Tensor Canonical Correlation Analysis for Action Classification

Tae-Kyun Kim, Shu-Fai Wong, Roberto Cipolla
Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, CB2 1PZ, UK

Abstract

We introduce a new framework, namely Tensor Canonical Correlation Analysis (TCCA) which is an extension of classical Canonical Correlation Analysis (CCA) to multidimensional data arrays (or tensors) and apply this for action/gesture classification in videos. By Tensor CCA, joint space-time linear relationships of two video volumes are inspected to yield flexible and descriptive similarity features of the two videos. The TCCA features are combined with a discriminative feature selection scheme and a Nearest Neighbor classifier for action classification. In addition, we propose a time-efficient action detection method based on dynamic learning of subspaces for Tensor CCA for the case that actions are not aligned in the space-time domain. The proposed method delivered significantly better accuracy and comparable detection speed over state-of-the-art methods on the KTH action data set as well as self-recorded hand gesture data sets.

1. Introduction

Many previous studies have been carried out to categorize human action and gesture classes in videos. Traditional approaches based on explicit motion estimation require optical flow computation or feature tracking, which is a hard problem in practice. Some recent work has analyzed human actions directly in the space-time volume without explicit motion estimation [1, 4, 8, 7]. Motion history images and the space-time local gradients are used to represent video data in [4, 8] and [1] respectively, having the benefits of being able to analyze quite complex and low-resolution dynamic scenes. However, both representations convey only partial data of the space-time information (mainly motion data) and are unreliable in cases of motion discontinuities and motion aliasing. Also, the method in [1] has the drawback of requiring to manually set the positions of local space-time patches. Importantly, it has been noted that spatial information contains cues as important as dynamic information for human action classification [2]. In the study, actions are represented as space-time shapes by the silhouette images and the Poisson equation. However, it assumes that silhouettes are extracted from video. Furthermore, as noted in [2], the silhouette images may not be sufficient to represent complex spatial information.

There are other important action recognition methods which are based on space-time interest points and visual code words [3, 6, 5]. The histogram representations are combined with either a Support Vector Machine (SVM) [6, 5] or a probabilistic model [3]. Although they have yielded good accuracy, mainly due to the high discrimination power of individual local space-time descriptors, they do not encode global space-time shape information. Their performance also highly depends on proper setting of the parameters of the space-time interest points and the code book.

In this paper, a statistical framework of extracting similarity features of two videos is proposed for human action/gesture categorization. We extend the classical canonical correlation analysis - a standard tool for inspecting linear relationships between two sets of vectors [9, 11] - into that of multi-dimensional data arrays (or high-order tensors) for analyzing the similarity of video data/space-time volumes. Note the framework itself is general and may be applied to other tasks requiring matching of various tensor data. The recent work (not published as a full paper) [12], which was studied independently of our work, also presents a concept of Tensor Canonical Correlation Analysis (TCCA) and backs up our new ideas. The originality of this paper is advocated not only by the new TCCA framework but also by new applications of CCA to action classification and efficient action detection algorithms.

This work was motivated by our previous success [16], where Canonical Correlation Analysis (CCA) is adopted to measure the similarity of any two image sets for robust object recognition. Image sets are collected either from a video or multiple still shots of objects. Each image in the two sets is vectorized and CCA is applied to the two sets of vectors. Recognition is performed based on canonical correlations, where higher canonical correlations indicate higher similarity of two given image sets. The CCA based method yielded much higher recognition rates than the traditional set-similarity measures e.g. Kullback

Leibler-Divergence (KLD). KLD-based matching is highly subjective to simple transformations of data (e.g. global intensity changes and variances), which are clearly irrelevant for classification, resulting in poor generalization to novel data. A key of CCA over traditional methods is its affine invariance in matching, which allows for great flexibility yet keeps sufficient discriminative information. The geometrical interpretation of CCA is related to the angle between two hyper-planes (or linear subspaces). Canonical correlations are the cosine of the principal angles and smaller angular planes are thought to be more alike. It is well known that object images are class-wise well-constrained to lie on low-dimensional subspaces or hyper-planes. This subspace-based matching effectively gives affine-invariance, i.e. invariant matching of the image sets to the pattern variations subject to the subspaces. For more details, refer to [16].

Despite the success of CCA in image-set comparison, the CCA is still insufficient for video classification as a video is more than simply a set of images. The previous method does not encode any temporal information of videos. The new tensor canonical correlation features have many favorable characteristics :

- TCCA yields affine-invariant similarity features of global space-time volumes.
- TCCA does not involve any significant tuning parameters.
- TCCA framework can be partitioned into sub-CCAs. The previous works on object recognition [16] based on image sets can be seen as a sub-problem of this framework.

The quality of TCCA features is demonstrated in terms of action classification accuracy being combined with a simple feature selection scheme and Nearest Neighbor (NN) classification. Additionally, time-efficient detection of a target video is proposed by incrementally learning the space-time subspaces for TCCA.

The rest of the paper is organized as follows: Backgrounds and notations are given in Section 2 and the framework and the solution for tensor CCA in Section 3. Section 4 and 5 are for the discriminative feature selection and the action detection method respectively. The experimental results are shown in Section 6 and we conclude in Section 7.

