HUMAN ACTION RECOGNITION BY FUSING DEEP FEATURES WITH GLOBALITY LOCALITY PRESERVING CANONICAL CORRELATION ANALYSIS

Nour El Din El Madany, Yifeng He, and Ling Guan

Department of Electrical and Computer Engineering , Ryerson University, Toronto, Ontario, Canada

ABSTRACT

This paper proposes a novel Globality Locality Preserving Canonical Correlation Analysis (GLPCCA) for multiview learning. The proposed GLPCCA can preserve the global and local structures. Furthermore, we present a human action recognition framework by using GLPCCA to fuse depth and RGB modalities, which include the proposed Hierarchical Pyramid of Depth Motion Map Deep Convolutional Neural Network (HP-DMM-CNN) for the depth images, and Optical flow CNN for the RGB videos. The proposed framework was evaluated using two datasets, UTD Multimodal Human Action Dataset (UTD-MHAD) and SBU Kinect Interaction data set.The experimental results demonstrated that the proposed GLPCCA can achieve a higher average accuracy compared to several existing methods.

***Index Terms*—** Human Action Recognition, Convolutional Neural Network, Multimodal Fusion

1. INTRODUCTION

Human action recognition has been gaining increasing interest due to its wide range of applications including human-computer interaction, and smart surveillance. The release of the low cost depth sensors like Kinect, directed the researchers in computer vision to address depth based action recognition. Depth cameras have several advantages over RGB cameras. First, depth maps are insensitive to variations in lighting condition. Second, depth maps provide better geometric cues which make the features extracted from depth map videos more discriminant than those from RGB videos.

Research on action recognition has spanned over nearly two decades. The work can be categorized into three main groups: *1) RGB based approaches*, *2) depth based approaches*, and *3) skeleton based approaches*. In the *RGB-based approaches*, researchers investigated action recognition from RGB videos only with different techniques. Some techniques were based on temporal template. In [1], Bobick *et al.* proposed to extract Motion Energy Image (MEI) and Motion History Image (MHI), which are mainly computed by accumulating the differences between human shape masks over each two consecutive frames. Gorelick *et al.* [2] presented a framework based on three dimensional shapes induced by the silhouettes in the space-time volume. Others attempted to capture the temporal information in a better manner. Laptev *et al.* [3] constructed a spatio-temporal volume stacking the silhouettes over a given sequence and introduced Harris 3D point detector to capture the interest points. Willems *et al.* [4] proposed Hessian 3D interest point detector and applied it in action recognition.

In the *depth-based approaches*, depth maps are used for action recognition. In [5], Yang *et al.* adopted a modified version of MHI named depth motion map (DMM). Wang *et al.* [6] extracted semi-local features called random occupancy pattern (ROP) features. In [7], Elmadany *et al.* introduced Hierarchical Pyramid of Depth Motion Map (HP-DMM) which can capture the temporal information in the depth motion maps .

In the *skeleton-based approaches*, action recognition is based only on skeleton. In [8], Yang *et al.* proposed a set of features based on the differences of joints to capture the posture of the skeleton. Others tried to describe the action by representing the trajectoury of the joints. For example, In [9], Hussein *et al.* represented the trajectory of the joints over the sequence using covariance descriptor. Also, Gowayyed *et al.* [10] proposed to represent the trajectory of joints over the sequence using histogram of oriented displacements.

Each modality can capture a certain kind of information that is likely to be complementary to the others. For example, some modalities capture the global information, while the others capture the local details of the action. Intuitively, integrating the information from multiple modalities via multiview learning is expected to improve the recognition performance. To facilitate this purpose, Hardon *et al.* [11] presented Canonical Correlation Analysis (CCA). Furthermore, Kernel Canonical Correlation Analysis (KCCA) [11] was introduced to reveal the nonlinear relation between two sets of data which CCA cannot discover. Extended from CCA, Multi-set Canonical Correlation Analysis (MCCA) was proposed [12] to deal with three or more sets of data. In [13], Locality Preserving Canonical Correlation Analysis (LPCCA) was proposed to capture the local structure of data samples and preserve the local structure in the new learned space. However, the learned space does not capture the global structure of the data samples.

In this paper, we propose a novel Globality Locality Pre-

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

serving Canonical Correlation Analysis (GLPCCA) for multiview learning, and applied it to human action recognition. The improvement of the recognition accuracy is due to the fact that the learned common feature space captures the structure in better way with the proposed GLPCCA.

