Tensor Canonical Correlation Analysis for Action Classification

Tae-Kyun Kim, Shu-Fai Wong and Roberto Cipolla

Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK.

This was accepted to appear in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, Minnesota, US, June 2007. This can be best viewed in color printing or electronic form.
Tensor Canonical Correlation Analysis for Action Classification

Tae-Kyun Kim, Shu-Fai Wong and Roberto Cipolla

Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK.

Abstract—This paper introduces a new statistical framework, namely Tensor Canonical Correlation Analysis (TCCA) which is an extension of classical Canonical Correlation Analysis (CCA) to multidimensional data arrays also called tensors. We apply this method to inspect multidimensional (joint space-time) linear relationships of two videos for human action and gesture classification. Tensor CCA avoids the difficult problem of explicit motion estimation and delivers maximally discriminative information conveyed in videos by taking into account the joint space-time domain of the video data. It is known that spatial information can be as important as temporal information for human action and gesture classification, although many methods have neglected the spatial domain. Tensor CCA provides a learnable set of flexible and descriptive similarity features of any two videos in the joint space-time domain. In this study the usefulness of our novel TCCA features has been examined in terms of action classification accuracy by combination with feature selection and a Nearest Neighbor classifier. In addition, we propose a time-efficient action detection and alignment method based on dynamic learning of subspaces for Tensor CCA for the case that actions are not aligned in the input space-time domain. The proposed method delivers significantly better accuracy at comparable detection speed over current state-of-the-art action recognition methods on the KTH action data set as well as on self-recorded hand gesture data sets.

Keywords—Action Classification, Gesture Recognition, Canonical Correlation Analysis, Tensor, Action Detection, Incremental Subspace Learning

1 Introduction

Video data is playing an increasingly important role in our lives owing to widespread digital capturing and storage devices. To handle the rapidly growing amount of video data, there is great demand for automatic video indexing and categorization. Automatic classification and localization of human actions and gestures as
key semantics of video data are very useful for various tasks such as video surveillance, object-level video summarization, and video retrieval.

Many methods for action categorization have been suggested in the past. Traditional approaches are based on the comparison of motion data requiring explicit motion estimation [4, 5]. The performance of such algorithms is highly dependent on the quality of the motion estimation, which is a hard problem in practice due to smooth surfaces, singularities, self-occlusions, appearance changes, and the aperture problem.

Some recent work has analyzed human actions directly in the space-time volume without explicit motion estimation [1, 6, 11, 10]. Motion history images and the space-time local gradients are used to represent video data in [6, 11] and [1] respectively, having the benefits of being able to analyze quite complex and low-resolution dynamic scenes. However, both representations only partially convey space-time information (mainly the motion data) and are unreliable in cases of motion discontinuities and motion aliasing. Additionally, the method in [1] involves manual setting up some important parameters such as the positions and scales 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 are insufficient 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, 9, 8, 7]. The local variations around the interest points are quantized by a code book then the histogram representations are combined with either a Support Vector Machine (SVM) [9, 8] or a probabilistic generative model [3]. Although they have yielded good accuracy to a certain degree, mainly due to the high discrimination power of individual local space-time descriptors, they exhibit ambiguity by ignoring global space-time shape information of action classes. Importantly, their performance is sensitive to the parameters of the space-time interest points and the code book, whose best setting is application or data dependent.

In this study, a novel statistical framework of extracting similarity features of two videos is proposed for human action/gesture categorization. We extend the classical canonical correlation analysis [13, 19] (see Section 2.1) - a standard tool for inspecting linear relationships between two sets of vectors- into that of multi-dimensional data arrays (also called high-order tensors) for analyzing the similarity of video data/space-time volumes. The proposed framework itself is general and may be applied to other tasks requiring matching of multi-dimensional data arrays (e.g. a single color image [15] and filter banks applied to a single image [16] can also yield a high-order tensor).

This work was motivated by our previous success [12], 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, containing changes in object appearance due to different lighting and pose. Each image in the two sets is vectorized and classical CCA applied to the two sets of vectors. Object recognition is performed based on canonical correlations, also called principal angles, where higher canonical correlations indicate higher similarity of two given image sets. The canonical correlation based method yielded much higher object recognition rates than the traditional set-similarity measures based on pdfs e.g. Kullback Leibler-Divergence (KLD) in [12]. 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 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 [12]. There are ample studies demonstrating the advantages of the subspace notion in visual recognition.

