Effective Fusing The Factored Matrices in Dual Tensors for Action Recognition

Chung-Yang Hsieh
National Chung Cheng University
Taiwan
k40964096@gmail.com

Wei-Yang Lin
National Chung Cheng University
Taiwan
wylin@cs.ccu.edu.tw

Abstract

In statistics, Canonical Correlation Analysis (CCA) is a kind of method to effectively analyze the correlation between two data. In recent years, some methods reported in literatures treated the action video sequences as the tensors, and calculated the similarity between two video sequences using CCA. Each of these tensors was unfolded into several matrices or vectors in the previous works. And estimated the similarity between two tensors via accumulating the canonical correlations on all of the pairs of the unfolded matrices or vectors. In this paper, we treat each of the unfolded matrices in a tensor as an individual, instead of accumulating the canonical correlations in a whole tensor, such that we can effectively use the characteristic of each unfolded matrix. We also propose an information fusion method to combine the similarities of each of the unfolded matrices between two tensors. Furthermore, we add the Histogram of Oriented Gradients (HOG) features to complement the tensor generated by the pure video sequence. Our method is validated on the UCF sports database, and the experimental result shows that the proposed method can compete with the state-of-the-art methods.

1 Introduction

In recent years, hardware advance promotes the development of human-computer interaction, the issue about action recognition has also been widely discussed and researched. There are many researches focus on sampling strategy, feature extraction and representation [1, 2, 3, 4, 5]. For the extracted features, using some transformations such as manifold or projection can further improve performance [6]. Furthermore, for action recognition, there are some literatures discuss that resize the video sequences as the 3^{rd} order tensors, and then evaluate the similarity of these tensors for classification [7, 8, 9, 10]. The tensor is usually unfolded into several matrices or vectors, for example, a 3^{rd} order tensor can be unfolded into three matrices. However, each of unfolded matrices in a tensor has different ability of discriminant for the classification problem. In this paper, we argue that utilizing the discriminant of various unfolded matrices effectively will be better than using the summation of all factored matrices directly.

For different classification problems, using some suitable features can improve performance effectively. For example, the geometric or color-based features for scene classification, and the Histograms of Oriented Gradients (HOG), Histograms of Optic Flow (HOF) for action recognition [5, 11, 12]. Therefore, in addition to using the resized video sequences as the tensors, we also treat the result of other feature extraction methods as another tensor, and let the complementary effect between tensors can further enhances the performance. In our experimental result, the HOG features can indeed effective complement for the resized video sequences, and thus enhance the performance.

The rest of this paper is organized as follows: The proposed method is described in detail in Section 2. In Section 3, we compare proposed method with the state-of-the-art techniques. Finally, conclusions are given in Section 4.

2 Methodology

In this section, we will present a novel approach for recognizing human action in video sequence. Our approach is inspired by some of the tensor-based approaches [8, 9, 7], but using an enhanced method to effectively combine the similarity measures between two tensors. The flow chart of our proposed method is shown in Figure 1 and the corresponding details are described in the following subsections.

2.1 Preprocessing

Given an input video sequence, we firstly transform its pixel values from the original RGB color space into gray-level format. Since video sequences may have various frame sizes and durations (i.e., number of frames in a video sequence), we also perform resizing in spatial domain and resampling in temporal domain to produce normalized raw data. The resulting normalized raw data can be represented as third order tensors in $\mathsf{R}^{W\times H\times L}$, where W, H, and L are the width, height, and length of a tensor, respectively. Simply stated, the preprocessing step converts video sequences into tensor volumes with equal size.

2.2 Low-level feature

Most of the existing tensor-based approaches for action recognition only utilize tensors constructed from raw data (i.e., pixel values). In other words, these approaches do not explicitly take feature extraction into consideration. However, an appropriate feature extraction could yield a new representation of raw data that makes the recognition performance better. Moreover, the complementarity between raw data and extracted features may provide a synergistic effect to further improve classification accuracy.