The rest of this paper is organized as follows. Section 2 presents details of the proposed GLPCCA. The new human action recognition framework is described in Section 3. The experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. THE PROPOSED GLPCCA

2.1. Review on Canonical Correlation Analysis

Canonical correlation analysis (CCA) aims at maximizing the correlation between two different sets [11]. Consider two sets of N zero mean variables $X \in \Re^{p\times N}$and $Y \in \Re^{q\times N}$, p and q are the dimensions of feature samples in X and Y, respectively. CCA aims at learning the projection basis functions which maximize the correlation between the projected samples. The CCA is formalized as follows:

$$\begin{aligned} \underset{W_x,W_y}{\text{maximize}} \quad & Trace(W_x^T C_{xy} W_y) \\ \text{subject to} \quad & W_x^T C_{xx} W_x = I; W_y^T C_{yy} W_y = I \end{aligned} \tag{1}$$

where $C_{xy} = XY^T$ is the cross-correlation matrix between the two sets, and C_{xx} and C_{yy} are the auto-correlation matrices of X and Y. $W_x^T \in \Re^{p\times l}$ and $W_y^T \in \Re^{q\times l}$ are the projection matrices for X and Y, respectively and l is the number of projected dimensions. As proved in [11], Eq. (1) is formalized as generalized eigenvalue decomposition problem as follows:

$$C_{xy} C_{yy}^{-1} C_{xy}^T W_x = \lambda^2 C_{xx} W_x \tag{2}$$

After W_x is obtained, W_y is derived by solving $C_{yy}^{-1} C_{xy}^T W_x/\lambda$. It is worth noting that CCA does not utilize any label information. As a consequence, the projected data samples using the derived projection directions W_x and W_y are not separated well. In other words, the learned space is not discriminative enough.

2.2. Review on Globality locality Preserving Projections

The core concept of Locality Preserving Projections (LPP) algorithms [14] is to encode the manifold structure of the samples using a laplacian graph. The laplacian graph can be constructed in supervised or unsupervised way. For supervised way, the laplacian graph captures only intra-class structure and ignores the inter-class structure which is important in distinguishing between classes. To overcome this problem, Globality Locality Preserving Projections (GLPP) was introduced [15]. GLPP captures the intra-class and inter-class structures of data samples. Let $X \in \Re^{p\times N}$ be set of data samples. GLPP aims to learn the projection matrix W that transfer the data samples into a subspace, where the geometric structure of the projected data samples is well preserved. Let $U = [u_1, ...u_i, ...u_k]$ denotes the mean space of data samples, where u_i is the mean of the class i data sample, and k is the number of classes. U is used for preserving the globality structure of the whole data. Let X_c subset of data samples which belongs to the c^{th} class. A reasonable criterion for revealing the real relationship among data samples is to learn intra-class and inter-class laplacian graphs, which are formulated as follows

$$\Phi_{inter} = \frac{1}{2}\sum_{i,j}(W^T u_i - W^T u_j)^2 B_{ij} \tag{3}$$

$$\Phi_{intra} = \frac{1}{2}\sum_{c\in C}\sum_{i,j\in c}(W^T x_i - W^T x_j)^2 S_{ij} \tag{4}$$

where C is the set of classes and S_{ij} and B_{ij} are the the adjacency weight matrices of intra-class and inter-class data samples, defined as follows

$$S_{ij} = \begin{cases} exp(-\|x_i - x_j\|^2/t_S) & i,j \in c, c \in C, i \neq j \\ 0 & otherwise \end{cases} \tag{5}$$

$$B_{ij} = \begin{cases} exp(-\|u_i - u_j\|^2/t_B) & i \neq j \\ 0 & otherwise \end{cases} \tag{6}$$

where t_S and t_B are a parameter generally taken as the mean square distance of data samples used in calculation of S_{ij} and B_{ij}. GLPP aims to minimize the following objective function

$$\begin{aligned} \psi = \Phi_{inter} + \beta\Phi_{intra} = \frac{1}{2}\sum_{i,j}(W^T u_i - W^T u_j)^2 B_{ij} + \\ \frac{1}{2}\beta\sum_{c\in C}\sum_{i,j\in c}(W^T x_i - W^T x_j)^2 S_{ij} \end{aligned} \tag{7}$$