Despite its success CCA is still insufficient for action/gesture classification tasks by simply representing a video as a set of images. CCA on these sets of images does not encode any temporal ordering information. We propose the idea of Tensor Canonical Correlation Analysis (TCCA), and its novel application to action classification in this paper. The new canonical correlation features have many favorable characteristics:

- CCA of videos yields a novel set of similarity features reflecting the joint spatial and temporal domains in video.
- The new features are flexible up to affine transformation for possible data variations in each domain.
- Action is analyzed as global space-time volumes, avoiding the challenging problems of explicit motion estimation.
- The proposed TCCA does not involve any tuning parameters.
- The Tensor CCA framework can be partitioned into sub-CCAs as each canonical correlation explains different aspects of the multi-dimensional data arrays. For example, previous work on object recognition [12, 22, 21] based on image sets can be seen as a sub-problem of this framework. Furthermore, comparison of a single image to a set of images, or comparison of a single to a single image can be also done within the tensor 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 proposed method significantly outperformed the state-of-the-art action classification methods in accuracy on the KTH data set [9] as well as our own hand-gesture data set. Also, the proposed detection method could yield economical computations of TCCA for the case when the action is not aligned in the space-time domain, delivering comparable detection speed to the state-of-the-art method in [1].

The rest of the chapter is organized as follows: Backgrounds and notations are given in Section 2 and the framework and the solution for tensor CCA are given in Section 3. Section 4 and 5 are devoted to the proposed 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) [23], Canonical Correlation Analysis (CCA) has been a standard tool for inspecting linear relationships between two random variables (or two sets of vectors) [19]. Given two random vectors \( x \in \mathbb{R}^{m_1}, y \in \mathbb{R}^{m_2} \), a pair of transformations \( u, v \), called canonical transformations, is found to maximize the correlation of the two vectors \( x' = u^T x \), \( y' = v^T y \) as

\[
\rho = \max_{u, v} \frac{E[x'y'^T]}{\sqrt{E[x'x'^T]E[y'y'^T]}} = \frac{u^T C_{xy} v}{\sqrt{u^T C_{xx} uu^T C_{yy} v}}
\]

where \( \rho \) is called the canonical correlation and multiple canonical correlations \( \rho_1, \ldots, \rho_d \) where \( d < \min(m_1, m_2) \) are defined by the next pairs of \( u, v \) which are orthogonal to the previous ones. A probabilistic version of CCA, which was recently developed in [13], gives another viewpoint. As shown in Figure 1, the model reveals how well two random variables \( x, y \) are represented by a common source (latent) variable \( z \in \mathbb{R}^d \) with the two likelihoods \( p(x|z), p(y|z) \), which comprises affine transformations w.r.t. the input variables \( x, y \) respectively. The maximum likelihood estimation on this model leads to the canonical transformations \( U = [u_1, \ldots, u_d], V = [v_1, \ldots, v_d] \) and the associated canonical correlations \( \rho_1, \ldots, \rho_d \), which are equivalent to those of the standard CCA. See [13] 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
Probabilistic Canonical Correlation Analysis tells how well two random variables $x, y$ are represented by a common source variable $z$.

Vectors is its affine invariance in matching, which gives robustness with respect to intra-class data variations as discussed in introduction. That is, canonical correlations are invariant to affine transformations w.r.t. inputs, i.e. $Ax + b, Cy + d$ for arbitrary $A \in \mathbb{R}^{m_1 \times m_1}, b \in \mathbb{R}^{m_1}, C \in \mathbb{R}^{m_2 \times m_2}, d \in \mathbb{R}^{m_2}$. This proof is straightforward from (1) as $C_{xy}, C_{xx}, C_{yy}$ are covariance matrices and are multiplied by arbitrary transformations $u, v$.

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

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

This matrix notation of CCA will be exploited to explain the proposed tensor CCA with the tensor notations in the following section.

