LETTER

Learning a Similarity Constrained Discriminative Kernel Dictionary from Concatenated Low-Rank Features for Action Recognition

Shijian HUANG[†,††a)], **Junyong YE**[†b)], **Tongqing WANG**[†], **Li JIANG**[†††], **Changyuan XING**[††††], ***Nonmembers***, ***and*** **Yang LI**[†], ***Student Member***

SUMMARY Traditional low-rank feature lose the temporal information among action sequence. To obtain the temporal information, we split an action video into multiple action subsequences and concatenate all the low-rank features of subsequences according to their time order. Then we recognize actions by learning a novel dictionary model from concatenated low-rank features. However, traditional dictionary learning models usually neglect the similarity among the coding coefficients and have bad performance in dealing with non-linearly separable data. To overcome these shortcomings, we present a novel similarity constrained discriminative kernel dictionary learning for action recognition. The effectiveness of the proposed method is verified on three benchmarks, and the experimental results show the promising results of our method for action recognition.

key words: *low-rank feature, similarity constraint, kernel method, human action recognition*

1. Introduction

Human action recognition is a very active research topic with many important applications. Many approaches have been proposed for action recognition [1], [2]. However, many traditional methods usually depend on accurate actor segmentation, body tracking or interest point detection. The low-rank feature proposed in our previous work [3] can well avoids these intermediate steps. Assuming that there is an action sequence, we first reshape each frame into a column vector. Then, the all column vectors will be stacked as the columns of a data matrix. Finally, the data matrix will be decomposed by robust principal component analysis to obtain the low-rank feature of action sequence. The obtained low-rank feature can well capture the motion information and remove action background. However the manner of extracting low-rank feature in the work [3] neglects the temporal information among the action sequence. To overcame this shortcoming, we split an entire action sequence into multiple overlapping subsequences, and concatenate the low-rank features of all subsequences according to their time order. The concatenated low-rank features naturally fit the dictionary learning, thus we present a novel dictionary learning model from concatenated low-rank features to recognize actions.

Recently, many researchers have given attention to dictionary learning [4]–[6]. However, traditional dictionary learning usually neglects the similarity among the coding coefficients and have poor performances in dealing with non-linearly separable data. To overcame these shortcomings, we present a novel similarity constrained discriminative kernel dictionary learning (SCDKDL). We test our method on KTH, UCF Sports and HMDB51 datasets, and experimental results demonstrate the competitive performance of our proposed method.

The remainder of the letter is organized as follows: In Sect. 2, we introduce our proposed approach. In Sect. 3, we verify the proposed method on three benchmarks. Finally, we conclude this letter in Sect. 4.

2. Proposed Approach

2.1 Concatenated Low-Rank Features

We first split an action sequence into multiple overlapping subsequences according to a overlapping ratio θ. Afterwards, we extract the low-rank features of all subsequences. Then, each low-rank feature will be represented as a feature vector by accumulated edge distribution histogram (AEDH) descriptor [3]. The AEDH is specifically designed to describe the low-rank feature, which counts the edge distribution of low-rank image transformed from low-rank feature. Finally, we concatenate all feature vectors into a feature matrix according to their time order. Figure 1 present the illustration of constructing a feature matrix with temporal information.

2.2 Proposed Dictionary Learning Model

The similarity constrained can force the coding coefficients from a same category as similar as possible and that from

Manuscript received July 9, 2015.
Manuscript revised September 17, 2015.
Manuscript publicized November 16, 2015.

