# Learning Cross-domain Dictionary Pairs for Human Action Recognition

Bingbing Zhang[1, a],Dongcheng Shi[1, b], Kang Ni[1, c] and Chao Liang[1, d]

[1] Changchun University of Technology, Jilin,130012,China

[a]icyzbb123@163.com,[b]1019625929@qq.com, [c]9500467@qq.com, [d]21260914@qq.com

**Keywords:** Human action recognition, Local motion pattern, Dictionary learning.

**Abstract.** This paper present a cross domain dictionary learning way, via the introduction of auxiliary domain,as the extra knowledge, the intra class diversity of the original training set (also known as the targetdomain) is effectively enhanced.Firstly,use local motion pattern feature as a low-level feature descriptor, and then through a cross domain reconstructive dictionary pairlearning, which brings the original target data and the auxiliary domain data into the same feature space to get correspondingsparse codes of each human action categories.Finally, classification and recognition is carried on thehuman action representation. Usingthe UCF YouTubedataset as theoriginal training setand the HMDB51 data set asthe auxiliary data set,the recognition rate of the proposed framework is significantly improvedon the UCF YouTube dataset.

## Introduction

In the past few years,human action recognition has been a hot topic in the field of computer vision. Due to the clutteredbackground, the geometric and photometric changes ofthe target,the application in real world is also a big challenge.

The low-level human action recognition is the basis and the first step of the human behavior Analysis.Generally,the process of the recognition isconsist of two major parts:theaction video representation and recognition.In the step of the description of the features, the local feature description of the human movement target, such as the spatial and temporal key pointshappens in the video contain important information that necessary for the analysis of human behavior. C. Harris and M. J. Stephens[1] proposed the classic corner detection method, the spatial and temporal characteristic expression of the moving objects can be well expressed. Laptev[2] expand the Harris corner detection[3]to the 3D space, which is also a kind of space-time interest points (STIP). We use the LMP descriptor[3],which is expansion of the STIP, to get more useful information about the movement ofthe target. Su et al[4]proposed the semantic features and Yao[5] proposed pose estimation feature. In these works, it is supposed that all test set and the training set in the same feature space and identically distributed. But in the real video surveillance, it cannot be always guaranteed. Insufficient training data, i.e. each action class training only an action template will lead to the reduction of therecognition rate, such as Cao's[6] and Liu's[7] algorithm there are similar problems. In the process of learning of the training set, to solve such problems, we divides the original training set into the target domain and theauxiliarydomain, learning a construction dictionary pair, bring the target domain data and the auxiliary domain data into the same feature space.

The remainder of this paper is organized as follows: section 2 discusses the work of action video representation before the cross-domain dictionary learning and the recognition. Section 3,we discuss related dictionary learning techniques and then introduce the cross domain dictionary learning method.Experimental results are presented in Section 4.We conclude the paper in Section5. The flow-chat of the algorithm is shown in Fig.1.

 

