. Besides, the VQ process ignores the relationships between different bases. Hence nonlinear kernel projection is required to make up such information loss. On the other side, as shown in Fig. 2 (c) in LLC, each descriptor is more accurately represented by multiple bases, and LLC code captures the correlations between similar descriptors by sharing bases.
- Local smooth sparsity. Similar to LLC, SC also achieves less reconstruction error by using multiple bases. Nevertheless, the regularization term of ℓ_1 norm in SC is not smooth. As shown in Fig. 2 (b), due to the over-completeness of the codebook, the SC process might select quite different bases for similar patches to

favor sparsity, thus losing correlations between codes.

The reason why WLLC+BoF outperforms LLC+BoF mainly lies in that using the weight information of codewords, the effect of noisy codewords upon the resulting codes can be reduced. Specifically, during the stage of building a codebook, the space of training samples is divided into lots of cluster centers (codewords). The number of samples assigned to different centers usually are various. If the sample number of a center is much less than the others, the center could be a noisy codeword, and its contribution to the coding should be reduced to enhance the discriminative power of the resulting codes. In WLLC, the influence of the noisy codewords is oppressed, while, in LLC, all the selected bases, including noisy bases, makes equal contribution to the coding. As a result, our codes have more discriminative power, and the recognition system based on WLLC+BoF achieves higher performance.

It can be observed in (3) that LLC algorithm ignores the weight information of codewords. More specifically, given a sample **x**, our mission is to find a reconstruction coefficient vector $\mathbf{z} = [z_1, \dots, z_K]$, where z_k is coefficient component related to the base vector \mathbf{b}_k . In (3), z_k is decided by two factors: the distance dist(\mathbf{x}, \mathbf{b}_k) and the similarity sim(\mathbf{x}, \mathbf{b}_k). The value dist(\mathbf{x}, \mathbf{b}_k) decides whether the base vector \mathbf{b}_k is selected as nearest bases for reconstructing \mathbf{x} , and sim(\mathbf{x}, \mathbf{b}_k) determines the value assigned to z_k .

In LLC scheme, the weight of every codeword is considered as same, and everyone makes equal contribution to reconstruction task, whatever their weight are. In our opinion, however, compared to the low-weighted codewords, the high-weighted ones should make more contribution to LLC coding stage. In order to produce more discriminative power of LLC code, we proposed a improved version of LLC named WLLC method that utilizes the weight information in codebook. The proposed WLLC is given as follows:

1. Given a sample data $\mathbf{x} \in \mathfrak{R}^d$, $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{N_s}$ denote N_s selected neighboring bases for reconstructing \mathbf{x} ; and $w_1, w_2, \cdots, w_{N_s}$ denote their corresponding weights. Weight w_n of codeword \mathbf{b}_n is defined by

$$w_n = \frac{\text{training samples assigned to codeword } \mathbf{b}_n}{\text{total training samples for building } \mathbf{B}}$$
(4)

- 2. Once the neighboring bases surrounding **x** is determined, their weights are normalized and denoted as $\overline{w}_1, \overline{w}_2, \cdots, \overline{w}_{N_s}$.
- 3. Finally, weight information of codewords are fused into weighted coefficient vector $\overline{\mathbf{z}} = [\overline{z}_1, \overline{z}_2, \cdots, \overline{z}_{N_s}]$, where $\overline{z}_n = z_n . \overline{w}_n$.

In normalization stage, to achieve the property of scaleinvariance, the codeword weights should be normalized. Besides the simplest normalization method – sum normalization (ℓ_1 normalization)[8] that is widely used in image processing, entropy normalization and exponent normalization are utilized to normalize the codeword weight

 Table 1
 The ARRE on the KTH, UCF Sports datasets in the training stage. For LLC and WLLC, the number of selected bases is set as 5. The codebook size is set as 500 for the two datasets.

Methods	KTH(%)	UCF Sports(%)
BoF+LLC	0.0015	0.0016
BoF+WLLC (sum norm.)	0.42	0.46
BoF+WLLC (entropy norm.)	0.25	0.21
BoF+WLLC (exponent norm.)	0.21	0.28

 $\mathbf{w} = [w_1, \cdots, w_K], 0 < w_n < 1:$

- sum normalization: $\overline{w}_n = w_n / \sum_j w_j$.
- entropy normalization: $\overline{w}_n = (-w_n).ln(w_n)/\sum_j(-w_j).$ $ln(w_j)$, where $(-w_n).ln(w_n)$ is the entropy of codeword weight $w_n, 0 < w_n < 1$.
- exponent normalization: $\overline{w}_n = \exp(w_n) / \sum_j \exp(w_j)$.

