MULTIPLE INSTANCE DISCRIMINATIVE DICTIONARY LEARNING FOR ACTION RECOGNITION

Hongyang Li[1], *Jun Chen*[1,2], *Zengmin Xu*[1], *Huafeng Chen*[1], *Ruimin Hu*[1,2]

[1]National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
[2]Collaborative Innovation Center of Geospatial Technology, China

ABSTRACT

Action recognition from video is a prominent research area in computer vision, with far-reaching applications. Current state-of-the-art action recognition methods is Fisher Vector (FV) coding model based on spatio-temporal local features. Though high dimensional local features have more representative, the high dimensions are challenge for the dictionary learning of FV model. This paper proposes a Multiple Instance Discriminative Dictionary Learning (MIDDL) method for action recognition. We introduce cross-validation method in multiple instance learning procedure, which prevents training from prematurely locking onto erroneous initial instances. In order to balance the positive instance number between positive bags, only the top ranked instances are labeled as positive in the step of iterative training classifiers. Taking these classifiers as discriminative visual words, we get the video global representation based on classifier response. The experimental results demonstrate the effectiveness of applying the learned discriminative classifiers as visual word on challenging action data sets, i.e. UCF50 and HMDB51.

Index Terms— multiple instance learning, discriminative dictionary, weakly supervised learning, action recognition

1. INTRODUCTION

Action recognition or classification is a prominent research area in computer vision, which can be applied to many applications such as video surveillance, human-computer interaction, human behavior understanding, etc. Though significant progresses have been made[1, 2], action recognition still remains a challenging task due to intra-class variations, background complexity, high-dimensional feature description, and other difficulties[3].

Local features are pooled and normalized to a vector as the video global representation in action recognition . A local feature vector is used to describe the local characteristics of the local space volume, which is composed of the pixels around the feature points. The improved dense trajectories (IDT) feature [1] combines trajectory shape descriptor, HOG/HOF[4] and MBH[5], which is superior to other handcrafted feature in the most challenging video data set. The deep-learned features, such as TDD[6] and C3D[7] feature, also achieve superior performance. We use the IDT vector to describe the local spatio-temporal volume of the action video.

More recently, many efforts have focused on developing discriminative dictionary for image object recognition or video action recognition.Taking the feature set in a video as a feature bag, and the video label as bag label, the distinguishing feature learning problem can be converted into a Multiple Instance Learning (MIL) problem.The straightforward way to use MIL for dictionary learning is first to learn a classifier for each object category, then use the classifier to select positive instances, finally build the dictionary by using k-means clustering algorithm to cluster these samples[8]. Sapienza et al. train one discriminative classifier for every action category by mi-SVM[9] algorithm for action detection, it is not directly used to construct a discriminative dictionary [10] . M^3IC formulates a novel maximum margin multiple instance clustering problem for the MIL task[11]. M^4I4 inspired by M^3IC propose a two-layer structure for action recognition to automatically exploit a mid-level "acton" representation[12]. These weakly-supervised actons are learned via a max-margin multi-channel MIL framework, which can capture multiple mid-level action concepts simultaneously. Wang at el. propose a MIL strategy for dictionary learning, that each code is represented by a linear SVM classifier[13].

In this paper, a new dictionary learning method based on MIL is proposed for feature encoding. We take the high dimension local features from one video as an instances bag. We introduce cross-validation method in MIL procedure, which prevents training from prematurely locking onto erroneous initial instances. In every iterative learning step, we restrict the max number for one positive bag, i.e. we only select top rank instances from every positive bag. At the same time, the negative instances labeled by classifier in positive bags do not participate in the iterative training. After learning one classifier, we filter positive instances whose scores is larger the threshold from all positive bags. Repeating this MIL process, we get multiple classifiers for one action category.

The research was supported by the NSFC(61170023, 61231015), the Hubei NSF(2014CFB712), the Internet of Things Development Funding Project of Ministry of industry in 2013(No. 25).

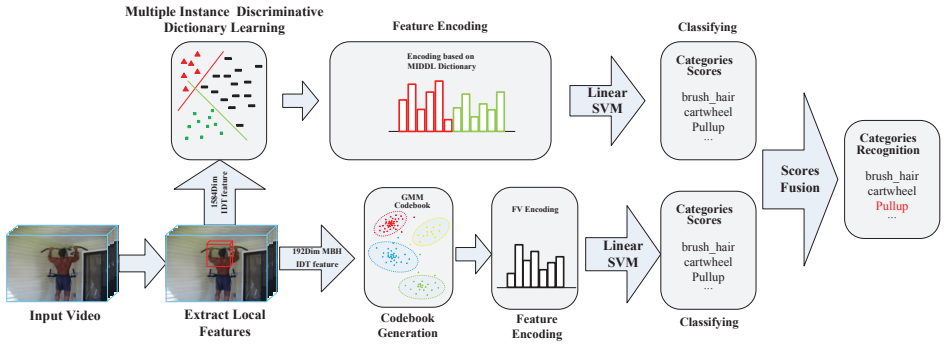

Fig. 1. An overview of the action recognition procedure.