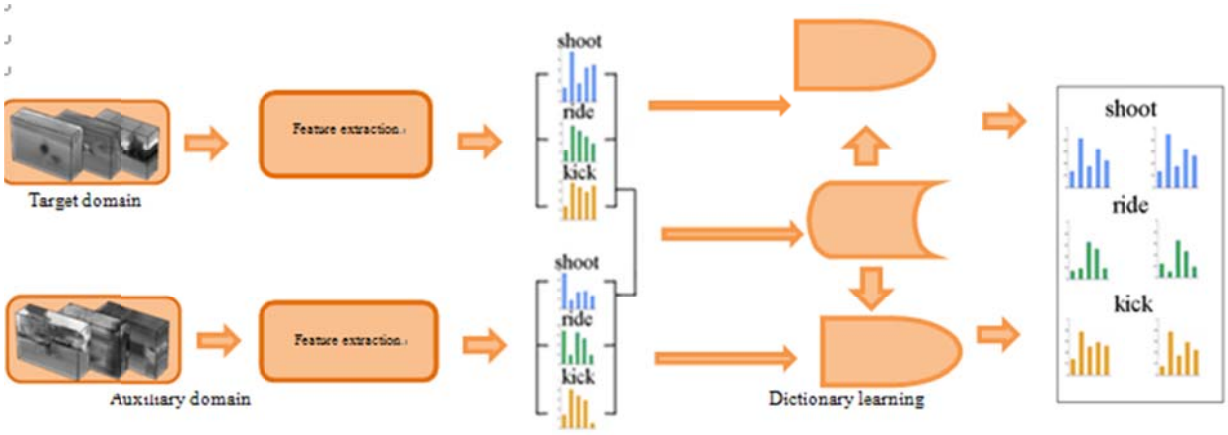


Fig.1 Flow chart of the algorithm

**The Local Motion Pattern Description**

We adopt the local motion pattern description as the low-level action video representation. Consider a video sequence $V(x, y, z)$ consisting of $f$ frames. Then it is first partitioned into S segments: $V = [V_1, V_2, \ldots V_S]$.Each segment contains $l = f/S$ consecutive frames.

**The process of the extraction .**We employ a 2D key points[8] detector in order to extract spatially structures and locate key points at the first frame of every segment. Through observing the temporal change of the key points over the remaining $(l - 1)$ frames,we can get the temporal information. Then the small patches of dimension$(y \times y \times b)$ are extracted around the key points in every segments. We also choose the Cuboid[14] descriptors, since it is widely popular and generates a good number of features.We set y=24, the results of the extraction in three consecutive frames are shown in Fig.2.

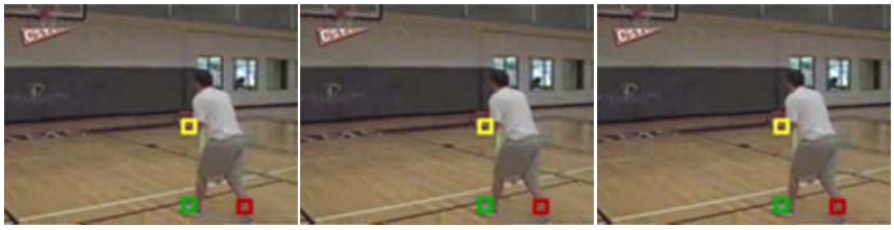

Fig.2 The LMP description extraction of three consecutive frames in UCF YouTube dataset

**The process of the computation of the LMP description.**In the section 2.1,small patches of dimension$(y \times y \times b)$are extracted around the key points in every segments.Firstly,2D Gaussian blurring is performed to each cube we capture above.Let us denote a blurred cube as $v \in R^{y \times y \times m}$.Secondly,The second(variance$(M_2)$,third(skewness,$(M_3)$,and fourth(kurtosis, $(M_4)$ central moments are computed for each pixel along the temporal direction.The moment matrix $M_r$,r={2. 3. 4},associated with v as follows:

$$M_r = [m_{ij}] \quad i,j1,2,\ldots,y(1)$$

Where

$$m_{ij} \sum_{t=1}^{b} (V_{ijt})^r \qquad (2)$$

Here,$V_{ijt}$ is the pixel value of the $t$th patch at location{i,j}.Through transforming every matrix into vectors, the LMP descriptors are formed. The process of computing the LMP descriptors is illustrated in Fig.3.The following vector m is an LMP descriptor. LMP feature vector for a patch of size$(24\times 24)$is of dimension[$1728\times 1$].