where β is a positive parameter to control the contribution of preserving intra-class and inter-class structure. Eq. (7) can be rewritten as follows

$$\begin{aligned} \psi = W^T U(G-B)U^T W + \\ \beta\sum_{c\in C} W^T X_c(D-S)X_c^T W \end{aligned} \tag{8}$$

where G is a diagonal matrices and its entries are column sum of B, $G_{ii} = \sum_j B_{ij}$. D is a diagonal matrices and its entries are column sum of S, $D_{ii} = \sum_j S_{ij}$. Eq. 8 can be reduced to

$$\psi = W^T(UKU^T + \beta\sum_{c\in C} X_c L X_c^T)W \tag{9}$$

where $K = G - B$ and $L = D - S$ are the graph laplacian of the intra-class and inter-class, respectively. The problem can be written as follows

$$\min_{W} \quad W^T AW \tag{10}$$

where $A = UKU^T + \beta \sum_{c \in C} XLX^T$ is a positive semi definite matrix. The above problem can be solved using eigenvalue decomposition.

2.3. The Proposed Globality Locality Preserving Canonical Correlation Analysis (GLPCCA)

Instead of considering the local structure or the intra-class structure during establishing correspondences among samples in locality preserving CCA (LPCCA),Globality Locality Preserving Canonical Correlation Analysis (GLPCCA) preserves both the local and global structures during establishing the correspondences. In other words, GLPCCA is a modified CCA which incorporates the local and global information and can be formulated as follows

$$\begin{aligned} \max_{W_x, W_y} \quad & Trace(W_x^T(U_x K_{xy} U_y^T + \beta \sum_{c \in C} X_c L_{xy} Y_c^T) W_y) \\ \text{subject to} \quad & W_x^T(U_x K_{xx} U_x^T + \beta X_c L_{xx} X_c^T) W_x = I; \\ & W_y^T(U_y K_{yy} U_y^T + \beta Y_c L_{yy} Y_c^T) W_y = I; \end{aligned} \tag{11}$$

where $U_x = [u_1^x, ..u_i^x, ..u_k^x]$ and $U_y = [u_1^y, ..u_i^y, ..u_k^y]$ are the mean class for X and Y, respectively. $L_{xx} = D_{xx} - S_x \circ S_x$, $L_{yy} = D_{yy} - S_y \circ S_y$, and $L_{xy} = D_{xy} - S_x \circ S_y$ where the symbol $\circ$ indicates element by element multiplication between two matrices, S_x can be calculated using Eq. (5), and $D_{xx}(D_{yy}, D_{xy})$ is a diagonal matrix of size $n \times n$, and its i^{th} diagonal element is the sum of the i^{th} row or column in $S_x \circ S_x(S_y \circ S_y,\ S_x \circ S_y)$. K_{xy}, K_{xx}, and K_{yy} are calculated in a similar manner to the way L_{xy}, L_{xx}, and L_{yy} are calculated. To obtain the projection matrices W_x^T, and W_y^T, the optimization problem Eq. (11) is converted to generalized eigenvalue decomposition as follows

$$\begin{pmatrix} & A_{xy} \\ A_{xy}^T & \end{pmatrix} \begin{pmatrix} W_x \\ W_y \end{pmatrix} = \lambda \begin{pmatrix} A_{xx} & \\ & A_{yy} \end{pmatrix} \begin{pmatrix} W_x \\ W_y \end{pmatrix} \tag{12}$$

where $A_{xy} = U_x K_{xy} U_y^T + \beta X L_{xy} Y^T$, $A_{xx} = U_x K_{xx} U_x^T + \beta \sum_{c \in C} X_c L_{xx} X_c^T$, and $A_{yy} = U_y K_{yy} U_y^T + \beta \sum_{c \in C} Y_c L_{yy} Y_c^T$ are positive semi definite matrices. Once the projection matrices pairs W_x and W_y are obtained, the dimensionality reduction can be performed by $W_x^T X$, and $W_y^T Y$. GLPCCA attempts to preserve both local and global structure of data samples from the two sets X and Y in the new space as in the original space. In our work, β is chosen as [15].