### 2.2 Multilinear Algebra and Notations

This section briefly introduces some notation and concepts of multilinear algebra [14, 18], or higher-order tensors for the proposed tensor CCA. A third-order tensor which has the three modes of dimensions $I, J, K$ is denoted by $A = (A)_{ijk} \in \mathbb{R}^{I \times J \times K}$. The inner product of any two tensors is defined as $\langle A, B \rangle = \sum_{i,j,k} (A)_{ijk} (B)_{ijk}$. The $mode$-$j$ vectors are the column vectors of matrix $A_{(j)} \in \mathbb{R}^{J \times (IK)}$ and the $j$-mode
product of a tensor $\mathcal{A}$ by a matrix $U \in \mathbb{R}^{J \times N}$ is

$$(\mathcal{B})_{ink} \in \mathbb{R}^{I \times N \times K} = (\mathcal{A} \times_j U)_{ink} = \sum_j (\mathcal{A})_{ijk} u_{jn}$$

(3)

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

3 Tensor Canonical Correlation Analysis

3.1 Joint and Single-shared-mode TCCA

Many previous studies [14, 15, 16, 17] 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). The ‘axes’, which correspond to ‘modes’ in tensor theory, will be used together.

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 $u, v$ 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. $(u_k, v_k)$) in $\Phi = \{ (u_k, v_k), (u_j, v_j), (u_i, v_i) \}$ is found to maximize the inner product of the respective canonical objects (e.g. $\mathcal{X} \times_k u_k, \mathcal{Y} \times_k v_k$) for the $IJ, IK, JK$ joint-shared-modes respectively. Then, the overall process of TCCA can be written as the optimiza-
Figure 2: Conceptual drawing of Tensor CCA. Joint-shared-mode TCCA (top) and single-shared-mode TCCA (bottom) 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 top or lines in bottom figure).

The problem of the canonical transformations \(\Phi\) to maximize the inner product of the canonical tensors \(X',Y'\) which are obtained from the three pairs of canonical objects by

\[
\rho = \max_{\Phi} \langle X', Y' \rangle, \quad \text{where}
\]

\[
(X')_{ijk} = (X \times_k u_k)_{ij}(X \times_j u_j)_{ik}(X \times_i u_i)_{jk}
\]

\[
(Y')_{ijk} = (Y \times_k v_k)_{ij}(Y \times_j v_j)_{ik}(Y \times_i v_i)_{jk}
\]

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

$$\Psi = \{((u_j^1, v_j^1), (u_k^1, v_k^1)), ((u_i^2, v_i^2), (u_k^2, v_k^2)), ((u_i^3, v_i^3), (u_j^3, v_j^3))\}$$

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

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

$$\mathcal{X}_{ijk} = (X \times_j u_j^1 \times_k u_k^1)(X \times_i u_i^2 \times_k u_k^2)(X \times_i u_i^3 \times_j u_j^3)$$

$$\mathcal{Y}_{ijk} = (Y \times_j v_j^1 \times_k v_k^1)(Y \times_i v_i^2 \times_k v_k^2)(Y \times_i v_i^3 \times_j v_j^3)$$

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

Note that both joint-shared-mode and single-shared-mode TCCA are a natural generalization of the standard CCA to third-order tensors. At this point, it may be worthwhile to compare the proposed two types of TCCA. The comparison is given with regard to balance between flexibility and descriptive powers of the canonical correlation features. Generally the single-shared-mode has more flexible (by two pairs of free transformations) and less data-descriptive features in matching. The question of which type is better may be dependent on the application. Importantly, in the tasks of action/gesture classification, we have observed that the joint-shared-mode TCCA delivers generally more discriminative features than the single-shared-mode TCCA (see Section 6). 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. The previous results [12] also agree with this observation. The CCA applied to object recognition with image sets [12] is identical to the IJ joint-shared-mode of the tensor CCA framework of this paper. A novel alternating solution for our tensor CCA is given in the following section.

### 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 fea-
tures as the inner products of the canonical objects. This is done by performing the SVD method for CCA [24] 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 \( \{(U_j, V_j), (U_k, V_k)\} \) which contain \( \{(u_j, v_j \in \mathbb{R}^J), (u_k, v_k \in \mathbb{R}^K)\} \) in their columns, yielding the \( d \) canonical correlations \( (\rho_1, \ldots, \rho_d) \) where \( d < \min(K, J) \) for given two data tensors, \( X, Y \in \mathbb{R}^{I \times J \times K} \). The solution is obtained by alternating the SVD method to maximize

\[
\max_{U_j, V_j, U_k, V_k} \langle X \times_j U_j \times_k U_k, Y \times_j V_j \times_k V_k \rangle.
\]