2. Backgrounds and Notations

2.1. Canonical Correlation Analysis

Since Hotelling (1936), Canonical Correlation Analysis (CCA) has been a standard tool for inspecting linear relationships between two random variables (or two sets of vectors) [11]. Given two random vectors $\mathbf{x} \in \mathbb{R}^{m_1}, \mathbf{y} \in \mathbb{R}^{m_2}$,

Figure 1. **Probabilistic Canonical Correlation Analysis** *tells how well two random variables* $\mathbf{x}, \mathbf{y}$ *are represented by a a common source variable* $\mathbf{z}$ *[9].*

a pair of transformations $\mathbf{u}, \mathbf{v}$, called canonical transformations, is found to maximize the correlation of the two vectors $\mathbf{x}' = \mathbf{u}^T\mathbf{x}, \mathbf{y}' = \mathbf{v}^T\mathbf{y}$ as

$$\rho = \max_{\mathbf{u},\mathbf{v}} \frac{E[\mathbf{x}'\mathbf{y}'^T]}{\sqrt{E[\mathbf{x}'\mathbf{x}'^T]E[\mathbf{y}'\mathbf{y}'^T]}} = \frac{\mathbf{u}^T\mathbf{C}_{\mathbf{xy}}\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{C}_{\mathbf{xx}}\mathbf{u}\mathbf{v}^T\mathbf{C}_{\mathbf{yy}}\mathbf{v}}} \tag{1}$$

where ρ is called the canonical correlation and multiple canonical correlations $\rho_1, ...\rho_d$ where $d < min(m_1, m_2)$ are defined by the next pairs of $\mathbf{u}, \mathbf{v}$ which are orthogonal to the previous ones. A probabilistic version of CCA [9] gives another viewpoint. As shown in Figure 1, the model reveals how well two random variables $\mathbf{x}, \mathbf{y}$ are represented by a common source (latent) variable $\mathbf{z} \in \mathbb{R}^d$ with the two likelihoods $p(\mathbf{x}|\mathbf{z}), p(\mathbf{y}|\mathbf{z})$, which comprises affine transformations w.r.t. the input variables $\mathbf{x}, \mathbf{y}$ respectively. The maximum likelihood estimation on this model leads to the canonical transformations $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_d], \mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_d]$ and the associated canonical correlations $\rho_1, ..., \rho_d$, which are equivalent to those of the standard CCA. See [9] for more details. Intuitively, the first pair of canonical transformations corresponds to the most similar direction of variation of the two data sets and the next pairs represent other directions of similar variations. Canonical correlations reveals the degree of matching of the two sets in each canonical directions.

Affine-invariance of CCA. A key of using CCA for high-dimensional random vectors is its affine invariance in matching, which gives robustness with respect to intra-class data variations as discussed above. Canonical correlations are invariant to affine transformations w.r.t. inputs, i.e. $\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{C}\mathbf{y} + \mathbf{d}$ for arbitrary $\mathbf{A} \in \mathbb{R}^{m_1 \times m_1}, \mathbf{b} \in \mathbb{R}^{m_1}, \mathbf{C} \in \mathbb{R}^{m_2 \times m_2}, \mathbf{d} \in \mathbb{R}^{m_2}$. This proof is straightforward from (1) as $\mathbf{C}_{\mathbf{xy}}, \mathbf{C}_{\mathbf{xx}}, \mathbf{C}_{\mathbf{yy}}$ are covariance matrices and are multiplied by arbitrary transformations $\mathbf{u}, \mathbf{v}$.

Matrix notations for Tensor CCA. Given two data sets as matrices $\mathbf{X} \in \mathbb{R}^{N \times m_1}$, $\mathbf{Y} \in \mathbb{R}^{N \times m_2}$, canonical correlations are found by the pairs of directions $\mathbf{u}, \mathbf{v}$. The canonical transformations $\mathbf{u}, \mathbf{v}$ are considered to have unit size hereinafter. The random vectors $\mathbf{x}, \mathbf{y}$ in (1) correspond to

the rows of the matrices $\mathbf{X}, \mathbf{Y}$ assuming $N \gg m_1, m_2$. The standard CCA can be written as

$$\rho = \max_{\mathbf{u},\mathbf{v}} \mathbf{X'}^T \mathbf{Y'}, \text{ where } \mathbf{X'} = \mathbf{X}\mathbf{u}, \mathbf{Y'} = \mathbf{Y}\mathbf{v}. \quad (2)$$

This matrix notation of CCA is useful to describe the proposed tensor CCA with the tensor notations in the following section.

2.2. Multilinear Algebra and Notations

This section briefly introduces useful notations and concepts of multilinear algebra [10]. A third-order tensor which has the three modes of dimensions I, J, K is denoted by $\mathcal{A} = (\mathcal{A})_{ijk} \in \mathbb{R}^{I \times J \times K}$. The inner product of any two tensors is defined as $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k} (\mathcal{A})_{ijk} (\mathcal{B})_{ijk}$. The *mode*-j vectors are the column vectors of matrix $\mathbf{A}_{(j)} \in \mathbb{R}^{J \times (IK)}$ and the j-mode product of a tensor $\mathcal{A}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J \times N}$ is

$$(\mathcal{B})_{ink} \in \mathbb{R}^{I \times N \times K} = (\mathcal{A} \times_j \mathbf{U})_{ink} = \Sigma_j (\mathcal{A})_{ijk} \mathbf{u}_{jn} \quad (3)$$

The j-mode product in terms of j-mode vector matrices is $\mathbf{B}_{(j)} = \mathbf{U}\mathbf{A}_{(j)}$.

3. Tensor Canonical Correlation Analysis

3.1. Joint and Single-shared-mode TCCA

Many previous studies have dealt with tensor data in its original form to consider multi-dimensional relationships of the data and to avoid *curse of dimensionality* when the multi-dimensional data array are simply vectorized. We generalize the canonical correlation analysis of two sets of vectors into that of two higher-order tensors having multiple shared modes (or *axes*).