[†]The authors are with the Key Laboratory of Optoelectronic Technology and Systems of the Ministry of Education, Chongqing University, Chongqing, China.
[††]The author is with the School of Electronic Information Engineering, Yangtze Normal University, Fuling, Chongqing, China.
[†††]The author is with the College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China.
[††††]The author is with the School of Computer Engineering, Yangtze Normal University, Fuling, Chongqing, China.
a) E-mail: huangshijian@cqu.edu.cn
b) E-mail: ygyocr@cqu.edu.cn
DOI: 10.1587/transinf.2015EDL8148

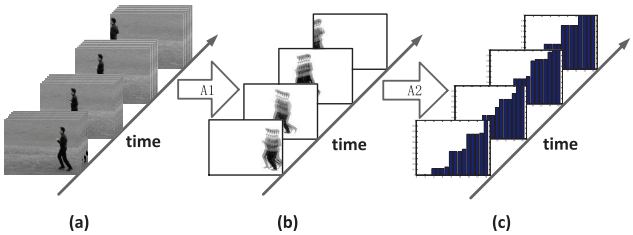

Fig. 1 Illustration of concatenating low-rank feature. (a) Split original action sequence. (b) Concatenated low-rank features. (c) Corresponding features matrix. Arrow A1 denotes low-rank feature extraction, and arrow A2 denotes AEDH representation.

different categories as different as possible, which is conducive to improve the classification performance. Thus, we add the similarity constrained into our dictionary learning model and employ kernel method to enhance the ability of our model in processing non-linearly separable data.

Let $X = [X_1, X_2, \ldots, X_C]$ be a set of training samples with C action categories, where $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}]$ and N_i is the number of training samples of the ith action category. The $x_{i,j} \in R^{P\times Q}$ is the feature matrix of the jth training sample in the ith category, where P is the length of the feature vector of each low-rank feature and Q is the number of the split subsequences. Our goal is learning a shared dictionary $D \in R^{P\times K}$. To facilitate optimization, we further use the sparse dictionary model [7] (i.e., $D = D_0U$), where $D_0 \in R^{P\times K}$ denotes a predefined base dictionary and $U = [u_1, u_2, \ldots, u_K] \in R^{K\times K}$ is a representation matrix.

Let $\phi : L \mapsto H$ denote a non-linear mapping from low dimensional space into a higher dimensional space. Then the mapped X_i can be denoted as $\phi(X_i)$. The mapped dictionary can be denoted as $\phi(D) = \phi(D_0)U$. We use the Gaussian kernel to compute the dot product of features x_i and x_j in the high-dimensional space, i.e., $\kappa(x_i, x_j) = \phi(x_i)^T\phi(x_j) = exp(-\lambda\|x_i - x_j\|^2)$. Finally, our proposed dictionary learning model can be formulated as:

$$\begin{aligned} < U^*, A^*, W^* > \; &= \arg\min_{D,A,W} \frac{1}{2}\sum_{i=1}^{C}\|\phi(X_i) - \phi(D_0)UA_i\|_F^2 \\ &+ \frac{\alpha}{2}S(B) + \frac{\beta}{2}\|W\|_F^2 \\ &+ \frac{\gamma}{2}\sum_{i=1}^{C}\|Y_i - WB_i\|_F^2 \\ s.t.\ \ &\|\phi(D_0)u_k\|_2^2 = 1, \quad \forall k = 1, 2, \ldots, K \end{aligned} \tag{1}$$

where $S(B)$ is defined as:

$$\begin{aligned} S(B) \; = \; &\sum_i\sum_{m\neq i}\sum_j\sum_n \|b_{i,j}^T b_{m,n}\|_2^2 \\ &- \sum_i\sum_j\sum_n \|b_{i,j}^T b_{i,n}\|_2^2 \end{aligned} \tag{2}$$

In Eq. (1), $A_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,N_i}] \in R^{K\times QN_i}$ are the coding coefficient matrix of X_i. $B = [B_1, B_2, \ldots, B_C]$, $B_i = [b_{i,1}, b_{i,2}, \ldots, b_{i,N_i}] \in R^{Q\times N_i}$. $b_{i,j} = a_{i,j}^T I_K \in R^Q$, where I_K is a vector of length K, and its each element is equal to $1/K$. $\|W\|_F^2$ is a regularization penalty term. $\|Y_i - WB_i\|_F^2$ is a classification error term, where $Y_i = [y_{i,1}, y_{i,2}, \ldots, y_{i,N_i}] \in R^{C\times N_i}$, $y_{i,j} = [0, 0, \ldots, 1, \ldots, 0, 0]^T$ (the nonzero position indicates the class). $W \in R^{C\times Q}$ is a linear classifier. α, β and γ are three scalars controlling the relative contribution of the corresponding terms. Note that we constraint B instead of A in Eq. (2). The reason is that the B is directly related to the training of the classifier W.