2. PROPOSED DICTIONARY LEARNING METHOD

In this section, we elaborate the proposed discriminative dictionary framework based on MIL. We train multiple classifiers for each action category by iterative MIL.

2.1. Multiple Instance Learning Model

First, we have a brief review of the standard MIL model. In MIL, the labeling information is significantly weakened as the labels are assigned only to the bags with latent instance level labels. Given bags set $\{X_1, X_2, \ldots, X_I, \ldots\}$ labelled as $Y_I \in \{+1, -1\}$, each bag X_I contains a number of instances $X_I = \{x_1, x_2, \ldots, x_i, \ldots\}$. We assume that there is at least one positive instance in each positive bag, no positive instance in any negative bag.The problem of MIL is to find a classifier, which can be used to distinguish the instance in the unknown bag. If there are positive instance in a bag, this bag is labelled positive, otherwise it is negative. Obviously this classifier to detect the positive instance has strong discrimination ability, and there are some similar instances among the positive bags, which are difference from all instances in negative bag. The goal of the classifier is to distinguish the instance from the bag and the classical MIL based on SVM model is the mi-SVM [9]. The objective optimization function of mi-SVM is as below.

$$\min_{\{y_i\}} \min_{w,b,\xi} \frac{1}{2}||w||^2 + \lambda \sum_i \xi_i$$
$$s.t. \forall i : y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, y_i \in \{-1, +1\}$$
$$\forall Y_I = +1 : \sum_i \frac{y_i + 1}{2} \geq 1$$
$$\forall Y_I = -1 : y_i = -1$$

Y_I indicates the label of bagX_I , x_i is the instance in bag$X$, y_i indicates the label of instancex_i . Constraint 1 said classifier should be separated from the region of positive and negative examples; constraint 2 indicates positive bag has at least one positive instance; constraint 3 said no positive sample package; the objective function said separates the boundary to maximize to meet the conditions of the above constraints.

Initially all the instances are assumed to have the bag label. The mi-SVM uses heuristic search method, through multiple iterations to obtain the local optimal solution. In each iteration step, it uses the learned classifier to select the positive instances in positive bag as the new positive instances.

2.2. Multiple Instance Dictionary Learning Method

The sets of local features extracted from each video is taken as a instance bag. For multiple action categories recognition problem, we use "one-vs-rest" policy, that the bags from one action category are labelled as positive, the remaining bags are labelled negative. Now we can learn the best classifier which can detect the most discriminative instances in each positive bag.

Inspired by mi-SVM method, a new MIL algorithm is proposed, that the algorithm process is shown in Algorithm 1. First, the positive and negative bags are split into two sub-bags respectively, so the cross validation can be used to carry out the iterative learning (STEP 1). All positive instances are labeled as positive, the negative instances are labeled as negative in the initial step, and the classifier is linear SVM. In the iterative learning process, the classifier learned from one sub-bags is used to detect positive instances in the other sub-bags.

For each positive sub-bag, we sort the instances scores in descending order, and only select the highest score of Top-K positive instances to update the training positive sample (STEP 7). If the number of positive instances from one positive bag is less than TopK, all the predicted positive instance in this bag are selected. When there is no predicted positive instance in the positive bag, the highest score instance is labeled as positive(STEP12). It is to ensure that at least one positive instance in each bag. In addition to the selected instances, other instances in the positive bags do not participate in the training. During the whole training process, the neg-

Algorithm 1 Multiple Instance Discriminative Classifier Learning Algorithm.

Input: postitive set $Pos = \{X_1^+, X_2^+, \ldots, X_N^+\}$,
negative set $Neg = \{X_1^-, X_2^-, \ldots, X_M^-\}$,
Output: One SVM classifier (w, b)

```
Split Positive Set Pos = {Pos_1, Pos_2},
Split Negative Set Neg = {Neg_1, Neg_2}
Assign SVM penental cofficient : λ, Maximum Number of Selecting Positive from One Bag: TopK, Maximum number of iterations: T
for i = 1 → T do
    (w, b) ← SVMTraining(Pos_1, Neg_1)
    Pos_1 = ∅
    for j = 1 → |Pos_2| do
        L_j^+ = {x_k | SelectTopInstances(Pos_2j, TopK), s.t. s_k > 0, s_k = w^T x_k + b, x_k ∈ Pos_2j}
        if |L_j^+| > 0 then
            Pos_1 ← Pos_1 + L_j^+
        end if
        if |L_j^+| == 0 then
            Pos_1 ← Pos_1 + SelectHighestScore(L_j^+)
        end if
    end for
    swap(Pos_1, Pos_2),swap(Neg_1, Neg_2)
end for
use the distances to generate the ranking list.
```

ative instances is not updated, i.e. the negative instances are completely same as initial negative instances. We repeat this process until the number of iterations reach to the maximum or the positive instance is no longer changed.