A single channel video volume is represented as a third-order tensor denoted by $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$, which has the three modes, i.e. axes of space (X and Y) and time (T). We assume that every video volume has the uniform size of $I \times J \times K$. Thus the third-order tensors can *share* any single mode or multiple modes. Note that the canonical transformations are applied to the modes which are not shared. For e.g. in (2), classical CCA applies the canonical transformations $\mathbf{u}, \mathbf{v}$ to the modes in $\mathbb{R}^{m_1}, \mathbb{R}^{m_2}$ respectively, having a shared mode in $\mathbb{R}^N$. The proposed Tensor CCA (TCCA) consists of the different architectures according to the number of the shared modes. The joint-shared-mode TCCA allows any two modes (i.e. a section of video) to be shared and applies the canonical transformation to the remaining single mode, while the single-shared-mode TCCA shares any single mode (i.e. a scan line of video) and applies the canonical transformations to the two remaining modes. See Figure 2 for the concept of the proposed two types of TCCA.

The proposed TCCA for two videos is conceptually seen as the aggregation of many different canonical correlation analyses, which are for two sets of XY sections (i.e. images), two sets of XT or YT sections (in the joint-shared-mode), or sets of X,Y or T scan lines (in the single-shared-mode) of the videos.

Joint-shared-mode TCCA. Given two tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K}$, the joint-shared-mode TCCA consists of three sub-analyses. In each sub-analysis, one pair of canonical directions is found to maximize the inner product of the output tensors (called **canonical objects**) by the mode product of the two data tensors by the pair of the canonical transformations. That is, the single pair (for e.g. $(\mathbf{u}_k, \mathbf{v}_k)$) in $\Phi = \{(\mathbf{u}_k, \mathbf{v}_k), (\mathbf{u}_j, \mathbf{v}_j), (\mathbf{u}_i, \mathbf{v}_i)\}$ is found to maximize the inner product of the respective canonical objects (e.g. $\mathcal{X} \times_k \mathbf{u}_k, \mathcal{Y} \times_k \mathbf{v}_k$) for the IJ, IK, JK joint-shared-modes respectively. Then, the overall process of TCCA can be written as the optimization problem of the canonical transformations Φ to maximize the inner product of the canonical tensors $\mathcal{X}', \mathcal{Y}'$ which are obtained from the three pairs of canonical objects by

$$\rho = \max_{\Phi} \langle \mathcal{X}', \mathcal{Y}' \rangle, \qquad \text{where} \quad (4)$$

$$(\mathcal{X}')_{ijk} = (\mathcal{X} \times_k \mathbf{u}_k)_{ij} (\mathcal{X} \times_j \mathbf{u}_j)_{ik} (\mathcal{X} \times_i \mathbf{u}_i)_{jk}$$
$$(\mathcal{Y}')_{ijk} = (\mathcal{Y} \times_k \mathbf{v}_k)_{ij} (\mathcal{Y} \times_j \mathbf{v}_j)_{ik} (\mathcal{Y} \times_i \mathbf{v}_i)_{jk}$$

and $\langle , \rangle$ denotes the inner product of tensors defined in Section 2.2. Note the mode product of the tensor by the single canonical transformation yields a matrix, a plane as the canonical object. Similar to classical CCA, multiple tensor canonical correlations $\rho_1, ..., \rho_d$ are defined by the orthogonal sets of the canonical directions.

Single-shared-mode TCCA. Similarly, the single-shared-mode tensor CCA is defined as the inner product of the canonical tensors comprising of the three canonical objects. The two pairs of the transformations in $\Psi = [\{(\mathbf{u}_j^1, \mathbf{v}_j^1), (\mathbf{u}_k^1, \mathbf{v}_k^1)\}, \{(\mathbf{u}_i^2, \mathbf{v}_i^2), (\mathbf{u}_k^2, \mathbf{v}_k^2)\}, \{(\mathbf{u}_i^3, \mathbf{v}_i^3), (\mathbf{u}_j^3, \mathbf{v}_j^3)\}]$ are found to maximize the inner product of the resulting canonical objects, by the mode product of the data tensors by the two pairs of the canonical transformations, for the I, J, K single-shared-modes. The tensor canonical correlations are

$$\rho = \max_{\Psi} \langle \mathcal{X}', \mathcal{Y}' \rangle, \qquad \text{where} \quad (5)$$

$$(\mathcal{X}')_{ijk} = (\mathcal{X} \times_j \mathbf{u}_j^1 \times_k \mathbf{u}_k^1)_i (\mathcal{X} \times_i \mathbf{u}_i^2 \times_k \mathbf{u}_k^2)_j (\mathcal{X} \times_i \mathbf{u}_i^3 \times_j \mathbf{u}_j^3)_k$$
$$(\mathcal{Y}')_{ijk} = (\mathcal{Y} \times_j \mathbf{v}_j^1 \times_k \mathbf{v}_k^1)_i (\mathcal{Y} \times_i \mathbf{v}_i^2 \times_k \mathbf{v}_k^2)_j (\mathcal{Y} \times_i \mathbf{v}_i^3 \times_j \mathbf{v}_j^3)_k$$

The canonical objects here are the vectors and the canonical tensors are given by the outer product of the three vectors.

