WEAKLY SUPERVISED CROSS-VIEW ACTION RECOGNITION VIA SEQUENTIAL MOTION ACCUMULATION

Yi Liu[1], *Lei Qin*[2], *Zhongwei Cheng*[1], *Yanhao Zhang*[3], *Weigang Zhang*[3], *Qingming Huang*[1,2]

[1]University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Key Lab. of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, China
[3]Harbin Institute of Technology, Harbin, 150001, China
{yiliu,lqin,zwcheng,yhzhang,wgzhang,qmhuang}@jdl.ac.cn

ABSTRACT

In real application scenarios, the visual observations of the same type of action vary significantly from one view to another. This paper addresses the action recognition problem under the view changes, especially when no labels are available in the target view. A novel feature, called Sequential Motion Accumulation (SMA), is proposed to characterize actions. The SMA descriptor depicts the temporal structure of motion property to explore the distinguishing action characteristics and their invariances across views. Moreover, we propose a weakly supervised categorization approach to generate target-view categorical prior for learning a cross-view metric, which can further improve the recognition accuracy of the SMA descriptor. Our method is verified on the multi-view IXMAS dataset, and it achieves superior performance compared with the state-of-the-art methods.

Index Terms— Action recognition, Cross view, Sequential motion accumulation

1. INTRODUCTION

Human action recognition is one of the most active research areas in computer vision. It is central to many applications, including visual surveillance, video indexing/retrieval, and human-computer interaction [1, 2]. However, it remains challenging to recognize actions from different views, due to the drastic visual variances caused by the changes of viewpoint.

The objective of cross-view action recognition is transferring action model learned in one view (source view) to another different view (target view). According to [3], this task can be categorized into three modes: the first mode is *correspondence mode*, which requires unlabeled action instances observed simultaneously in both source and target views. The second mode relaxes instance-to-instance correspondence but requires partially labeled samples in the target view, which is referred as the *partially labeled mode*. In contrast, the third mode referred as *unlabeled mode* contains no label information in the target view. The supervision becomes weaker from the first mode to the third, while the difficulty of task goes greater. In this paper, we focus on the *unlabeled mode*.

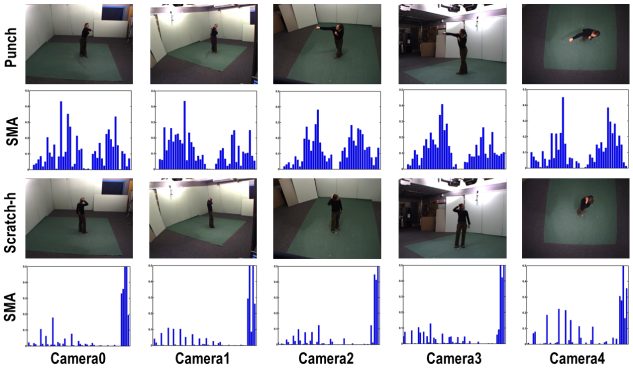


Fig. 1. Examples from the IXMAS multi-view dataset [7] that illustrate the visual variances across different views, and the discriminability and view invariance of the SMA descriptor. Rows 1 and 3 show two actions observed from different angles, while rows 2 and 4 represent the corresponding SMA descriptors, respectively.

Recent research efforts for cross-view action recognition mostly address the *correspondence mode* and the *partially labeled mode*. These works have been made along two directions: one focuses on viewpoint invariant features, and the other tends to reduce the gap between viewpoints via transfer learning. Many of the view-invariant features are based on trajectories extracted from human bodies that require accurately tracking body parts, joints or landmarks under different viewpoints [4–6]. Other feature descriptors based on silhouettes [7] and temporal self-similarities [8] are also widely used. Weinland et al. [7] used an exemplar-based HMM to fully reconstruct the 3D models of human actions by silhouettes from multiple cameras. Junejo et al. [8] proposed an action descriptor based on temporal self-similarity matrix. It characterizes the difference between frames by computing

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61133003, 61332016, 61202234 and 61202322.

distances between all pairs of extracted features. Its promising performance indicates that the temporal information is robust to the changes of viewpoint. Our approach also considers temporal structure of observations, but based on motion property. The other direction of cross-view action recognition relaxes structural constraints and transfers the correspondences across the actions of different views. Liu et al. [9] proposed a view knowledge transfer learning framework to explore higher-level features which can bridge the semantic gap across view-dependent vocabularies. Li et al. [3] proposed a novel notion of 'virtual views' that associate action descriptors from source view and target view with linear transformations. As mentioned previously, these learning methods rely on corresponding or label information in the target view, with stronger supervision than our approach.