Through the above MIL learning process, we get one discriminative classifier. It is not enough that one discriminative classifier for each positive bag, so we repeat this learning process to get more classifiers. In the process of training the next discriminative classifier, we detect all the positive instances from all positive bags and filter out these instances whose score are higher the threshold, that can make the difference between the new learning classifier and the classifier obtained previously. Since the classifiers are trained on discriminative instances, each classifier can better detect the discriminative features for current category. Repeating this learning process C times, we get C classifiers for one category.

For multiple action recognition problem, we use "one-vs-rest" policy, that each time one action category videos are taken as positive bags, the rest of the videos as negative bags. Through the MIL dictionary learning, we can learn C classifiers for each action category. If there is N action categories, we get $C * N$ discriminative classifiers which are combined to be a visual dictionary.

Fig. 2. Sample frames from different actions datasets. (a) UCF50 (b)HMDB51

2.3. Video Representation Based on Classifiers

Give a action video , we extract local feature vector set $X = \{x_i\}_{i=1}^n \in R^d$, and use MIL dictionary learning method to get the SVM classifier set $(W, B) = \{(w_j, b_j)\}_{j=1}^m \in (R^d, R)$. Then, we calculate the response score of each local feature vector in the classifier $f_i = W^T x_i + B, f_i \in R^m$. We map the feature response score to the real number between $0-1$ by the sigmoid function $\{s_{ij} = 1/(1 + exp(-f_{ij})\}_{j=1}^m$, s_{ij} expresses the response value of x_i on the classifier (w_j, b_j). Finally, we use max pooling and $l2$-normalization to obtain the video global representation vector.

3. EXPERIMENTS

We evaluate performance of the proposed method for action recognition on two action data sets, and compare it with previous methods in literature. These data sets are the most challenging data sets in recently . The performance of the action recognition is evaluated by the average precise.

3.1. The Data Set

The **UCF50** dataset[14] has 50 action categories and contains 6,618 videos, consisting of realistic videos taken from YouTube ranging from general sports to daily life exercises.The dataset are divided into 25 folds and we follow the recommended 25-folds cross-validation to report the performance.

The **HMDB51** dataset[15] collects video clips in abundant source, both from movies and Internet, there are 6,766 videos and 51 action categories in total .We follow the original protocol using three train-test splits . For every class and split, there are 70 videos for training and 30 videos for testing. We report average accuracy over the three splits as performance measure on the original videos.

3.2. Experiment Detail

We extract two scales IDT local features, that one is the baseline scale, the other is two times spatial scale the standard scale.The baseline scale feature parameters of standard IDT is same as the literature[1]: the size of the volume is 32×32 pixels and 15 frames. To embed structure information in the rep-

resentation, the volume is subdivided into a spatio-temporal grid of size $2 \times 2 \times 3$, with a dense sampling step size of 5 pixels.Thus, we could get the 30 dimension trajectory shape descriptor, 96 dimension HOG image gradient descriptors, 108 dimension HOF descriptors of optical flow distribution, and 96 dimension MBHx and 96 dimension MBHy descriptor which are used to describe the distribution information of optical flows difference in x,y aspects . We concatenate HOG, HOF, MBG feature descriptors into a single feature descriptor of which dimension is 396, this type features are used to fisher vector encode. The dimension of the other expansion spatial scale IDT is $2 * 2 * 396 = 1584$. We first reduce the descriptor dimensionality by a factor of two using Principal Component Analysis (PCA), then input to MIDDL model.

We random sample 2,000 local features from each training video to construct a positive bag, and select 10,000 features from each action category to construct negative instances set. The default value of maximum number of selecting positive from one positive bag is 5, the default value of maximum number of iterations is 20, the default number of classifier for each class is 11. After encoding , we perform $L2-$normalization on each channel respectively and concatenate them to a single vector as the video global representation. Finally we use the linear SVM to do classification.

3.3. Experimental Results and Analyses

First of all, we compared our MIL method to the classical MI-SVM and mi-SVM on ten action categories from data set HMDB51. Table 1 shows the efficient of dictionary training by different algorithms with the same parameters . It can be seen that , the cross validation and maximum suppression can improve performance to classical MIL based on SVM, that is to say we get more discriminative classifier. In fact, when TopK value equal 1, our method is similar to MI-SVM; When TopK value equal ∞, our method is simliar to mi-SVM.