2.3 Optimization Method

The SCDKDL model can be solved in an iterative manner. Specifically, we iteratively minimize the objection function over the atom representation matrix U, coding coefficients matrix A, and linear classifier W until the final optimal values of them are obtained.

We first initialize U and A. Specifically, we group all the low-rank features of an action dataset into K clusters by using k-means algorithm, and set the K centres as the initial base dictionary D_0. U can be initialized as an identity matrix $I_{K\times K}$, and A can be initialized as:

$$< A^* >= \arg\min_A\{f_1(A) = \frac{1}{2}\|\phi(X) - \phi(D_0)UA\|_F^2\} \tag{3}$$

Let the partial derivative of $f_1(A)$ equal zero, i.e., $\frac{\partial f_1(A)}{\partial A} = -U^T\phi^T(D_0)(\phi(X) - \phi(D_0)UA) = 0$. The initial A^0 can be computed as $A^0 = (U^T\kappa(D_0, D_0U)^{-1}U^T\kappa(D_0, X)$.

Then we fix U and A, and update W. Here, Eq. (1) can be rewritten as:

$$< W^* >= \arg\min_W\{f_2(W) = \frac{\beta}{2}\|W\|_F^2 + \frac{\gamma}{2}\|Y - WB\|_F^2\} \tag{4}$$

Let $\frac{\partial f_2(W)}{\partial W}$ equal zero, i.e., $\frac{\partial f_2(W)}{\partial W} = \beta W - \gamma(Y - WB)B^T = 0$. The optimal $W^* = \gamma YB^T(\beta I_{K\times K} + \gamma BB^T)^{-1}$.

Afterwards, we update A atom by atom with U and W fixed. Here, Eq. (1) can be rewritten as:

$$\begin{aligned} < a_{i,j}^* > \; &= \arg\min_{a_{i,j}} f_3(a_{i,j}) \\ &= \arg\min_{a_{i,j}} \frac{1}{2}\|\phi(x_{i.j}) - \phi(D_0)Ua_{i,j}\|_F^2 \\ &+ \frac{\alpha}{2}\sum_{m\neq i}\sum_n \|I_K^T a_{i,j}a_{m,n}^T I_K\|_2^2 \\ &- \frac{\alpha}{2}\sum_n \|I_K^T a_{i,j}a_{i,n}^T I_K\|_2^2 \\ &+ \frac{\gamma}{2}\|y_{i,j} - Wa_{i,j}^T I_K\|_2^2 \end{aligned} \tag{5}$$

The partial derivative of $f_3(a_{i,j})$ can be computed as:

$$\begin{aligned} \frac{\partial f_3(a_{i,j})}{\partial a_{i,j}} \; &= -U^T\kappa(D_0, x_{i,j}) + U^T\kappa(D_0, D_0)Ua_{i,j} \\ &+ \alpha\sum_{m\neq i}\sum_n I_K(I_K^T a_{i,j}a_{m,n}^T I_K)I_K^T a_{m,n} \\ &- \alpha\sum_n I_K(I_K^T a_{i,j}a_{i,n}^T I_K)I_K^T a_{i,n} \\ &- \gamma I_K(y_{i,j} - Wa_{i,j}^T I_K)^T W \end{aligned} \tag{6}$$

We update $a_{i,j}$ via $a_{i,j}^{(t+1)} = a_{i,j}^{(t)} - \rho \frac{\partial f_3(a_{i,j})}{\partial a_{i,j}}$, where ρ is a step length. In the solving process, we use two termination criteria, i.e., the algorithm runs out the maximum iteration number T or $\|a_{i,j}^{(t+1)} - a_{i,j}^{(t)}\|_F^2 \leq \epsilon$ (ϵ is a preset threshold).