Here, we perform feature extraction on the normalized raw data using the HOG descriptor. In particular, for every resized image frame within a normalized raw datum, we firstly calculate intensity gradient at each pixel location. We can accumulate the orientations of

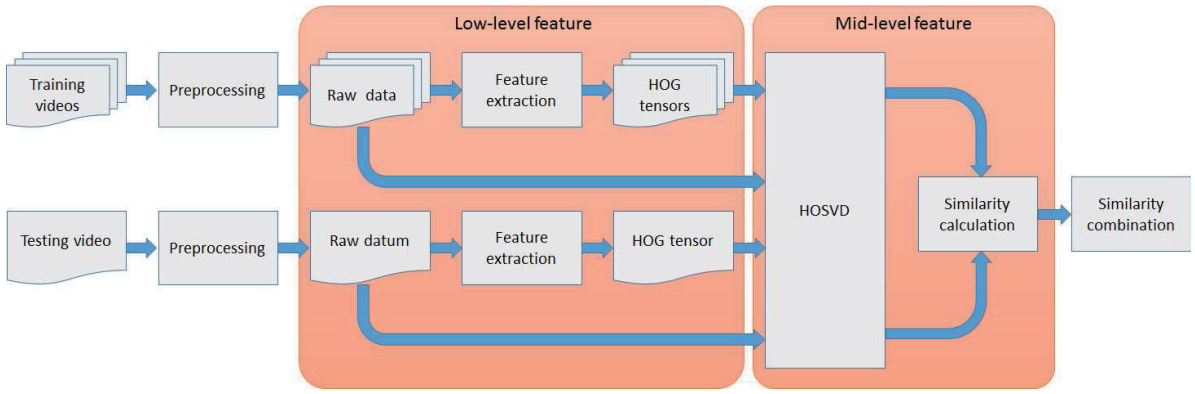

Figure 1: Flowchart of the proposed action recognition method.

gradient in each cell into a histogram. Then, orientation histograms from several cells are composed into a block histogram. The main idea behind the HOG descriptor is to represent the trend of gradient orientation within a local image area. This descriptor has been found very useful in representing the shape of human body [13]. By calculating HOG descriptors from the normalized raw data, we can generate fourth order tensors in $\mathsf{R}^{W' \times H' \times B \times L}$, where W and H denote the numbers of regions on the horizontal and vertical directions respectively, L denotes the length of a normalized video sequence, and B denotes the number of bins in the HOG descriptor.

2.3 Mid-level feature

We consider the raw data tensor and HOG tensor obtained in the previous step as low-level features. The details of mid-level feature extraction and similarity calculation are described in the following two subsections.

2.3.1 HOSVD

The aim of using the HOSVD algorithm is to unfold a higher order tensor into 2D matrices. Given an N^{th} order tensor $\mathcal{A} \in \mathsf{R}^{s_1 \times s_2 \times \ldots \times s_N}$, it can be unfolded into a set of 2D matrices $\{\mathbf{A}_{(k)}\}_{k=1}^{N}$ where the size of $\mathbf{A}_{(k)}$ is $s_k \times \prod_{i=k} s_i$ [14].

Similar to Singular Value Decomposition (SVD), an unfolded matrix can be factored using HOSVD. The HOSVD of an unfolded matrix $\mathbf{A}_{(k)}$ is a factorization as follows:

$$\mathbf{A}_{(k)} = \mathbf{U}_{(k)} \mathbf{\Sigma}_{(k)} \mathbf{V}_{(k)}^T, \tag{1}$$

where $\mathbf{\Sigma}_{(k)}$ is a diagonal matrix, $\mathbf{U}_{(k)}$ and $\mathbf{V}_{(k)}$ are the orthogonal matrices spanning the column space and row space of $\mathbf{A}_{(k)}$, respectively. As indicated in a previous study [7], factored matrix $\mathbf{V}_{(k)}$ can be considered as a point in a Grassmann manifold because it is the orthogonal matrix spanning the row space associated with nonzero singular values of $\mathbf{A}_{(k)}$. Consequently, factored matrix $\mathbf{V}_{(k)}$ represents the mapping of a tensor from low-level feature space to a mid-level feature space (i.e., Grassmann manifold). We can then compute similarity in mid-level feature spaces for tensor classification. For a raw data tensor, we will obtain three factored matrices as its representations in mid-level feature spaces. Similarly, we will obtain four factored matrices as the representation of an HOG tensor in mid-level feature spaces.