Given a random guess for \( U_j, V_j \), the input tensors \( X, Y \) are projected as \( \tilde{X} = X \times_j U_j, \tilde{Y} = Y \times_j V_j \). Then, the best pair of \( U_k^*, V_k^* \) which maximizes \( \langle \tilde{X} \times_k U_k, \tilde{Y} \times_k V_k \rangle \) are found. Letting

\[
\tilde{X} \leftarrow \tilde{X} \times_k U_k^*, \quad \tilde{Y} \leftarrow \tilde{Y} \times_k V_k^*,
\]

then the pair of \( U_j^*, V_j^* \) are found to maximize \( \langle \tilde{X} \times_j U_j, \tilde{Y} \times_j V_j \rangle \). Let

\[
\tilde{X} \leftarrow \tilde{X} \times_j U_j^*, \quad \tilde{Y} \leftarrow \tilde{Y} \times_j V_j^*.
\]

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

**SVD method for CCA** [24] is embedded into the proposed alternating solution. First, the tensor-to-matrix and the matrix-to-tensor conversion is defined as

\[
A \in \mathbb{R}^{I \times J \times K} \leftrightarrow A_{(ij)} \in \mathbb{R}^{(I,J) \times K}
\]

where \( A_{(ij)} \) 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 \( A \). Let \( \tilde{X} \rightarrow \tilde{X}_{(ij)} \) and \( \tilde{Y} \rightarrow \tilde{Y}_{(ij)} \) in (7). If \( P_{(ij)}^1, P_{(ij)}^2 \) denote two orthogonal basis matrices of \( \tilde{X}_{(ij)}, \tilde{Y}_{(ij)} \) respectively, canonical correlations are obtained as singular values of \( (P^1)^T P^2 \) by

\[
(P^1)^T P^2 = Q_1 \Lambda Q_2^T, \quad \Lambda = diag(\rho_1, \ldots, \rho_K).
\]

The solutions for the mode products in (7) are accordingly given as \( \tilde{X} \times_k U_k^* \leftarrow G_{(ij)}^1, \tilde{Y} \times_k V_k^* \leftarrow G_{(ij)}^2 \) where \( G_{(ij)}^1 = P^1 Q_1, G_{(ij)}^2 = P^2 Q_2 \). The solutions for (8) are similarly found by converting the tensors into the matrix representations s.t. \( \tilde{X} \rightarrow \tilde{X}_{(ik)}, \tilde{Y} \rightarrow \tilde{Y}_{(ik)} \). When it converges, \( d \) canonical correlations are obtained from the first \( d \) correlations of either \( (\rho_1, \ldots, \rho_K) \) or \( (\rho_1, \ldots, \rho_J) \), where \( d < \min(K, J) \).
Discriminative Feature Selection for TCCA

Figure 3: Example of Canonical Objects. Given two different lighting 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.

The canonical transformations, for e.g. in (7), are also obtained by

$$U_k^* = (\tilde{X}^T_{(ij)} \tilde{X}_{(ij)})^{-1} \tilde{X}^T_{(ij)} P^1 Q_1$$

$$V_k^* = (\tilde{Y}^T_{(ij)} \tilde{Y}_{(ij)})^{-1} \tilde{Y}^T_{(ij)} P^2 Q_2$$

All other component processes of TCCA can be similarly carried out, delivering the $6 \times d$ canonical correlation features in total. 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 cor-
relation on each of the canonical components are used as similarity measures for recognition.

In general, each canonical correlation feature carries a different amount of discrimination 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 $\rho_i, i = 1, ..., 6 \times d$ 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 $M(\rho_i) = \text{sign} [\rho_i - C]$ and aggregate the weak learners using the AdaBoost algorithm [25]. In an iterative update scheme classifier performance is optimized on the training data to yield the final strong classifier by

$$M(\rho) = \text{sign} \left[ \sum_{i=1}^{M} w_{L(i)} M(\rho_{L(i)}) - \frac{1}{2} \sum_{i=1}^{M} w_{L(i)} \right]$$

(11)

where $w$ contains the weights and $L$ the list of the selected features. The list of the features learnt by Adaboost is exploited to select the total number of features to use for classification. Nearest Neighbor (NN) classification in terms of the sum of the canonical correlations chosen is performed to categorize a new test video.

Provided a sufficient training data set reflecting various changes of action classes, the discriminative transformation [12] may be an alternative discriminative learning of canonical correlation features for TCCA.

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 still computationally demanding because every possible position and scale of the input space 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 mainly explain the method to further speed up the joint-shared-mode TCCA for action detection by incrementally learning the required subspaces. The following section gives a brief introduction of prior work on incremental Principal Component Analysis [26, 27].
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.