Finally, we update U atom by atom with W and A fixed. Here, Eq. (1) can be rewritten as:

$$< U^* >= \arg\min_U \frac{1}{2}\|\phi(X) - \phi(D_0)UA\|_F^2$$
$$s.t. \quad \|\phi(D_0)u_k\|_2^2 = 1, \quad \forall k = 1, 2, \ldots, K \tag{7}$$

We define an intermediate variable $\phi(\widetilde{X}) = \phi(X) - \phi(D_0)\widetilde{u_k}\widetilde{A_k}$, where $\widetilde{u_k}$ denotes U discarding the kth column and $\widetilde{A_k}$ is A discarding the kth row. Then Eq. (7) can be rewritten as:

$$< u_k^* >= \arg\min_{u_k}\{f_4(u_k) = \frac{1}{2}\|\phi(\widetilde{X}) - \phi(D_0)u_k a_k\|_F^2\}$$
$$s.t. \quad \|\phi(D_0)u_k\|_2^2 = 1, \quad \forall k = 1, 2, \ldots, K \tag{8}$$

Where a_k denotes the kth row of A. Let its partial derivative equal to zero, i.e., $\frac{\partial f_4(u_k)}{\partial u_k} = -\phi^T(D_0)(\phi(\widetilde{X}) - \phi(D_0)u_k a_k a_k^T = 0$, then the optimal u_k can be computed as $u_k = \frac{1}{a_k a_k^T}\kappa^{-1}(D_0, D_0)\kappa(D_0, \widetilde{X})a_k^T$. Considering the constraint term $\|\phi(D_0)u_k\|_2^2 = 1$, we finally update $u_k^* = \frac{u_k}{\|\widehat{u_k}\|_2}$, where $\|\widehat{u_k}\|_2 = \sqrt{u_k^T\kappa(D_0, D_0)u_k}$. To maintain the consistent values of $\|\phi(\widetilde{X}) - \phi(D_0)u_k a_k\|_F^2$ and $\|Y - WB\|_F^2$, we also update $a_k^* = a_k\|\widehat{u_k}\|_2$ and $W^* = \frac{W}{\|\widehat{u_k}\|_2}$ simultaneously.

2.4 Recognition Protocol

Given a new test video v_t, we first extract its concatenated low-rank features x_t, and compute the corresponding coding coefficient matrix a_t by $a_t = (U^T\kappa(D_0, D_0)U)^{-1}U^T\kappa(D_0, x_t)$. Then we further pool the a_t into $b_t = a_t^T I_K$. Finally, we assign v_t to the object class by:

$$identity(v_t) = \arg\max_i(l_i), \quad i = 1, 2, \ldots, C \tag{9}$$

where $l_i \in l$, and $l = Wb_t$ is the class label vector.

3. Experimental Results

In this section, We evaluate the performance of the proposed method on KTH and UCF Sports datasets in a leave-one-out cross validation manner, and on HMDB51 dataset in the manner as the method [10].

3.1 Parameter Settings

In our model, the length of a single low-rank feature is normalized to 200 (i.e., $P = 200$). The parameter of Gaussian kernel λ is set to 0.05. The coefficients α, β and γ in Eq. (1) are set to 0.25, 0.1 and 0.3 respectively. In updating A, the step length ρ, the maximum iteration number T and the threshold ϵ are set to 0.5, 30 and 0.001 respectively.

Besides, for KTH, the parameters θ, K and Q are set to 0.3, 1600 and 100 respectively. For UCF Sports, the parameters θ, K and Q are set to 0.4, 1000 and 60 respectively. For HMDB51, the parameters θ, K and Q are set to 0.4, 1200 and 80 respectively.