$$m = [m_2 m_3 m_4]^T \quad (3)$$

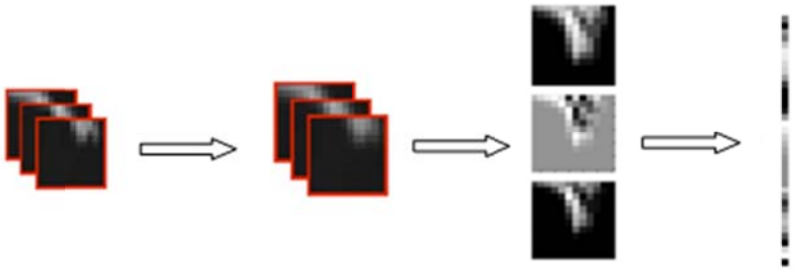

Fig.3 Conversion of a cube to an LMP descriptor

**Cross-domain Dictionary Learning**

**Dictionary Learning.**Let$\mathbf{Y_t} = [\mathbf{y_t^1}, \mathbf{y_t^2}, \dots, \mathbf{y_t^N}]$be the set of target domain n-dimensional input signals. $\mathbf{Y_t}$ can be reconstructed by learning the reconstructive dictionary.Considering the reconstruction error,the transformation can be formulated as:

$$y = D_t x_t + E(x) \quad (4)$$

where $E(x)$ represent the reconstruction， $D_t = [d_t^1, d_t^2, \dots d_t^N]$ is the dictionary of target domain, $x_t = [x_t^1, x_t^2, \dots x_t^N]$ is a set of sparse codes.The way of learning the reconstructive dictionary to get the the sparse representation can be solved by following optimal question.

$$< D_t, X_t >= arg \min_{D_t X_t} \| Y_t - D_t X_t \|_2^2, s.t. \forall i, \| x_t^i \|_0 \leq T \quad (5)$$

Generally,the number of the dictionary atoms is larger than N to guarantee the dictionaryis over-completed.T is the constraint factor that limits the number of non-zero elements in the sparse codes.

**The importance of the way of dictionary learning.**The choice of a method for dictionary learning critically determines the performance of sparse representation.The K-SVD algorithm[9]is a popular and efficient dictionary learningmethod to solve the optimal problem to abstain the dictionary.

But the general algorithm is not considering the characteristics of the training set sample. The performance of sparse representation is poor when the data is not matched.In order to reduce such impact on the recognition, in the dictionary learning step, by the learning of the target domain data set and the auxiliary data set, expanding intra class diversity. This kind of dictionary learning method is so called cross domain dictionary learning.

**The computation process of the CDDL Algorithm**

**The optimal problem.**$\mathbf{Y_t}$ represents L n-dimensional target domain patterns of one human action categories and$\mathbf{Y_s}$ represent M n-dimensional source domain patterns， We need learn one reconstructive dictionary pair to guarantee the global smoothness， So the problem can be solve through following optimization problem:

$$< D_t, D_s, X_t, X_s >= \arg\min_{D_t, D_s, X_t, \emptyset} \| Y_t - D_t X_t \|_2^2 + \| Y_t - D_t X_t \|_2^2 + \emptyset([X_t X_s])$$
$$s.t. \forall i, [\| x_t^i \|_0, \| x_s^i \|_0] \leq T \quad (6)$$

where$\emptyset(\cdot)$ is the distances of similar cross-domain instance of the same category， $D_t = [d_t^1, d_t^2 \dots d_t^N] \in R^{n \times N}$is the learned target domain dictionary， $X_t = [x_t^1, x_t^2 \dots x_t^N] \in R^{N \times L}$ is a set of sparse codes of the target domain。$D_s = [d_s^1, d_s^2 \dots d_s^N] \in R^{n \times N}$ is thelearned soure dictionary， $X_s = [x_s^1, x_s^2 \dots x_s^N] \in R^{N \times M}$ is a set of sparse codes of the source domain。The value of N is designed larger than M to make sure the dictionary is overcompleted.

**Solve the optimal problem.** Make sure that the numbers of the dictionary atoms of $D_t$ and $D_s$ are the same. According to the paper[10],we rewrite the objective function above:

$$< D_t, D_s, X_t, \emptyset, W >= arg \min_{D_t, D_s, X_t, \emptyset, W}$$
$$\| (Y_t Y_s A^T \sqrt{\alpha} Q \sqrt{\beta} H)^T - (D_t D_s \sqrt{\alpha} \emptyset \sqrt{\beta} W)^T \|_2^2$$
$$s.t. \forall i, \| x_t^i \|_0 \leq T \quad (7)$$

whereW are the coefficients of the linear classifier,H are the class labels of the target domain,Q are the target discriminative sparse codes, $\alpha$ and $\beta$can control the contribution of the Q.