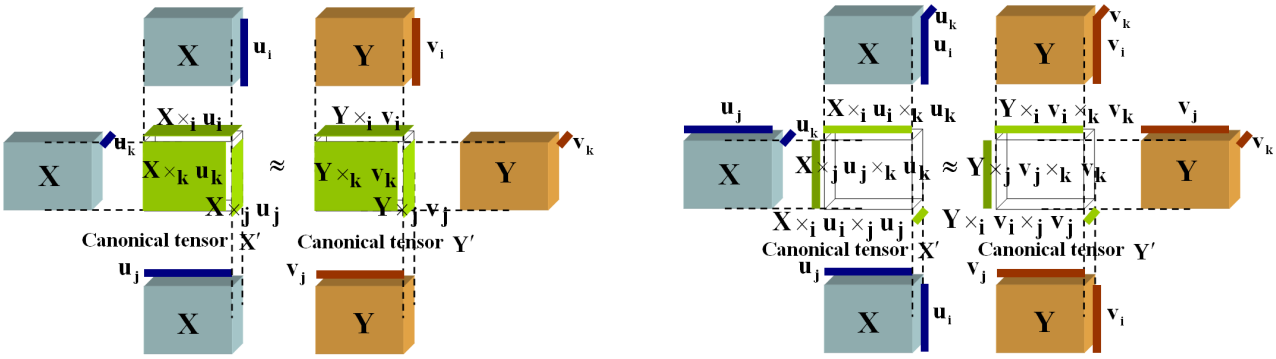

Figure 2. **Conceptual drawing of Tensor CCA.** *Joint-shared-mode TCCA (left) and single-shared-mode TCCA (right) of two video volumes (**X,Y**) are defined as the inner product of the canonical tensors (two middle cuboids in each figure), which are obtained by finding the respective pairs of canonical transformations (**u,v**) and canonical objects (green planes in left or lines in right figure).*

Interestingly, in the tasks of action/gesture classification, we have observed that the joint-shared-mode TCCA delivers more discriminative features than the single-shared-mode TCCA, maybe due to the good balance between the flexibility and the descriptive powers of the features in the joint-shared space. Generally the single-shared-mode has more flexible (by two pairs of free transformations) and less data-descriptive features in matching. The plane-like canonical objects in the joint-shared-mode seem to maintain sufficient discriminative information of action video data while giving robustness in matching. Note that only a single-shared-mode was considered in [12] (similarly to the proposed single-shared-mode TCCA). The previous results [16] also agree with this observation. The CCA applied to object recognition with image sets is identical to the IJ joint-shared-mode of the tensor CCA framework of this paper.

3.2. Alternating Solution

A solution for both types of TCCA is proposed in a so-called *divide-and-conquer* manner. Each independent process is associated with the respective canonical objects and canonical transformations and also yields the canonical correlation features as the inner products of the canonical objects. This is done by performing the SVD method for CCA [13] a single time (for the joint-shared-mode TCCA) or several times alternatively (for the single-shared-mode TCCA). This section is devoted to explain the solution for the I single-shared-mode for example. This involves the orthogonal sets of canonical directions $\{(\mathbf{U}_j, \mathbf{V}_j), (\mathbf{U}_k, \mathbf{V}_k)\}$ which contain $\{(\mathbf{u}_j, \mathbf{v}_j \in \mathbb{R}^J), (\mathbf{u}_k, \mathbf{v}_k \in \mathbb{R}^K)\}$ in their columns, yielding the d canonical correlations $(\rho_1, ...\rho_d)$ where $d < min(K, J)$ for given two data tensors, $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K}$. The solution is obtained by alternating the SVD method to maximize

$$\max_{\mathbf{U}_j, \mathbf{V}_j, \mathbf{U}_k, \mathbf{V}_k} \langle \mathcal{X} \times_j \mathbf{U}_j \times_k \mathbf{U}_k,\ \mathcal{Y} \times_j \mathbf{V}_j \times_k \mathbf{V}_k \rangle. \quad (6)$$

Given a random guess for $\mathbf{U}_j, \mathbf{V}_j$, the input tensors $\mathcal{X}, \mathcal{Y}$ are projected as $\widetilde{\mathcal{X}} = \mathcal{X} \times_j \mathbf{U}_j$, $\widetilde{\mathcal{Y}} = \mathcal{Y} \times_j \mathbf{V}_j$. Then, the best pair of $\mathbf{U}_k^*, \mathbf{V}_k^*$ which maximizes $\langle \widetilde{\mathcal{X}} \times_k \mathbf{U}_k, \widetilde{\mathcal{Y}} \times_k \mathbf{V}_k \rangle$ are found. Letting

$$\widetilde{\mathcal{X}} \leftarrow \widetilde{\mathcal{X}} \times_k \mathbf{U}_k^*, \quad \widetilde{\mathcal{Y}} \leftarrow \widetilde{\mathcal{Y}} \times_k \mathbf{V}_k^*, \quad (7)$$

then the pair of $\mathbf{U}_j^*, \mathbf{V}_j^*$ are found to maximize $\langle \widetilde{\mathcal{X}} \times_j \mathbf{U}_j, \widetilde{\mathcal{Y}} \times_j \mathbf{V}_j \rangle$. Let

$$\widetilde{\mathcal{X}} \leftarrow \widetilde{\mathcal{X}} \times_j \mathbf{U}_j^*, \quad \text{and} \quad \widetilde{\mathcal{Y}} \leftarrow \widetilde{\mathcal{Y}} \times_j \mathbf{V}_j^* \quad (8)$$

and repeat the procedures (7) and (8) until convergence. The solutions for the steps (7), (8) are obtained as follows:

SVD method for CCA [13] is embedded into the proposed alternating solution. First, the tensor-to-matrix and the matrix-to-tensor conversion is defined as

$$\mathcal{A} \in \mathbb{R}^{I \times J \times K} \quad \longleftrightarrow \quad \mathbf{A}_{(ij)} \in \mathbb{R}^{(IJ) \times K} \quad (9)$$

where $\mathbf{A}_{(ij)}$ is a matrix which has K column vectors in $\mathbb{R}^{I \times J}$ which are obtained by concatenating all elements of the IJ planes of the tensor $\mathcal{A}$. Let $\widetilde{\mathcal{X}} \rightarrow \widetilde{\mathbf{X}}_{(ij)}$ and $\widetilde{\mathcal{Y}} = \widetilde{\mathbf{Y}}_{(ij)}$ in (7). If $\mathbf{P}_{(ij)}^1, \mathbf{P}_{(ij)}^2$ denote two orthogonal basis matrices of $\widetilde{\mathbf{X}}_{(ij)}, \widetilde{\mathbf{Y}}_{(ij)}$ respectively, canonical correlations are obtained as singular values of $(\mathbf{P}^1)^T \mathbf{P}^2$ by

$$(\mathbf{P}^1)^T \mathbf{P}^2 = \mathbf{Q}_1 \mathbf{\Lambda} \mathbf{Q}_2^T, \quad \mathbf{\Lambda} = diag(\rho_1, ...\rho_K). \quad (10)$$

The solutions for the mode products in (7) are given as $\widetilde{\mathcal{X}} \times_k \mathbf{U}_k^* \leftarrow \mathbf{G}_{(ij)}^1$, $\widetilde{\mathcal{Y}} \times_k \mathbf{V}_k^* \leftarrow \mathbf{G}_{(ij)}^2$ accordingly where $\mathbf{G}_{(ij)}^1 = \mathbf{P}^1 \mathbf{Q}_1$, $\mathbf{G}_{(ij)}^2 = \mathbf{P}^2 \mathbf{Q}_2$. The solutions for (8) are similarly found by converting the tensors into the matrix representations s.t. $\widetilde{\mathcal{X}} \rightarrow \widetilde{\mathbf{X}}_{(ik)}$, $\widetilde{\mathcal{Y}} \rightarrow \widetilde{\mathbf{Y}}_{(ik)}$. When it converges, d canonical correlations are obtained from the first d correlations of either $(\rho_1, ...\rho_K)$ or $(\rho_1, ...\rho_J)$, where $d < min(K, J)$.

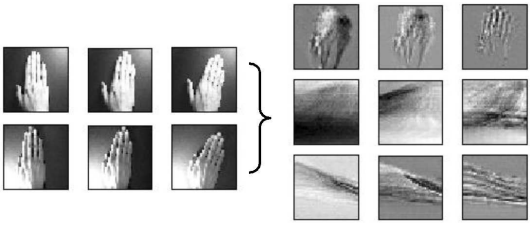

Figure 3. ***Example of Canonical Objects.*** *Given two sequences of the same hand gesture class (the left two rows), the first three canonical objects of the IJ, IK, JK joint-shared-mode are shown in the top, middle, bottom row respectively. The different canonical objects explains data similarity in different data dimensions.*

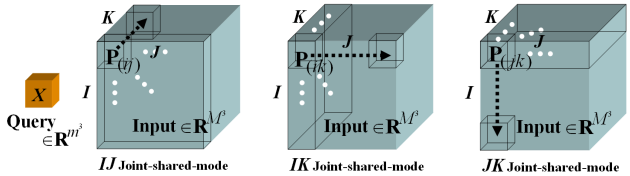

Figure 4. **Detection Scheme.** *A query video is searched in a large volume input video. TCCA between the query and every possible volume of the input video can be speeded-up by dynamically learning the three subspaces of all the volumes (cuboids) for the IJ, IK, JK joint-shared-mode TCCA. While moving the initial slices along one axis, subspaces of every small volume are dynamically computed from those of the initial slices.*

The J and K single-shared-mode TCCA are performed in the same alternating fashion, while the IJ, IK, JK joint-shared-mode TCCA by performing the SVD method a single time without iterations.

4. Discriminative Feature Selection for TCCA

By the proposed tensor CCA, we have obtained $6 \times d$ canonical correlation features in total. (Each of the joint-shared-mode and single-shared-mode has 3 different CCA processes and each CCA process yields d features). Intuitively, each feature delivers different data semantics in explaining the data similarity. For example in Figure 3, the canonical objects computed for the two hand gesture sequences of the same class are visualized. One of each pair of canonical objects is only shown here, as the other is very much alike. The canonical objects of the IJ joint-shared-mode show the common spatial components of the two given videos. The canonical transformations applied to the K axis (time axis) deliver the spatial component which is independent of temporal information, e.g. temporal ordering of the video frames. The different canonical objects of this mode seem to capture different spatial variations of the data. Similarly, the canonical objects of the IK, JK joint-shared-mode reveal the common components of the two videos in the joint space-time domain. Canonical correlations indicating the degree of the data correlation on each of the canonical components are used as similarity measures for recognition.

In general, each canonical correlation feature carries a different amount of discriminative information for video classification depending on applications. A discriminative feature selection scheme is proposed to select useful tensor canonical correlation features. First, the intra-class and inter-class feature sets (i.e. canonical correlations ρ_i, $i = 1, ..., 6 \times d$ computed from any pair of videos) are generated from the training data comprising of several class examples. We use each tensor CCA feature to build simple weak classifiers $\mathcal{M}(\rho_i) = sign\left[\rho_i - C\right]$ and aggregate the weak learners using the AdaBoost algorithm [14]. In an iterative update scheme classifier performance is optimized on the training data to yield the final strong classifier with the weights and the list of the selected features. Nearest Neighbor (NN) classification in terms of the sum of the canonical correlations chosen from the list is performed to categorize a new test video.