3.2 Experiment Results on Three Datasets

Figures 2, 3 and 4 show the confusion matrixs on KTH, UCF Sports and HMDB51 datasets respectively. On KTH, the main confusion occurs between jogging and running. On UCF Sports, the running and skateboarding are relatively easy to confuse. On HMDB51, because there are too many action categories, we can only observe the rough confusion among action categories. Finally, with our approach, the overall average accuracies are 98.67% on KTH, 94.67% on

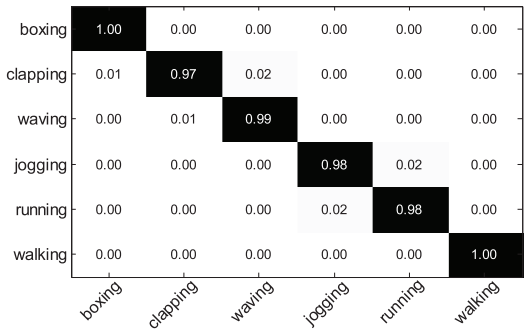

Fig. 2 Confusion matrix on KTH dataset

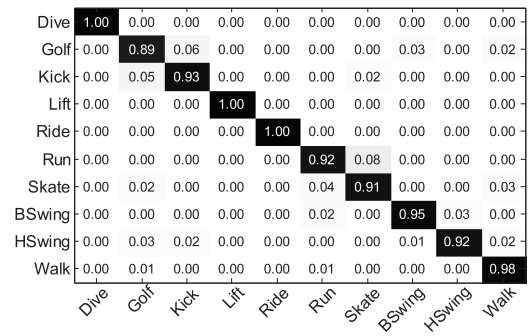

Fig. 3 Confusion matrix on UCF Sports dataset

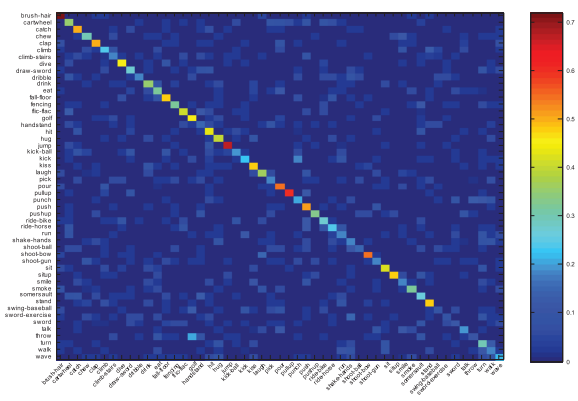

Fig. 4 Confusion matrix on HMDB51 dataset

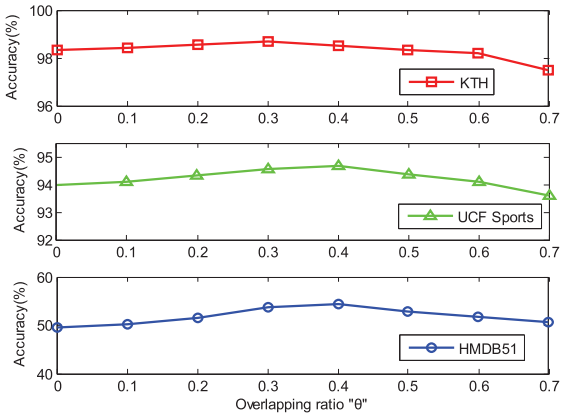

Fig. 5 Performance comparison of different overlapping ratios

Table 1 Performances (%) of our different models on three datasets

Our model	KTH	UCF Sports	HMDB51
Without similarity constraint	92.81	89.33	46.51
Without kernel method	95.65	90	48.37
Without both	89.53	84	45.13
Complete model	**98.67**	**94.67**	**54.29**

Table 2 Performances (%) of different methods on KTH and UCF Sports

Method	Year	KTH	UCF Sports
Zhu et al. [4]	2010	94.92	84.33
Zhang et al. [5]	2012	95.6	87.33
Wang et al. [8]	2013	94.2	88
Zhang et al. [6]	2014	93.8	86.7
Li et al. [9]	2014	96.33	92
Huang et al. [3]	2015	97.32	92.67
Ours		**98.67**	**94.67**