where $(-w_n).ln(w_n)$ is the entropy of codeword weight w_n . In information theory, Shannon entropy [26] is calculated to measure the amount of information taken by a symbol. Here, using Shannon entropy, the weight of a codeword is transformed into the amount of information. We use exponential function exp(x) rather than Gaussian function $\exp(-x^2/\sigma^2)$, because when the input x changes slightly, Gaussian function produces the output with significantly changes, and the parameter σ has a great impact on the Gaussian function and its value is hard to be set. In contrast to the Gaussian function, exponential function keeps the smoother output when the input changes. In addition, the reconstruction error $\|\mathbf{x}_i - \mathbf{B}\overline{\mathbf{z}}_i\|_2$ caused by the normalization should be kept lower level to guarantee the corrected codes with high quality. The average relative reconstruction error (ARRE) is defined as follows:

$$ARRE = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\|\mathbf{x}_i - \mathbf{B}\overline{\mathbf{z}}_i\|_2}{\|\mathbf{x}_i\|_2} \right)$$
(5)

where \mathbf{x}_i denotes the *i*-th training sample; $\overline{\mathbf{z}}_i$ denotes the normalized code of \mathbf{x}_i ; **B** is the codebook; *N* is the number of training samples. In LLC, $\overline{\mathbf{z}}_i = \mathbf{z}_i$. Table 1 presents the ARRE in the training stage on the KTH and UCF Sports datasets. It can be seen that compared with the ARRE of LLC, the ARRE increases considerably after normalization. Nevertheless, in contrast to the improvement on recognition accuracy, such ARRE is acceptable.

4. Experiments and Analysis

4.1 Experimental Datasets

We evaluated our approach on two widely used datasets: KTH dataset and UCF sports dataset.

 The KTH dataset contains six types of human action examples (i.e., boxing, hand clapping, hand waving, jogging, running, and walking) are performed by 25 different subjects. Each action is performed in four scenarios: indoors, outdoors, outdoors with scale variation, and outdoors with different clothes. It contains





Fig. 3 Examples from the two public datasets: (a) KTH dataset. (b) UCF sports dataset.



Fig. 4 Both the number of nearest bases and weight normalization methods can effect the discriminative power of WLLC method. (a) Recognition result on the KTH dataset. (b) Recognition result on the UCF Sports dataset.

600 low-resolution video sequences (160×120 pixels). Examples of this dataset are shown in Fig. 3 (a).

• The UCF sports dataset includes a set of 150 videos, which are collected from various broadcast sports channels such as BBC and ESPN. It contains 10 different actions: diving, golf swing, horse riding, kicking, lifting, running, skating, swing bar, swing floor, and walking. This dataset is challenging for a wide range of scenarios and viewpoints. Examples of this datasets are presented in Fig. 3 (b).

4.2 Experimental Setup

In all experiment, spatio-temporal interest points (STIPs) are detected by using Dollar detector proposed in [12], and cuboids is adopted to extract ST local features, and HOG+HOF[13] is adopted to describe these features.

 Table 2
 The average recognition rate of BoF model with VQ, LLC, and

 WLLC on the KTH, UCF Sports datasets. For LLC and WLLC, the number of selected bases is set as 5.

Methods	KTH(%)	UCF Sports(%)
BoF+VQ	91.3	82.8
BoF+LLC	94.5	88.9
BoF+WLLC (sum norm.)	94.7	89.5
BoF+WLLC (entropy norm.)	95.1	90.7
BoF+WLLC (exponent norm.)	95.5	91.1

Table 3Performance comparison with other systems.

Methods	Year	KTH(%)	UCF Sports(%)
Zhu et al. [15]	2010	94.9	84.3
Wu et al. [16]	2011	94.5	91.3
Guha et al. [17]	2012	_	91.1
Bregonzio et al. [18]	2012	94.3	_
Saghafi et al. [19]	2012	92.6	_
BoF+WLLC (sum norm.)		94.7	89.5
BoF+WLLC (entropy norm.)		95.1	90.7
BoF+WLLC (exponent norm.)		95.5	91.1

The multi-scale version Dollar detector, whose spatial scale $\tau = [1.2, 1.3, 1.4, 1.5]$ and temporal scale $\omega = [0.4, 0.45, 0.5, 0.55]$, is used to extract STIPs. The codebook is constructed with k-means algorithm at Euclidean distance as metric over local mixed features, and its size is set as 500 (K = 500). To compare the recognition performance based on different coding strategies, VQ, LLC and WLLC are used. Specifically, local features are encoded as corresponding coefficient vectors by coding algorithm, next, a coefficient histogram is obtained by average-pooling method [14] over coefficient vectors from one action video. As a result, an action video is represented as a coefficient histogram. In classification, a linear support vector machine (SVM) is employed to classify over these ℓ_2 normalized coefficient histograms.

