Online Action Recognition based on Incremental Learning of Weighted Covariance Descriptors

Chang Tang1,2, Wanqing Li2, Pichao Wang2
1School of Computer Science, China University of Geosciences, China; 2Advanced Multimedia Research Lab, University of Wollongong, Australia.
tangchang@tju.edu.cn

Abstract

Online action recognition aims to recognize actions from unsegmented streams of data in a continuous manner. One of the challenges in online recognition is the accumulation of evidence for decision making. This paper presents a fast and efficient online method to recognize actions from a stream of noisy skeleton data. The method adopts a covariance descriptor calculated from skeleton data and is based on a novel method developed for incrementally learning the covariance descriptors, referred to as weighted covariance descriptors, so that past frames have less contributions to the descriptor and current frames and informative frames such as key frames contribute more towards the descriptor. The online recognition is achieved using an efficient nearest neighbour search against a set of trained actions. Experimental results on MSRC-12 Kinect Gesture dataset and our newly collocated online action recognition dataset have demonstrated the efficacy of the proposed method.

1. Introduction

Human action recognition is an active research topic in computer vision because of the wide range of its potential application, viz. surveillance, video games, video indexing and search, and human-robot interaction. In the last few decades many approaches have been proposed to recognize actions from monocular or RGB video sequences [2]. However, these methods face the difficulties posed by changes in illumination, variations in viewpoint, occlusion and cluttered background. Perhaps more importantly, these methods are somewhat impaired by the loss of 3D information in conventional video.

Since the release of low-cost RGB-D sensors such as Microsoft Kinect sensors, much effort and advances have been made on action recognition from depth maps. Compared with RGB data, depth maps have several advantages for action recognition. Firstly, depth data can provide additional body shape and structure information, and skeleton joints can be easily recovered from a single depth map. Secondly, depth sensors are insensitive to illumination. However, the depth maps suffer from high level of noise when they are captured by a commodity RGB-D camera.

Many methods have been proposed for recognizing actions from depth or skeleton data [1, 25]. These methods are often based on different forms of input or derived data such as depth-map-based [47], skeleton joints [42, 49] or body parts [35], cloud points [44, 51], local interest points [45], and surface normals [28, 46]. Most of them are typically concerned with classifying actions from segmented sequences of input data, each corresponding to one action. They assume that all the instances, training or testing, are temporally segmented beforehand and recognition is carried out offline. This assumption is usually not valid when data are streamed in real-time and recognition has to be done online. In this paper, we propose an online action recognition method from skeleton data based on weighted covariance descriptors. The method assumes that segmented and label action instances are available for training and recognition has to be carried out online. To facilitate the online-recognition, a method is developed for incrementally learning the covariance descriptors so that past frames have less contributions to the descriptor and current frames and informative frames such as key frames contribute more towards the descriptor. This weighting scheme has to a certain extent encoded some temporal information in the covariance descriptor. Experimental results on MSRC-12 Kinect Gesture dataset and our newly collocated online action recognition dataset have demonstrated the efficacy of the proposed method.

2. Proposed Method

Suppose we have $L$ possible action classes and $M$ segmented training action instances, each training action instance corresponds to one of the $L$ action class. The action-label set can be denoted by $\mathcal{L} = \{1\}_{l=1}^L$. Let $\{M^l_n\}_{n=1}^{N_l}$
denote the set of \( N_l \) single-action instances of class \( l \), so \( N_1 + \cdots + N_L = M \). Given a test video sequence \( V \) with an unknown order and number of actions \( I \), its unknown label sequence is represented as \( Z = (z_1, \ldots, z_t, \ldots, z_I) \), \( z_t \in L \); the boundaries between two consecutive actions are also unknown. In practice, we cannot access the whole test video sequence at once; only one frame of the test streaming video is available at time \( t \). The task of online action recognition is to decide what action class the human is performing at any time \( t \) using sufficient information of the previous frames before \( t \), and find a partition between two consecutive different single actions. If we have sufficient information at time \( t \) to make a decision, we determine that the subject is performing an action \( l \), otherwise we continue accumulating information from time \( t + 1 \).

2.1. The Covariance Descriptor

Let \( S = [s_1, s_2, \cdots, s_n] \) be the data matrix composed of \( n \) feature vectors, each feature vector containing \( d \) feature variables. The data matrix can be represented by the \( d \times d \) sample covariance matrix, \( C \), as follows:

\[
C = \frac{1}{n-1} \sum_{i=1}^{n} (s_i - \bar{s})(s_i - \bar{s})^T ,
\]  

where \( \bar{s} \) is the mean of the feature vectors. The diagonal entries of the covariance matrix represent the variance of each individual feature variable, while the non-diagonal entries are their respective correlations. Our motivation for using covariance as feature descriptor can be summarized as follows: the covariance matrix is very simple and fast to compute and incrementally update, thus meeting the requirement of online recognition very well; the covariance matrix can capture information about the shape of the joint probability distribution of the set of feature variables. The feature variables of different actions generally have different distributions, so covariance matrices can discriminate different actions.