In this paper, we propose a Sequential Motion Accumulation descriptor for action representation and a weakly supervised categorization method for cross-view metric learning. The main contributions of this work are two-fold:

- Firstly, we represent actions using the temporal order of motion property, which can overcome the huge visual variances caused by view changes. Examples of the SMA descriptor is illustrated in Fig. 1. This figure intuitively demonstrates that the visual observations vary significantly across different views. However, the SMA descriptor is robust to view variations and discriminative among different actions.
- Secondly, we propose a weakly supervised categorization method. This method utilizes the view invariance of the SMA descriptor to generate target-view categorical prior for learning a cross-view metric. It can filter out fluctuations among actions of different views while retaining sufficient discrimination. Additionally, this method does not require prior knowledge in the target view, and better satisfies the practical application requirements.

In Section 2, we describe the details of the SMA descriptor, followed by the weakly supervised categorization in Section 3 and the experimental results in Section 4.

2. SEQUENTIAL MOTION ACCUMULATION

Sequential Motion Accumulation (SMA) descriptor is based on the spatio-temporal cuboid detector [10]. The interest point detection is operated on a stack of images denoted by $I(x, y, t)$. The response function of this interest point detector is calculated by application of separable linear filters as:

$$r = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where g is a 2D Gaussian smoothing kernel applied spatially, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally. This detector responds strongly to regions with high motion intensity, including spatio-temporal corners.

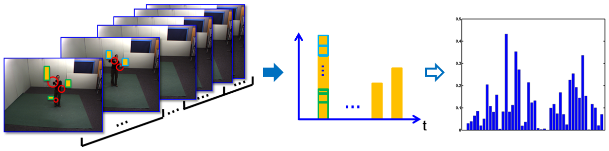


Fig. 2. SMA descriptor extraction. Each bin in the middle histogram corresponds to the accumulation of motion intensities in one temporal segment in the left part. The right chart indicates the normalized histogram (the final feature).

We utilize this detector to detect interest points in videos, and obtain the corresponding motion intensity responses. Then the temporal structure of the motion intensity is used to describe actions. Due to the temporal order being diverse for different actions but similar across different views, the temporal structure is informative for action classification and robust to view variations. As mentioned in [11–13], an action can be defined as a finite sequence of segments over a finite period of time. For instance, if "left leg moves forward" and "right leg moves forward" are two individual segments, then the walking behavior is a finite sequence of those segments (e.g. "left leg moves forward → right leg moves forward → left leg moves forward ..."). The temporal order of these segments remains the same, no matter which viewpoint the actions are captured from. In our method, we decompose each action video into temporal segments, so a video becomes a finite sequence of individual parts. Then we accumulate motion intensities in each part, obtaining motion property descriptor for each temporal segment. The SMA is the concatenation of those accumulations (as shown in Fig. 2). Let S_i denote the i-th segment of an action. The SMA descriptor is defined as:

$$\begin{gathered} f = (f_{S_1}, f_{S_2}, ..., f_{S_n}) \\ f_{S_i} = \sum_{\mathrm{P}_j \in S_i} r_{P_j} \Big/ \sum_{l=1}^{n} f_{S_l} \end{gathered} \tag{2}$$

where r_{P_j} is the motion intensity of an interest point P_j, computed by Eqn. 1. The parameter n denotes the number of segments, which is fixed for action videos to guarantee the same dimension of the features. The SMA descriptor with larger n maintains more distinguishing characteristics, but introduces more fluctuations due to the differences of video length and movement speed between various actions. So the setting of n is a tradeoff between reserving action separable characteristics and reducing noise. We solve this problem by describing temporal structure with multiple numbers of motion segments.

3. WEAKLY SUPERVISED CATEGORIZATION

As mentioned previously, our approach addresses the *unlabeled mode* of cross-view action recognition. For this purpose, we provide a weakly supervised categorization method

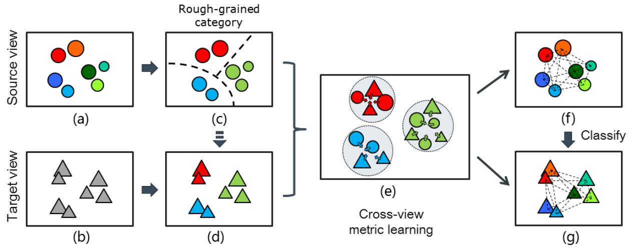


Fig. 3. The process of weakly supervised categorization. To explore distinguishing characteristics in the unlabeled target view (b), we first discover rough-grained categories in the source view (a)→(c). Then classification model based on the source-view examples is used to predict target-view examples (b)$\xrightarrow{(c)}$(d). As a result, we can learn a cross-view metric based on the categorical information of both the source and target views (e). Finally, we train on source-view actions and recognize target-view actions with this cross-view metric (f)→(g).