3. HUMAN ACTION RECOGNITION USING THE PROPOSED GLPCCA

Building reliable human action recognition system is affected greatly by the features which represent the action. We examine the use of three types of feature sets which are depth maps and RGB videos. The proposed human action recognition framework uses Hierarchical Pyramid of DMM and Convolutional neural network (HP-DMM-CNN) for depth maps, and Optical Flow Convolutional Neural Network (Optical flow CNN) for RGB videos.

3.1. Depth Maps

Depth videos are important in action recognition because their ability to capture 3D action structure effectively. Here, we propose the Hierarchical Pyramid DMM Deep Convolutional Neural Network (HP-DMM-CNN) to extract fetaures from depth map videos. Based on DMM initially proposed by Yang *et al.* [5], HP-DMM is capable of representing the changes in the human motions in a better way. In DMM, each depth frame is projected onto three Cartesian planes for the purpose of representing the 3D structure of motion. The projections form three projected maps denoted as DM_F, DM_S, and DM_T representing the front view, the side view, and the top view, respectively. For each view, the absolute difference between two consecutive projected maps is computed. Then, differences are accumulated over the time to form the DMM. DMM computes the motion energy over the video sequence by stacking the motion energy over the duration of the video. However, DMM faces two challenges. First, DMM fails to capture the temporal order of motions over time, making it difficult to recognize two actions with reverse temporal orders, such as push and pull. Second, DMM suffers from motion self-occlusion. Since DMM accumulates the motion differences over the video, the most recent motion may overwrite the previous motion occurring at the same location. This problem leads to information loss which has a negative impact on recognition accuracy.

To construct HP-DMM, the video sequence is divided into multiple partitions. HP-DMM has the ability to capture more detailed information about the action and fine changes of the human movements due to capturing the sub actions within the video sequence. In our work, we employ a 2-level pyramid with three partitions which showed good results [7]. Finally, Deep CNN [16] is used to describe the local shape of each $DMM_{F,S,T}$ in each partition. The architecture of Deep CNN follows [16]. The output from the second fully connected layer is used as a descriptor for each $DMM_{F,S,T}$ in each partition. In this work, we used the publicly available $VGG - f$ pretrained on ImageNet ILSVRC-2012 challenge dataset.

3.2. RGB Videos

In our work, optical flow CNN is used to capture motion. At each frame, descriptor is computed at each frame. The final descriptor is the fusion of all time descriptor. For each two successive frames optical flow is computed as in [17]. Then, we used optical flow CNN to extract the descriptor for each optical flow frame f_t. The architecture of Deep CNN is identical to [18]. In this work, we used the publicly available $VGG-f$ pretrained on UCF101 dataset. The output from the second fully connected layer is used as a descriptor for each optical flow frame. The descriptor Des is computed as follow

$$Des = [\min_{1<t\leq T}(f_t) \min_{1<t\leq T}(f_t - f_{t-1}) \max_{1<t\leq T}(f_t) \max_{1<t\leq T}(f_t - f_{t-1})]^T \tag{13}$$

where T is the number of frames per video and f_t is the output of second fully connected layer at time t.

3.3. Fusion and Classification

After the shared space is learned using GLPCCA, the final descriptor F can be written as follows:

$$F = \begin{bmatrix} W_1'X_1 & W_2'X_2 \end{bmatrix} \tag{14}$$

where W_i is the projection matrix for the i^{th} feature set X_i. Finally, Linear SVM is adopted as a classifier.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed GLPCCA for human action recognition using two datasets, UTD Multimodal Human Action Dataset [19] and SBU Kinect Interaction dataset [20].

4.1. Experiments using UTD Multimodal Human Action Dataset

UTD Multimodal Human Action Dataset (UTD-MHAD) [19] is composed of 27 actions performed by 8 subjects. Each subject performed the action 4 times. The dataset contains 861 action sequences. We followed the experiment settings in [19], where a half of the subjects were used in training and another half for testing. First, we look into the recognition accuracy for each modality in UTD-MHAD dataset. The recognition performance for HP-DMM-CNN, and Optical flow CNN are first computed. Optical flow CNN and HP-DMM-CNN achieve recognition accuracy of 82.56 % and 82 %. It is noticed from Table 1 that GLPCCA has the highest recognition accuracy of 93.26 %, among state of the art multi-view techniques compared.