2.2. Overview of the method

In this section, we will give the details of our proposed method. There are two main parts in our work: offline training phase and online recognition phase.

During the training phase, we use each of the labeled training instances to construct a covariance matrix. In order to make the covariance matrices more discriminative, we use the symmetric positive definite (SPD) matrix dimensionality reduction method [15] to learn a projection matrix. The original covariance matrices are then projected onto a low dimensional space but more discriminative. As described in [15], the projection matrix learning can be expressed as the following minimization problem

\[
P^* = \arg \min_{P \in \mathbb{R}^{n \times m}} \sum_{i,j} A_{ij} \delta^2 (P^T X_i P, P^T X_j P) \\
\text{s.t. } P^T P = I_m
\]  

where \( X_i \) and \( X_j \) are any two different covariance matrices in the training set. \( A_{ij} \) is a real symmetric affinity matrix which encode the structure of the original data. \( n \) is the original covariance matrix dimension and \( m \) is the projected covariance matrix dimension. \( I_m \) is an \( m \times m \) identity matrix. \( \delta \) is the Stein metric function [29] or AIRM metric function [29]. In our experiment, we use the Stein metric for training and testing. For any two SPD covariance matrices \( X \) and \( Y \), their Stein metric is defined as

\[
\delta^2 (X, Y) = \ln \det \left( \frac{X + Y}{2} \right) - \frac{1}{2} \ln \det (XY).
\]  

There are two advantages in using the projection method. First it tends to render the covariance matrices more discriminative. Second, it projects the high dimensional matrix into a low dimensional matrix, so the subsequent estimation of the distance between the two matrices is computationally efficient. During the online recognition phase, a continuous video stream which is composed by some actions with unknown order and labels is given frame by frame; the online action labels are provided by the training labels. At time \( t \), we use the previous \( t - 1 \) frames and current frame to construct a covariance matrix \( C_t \). Then project \( C_t \) to the low dimensional space using \( P \) and compare the distance between all the projected training covariance matrices and \( C_t \). For each training class \( l \), we have \( N_l \) distances represented as \( \{d_{C_t, l_{i_1}}, \ldots, d_{C_t, l_{N_l}}\} \). The final distance between \( C_t \) and class \( l \) is given by

\[
d_{C_t, l} = \min \{d_{C_t, l_{i_1}}, \ldots, d_{C_t, l_{N_l}}\}.
\]  

Then we have \( L \) distances for each frame as time progresses. At time \( t \), we use the \( L \) distances and their standard deviation to decide what action is being performed or determine whether there is a boundary between two consecutive
actions. At the beginning, we first initialize a covariance matrix using the first $t_0$ frames, and give an action label $l_{t_0}$ by the criterion

$$l_{t_0} = \arg \min_l \{ d_{C_{t_0}, l} \}_{l=1}^L.$$

(5)

From time $t_0 + 1$, if the standard deviation value of the $L$ distances is a local minima and a new action is detected by Eq.5 action change takes place. Otherwise, no action change takes place. It is convenient to use the standard deviation of the $L$ distances to decide whether there exist a boundary between two actions. For instance, when the estimated distance at a given time is the minimum among others and the standard deviation is also large, this indicates that some specific action is being performed and there is no action change. In contrast, if a new action starts and an old action ends, there exists a transitional stage, so all the estimated distances are similar and the standard deviation is relatively low. In [11] and [7], a similar method is used on the SVM scores. Fig.1 shows a segment of minimum distances between online covariance matrix and training covariance matrices of each action class. Fig.2 shows the standard deviation of the distances in Fig.1. As can be seen, when it comes to an action change, the standard deviation value goes to a local minima.

### 2.2.1 Incremental leaning of weighted covariance matrices

In previous methods, feature descriptors extracted from one video frame or video segments have been weighted equally. However, the contribution of each frame to action recognition varies as some frames are more discriminative than others. In this regard, the important frame should be weighted proportionally higher (frame-based weighting). Furthermore, during the online action recognition, frames nearer the current frame provide more update information than past frames (temporal weighting). Our proposed method incorporates these two weighting schemes in an efficient algorithm to update the covariance descriptor. Specifically, we assign two kinds of weights to each video frame - frame weight and temporal weight. There are six considerations in assigning the weights for each frame: 1) the frame weight of each sample should vary with there importance to recognize an action; 2) the temporal weight of each frame should vary over time $t$; 3) the frame weight of each video frame should not vary over time; 4) the frames from the current time should have higher time weights than previous frames; 5) the weights should not affect the ability to incrementally obtain the new covariance matrix when a new frame feature vector is available; 6) the weights should not affect the fast covariance computation.

### References

[1] J. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[3] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

[4] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(2):411–421, 2007.

[5] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[6] N. P. Cuntoor and R. Chellappa. Key frame-based activity representation using antieigenvalues. In *Computer Vision–ACCV 2006*, pages 499–508. Springer, 2006.

[7] R. De Rosa, N. Cesa-Bianchi, I. Gori, and F. Cuzzolin. Online action recognition via nonparametric incremental learning. In *BMVC*. BMVA Press, 2014.