that utilizes the view invariance of the SMA descriptor to discover and use target-view prior knowledge. This method has two parts: The first part constructs rough-grained category to explore categorical information in the target view, and the second one learns a cross-view metric for action classification. The process of weakly supervised categorization is illustrated in Fig. 3.

Rough-grained category construction. Let $\{C_i\}_{i=1}^{k}$ denote the real class labels while $\{V_j\}_{j=1}^{m}$ denote the rough-grained categories and μ_j is the mean of instance vectors in V_j. We construct a rough-grained classification matrix $A \in \mathbb{R}^{k\times m}$, that each row corresponds to a real class and each column corresponds to a rough-grained category. The element a_{ij} counts the number of instances which belong to C_i that are classified into V_j, and b_i denotes the maximum a_{ij} in the i-th row of A. Each C_i is classified into V_x, that x is the column of b_i. As the rough-grained category is the cluster of the real class labels, instances of one real class should be classified into the same rough-grained category mostly to guarantee the accuracy. It means we should maximum the sum of b_i for each row of A. The optimization problem can be formulated by:

$$\begin{aligned} \max\ score &= \sum_{i=1}^{k} b_i \\ b_i &= \max a_{ij}\ ;\ \forall j \in [1,m] \end{aligned} \qquad (3)$$

The detailed algorithm of rough-grained category construction is listed in Algorithm 1. Once the optimal $\{V_j\}_{j=1}^{m}$ are found, the target-view instances can be rough-grained labeled (as shown in Fig. 3 (b)$\xrightarrow{(c)}$(d)).

Cross-view metric learning. With the rough-grained labeled source-view and target-view observations, we utilize large margin nearest neighbor (LMNN) [14] to learn a discriminative cross-view metric. LMNN is a Mahanalobis distance metric, which can filter out fluctuations between observations, and capitalize on useful statistical regularities in both the source and target views. So utilizing LMNN instead of Euclidean distances to measure the dissimilarities between instances from different viewpoints is a powerful method for learning classification model.

Algorithm 1: Rough-grained Category Construction

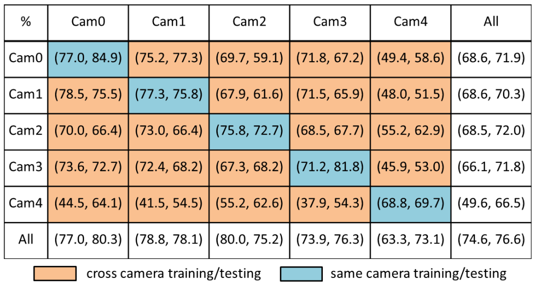
```
Input: Initial rough-grained category number $m_{\_ini}$
       Maximum iteration $Maxiter$
Output: Rough-grained categories $\{V_j\}_{j=1}^{m}$
Initialize set $level = 1$, $score = 0$;
    kmeans with $m_{\_ini}$ cluster centers;
while $iteration \leq Maxiter$ do
    update $A$ with $\{\mu_j\}_{j=1}^{m}$ as cluster centers;
    compute $score$ according to Eqn. 3 ;
    if $score\ descend$ then
        $level = level + 1$;
        back to previous $\{V_j\}_{j=1}^{m}$;
    else
        $level = 1$;
    refine $\{V_j\}_{j=1}^{m}$ by classifing $C_p$ into $V_q$, that $p$ is
    the row of the $level$-th $\min_{1\leq i\leq k}(b_i)$ and $q$ is the
    column of the second $\max_{1\leq j\leq m}(a_{pj})$;
    $iteration = iteration + 1$;
```

Classifier at different level is learned with different number m of rough-grained categories. As each classifier only provides partial knowledge, we consider multi-level weakly supervised categorization and utilize max voting to effectively fuse all level predictions. A new data instance is classified to the category obtained from the majority vote of classifiers. It is a simple yet powerful method to combine the discriminant capability of multiple models.