Table 3 Performances (%) of different methods on HMDB51

Method	Year	HMDB51
Kuehne et al. [10]	2011	22.83
Jiang et al. [11]	2012	40.7
Wang et al. [12]	2013	**57.2**
Li et al. [9]	2014	29.6
Wu et al. [13]	2014	47.1
Huang et al. [3]	2015	49.71
Ours		54.29

UCF Sports and 54.29% on HMDB51.

3.3 Comparison Experiments

Figure 5 compares the performances of different overlapping ratios on three datasets, which show the effectiveness of the optimal overlapping ratio θ (i.e., $\theta = 0.3$ for KTH, $\theta = 0.4$ for UCF Sports and HMDB51).

Table 1 lists the comparison results of our different models, which demonstrate the effectiveness of our complete model. Tables 2 and 3 present the comparison of our method with state-of-the-art results on the three datasets. It can be observed that our method outperforms the compared methods besides the method [12] on HMDB51, and achieves the best performances on KTH and UCF Sports datasets.

4. Conclusion

In this letter, to obtain the temporal information of low-rank feature and overcome the shortcomings of traditional dictionary learning model, we present learning a similarity constrained discriminative kernel dictionary from concatenated low-rank features for action recognition. Experimental results on three benchmarks demonstrate the effectiveness of our method. In the future, we will investigate novel dictionary learning models to recognize more complex actions.

Acknowledgement

This work was supported by the Scientific and Technological Research Program of Chongqing Municipal Education Commission of China under Grant KJ1401207, and Fundamental Research Funds for the Central Universities of China under Grant 106112013CDJZR120014.

References

[1] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," IEEE Trans. Circuits Syst. Video Technol., vol.23, no.2, pp.236–243, 2013.
[2] M.A.R. Ahad, J.K. Tan, H. Kim, and S. Ishikawa, "Motion history image: its variants and applications," Mach. Vis. Appl., vol.23, no.2, pp.255–281, 2012.
[3] S. Huang, J. Ye, T. Wang, L. Jiang, G. Wu, and Y. Li, "Extracting refined low-rank features of robust PCA for human action recognition," Arab. J. Sci. Eng., vol.40, no.5, pp.1427–1441, 2015.
[4] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," Computer Vision – ACCV 2010, Lecture Notes in Computer Science, vol.6493, pp.660–671, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
[5] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Action recognition using context-constrained linear coding," IEEE Signal Process. Lett., vol.19, no.7, pp.439–442, 2012.
[6] S. Zhang, H. Yao, X. Sun, K. Wang, J. Zhang, X. Lu, and Y. Zhang, "Action recognition based on overcomplete independent components analysis," Inf. Sci., vol.281, pp.635–647, 2014.
[7] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: learning sparse dictionaries for sparse signal approximation," IEEE Transac. on Signal Process., vol.58, no.3, pp.1553–1564, 2010.
[8] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," Int. J. Comput. Vis., vol.103, no.1, pp.60–79, 2013.
[9] Y. Li, J. Ye, T. Wang, and S. Huang, "Augmenting bag-of-words: A robust contextual representation of spatiotemporal interest points for action recognition," Vis. Comput. (2014). DOI:10.1007/s00371-014-1020-8.
[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," Proc. 2011 IEEE Int. Conf. on Comput. Vis., pp.2556–2563, 2011.
[11] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," European Conference on Computer Vision, vol.7576, pp.425–438, 2012.
[12] H. Wang and C. Schmid, "Action recognition with improved trajectories," Proc. 2013 IEEE Int. Conf. on Comput. Vis, pp.3551–3558, 2013.
[13] J. Wu, D. Hu, and F. Chen, "Action recognition by hidden temporal models," Vis. Comput., vol.30, no.12, pp.1395–1404, 2014.