The leave-one-out cross-validation (LOOCV) is used to evaluate the performance of our algorithm. For the KTH dataset, local features from all videos of one subject are used to construct a codebook by *k*-means clustering algorithm. For each LOO run, we learn a model from the videos of 24 subjects, test the videos of the remaining subject. The recognition rate is the average value of the 25 runs. For UCF sports, features from 20 videos (2 videos selected from each action, 10 class actions) are used to build a codebook by *k*means. In each LOO, one video of each class is randomly selected as test data, the other videos are treated as training data. 100 LOO runs are carried out. The recognition rate is the average value of the 100 runs.

4.3 Experimental Results

In Table 2, we can see that action recognition systems based on the proposed WLLCs achieve better performance than that based on VQ and LLC, since the codeword weight information is considered during coding stage, and that exponent normalization method is reasonable than both entropy and sum normalization.

From Table 3, it is clear that recognition system based

on BOF+WLLC(exponent norm.) outperforms the classical systems on the two datasets. And the accuracies of recognition systems based on BOF+WLLC(exponent norm.)/(sum norm.) slightly drop.

5. Conclusion

The classical LLC algorithm ignores the weight information of codewords during coding process. To improve the coding performance of classical LLC algorithm, this paper proposed the WLLC algorithm, which takes account of such weight information. Experiment on the KTH and UCF Sports datasets shows that action recognition system based on WLLC outperforms that based on VQ and LLC methods.

Acknowledgements

This research is supported by National Nature Science Foundation of China (Grant no.61271288).

References

- J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," Proc. IEEE International Conference on Computer Vision, vol.2, pp.1470–1477, 2003.
- [2] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, pp.1294–1309, 2009.
- [3] J. van Gemert, C. Veenman, A. Smeulders, et al, "Visual word ambiguity," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.7, pp.1271–1283, 2010.
- [4] A. Coates and A.Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," Proc. 28th International Conference on Machine Learning (ICML), pp.921–928, 2011.
- [5] R. Rigamonti, M.A. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?" IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp.1545–1552, 2011.
- [6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," CVPR, pp.1794–1801, 2009.
- [7] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," Advances in Neural Information Processing Systems, pp.2223–2231, 2009.
- [8] J. Wang, J. Yang, and K. Yu, "Locality-constrained linear coding for image classification," CVPR, pp.3360–3367, 2010.
- [9] L. Liu, L. Wang, and X. Liu, "In defense of softassignment coding," in ICCV, 2011.
- [10] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image

classification," CVPR, pp.1753-1760, 2011.

- [11] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," CVPR, pp.2415–2422, 2008.
- [12] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," Proc. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.452–460, Oct. 2005.
- [13] L. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," CVPR, pp.3222–3229, 2008.
- [14] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," CVPR, pp.1794–1801, 2009.
- [15] Y. Zhu, X. Zhao, and Y. Fu, "Sparse coding on local spatialtemporal volumes for human action recognition," Proc. Computer Vision, pp.342–352, Berlin, Germany, 2010.
- [16] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.489–496, 2011
- [17] T. Guha and R.K. Ward, "Learning sparse representations for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.8, pp.1576–1588, 2012.
- [18] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," Pattern Recognit., vol.45, no.3, pp.1220–1234, March 2012.
- [19] K. Saghafi and D. Rajan, "Human action recognition using Posebased discriminant embedding," Signal Process., vol.27, pp.96–111, 2012.
- [20] X. Deng, X. Liu, and M. Song, "LF-EME: Local features with elastic manifold embedding for human action recognition," Neurocomputing, vol.99, no.1, pp.144–153, 2013.
- [21] B.A. Olshausen and D.J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, vol.381, no.6583, pp.607–609, 1996.
- [22] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323– 2326, 2000.
- [23] X. Zhu, Z. Yang, and J. Tsien, "Statistics of natural action structures and human action recognition," J. Vision, vol.12, no.9, pp.834–834, 2012.
- [24] B. Leibe, A. Leonardis, and B. Schiele, "Robust objector detection with interleaved categorization and segmentation," Int. J. Comput. Vis., vol.77, no.1-3, pp.259–289, 2008.
- [25] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, Issue. 11, pp.2188–2202, 2011.
- [26] C.E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J., vol.27, no.3, pp.379–423, 1948.