4. EXPERIMENTS AND RESULTS

We evaluate our approach on the IXMAS multi-view action dataset [7], which contains eleven daily-life actions, such as *check watch*, *turn around*, and *pick up*. Each action is performed three times by twelve actors and recorded simultaneously from five different views: four side views and one top view.

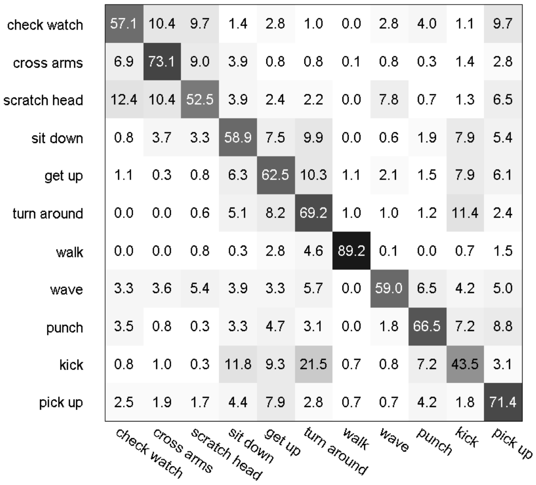
We extract at most 300 cuboids with motion intensity from each video. To solve the issue of temporal synchronization, numbers of temporal motion segments are set to 4, 16 and 64, according to diverse video lengths. Then we concatenate these temporal features into a multi-scale temporal representation. In this way, each video is described by a normalized 84-dimensional vector, which is used for weakly



%	Cam0	Cam1	Cam2	Cam3	Cam4	All
Cam0	(77.0, 84.9)	(75.2, 77.3)	(69.7, 59.1)	(71.8, 67.2)	(49.4, 58.6)	(68.6, 71.9)
Cam1	(78.5, 75.5)	(77.3, 75.8)	(67.9, 61.6)	(71.5, 65.9)	(48.0, 51.5)	(68.6, 70.3)
Cam2	(70.0, 66.4)	(73.0, 66.4)	(75.8, 72.7)	(68.5, 67.7)	(55.2, 62.9)	(68.5, 72.0)
Cam3	(73.6, 72.7)	(72.4, 68.2)	(67.3, 68.2)	(71.2, 81.8)	(45.9, 53.0)	(66.1, 71.8)
Cam4	(44.5, 64.1)	(41.5, 54.5)	(55.2, 62.6)	(37.9, 54.3)	(68.8, 69.7)	(49.6, 66.5)
All	(77.0, 80.3)	(78.8, 78.1)	(80.0, 75.2)	(73.9, 76.3)	(63.3, 73.1)	(74.6, 76.6)

🟧 cross camera training/testing 🟦 same camera training/testing

Fig. 4. Cross-view recognition accuracy on the IXMAS dataset in the *unlabeled mode*. Each row is a source view and each column is a target view. The two numbers in a tuple are the average recognition accuracy of SSM-hog-of [8] and our method, respectively. (best viewed in color)



	check watch	cross arms	scratch head	sit down	get up	turn around	walk	wave	punch	kick	pick up
check watch	57.1	10.4	9.7	1.4	2.8	1.0	0.0	2.8	4.0	1.1	9.7
cross arms	6.9	73.1	9.0	3.9	0.8	0.8	0.1	0.8	0.3	1.4	2.8
scratch head	12.4	10.4	52.5	3.9	2.4	2.2	0.0	7.8	0.7	1.3	6.5
sit down	0.8	3.7	3.3	58.9	7.5	9.9	0.0	0.6	1.9	7.9	5.4
get up	1.1	0.3	0.8	6.3	62.5	10.3	1.1	2.1	1.5	7.9	6.1
turn around	0.0	0.0	0.6	5.1	8.2	69.2	1.0	1.0	1.2	11.4	2.4
walk	0.0	0.0	0.8	0.3	2.8	4.6	89.2	0.1	0.0	0.7	1.5
wave	3.3	3.6	5.4	3.9	3.3	5.7	0.0	59.0	6.5	4.2	5.0
punch	3.5	0.8	0.3	3.3	4.7	3.1	0.0	1.8	66.5	7.2	8.8
kick	0.8	1.0	0.3	11.8	9.3	21.5	0.7	0.8	7.2	43.5	3.1
pick up	2.5	1.9	1.7	4.4	7.9	2.8	0.7	0.7	4.2	1.8	71.4

Fig. 5. Confusion matrix for all classes of averaged over all cross-camera setups in Fig. 4.