2.3.2 Similarity in mid-level feature space

In a mid-level feature space, we use CCA to measure the similarity between two factored matrices. It is well know in statistics that CCA is a way of evaluating correlation between two sets of random variables. Given two random vectors $\mathbf{x} \in \mathsf{R}^m$ and $\mathbf{y} \in \mathsf{R}^n$, the goal of CCA is to find two vectors $\mathbf{u} \in \mathsf{R}^m$ and $\mathbf{v} \in \mathsf{R}^n$ which maximize the correlation of $\mathbf{u}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{y}$, i.e.,

$$\rho = \max_{\mathbf{u},\mathbf{v}} \mathrm{corr}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y}). \tag{2}$$

The vectors $\mathbf{u}$ and $\mathbf{v}$ are called canonical transformations and ρ is called canonical correlation. Suppose one can find a pair of canonical transformations $\mathbf{u}_1$ and $\mathbf{v}_1$ which achieve maximum correlation ρ_1. Then, one can also seek another pair of canonical transformations maximizing the same objective function (i.e., $\mathrm{corr}(\mathbf{u}^T\mathbf{x}, \mathbf{v}^T\mathbf{y})$) subject to the constraint that they are orthogonal to the previous ones. This procedure can be repeated up to $d = \min\{m, n\}$ times. As a result, one ends up having d pairs of canonical transformations $\{\mathbf{u}_i, \mathbf{v}_i\}_{i=1}^{d}$ and the corresponding canonical correlations $\{\rho_i\}_{i=1}^{d}$. These canonical correlations can be arranged into a vector $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_d]^T$.

Since a raw data tensor is unfolded into three matrices, the similarity between a pair of raw data tensors is represented by three sets of canonical correlations. Similarly, an HOG tensor is unfolded into four matrices. The similarity between a pair of HOG tensors is represented by four sets of canonical correlations. Totally, we have seven sets of canonical correlations as the similarity measurement between two video sequences. We discuss the issue of how to effectively combine these canonical correlations in the next section.

2.4 Similarity combination

The traditional method to evaluate the similarity between two tensors is to sum up all of the canonical correlations. In the classification step, a testing sample is evaluated with training samples using CCA, and it is classified as the label with the sample which has the

maximum summation of canonical correlations. However, different factored matrices should have different discriminant abilities. Our approach treats each factored matrix as an individual, rather than sums up the canonical correlations in a whole tensor. That is, each pair of factored matrices sum up the canonical correlations individually. Then, we combine these results from normalized raw data and HOG features and project the combined vectors to another space.

Let ψ be the 1-norm of the vector $\boldsymbol{\rho}$. So we have a new feature vector $\{\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6, \psi_7\}$ for each video sequence, where $\{\psi_1, \psi_2, \psi_3\}$ is obtained from the 3^{rd} tensor (normalized raw data) and $\{\psi_4, \psi_5, \psi_6, \psi_7\}$ is obtained from the 4^{th} tensor (HOG features). Figure 2 is the illustration of the combination and the projection processes. First, the similarities between a testing normalized raw data to t training normalized raw data are estimated using CCA. Similarly, the similarities between a testing HOG data to t training HOG data are also estimated using CCA too. Here we can get t similarity vectors $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_t$. In order to find a space which can maximize the difference of combined vectors, we use Principal Component Analysis (PCA) to find the best projection vectors. There are three benefits of using PCA. *1)* reduce the amount of data. *2)* it is an unsupervised learning method. *3)* project the data to the space as possible as to decentralize the data, this will help classification. We tread $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_t$ as the training samples and calculate the best projection space from them using PCA. At the same time, we generate the sample $\boldsymbol{\Psi}_T$ to represent the testing sample. The sample $\boldsymbol{\Psi}_T$ is defined as follows:

$$\boldsymbol{\Psi}_T = \{\max(\psi_{j,1}), \ldots, \max(\psi_{j,7}) \mid j = 1 \ldots t\}. \quad (3)$$