5.1 Review on Incremental Principal Component Analysis

An efficient update scheme of eigensubspaces has been developed when a new set of vectors is added to an existing data set. This is useful for applications such as object tracking and surveillance where training images are accumulated over time. Given two sets of data represented by eigenspace models $\{\mu_i, M_i, P_i, \Lambda_i\}_{i=1,2}$, where $\mu_i$ is the mean, $M_i$ the number of samples, $P_i$ the matrix of eigenvectors and $\Lambda_i$ the eigenvalue matrix of the $i$-th data set, the combined eigenspace model $\{\mu_3, M_3, P_3, \Lambda_3\}$ is computed. The eigenvector matrix $P_3$ can be represented by

$$P_3 = \Phi R = h([P_1, P_2, \mu_1 - \mu_2])R,$$

where $\Phi$ is the orthonormal column matrix spanning the entire combined data space, $R$ is a rotation matrix, and $h$ is a vector orthonormalization function. Using this representation, an original eigenproblem for $P_3$, $\Lambda_3$ is converted into a smaller eigenproblem as

$$S_{T,3} = P_3 \Lambda_3 P_3^T \Rightarrow \Phi^T S_{T,3} \Phi = R \Lambda_3 R.$$

Note the matrix $\Phi^T S_{T,3} \Phi$ has the reduced size $d_{T,1} + d_{T,2} + 1$, where $d_{T,1}, d_{T,2}$ are the number of the eigenvectors in $P_1$ and $P_2$ respectively. Thus the eigenanalysis here only takes $O((d_{T,1} + d_{T,2} + 1)^3)$ computations, whereas the eigenanalysis in the l.h.s. (13) requires $O(\min(N, M_3)^3)$, where $N$ is the dimension of the input data and $M_3$ is the total number of data points. Usually, $N, M_3 \gg d_{T,1} + d_{T,2} + 1$. 

13
5.2 Dynamic Subspace Learning for TCCA

The computational complexity of the joint-shared-mode TCCA in (10) depends on the computation of orthogonal basis matrices $P^1, P^2$ and the Singular Value Decomposition (SVD) of $(P^1)^TP^2$. The total complexity trebles this computation for the $I,J,IK,JK$ joint-shared-mode. From the theory of [24], 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 $P^1 \in \mathbb{R}^{N \times d}, P^2 \in \mathbb{R}^{N \times d}$ where $d$ is a usually small number, the complexity of the SVD of $(P^1)^TP^2$ taking $O(d^3)$ is relatively negligible. Given the respective three sets of eigenvectors of a query video, time-efficient detection can be performed by incrementally learning the three sets of eigenvectors, the space-time subspaces $P_{(ij)}, P_{(ik)}, P_{(jk)}$ of every possible volume (cuboid) of an input video for the $I,J,IK,JK$ 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 $P_{(ij)}, P_{(ik)}, P_{(jk)}$ 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 $I,J$ joint-shared-mode TCCA, the subspaces $P_{(ij)}$ of all cuboids in the initial $I,J$-slice of the input video are computed. Then, the subspaces of all next cuboids are dynamically computed, 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, $X^k$ is represented by the matrix $X_{(ij)}^k = \{x_{(ij)}^k, \ldots, x_{(ij)}^{k+m-1}\}$, where $X_{(ij)}^k$ is obtained by the tensor-to-matrix conversion defined in (9). The scatter matrix $S^k = (X_{(ij)}^k)(X_{(ij)}^k)^T$ is written w.r.t. the scatter matrix of the previous cuboid at $k-1$ as

$$S^k = S^{k-1} + (x_{(ij)}^{k+m-1})(x_{(ij)}^{k+m-1})^T - (x_{(ij)}^{k-1})(x_{(ij)}^{k-1})^T. \quad (14)$$

This involves both incremental and decremental learning. A new vector $x_{(ij)}^{k+m-1}$ is added and an existing vector $x_{(ij)}^{k-1}$ is removed from the $(k-1)$-th cuboid. The sufficient spanning set [26, 27] of the current scatter matrix can be $\Upsilon = h([P_{(ij)}^{k-1}, x_{(ij)}^{k+m-1}])$ where $h$ is a vector orthogonalization function and $P_{(ij)}^{k-1}$ is the $I,J$ subspace of the previous cuboid. The eigenvectors of the current scatter matrix can be the product of the sufficient spanning set by an arbitrary rotation matrix $R$ as $P_{(ij)}^{k} = \Upsilon R$. Therefore the original eigen-problem to solve is reduced to the much smaller eigen-problem by

$$S^k = P_{(ij)}^k \Lambda^k (P_{(ij)}^k)^T \Rightarrow \Upsilon^T S^k \Upsilon = R \Lambda^k R. \quad (15)$$

\[1\]The sufficient spanning set denotes an economical set of bases which can span most data energy, which helps to obtain a small eigen-problem to solve.
6 Experimental Results