Table 1. Comparison of Recognition Results

	cross camera	same camera	any-to-any
SMA+Wsc	**63.9**	77.0	**69.9**
SSM-hog-of [8]	61.8	74.0	64.3
SMA	59.0	68.5	64.9
STIP-hog-hof	42.4	**80.6**	50.0
Farhadi [15]	58.1	68.8	60.3
Weinland [7]	—	57.9	—

supervised categorization. To evaluate the performance of weakly supervised categorization, we conduct experiments on all possible numbers of rough-grained categories that range from 3 to 11, and the final result is a combination of all multi-level results obtained by k nearest neighbors classifier. For a better comparison, we follow the same *leave-one-out cross-validation* (LOOCV) setting, and make sure at least one person does not appear in the training and testing sets simultaneously for cross-view recognition.

Cross-view action recognition accuracy averaged over all categories and test subjects is shown in Fig. 4. This table illustrates that the cross-view performance of our method outperforms SSM-hog-of [8] for most combinations of training and testing cameras. It is interesting to note that our method performs much better when tested on top view (Cam 4) with learning on side views (Cam 0-3), and vice versa. As top view captures totally different visual information, we believe the performance on top view is more important for evaluating a cross-view action recognition method. The promising top-view results of our approach illustrate its robustness to huge view variances. Moreover, our approach obtains comparable results with the SSM-hog-of when the same or similar views are used for training and testing. It indicates that our feature is also discriminative among different actions. Note that the cross-view recognition accuracies are close to the performance of same-view recognition for most view combinations, which further verifies the view-invariance of our approach. Fig. 5 shows the confusion matrix corresponding to the average confusion computed for all cross-camera recognition setups in Fig. 4. The per-class cross-view recognition performances are promising for most classes.

The comparison of diverse cross-view action recognition methods for different combinations of source and target view setups is summarized in Table 1. Besides SSM-hog-of, we also compare with space-time interest points (STIP) [16] and other two learning methods. Table 1 shows that both the SMA descriptor and SMA with weakly supervised categorization outperform STIP in cross-view recognition. But the STIP gains better result in same-camera recognition. It illustrates that the SMA descriptor and weakly supervised categorization method are more robust to viewpoint variations, but somewhat reduce the discriminative power. [7] and [15] propose effective transfer learning approaches for cross-view action recognition. Our work outperforms all these view-invariant recognition techniques for different combinations of training and testing view setups. Especially, our method addresses the *unlabeled mode* without links or labels in the target view, which is more difficult than the cases presented in [7, 15].

5. CONCLUSIONS

In this paper, we address the problem of recognizing actions from an unlabeled view using instances extracted from other view. For this purpose, we have proposed a Sequential Motion Accumulation descriptor based on temporal information, which achieves significant improvement over features based on appearance for cross-view action recognition. It is able to produce further gains by weakly supervised categorization. This learning method discovers and uses categorical prior in the target view to learn a cross-view metric. Experiments are conducted on the IXMAS multi-view dataset and our method obtains promising results compared with alternative methods in the literature.

6. REFERENCES

[1] Jake K. Aggarwal and Quin Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[2] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[3] Ruonan Li and Todd Zickler, "Discriminative virtual views for cross-view action recognition," in *CVPR*, 2012, pp. 2855–2862.

[4] Cen Rao, Alper Yilmaz, and Mubarak Shah, "View-invariant representation and recognition of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 203–226, 2002.

[5] Yuping Shen and Hassan Foroosh, "View-invariant action recognition using fundamental ratios," in *CVPR*, 2008.

[6] Alper Yilmaz and Mubarak Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *ICCV*, 2005, pp. 150–157.

[7] Daniel Weinland, Edmond Boyer, and Rémi Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *ICCV*, 2007, pp. 1–7.

[8] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, 2011.

[9] Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese, "Cross-view action recognition via view knowledge transfer," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, October 2005.

[11] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proceedings of the 12th European Conference of Computer Vision (ECCV)*, Crete, Greece, September 2010.

[12] Feng Zhou, Fernando De la Torre, and Jessica K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion.," in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, pp. 582–596.

[13] Anwaar Ul Haq, Iqbal Gondal, and Manzur Murshed, "On temporal order invariance for view-invariant action recognition," in *IEEE Trans. Circuits Syst. Video Techn.*, 2013, pp. 203–211.

[14] K.Q. Weinberger and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[15] Ali Farhadi and Mostafa Kamali Tabrizi, "Learning to recognize activities from the wrong view point," in *ECCV*, 2008, pp. 154–166.

[16] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.