Due to the relationship of value ψ_k and similarity is positive correlation, so we use the maximum of each dimension from $\boldsymbol{\Psi}_1$ to $\boldsymbol{\Psi}_t$ to represent the testing samples. Finally, all samples $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_t$ and $\boldsymbol{\Psi}_T$ are projected to the space calculated from PCA. The projected mid-level feature vectors are $\widehat{\boldsymbol{\Psi}}_1, \ldots, \widehat{\boldsymbol{\Psi}}_t$ and $\widehat{\boldsymbol{\Psi}}_T$ where $\widehat{\boldsymbol{\Psi}}_1, \ldots, \widehat{\boldsymbol{\Psi}}_t$ denote the training mid-level feature vectors and $\widehat{\boldsymbol{\Psi}}_T$ denotes the testing mid-level feature vector.

3 Experiments

We test our approach on the UCF sports database [15]. The UCF sports database has 150 video sequences gathered from broadcast television channels such as BBC and ESPN. This database includes ten categories which are diving, golf swinging, kicking, lifting, riding horse, running, skateboarding, pommel-horse, high-bar swinging, and walking, respectively. Each video sequence is resized and resampled into a $32 \times 32 \times 64$ tensor; moreover, each video sequence is also resized and resampled into a $72 \times 72 \times 64$ video sequence for HOG feature extraction. The experiments on the UCF sports database are performed using leave-one-out cross-validation. The example frames of 5 action category from the UCF sports database are shown in Figure 3.

We compare the recognition performance of our proposed method with the state-of-the-art methods. In order to make a fair comparison with the past results, we follow exactly the same experimental procedure as

diving | golf swinging | kicking | lifting | riding horse

Figure 3: Some example frames of the UCF sports database.

Table 1: Comparison of recently reported results on the UCF sports database.

Author, year	Recognition rate
Raptis *et al.*, 2012 [11]	79.4%
Kim *et al.*, 2007 [8]	82%
Lui *et al.*, 2011 [10]	88%
Deng *et al.*, 2013 [6]	88.2%
Wang *et al.*, 2011 [4]	88.04%
O'Hara *et al.*, 2012 [17]	91.3%
Wu *et al.*, 2013 [12]	92.48%
Yuan *et al.*, 2013 [18]	92.67%
Our method	**92.67%**
Sadanand *et al.*, 2012 [19]	95%
Harandi *et al.*, 2013 [16]	96.6%

those described in the previous works [8, 9]. Classification accuracies of the state-of-the-art approaches on the UCF sports database are summarized in Table 1. Our proposed method achieves the recognition accuracy of 92.67%, which is among the top three results on the UCF sports database. It shows that our method achieves comparable performance as the recently proposed methods for action recognition. Although the best result in Table 1 is 96.6%, it is achieved using Auto Regressive and Moving Average (ARMA) modelling to improve recognition performance [16]. Their proposed method achieves recognition accuracy of 88.4% without using ARMA modeling.

4 Conclusions

This paper demonstrates that the discriminant of each matrix decomposed from the tensor is an important characteristic for classification. We propose to accumulate the similarity in each individual matrix decomposed from the tensor, then project the mid-level features to another suitable space for classification. Furthermore, we treat the HOG features as another tensor, and merge the resulting vectors in two tensors to make up the imperfect parts of using a single tensor for classification problem. According to the various classification problems, our proposed method is easy to combine with one or more suitable features and treats them as the triple or multiple tensors. Finally, we also verified that the two phases can really enhance the recognition rate on the UCF sports database.