The column-wise $L_2$ normalization is applied to D,the optimization problem above canbe solved using the K-SVD.Each dictionaryelement $d_k$ and its non-zero coefficient $x_t^k$ can be computed by

$$< d_k, d_s >= arg\ \min_{d_k, x_t^k} \quad \| E_k - d_k x_t^k \|_F^2,$$

s.t.$\forall i, \|$

$$x_t^i \|_0 \leq T \quad (8)$$

Where$E_k = Y - \sum_{i \neq k} d_i * x_t^i$.K-SVD is used as follows:

$$U \sum V = SVD(E_k)$$
$$\widetilde{d_k} = U(:, 1)$$
$$\tilde{x}_t^k = \sum(1,1) V(1, :) \quad (9)$$

Where $U(:, 1)$ indicates the firsrt column of U，$V(1, :)$ indicates the first raw of V.

## Classification

During the process of solving the optimization problem, $D_t$ , $D_S$ , $\phi$ and W are jointlynormalized.So they can't be directly applied to the construct the classification framework，According to paper[11] $D_t$,$D_S$,$\phi$ and W can be computed as:

$$\widetilde{D}_t = \{\frac{d_t^1}{\|d_t^1\|_2}, \frac{d_t^2}{\|d_t^2\|_2}, \cdots, \frac{d_t^K}{\|d_t^K\|_2}\}$$
$$\widetilde{D}_s = \{\frac{d_s^1}{\|d_s^1\|_2}, \frac{d_s^2}{\|d_s^2\|_2}, \cdots, \frac{d_s^K}{\|d_s^K\|_2}\}$$
$$\widetilde{\phi} = \{\frac{\phi_s^1}{\|\phi_s^1\|_2}, \frac{\phi_s^2}{\|\phi_s^2\|_2}, \cdots, \frac{\phi_s^K}{\|\phi_s^K\|_2}\}$$
$$\widetilde{W} = \{\frac{w^1}{\|w^1\|_2} \frac{w^2}{\|w^2\|_2}, \cdots, \frac{w^K}{\|w^K\|_2}\} \quad (10)$$

Given a target domain query sample $y_t^i$ ,its sparse representation $x_t^i$ can be computed through$\widetilde{D}_t$,with the linear classifier F(x: W),the label of $y_t^i$ can be decided as:

$$l_j = \arg\max_j (l_j = \widetilde{W} x_t^i). \quad (11)$$

## Experiments

To validate the effectiveness of our algorithm, experiments are carried out using two data sets, where the YouTube UCF data set is viewed as the target domain. the HMDB51data set is more a challenging real-world scenarios, it is viewed as the auxiliary domain. We choose the same action category from the HMDB51 data set and UCFYouTube data set, including a bike, diving, playing golf, jumping, hitting, riding, pitching these seven actions. Figure 3 and figure 4 are representative images of 2 data sets.In the UCF YouTube dataset as training random from all data classes have selected number of action for the 5/16/24 executor.Firstly, we computer the LMP descriptors from video data, Local-constrained Linear Coding[15]is applied to the low-level descriptors. And then carryout cross domain dictionary learning. Finally, the corresponding sparse representation isobtained for recognition. After the process of the training, we make use of Eq. 11 to decide the category of the test action, it's so called classification.

We compares with LLC[12], K-SVD[9], and LC-KSVD[13], in Table 1, the method of K-SVD and LC-KSVD dictionary learning are unsupervised, and ours is supervised one. The number of the executor in each action category is 5/16/24 respectively. We can discover that for many cases, knowledge transform the auxiliary domain into the target domain can cause certain performance. We can see that with the increase of the number of the executor, the recognition rateincrease. So the cross domain dictionary learning method is suitable forthe large data set recognition task.