Hand-Gesture Recognition. We acquired Cambridge-Gesture database\(^2\) consisting of 900 image sequences of 9 hand gesture classes, which are defined by 3 primitive shapes and motions. The matrices \(\Lambda^k, R\) are computed as the eigenvalue and eigenvector matrix of \(\Upsilon^T S_k^k \Upsilon\). The final eigenvectors are obtained as \(P^{k}_{(ij)} = \Upsilon R\) after removing the components in \(R\) corresponding to the least eigenvalues in \(\Lambda^k\), keeping the dimension of \(P^{k}_{(ij)}\) be \(R^{m^2 \times d^3}\).

Similarly, the subspaces \(P_{(ik)}, P_{(jk)}\) 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 a batch (i.e. the l.h.s. of (15)) and the proposed way (the r.h.s. of (15)). Provided the subspaces of every cuboid of a unit scale, efficient multi-scale search is also plausible by merging two or more subspaces also based on incremental subspace learning theory. The details will be available in our follow-up technical report.

\(^2\)The database is publicly available on request. Contact e-mails: tkk22@cam.ac.uk
hand shapes and 3 primitive motions (see Figure 5). Each class contains 100 image sequences (5 different illuminations ×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 × 20 × 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 training sequences. The alternating solution stably converged as shown in the left of Figure 6. 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 middle of Figure 6, it is shown that about the first 60 features contained most of the discriminatory 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.

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

<table>
<thead>
<tr>
<th>Number of features</th>
<th>Joint-mode</th>
<th>Dual-mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>05</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table 1: Accuracy Comparison** of the joint-shared-mode TCCA and dual-mode TCCA (using both joint and single-shared mode).
Table 2: Hand-gesture recognition accuracy (%) of the four different illumination sets.

<table>
<thead>
<tr>
<th>Methods</th>
<th>set1</th>
<th>set2</th>
<th>set3</th>
<th>set4</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>81</td>
<td>81</td>
<td>78</td>
<td>86</td>
<td>82±3.5</td>
</tr>
<tr>
<td>Niebles et al. [3]</td>
<td>70</td>
<td>57</td>
<td>68</td>
<td>71</td>
<td>66±6.1</td>
</tr>
<tr>
<td>Wong et al. [11]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>44</td>
</tr>
</tbody>
</table>

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

Figure 7: Confusion matrix of hand gesture recognition.

60 features in the Table 1.

Table 2 shows the recognition rates of the proposed TCCA, Niebles et al.’s method [3] (the probabilistic Latent Semantic Analysis (pLSA) with the space-time descriptors, 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 [11]). The original codes and the best settings of the parameters (e.g. the size parameters of the space-time descriptors and the size of the code book) were used in the evaluation for the 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. The proposed method is far better in correct identifica-
Table 3: Recognition accuracy (%) on the KTH action data set.

<table>
<thead>
<tr>
<th>Methods</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>95.33</td>
</tr>
<tr>
<td>Niebles et al. [3]</td>
<td>81.50</td>
</tr>
<tr>
<td>Dollar et al. [8]</td>
<td>81.17</td>
</tr>
<tr>
<td>Schuldt et al. [9]</td>
<td>71.72</td>
</tr>
<tr>
<td>Ke et al. [10]</td>
<td>62.96</td>
</tr>
</tbody>
</table>

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 data base [9]. 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 [8], Schuldt et al. [9]) or pLSA (Niebles et al. [3]). Ke et al. applied the spatio-temporal volumetric features [10]. 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 close to the manually setting (solid ones). The Figure 10 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
Figure 8: Example action detection results for 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.

Figure 9: Confusion matrix of TCCA method for the KTH data set.

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 using 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
Figure 10: Action detection result. (a) An example input video sequence of continuous hand clapping actions. (b) The detection result: all three correct hand clapping actions are detected at the highest three peaks, with the three intermediate actions at the three lower peaks.

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 simple video processing techniques for finding the focus of attention, e.g. by moving area segmentation. Other pre-processing methods could yield further speed-ups.

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 pairwise 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.
Bibliography