5. Action Detection by Tensor CCA

The proposed TCCA is time-efficient provided that actions or gestures are aligned in the space-time domain. However, searching non-aligned actions by TCCA in the three-dimensional (X,Y, and T) input space is computationally demanding because every possible position and scale of the input volume needs to be scanned. By observing that the joint-shared-mode TCCA does not require the iterations for the solutions and delivers sufficient discriminative power (See Table 1), time-efficient action detection can be done by sequentially applying joint-shared-mode TCCA followed by single-shared-mode TCCA. The joint-shared-mode TCCA can effectively filter out the majority of samples which are far from a query sample then the single-shared-mode TCCA is applied to only few candidates. In this section, we explain the method to further speed up the joint-shared-mode TCCA by incrementally learning the required subspaces based on the incremental PCA [15].

The computational complexity of the joint-shared-mode TCCA in (10) depends on the computation of orthogonal basis matrices $\mathbf{P}^1, \mathbf{P}^2$ and the Singular Value Decomposition (SVD) of $(\mathbf{P}^1)^T\mathbf{P}^2$. The total complexity trebles this computation for the IJ, IK, JK joint-shared-mode. From the theory of [13], the first few eigenvectors corresponding to most of the data energy, which are obtained by Principal Component Analysis, can be the orthogonal basis matrices. If $\mathbf{P}^1 \in \mathbb{R}^{N \times d}, \mathbf{P}^2 \in \mathbb{R}^{N \times d}$ where d is a usually small number, the complexity of the SVD of $(\mathbf{P}^1)^T\mathbf{P}^2$ taking $O(d^3)$ is relatively negligible. Given the respective three sets of eigenvectors of a query video, time-efficient scanning can be performed by incrementally learning the three sets of eigenvectors, the space-time subspaces

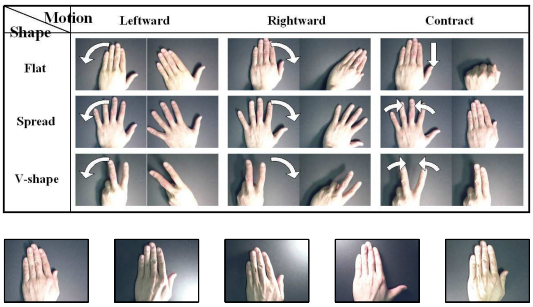

Figure 5. ***Hand-Gesture Database.*** *(top) 9 different gestures generated by 3 different shapes and 3 motions. (bottom) 5 different illumination conditions in the database.*

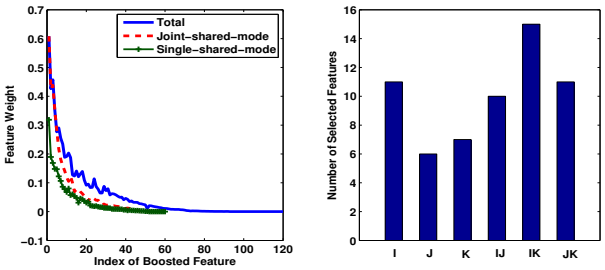

Figure 6. ***Feature Selection.*** *(left) The weights of TCCA features learnt by boosting. (right) The number of TCCA features chosen for the different shared-modes.*

$\mathbf{P}_{(ij)}, \mathbf{P}_{(ik)}, \mathbf{P}_{(jk)}$ of every possible volume (*cuboid*) of an input video for the IJ, IK, JK joint-shared-mode TCCA respectively. See Figure 4 for the concept. There are three separate steps which are carried out in same fashion, each of which is to compute one of $\mathbf{P}_{(ij)}, \mathbf{P}_{(ik)}, \mathbf{P}_{(jk)}$ of every possible volume of the input video. First, the subspaces of every cuboid of the initial slices of the input video are learnt, then the subspaces of all remaining cuboids are incrementally computed while moving the slices along one of the axes. For example, for the IJ joint-shared-mode TCCA, the subspaces $\mathbf{P}_{(ij)}$ of all cuboids in the initial IJ-slice of the input video are computed. Then, the subspaces of all next cuboids are dynamically computed from the previous subspaces, while pushing the initial cuboids along the K axis to the end as follows (for simplicity, let the size of the query video and input video be $\mathbb{R}^{m^3}, \mathbb{R}^{M^3}$ where $M \gg m$) :

The cuboid at k on the K axis, $\mathcal{X}^k$ is represented as the matrix $\mathbf{X}^k_{(ij)} = \{\mathbf{x}^k_{(ij)}, ..., \mathbf{x}^{k+m-1}_{(ij)}\}$ (See the definition (9)). The scatter matrix $\mathbf{S}^k = (\mathbf{X}^k_{(ij)})(\mathbf{X}^k_{(ij)})^T$ is written w.r.t. the scatter matrix of the previous cuboid at $k-1$ as $\mathbf{S}^k = \mathbf{S}^{k-1} + (\mathbf{x}^{k+m-1}_{(ij)})(\mathbf{x}^{k+m-1}_{(ij)})^T - (\mathbf{x}^{k-1}_{(ij)})(\mathbf{x}^{k-1}_{(ij)})^T$. This involves both incremental and decremental learning. A new vector $\mathbf{x}^{k+m-1}_{(ij)}$ is added and an existing vector $\mathbf{x}^{k-1}_{(ij)}$ is removed from the $(k-1)$-th cuboid. Based on

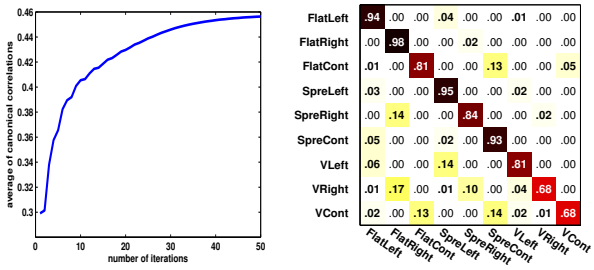

Figure 7. *(left) Convergence graph of the alternating solution for TCCA. (right) Confusion matrix of hand gesture recognition.*