4.2. SBU Kinect Interaction Dataset

SBU dataset was presented to recognize the interaction between two person interactions. SBU dataset is composed of 8 actions performed by 21 subjects. Note that in most interactions, one person is acting and the other person is reacting. we followed 5-fold cross-validation scheme, the same settings used in [20].

Table 1. Recognition accuracy comparison between the proposed methods GLPCCA and the other multi-view techniques on UTD-MHAD Dataset.

Multi-view Technique	RGB-Depth %
MvDA-Vc [21]	85.81
GMA [22]	85.12
CCA [11]	88.84
Cluster CCA [23]	88.84
LPCCA [13]	90.93
MCC-GP [24]	90.47
GLPCCA	**93.26**

We first look into the recognition accuracy of each modality. The recognition performances of HP-DMM-CNN, and optical flow CNN are 84.48 % and 82.81 %. To exploit the benefits of multi-view learning using GLPCCA for two modalities, experiments on the fusion of depth and RGB are conducted and the results are presented in Table 2 along with those obtained by MvDA-Vc, GMA, CCA, Cluster-CCA, LPCCA, MCC-GP. It is noticed from Table 2 that GLPCCA achieves the highest recognition accuracy of 90.1 %.

Table 2. Recognition accuracy comparison between the proposed methods GLPCCA and the other multi-view techniques on SBU Kinect Interaction Dataset.

Multi-view Technique	RGB-Depth %
MvDA-Vc [21]	89.06
GMA [22]	86.85
CCA [11]	88.43
Cluster CCA [13]	88.4
LPCCA [13]	87.9
MCC-GP [24]	88.3
GLPCCA	**90.1**

5. CONCLUSIONS

We proposed a novel approach for multi-view learning using GLPCCA. The proposed GLPCCA is applied to fuse multimodal features in human action recognition. The experimental results demonstrated that GLPCCA performances better in human action recognition than the other the methods compared in terms of average recognition accuracy.

6. REFERENCES

[1] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257–267, 2001.

[2] Gorelick, L., Blank, M. Shechtman, and R. E. Irani, M. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[3] I. Laptev and T. Lindeberg, "On space-time interest points," *International Journal of Computer Vision*, pp. 650–663, 2005.

[4] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatiotemporal interest point detector," in *European Conference on Computer Vision*, 2008.

[5] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *ACM International Conference on Multimedia (MM)*, 2012.

[6] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y.Wu, "Robust 3d action recognition with random occupancy patterns," in *European Conference on Computer Vision(ECCV)*, 2012.

[7] N. Elmadany, Y. He, and L. Guan, "Human action recognition using temporal hierarchical pyramid of depth motion map and keca," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2015.

[8] X. Yang and Y. Tian, "Eigenjoints-based action recognition using nave-bayes-nearest-neighbor," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012.

[9] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[10] M. Gowayyed, M. Torki, M. Hussein, and M. El-Saban, "Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[11] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis, an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[12] M.A. Hasan, "On multi-set canonical correlation analysis," in *International Joint Conference on Neural Networks*, 2009, pp. 1128–1133.

[13] T. Sun and S. Chen, "Locality preserving cca with applications to data visualization and pose estimation," *Journal Image and Vision Computing*, vol. 25, no. 5, 2007.

[14] X. He and P. Niyogi, "Locality preserving projections," in *The Annual Conference on Neural Information Processing Systems (NIPS)*, 2003.

[15] S. Huang, A. Elgammal, J. Lu, and D. Yang, "Cross-speed gait recogntion using speed invariant gait templates and globality-locality preserving projections," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 10, 2015.

[16] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convoltional networks," in *The British Machine Vision Conference (BMVC)*, 2014.

[17] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on theory for warping," in *The European Conference on Computer Vision (ECCV)*, 2004.

[18] G. Gkioxari and J. Malik, "Finding action tubes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *IEEE International Conference on Image Processing (ICIP)*, 2015.

[20] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[21] M. Kan, S. Shan, H. Zhang, and S. Lao gand X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, 2016.

[22] A. Sharma, A. Kumar, H. Daume III, and D. Jacobs, "Generalized multiview analysis: a discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[23] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[24] Y. Yuan and Q. Sun, "Multiset canonical correlations using globality preserving projections with applications to feature extraction and recognition," *IEEE Transactions on Neural Network and Learning Systems*, vol. 25, no. 6, 2014.