## Conclusions

Across domain dictionary learning method through the introduction of auxiliary domains is proposed, which effectively expand the target domain intra class diversity, improving the

classification accuracy of the recognition system.

Experiments based on UCF YouTube data set, through the comparison with the state-of-art algorithm such as LLC[12], K-SVD[9], and LC-KSVD[13] prove that the cross domain dictionary learning method, suitable for the amount of data is larger. And the use of the source domain is available.

Fig.4 images in UCF Fig.5 images in HMDB51

Table1 Quantities Comparison among the Algorithms

| methods | LLC | LLC | K-SVD | K-SVD | LC-KSVD | LC-KSVD | DCDDL |
|---|---|---|---|---|---|---|---|
| Learning way | | | Unsupervised | Unsupervised | supervised | supervised | supervised |
| auxiliary domain | no | yes | no | yes | no | yes | yes |
| 24 | 86.67% | 86.67% | 81.33% | 82.22% | 85.67% | 86.67% | **88.89%** |
| 16 | 70.17% | 70.88% | 63.97% | 63.96% | 72.03% | 72.08% | **73.05%** |
| 5 | 53.35% | 54.10% | 51.05% | 50.05% | 56.55% | 56.55% | **56.88%** |

## References


[1]Chris Harris,Mike Stephens, A combined corner and edge detector, the 4th Alvey Vision Conference. 1988, 147-151.

[2] Lena Gorelick,Moshe Blank,Eli Shechtman,Michal Irani,Ronen Basri, Actions assp-ace-time shapes,J. IEEE Trans on PAMI, 2007,29(12)2247–2253.

[3] Guha. T, Ward.R.K, Learning sparse representations for human action recognition, J. IEEE Trans on PAMI, 2012, 34(8)1576-1584.

[4]Su.Y, Jurie F, Improving image classification using semantic attributes,J.International Journal of Computer Vision, 2012, 100(5)1–19.

[5]Yao A, Gall J Van, L. G, Coupled action recognition and pose estimation from multip-le views, J. International Journal of Computer Vision, 2012, 100(5)16–37.

[6]Cao X, Wang Z, Yan P, Li X, Transfer learning for pedestrian detection, J.NeuroCo-mputing,2013,100(5)1–57.

[7]J Liu, J Luo,M Shah, Recognizing realistic actions from videos "in the wild", C. Co-nference onComputer Vision and Pattern Recognition,2009.

[8]Laptev I, On space-time interest points,J.International Journalof Computer Vision,2005, 64(2)107-123.

[9]M. Aharon, M Elad, and A Bruckstein, K-SVD: An algorithm for designing over comp-l ete dictionaries for sparse representation,J.2006, IEEE Transactions on Signal Processing, 54( 1)4311–4322.

[10]Mairal, F Bach, J Ponce, G Sapiro, AZisserman. Supervised dictionary learning,J. A-dva nces in Neural Information Processing Systems, 2009.

[11]Zhang.Q.,LiB. Discriminative K-SVD for dictionary learning in face recognition, CVPR, 2010.

[12]J. Wang., J. Yang. , K. Yu. , F. Lv. ,T.Huan. ,Y Gong, Locality-constrained linear codi ng for image classification, C. IEEE Conference on Computer Vision and Pattern Reco-gniti on,2010.

[13]Z. Jiang. , Z. Lin .,L. S. Davis., Learning a discriminative dictionary for sparse codi-ng via label consistent K-SVD,C. IEEE Conference on Computer Vision and Pattern Reco-gnit ion,2011.

[14]P. Dollar. , V. Rabaud. ,G Cottrell. , S. Belongie. , Behavior recognition via sparse

spatial-temporal features, C.Proc Second IEEE Joint Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance,2005,65-72.

[15]Wang. J. , Yang. J. ,Yu. K. , Lv. F. , huang. T. ,Gong. Y. ,Locality-constrained l-inear coding for image classification, CVPR,2009.