	Joint-mode				Dual-mode
Number of features	01	05	20	60	60
Accuracy (%)	52	72	76	76	81

Table 1. ***Accuracy Comparison*** *of the joint-shared-mode TCCA and dual-mode TCCA (using both joint and single-shared mode).*

the previous study on incremental PCA [15], the sufficient spanning set $\Upsilon = h([\mathbf{P}^{k-1}_{(ij)}, \mathbf{x}^{k+m-1}_{(ij)}])$, where h is a vector orthogonalization function and $\mathbf{P}^{k-1}_{(ij)}$ is the IJ subspace of the previous cuboid, can be efficiently exploited to compute the eigenvectors of the current scatter matrix, $\mathbf{P}^k_{(ij)}$. For the detailed computations, refer to [15].

Similarly, the subspaces $\mathbf{P}_{(ik)}, \mathbf{P}_{(jk)}$ for the IK, JK joint-shared-mode TCCA are computed by moving the all cuboids of the slices along the I, J axes respectively. By this way, the total complexity of learning of the three kinds of the subspaces of every cuboid is significantly reduced from $O(M^3 \times m^3)$ to $O(M^2 \times m^3 + M^3 \times d^3)$ as $M \gg m \gg d$. $O(m^3), O(d^3)$ are the complexity for solving eigen-problems in batch-mode and the proposed dynamic way. Efficient multi-scale search is similarly plausible by merging two or more cuboids.

6. Experimental Results

Hand-Gesture Recognition. We acquired *Cambridge-Gesture data base* consisting of 900 image sequences of 9 hand gesture classes, which are defined by 3 primitive hand shapes and 3 primitive motions (see Figure 5). Each class contains 100 image sequences (5 different illuminations$\times$10 arbitrary motions of 2 subjects). Each sequence was recorded in front of a fixed camera having roughly isolated gestures in space and time. All video sequences were uniformly resized into $20 \times 20 \times 20$ in our method. All training was performed on the data acquired in the single plain illumination setting (leftmost in Figure 5) while testing was done on the data acquired in the remaining settings.

The proposed alternating solution in Section 3.2 was performed to obtain the TCCA features of every pair of the

Methods	set1	set2	set3	set4	total
Our method	81	81	78	86	82±3.5
Niebles et al. [3]	70	57	68	71	66±6.1
Wong et al. [8]	-	-	-	-	44

Table 2. ***Hand-gesture recognition accuracy*** *(%) of the four different illumination sets.*

training sequences. The alternating solution stably converged as shown in the left of Figure 7. Feature selection was performed for the TCCA features based on the weights and the list of the features learnt from the AdaBoost method in Section 4. In the left of Figure 6, it is shown that about the first 60 features contained most of the discriminative information. Of the first 60 features, the number of the selected features is shown for the different shared-mode TCCA in the right of Figure 6. The joint-shared-mode (IJ, IK, JK) contributed more than the single-shared-mode (I, J, K) but both still kept many features in the selected feature set. From Table 1, the best accuracy of the joint-shared-mode was obtained by 20 - 60 features. This is easily reasoned when looking at the weight curve of the joint-shared-mode in Figure 6 where the weights of more than 20 features are non-significant. The dual-mode TCCA (using both joint and single-shared mode) with the same number of features improved the accuracy of the joint-shared mode by 5%. NN classification was performed for a new test sequence based on the selected TCCA features. Note that the performance of TCCA without any feature selection also delivered the best accuracy as shown at 60 features in the Table 1.

Table 2 shows the recognition rates of the proposed TCCA, Niebles et al.'s method [3], which exhibited the best action recognition accuracy among the state-of-the-arts in [3]), and Wong et al.'s method (Relevance Vector Machine (RVM) with the motion gradient orientation images [8]). The original codes and the best settings of the parameters were used in the evaluation for the two previous works. As shown in Table 2, the previous two methods yielded much poorer accuracy than our method. They often failed to identify the sequences of similar motion classes having different hand shapes, as they cannot explain the complex shape variations of those classes. Large intra-class variation in spatial alignment of the gesture sequences also caused the performance degradation, particularly for Wong et al.'s method which is based on global space-time volume analysis. Despite the rough alignment of the gestures, the proposed method is significantly superior to the previous methods by considering both spatial and temporal information of the gesture classes effectively. See Figure 7 for the confusion matrix of our method.

Action Categorization on KTH Data Set. We followed the experimental protocol of Niebles et al.'s work [3] on the KTH action data set, which is the largest public action

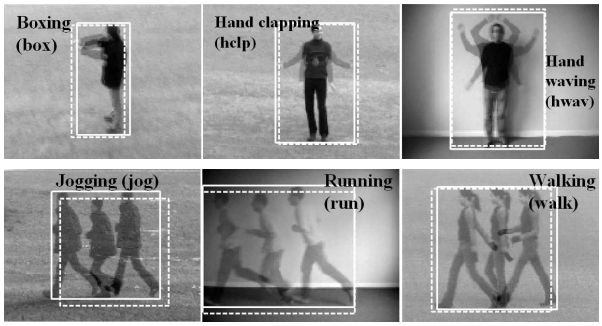

Figure 8. **Example videos of KTH data set.** *The bounding boxes (solid box for the manual setting, the dashed one for the automatic detection) indicate the spatial alignment and the superimposed images of the initial, intermediate and the last frames of each action show the temporal segmentation.*