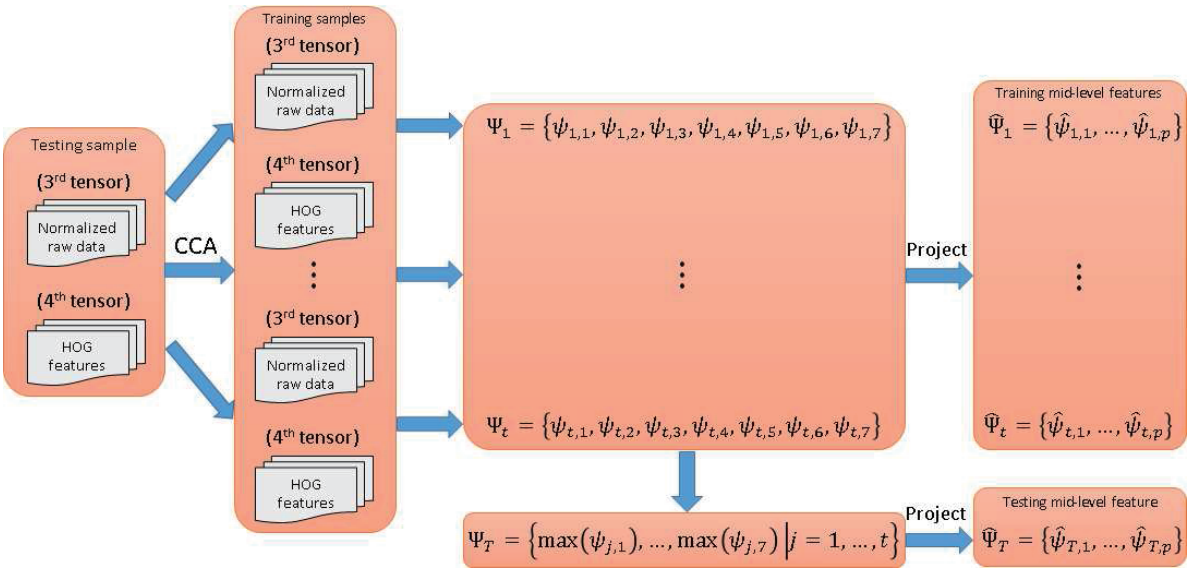

Figure 2: The combination process from low-level feature to mid-level feature

References

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *International Conference on Pattern Recognition*, vol. 3, pp. 32–36, IEEE, 2004.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, IEEE, 2005.

[3] F. Shi, E. Petriu, and R. Laganiere, "Sampling strategies for real-time action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2595–2602, IEEE, 2013.

[4] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3169–3176, IEEE, 2011.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[6] X. Deng, X. Liu, M. Song, J. Cheng, J. Bu, and C. Chen, "Lf-eme: Local features with elastic manifold embedding for human action recognition," *Neurocomputing*, vol. 99, pp. 144–153, 2013.

[7] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 833–839, IEEE, 2010.

[8] T.-K. Kim, K.-Y. K. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[9] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[10] Y. M. Lui and J. R. Beveridge, "Tangent bundle for human action recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, pp. 97–102, IEEE, 2011.

[11] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1242–1249, IEEE, 2012.

[12] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm.," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 23, no. 8, pp. 1422–1431, 2013.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, IEEE, 2005.

[14] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[15] J. A. Mikel D. Rodriguez and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Kernel analysis on grassmann manifolds for action recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1906–1915, 2013.

[17] S. O'Hara and B. A. Draper, "Scalable action recognition with a subspace forest," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1210–1217, IEEE, 2012.

[18] C. Yuan, W. Hu, G. Tian, S. Yang, and H. Wang, "Multi-task sparse learning with beta process prior for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 423–429, IEEE, 2013.

[19] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1234–1241, IEEE, 2012.