[8] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *Cybernetics, IEEE Transactions on*, 45(7):1340–1352, 2014.

[9] I. L. Dryden, A. Koloydenko, and D. Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pages 1102–1123, 2009.

[10] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr, and R. Sukthankar. Exploring the trade-off between
accuracy and observational latency in action recognition. International Journal of Computer Vision, 101(3):420–436, 2013.

[11] S. R. Fanello, I. Gori, G. Metta, and F. Odone. Keep it simple and sparse: Real-time action recognition. The Journal of Machine Learning Research, 14(1):2617–2640, 2013.

[12] D. A. Forsyth and J. Ponce. A modern approach. Computer Vision: A Modern Approach, 2003.

[13] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In SIGCHI, pages 1737–1746. ACM, 2012.

[14] K. Guo, P. Ishwar, and J. Konrad. Action recognition using sparse representation on covariance manifolds of optical flow. In AVSS, pages 188–195. IEEE, 2010.

[15] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In ECCV, pages 17–32. Springer, 2014.

[16] M. T. Harandi, C. Sanderson, A. Wiliem, and B. C. Lovell. Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures. In WACV, pages 433–439. IEEE, 2012.

[17] M. Hasan and A. K. Roy-Chowdhury. Continuous learning of human activity models using deep nets. In ECCV, pages 705–720. Springer, 2014.

[18] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on, 13(2):415–425, 2002.

[19] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In IJCAI, pages 2466–2472. AAAI Press, 2013.

[20] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 73–80. IEEE, 2013.

[21] X. Jiang, F. Zhong, Q. Peng, and X. Qin. Online robust action recognition based on a hierarchical model. The Visual Computer, 30(9):1021–1033, 2014.

[22] S. Kpotufe and F. Orabona. Regression-tree tuning in a streaming setting. In NIPS, pages 1788–1796, 2013.

[23] K. Kulkarni, G. Evangelidis, J. Cech, and R. Horrada. Continuous action recognition based on sequence alignment. International Journal of Computer Vision, 1–25, 2014.

[24] I. Kviatkovsky, E. Rivlin, and I. Shimshoni. Online action recognition using covariance of shape and motion. Computer Vision and Image Understanding, 129:15–26, 2014.

[25] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In CVPR Workshop, pages 9–14. IEEE, 2010.

[26] R. Minhas, A. A. Mohammed, and Q. J. Wu. Incremental learning in human action recognition based on snippets. Circuits and Systems for Video Technology, IEEE Transactions on, 22(11):1529–1541, 2012.

[27] L. Miranda, T. Vieira, D. Martínez, T. Lewiner, A. W. Vieira, and M. F. Campos. Online gesture recognition from pose kernel learning and decision forests. Pattern Recognition Letters, 39:65–73, 2014.

[28] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In CVPR, pages 716–723. IEEE, 2013.

[29] X. Pennecc, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. International Journal of Computer Vision, 66(1):41–66, 2006.

[30] G. R. Price. Extension of covariance selection mathematics. Annals of human genetics, 35(4):485–490, 1972.

[31] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In ECCV, pages 742–757. Springer, 2014.

[32] M. Raptis and L. Sigal. Poselet key-framing: A model for human activity recognition. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 2650–2657. IEEE, 2013.

[33] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In WACV, pages 103–110. IEEE, 2013.

[34] J. Shawe-Taylor and N. Cristianini. Kernel methods for pattern analysis. Cambridge university press, 2004.

[35] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1):116–124, 2013.

[36] J. W. Silverstein. Eigenvalues and eigenvectors of large dimensional sample covariance matrices. Contemporary Mathematics, 50:153–159, 1986.

[37] S. Sra. Positive definite matrices and the symmetric stein divergence. Technical report, 2011.

[38] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. Audio, Speech, and Language Processing, IEEE Transactions on, 14(5):1557–1565, 2006.
[39] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, pages 589–600. Springer, 2006.

[40] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(10):1713–1727, 2008.

[41] L. Unzueta, O. Mena, B. Sierra, and Á. Suescun. Kinetic pseudo-energy history for human dynamic gestures recognition. In *Articulated Motion and Deformable Objects*, pages 390–399. Springer, 2008.

[42] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, pages 588–595. IEEE, 2014.

[43] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos. On the improvement of human action recognition from depth map sequences using space–time occupancy patterns. *Pattern Recognition Letters*, 36:221–227, 2014.

[44] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*, pages 872–885. Springer, 2012.

[45] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, pages 2834–2841. IEEE, 2013.

[46] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, pages 804–811. IEEE, 2014.

[47] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACMM*, pages 1057–1060. ACM, 2012.

[48] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *ACCV*, pages 1–5, 2014.

[49] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, pages 2752–2759. IEEE, 2013.

[50] J. Zhang, L. Wang, L. Zhou, and W. Li. Learning discriminative stein kernel for spd matrices and its applications. *Neural Networks and Learning Systems, IEEE Transactions on*, 2015.

[51] X. Zhao, X. Li, C. Pang, Q. Z. Sheng, S. Wang, and M. Ye. Structured streaming skeleton–a new feature for online human gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1s):22, 2014.