Table 1. Performance comparison with other MIL methods for learning discriminative dictionary

method	396 dim	1584 dim
mi-SVM	68.6	73.5
MI-SVM	70.7	78.7
ours(**TopK=5)**	72.6	**85.4**
ours(TopK=10)	**74.3**	78.2
ours(TopK=20)	72.8	76.3
ours(TopK=50)	67.9	75.1

The results are obtained with the HOF, MBH feature descriptor on 10 action categories from HMDB51 data set. We set the TopK value of maximum number of selecting positive from each positive bag 5, 10, 20, 50 respectively.The experimental results show that when the parameters are too small, the insufficient of positive samples result in the deterioration of classifier discrimination ability. When the parameter is too large, it will lead to an increase in the number of fault positive samples, which also affect the ability of the classifier.

Besides, we investigate the classification performance w.r.t. different number of classifiers per category on 10 class actions from HMBD51 dataset. As shown in Table 2, we observe that the performance boosts along with increasing number of classifiers in our method, and achieves the optimum value at $C = 10$. Continued increase of the number will not improve accuracy further and even deteriorate performance. This implies that our MIL method can learn classifiers to detect rather discriminative local features for action recognition.

Table 2. The classification performance w.r.t. different number of classifier (i.e., parameter C) . The results are obtained on HMDB51 ten action classes.

C	1	3	5	7	9	11	13
AP	78.7	79.2	80.5	82.6	83.6	85.4	85.1

At last, we fuse two encoding method on score-level. For FV method, we use MBH feature and set the number of GMM 128. The result in Table 3 shows that the fusion of the score based on MBH FV and the score based on MIDDL encoding improves action recognition precise. This confirms that the FV encoding and the MIDDL encoding are complementary to each other, as the former directly models local features global distribution and the latter models salient features distribution.

Table 3. Action Recognition mAP(%) on the KTH, UCF50 and HMDB51 data sets.

UCF50		HMDB51	
Shi et al.[16]	83.3	Jain et al.[17]	52.1
Oneata et al.[18]	90.0	Oneata et al.[18]	54.8
Wang et al.[1]	**91.2**	**Wang et al.[1]**	**57.2**
our approach	**93.1**	**our approach**	**60.3**

4. CONCLUSION

We propose a discriminative dictionary learning method which improved the classical MIL based on SVM. The performance of the classifier is enhanced by using cross validation method and by limiting positive instances number in each positive package in iterative learning process, which decreased the errors accumulation caused by the erroneous judgments. The MIDDL encoding can effectively filter the interference features in the local features and make the global expression of the video more compact and more discriminative. The performance of action recognition algorithm proposed in this paper is compared on multiple datasets, the experiments validate the effectiveness of the algorithm.

5. REFERENCES

[1] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *International Conference on Computer Vision*, 2013.

[2] S. Sadanand and J.J. Corso, "Action bank: A high-level representation of activity in video," in *Conference on Computer Vision and Pattern Recognition*, 2012.

[3] J. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, 2011.

[4] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Conference on Computer Vision and Pattern Recognition*, 2008.

[5] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006.

[6] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep convolutional descriptors," in *Conference on Computer Vision and Pattern Recognition*, 2015.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *International Conference on Computer Vision*, 2015.

[8] Q.N. Li, J.J. Wu, and Z.W. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Conference on Computer Vision and Pattern Recognition*, 2013.

[9] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Neural Information Processing Systems Conference*, 2002.

[10] M. Sapienza, F. Cuzzolin, and P.H.S. Torr, "Learning discriminative space-time action parts from weakly labelled videos," *International Journal of Computer Vision*, vol. 110, no. 1, pp. 30–47, 2013.

[11] D. Zhang, F. Wang, L. Si, and T. Li, "M3ic: Maximum margin multiple instance clustering," in *International Joint Conference on Artificial Intelligence*, 2009.

[12] J. Zhu, B.Y. Wang, X.K. Yang, W.J. Zhang, and Z.W. Tu, "Action recognition with actons," in *International Conference on Computer Vision*, 2013.

[13] X.G. Wang, B.Y. Wang, X. Bai, W.Y. Liu, and Z.W Tu, "Max-margin multiple-instance dictionary learning," in *International Conference on Machine Learning*, 2013.

[14] K.K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *International Conference on Computer Vision*, 2011.

[16] F. Shi, E. Petriu, and R. Laganiere, "Sampling strategies for real-time action recognition," in *Conference on Computer Vision and Pattern Recognition*, 2013.

[17] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Conference on Computer Vision and Pattern Recognition*, 2013.

[18] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *International Conference on Computer Vision*, 2013.