Methods	(%)	Methods	(%)
Our method	95.33	Schuldt et al. [6]	71.72
Niebles et al. [3]	81.50	Ke et al. [7]	62.96
Dollar et al. [5]	81.17		

Table 3. ***Recognition accuracy (%) on the KTH action data set.***

data base [6]. The data set contains six types (boxing, hand clapping, hand waving, jogging, running and walking) of human actions performed by 25 subjects in 4 different scenarios. Leave-one-out cross-validation was performed to test the proposed method, i.e. for each run the videos of 24 subjects are exploited for training and the videos of the remaining subject is for testing. Some sample videos are shown in Figure 8 with the indication of the action alignment. In TCCA method, the aligned video sequences were uniformly resized to $20 \times 20 \times 20$. This space-time alignment of actions was manually done for accuracy comparison but can also be automatically achieved by the proposed detection scheme. See Table 3 for the accuracy comparison of several methods and Figure 9 for the confusion matrix of our method. The competing methods are based on histogram representations of the local space-time interest points with SVM (Dollar et al [5], Schuldt et al. [6]) or pLSA (Niebles et al. [3]). Ke et al. applied the spatio-temporal volumetric features [7]. While the previous methods delivered the accuracy around 60-80%, the proposed method achieved impressive accuracy at 95%. The previous methods lost important information in the global space-time shapes of actions resulting in ambiguity for more complex spatial variations of the action classes.

Action Detection on KTH Data Set. The action detection was performed by the training set consisting of the sequences of the five persons, which do not contain any testing persons. The scale (also the aspect ratio of axes) of actions were class-wise fixed. Figure 8 shows the proposed detection results by the dashed bounding boxes, which are

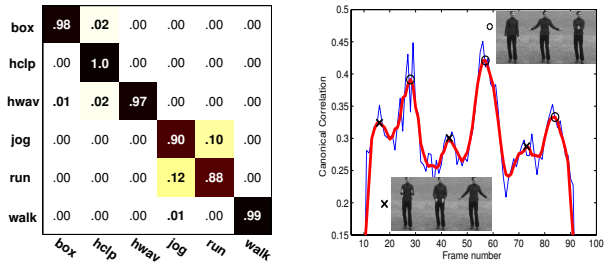

Figure 9. *(left) Confusion matrix of our method for the KTH data set. (right) The detection result for the input video which involves continuous hand clapping actions: all three correct hand clapping actions are detected at the highest three peaks, with the three intermediate actions at the three lower peaks.*

close to the manually setting (solid ones). The right of Figure 9 shows the detection results for the continuous hand clapping video, which comprises of the three correct unit clapping actions defined. The maximum canonical correlation value is shown for every frame of the input video. All three correct hand clapping actions are detected at the three highest peaks, with the three intermediate actions at the three lower peaks. The intermediate actions which exhibited local maxima between any two correct hand-clapping actions had different initial and end postures from those of the correct actions.

The detection speed differs for the size of input volume with respect to the size of query volume. The proposed detection method required about 136 seconds on average for the boxing and hand clapping action classes and about 19 seconds on average for the other four action classes on a Pentium 4 3GHz computer running non-optimized Matlab code. For example, the volume sizes of the input video and the query video for the hand clapping actions are $120 \times 160 \times 102$ and $92 \times 64 \times 19$ respectively. The dimension of the input video and query video was reduced by the factors $4.6, 3.2, 1$ (for the respective three dimensions). The obtained speed seems to be comparable to that of the state-of-the-art [1] and fast enough to be integrated into a real-time system if provided with a smaller search area either by manual selection or by some pre-processing techniques for finding the *focus of attention*, e.g. by moving area segmentation.

7. Conclusions

We proposed a novel Tensor Canonical Correlation Analysis (CCA) which can extract flexible and descriptive correlation features of two videos in the joint space-time domain. The proposed statistical framework yields a compact set of pair-wise features. The proposed features combined with the feature selection method and a NN classifier significantly improves the accuracy over current state-of-the-art action recognition methods. Additionally, the proposed detection scheme for Tensor CCA could yield time-efficient action detection or alignment in a larger volume input video.

Currently experiments on simultaneous detection and classification of multiple actions by TCCA are being carried out. Efficient multi-scale search by merging the space-time subspaces and will also be considered.

References

[1] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR*, 2005.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *CVPR*, 2005.

[3] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words, In *BMVC*, 2006.

[4] A. Bobick and J. Davis. The recognition of human movements using temporal templates. *PAMI*, 23(3):257–267, 2001.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[6] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.

[7] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[8] S-F. Wong and R. Cipolla. Real-time interpretation of hand motions using a sparse Bayesian classifier on motion gradient orientation images. In *BMVC*, 2005.

[9] F.R. Bach and M.I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. TR 688, University of California, Berkeley, 2005.

[10] M.A.O. Vasilescu and D. Terzopoulos. Multilinear Analysis of Image Ensembles: TensorFaces. In *ECCV*, 2002.

[11] D. Hardoon, S. Szedmak and J.S. Taylor Canonical correlation analysis; An overview with application to learning methods *Neural Computation*, 16(12):639–2664, 2004.

[12] R. Harshman. Generalization of Canonical Correlation to N-way Arrays. Poster at *Thirty-fourth Annual Meeting of the Statistical Society of Canada*, May 2006.

[13] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[14] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *2nd European Conference on Computational Learning Theory*, 1995.

[15] P. Hall, D. Marshall, and R. Martin. Merging and splitting eigenspace models. *PAMI*, 2000.

[16] T-K. Kim, J. Kittler, and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. *IEEE Trans. on PAMI*, Vol.29, No.6, 2007.