Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection

Tae-Kyun Kim and Roberto Cipolla, *Member*, *IEEE*

Abstract—This paper addresses a spatiotemporal pattern recognition problem. The main purpose of this study is to find a right representation and matching of action video volumes for categorization. A novel method is proposed to measure video-to-video volume similarity by extending Canonical Correlation Analysis (CCA), a principled tool to inspect linear relations between two sets of vectors, to that of two multiway data arrays (or tensors). The proposed method analyzes video volumes as inputs avoiding the difficult problem of explicit motion estimation required in traditional methods and provides a way of spatiotemporal pattern matching that is robust to intraclass variations of actions. The proposed matching is demonstrated for action classification by a simple Nearest Neighbor classifier. We, moreover, propose an automatic action detection method, which performs 3D window search over an input video with action exemplars. The search is speeded up by dynamic learning of subspaces in the proposed CCA. Experiments on a public action data set (KTH) and a self-recorded hand gesture data showed that the proposed method is significantly better than various state-of-the-art methods with respect to accuracy. Our method has low time complexity and does not require any major tuning parameters.

Index Terms—Action categorization, gesture recognition, canonical correlation analysis, tensor, action detection, incremental subspace learning, spatiotemporal pattern classification.

1 Introduction

The automatic classification and localization of human actions/gestures is useful for various applications such as video surveillance, human-computer interfaces, and object-level video summarization and retrieval. Broadly, relevant studies have either exploited explicit motion representation such as tracked trajectories of body parts [39], [9], [10], [3], [8] or directly analyzed space-time volumes [1], [7], [5]. Methods using tracked trajectories interpret actions *purely by motion information* and have tried to explicitly tackle main sources of variation in human motion, e.g., moving cameras, view point, and execution rate changes. However, obtaining the trajectory of body parts requires much human supervision for initialization. Recognition accuracy of this method is highly dependent on tracking in an unconstrained environment, which is a currently challenging topic of computer vision research. Active/passive markers on human bodies have been often used to reduce the complexity of the problem. A major problem with methods directly analyzing space-time volumes, on the other hand, is to find an efficient representation and matching of action videos, while at the same time avoiding the difficult problem of explicit motion representation. These methods, the so-called view-/or exemplar-based methods, make partial use of *both spatial and temporal information* delivering high recognition accuracy for a limited view. The methods in this category [1], [7], [5] are more suited to simple motions. Action is often discriminated from activity [12]: Action is an individual atomic unit of activity and activity is a series of actions in a predefined temporal order [44]. Whereas the trajectory-based approach is better suited to activity recognition by interpreting temporal transition, volume analysis methods are better suited to action recognition. This paper focuses on action (cf. activity) recognition methods that interpret video volumes without the use of trajectory estimation.

A number of recent works have analyzed human actions directly in space-time volumes. Video volume matching has been performed by utilizing dense optical flows [7], [6]. Optical flow estimation for dense, unconstrained, and nonrigid motion is, however, noisy and unreliable due to problems caused by smooth surfaces, self-occlusions, and appearance changes. The comparison of two video volumes has been achieved either by matching templates called motion history images [13], [8] or by measuring correlation of gradients of local space-time patches [1]. Motion history images as a holistic (cf. local) representation tend to be sensitive to changes in background and geometrical variation of actions. The method of local space-time patches [1] requires the manual setting of positions and scales of the local patches, whose optimal settings depend on the data. Silhouette images have been used [2], [4]. Feature vectors are extracted from silhouette images of action sequences and Poisson equation and the euclidian distance of the feature vectors is served as similarity of action sequences in [2]. As noted in [2], silhouettes are not always available and insufficient to represent complex spatial information.

One popular approach toward action recognition is based on spatiotemporal *bag-of-words* [5], [16], [15], [14]. Space-time interest points are detected in video volumes and local space-time variations around interest points are

- *T.-K. Kim is with Sidney Sussex College, University of Cambridge, Cambridge, CB2 3HU, UK. E-mail: tkk22@cam.ac.uk.*
- *R. Cipolla is with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK. E-mail: cipolla@cam.ac.uk.*

Manuscript received 2 July 2007; revised 5 Feb. 2008; accepted 2 June 2008; published online 12 June 2008.
Recommended for acceptance by S. Sclaroff.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-07-0400.
Digital Object Identifier no. 10.1109/TPAMI.2008.167.

described by histograms. Histogram representations are then combined with either Support Vector Machine (SVM) [16] or a probabilistic generative model [5]. Although they have yielded good accuracy, mainly due to the high discrimination power of individual local descriptors, they exhibit ambiguity by ignoring global space-time shape information. In spite of recent attempts [20], [19] to incorporate global information of action classes, there remains the difficulty of setting parameters of the space-time interest points as, again, these are application or data dependent.

Traditional classifiers may be applied to either vector or tensor representation of video volumes for action recognition tasks. Once a video volume is converted to a finite dimensional vector, applying classifiers, e.g., SVM or NN classifier, is straightforward. Concurrent studies have been carried out to classify tensors as an original form of imagery data without requiring vectorization. Ensembles of multilinear classifiers have been developed for the tensor data obtained from a color image [25] and the discriminant analysis method for the tensor data from a gray image using filter banks [26]. Corpora of motion capture data (obtained by infrared light markers) of multiple people and actions are analyzed as tensors for human motion synthesis and recognition [41]. There are, however, few previous works that analyze video volume tensors for action classification, except where SVM for tensor data has been proposed [43]. Both tensor classifiers [43] and traditional vector classifiers (aforementioned) directly exploit pixel statistics of holistic video volumes without useful feature extraction. They are, therefore, sensitive to spatiotemporal pattern variations of actions, thus exhibiting poor generalization on novel testing data under small sample size (see Section 6.1 for accuracy comparison of those methods).

Canonical correlation analysis (CCA). We have investigated a more principled and effective way of video volume matching. CCA, which has been, since Hotelling (1936), a standard tool for inspecting linear relations between two random variables (or two sets of vectors) [29], has more recently received increasing attention in computer vision literature (e.g., [45], [31], [32], [23], [21]). CCA has been applied to human gait recognition [3], where trajectories of joint angles of an articulated body are modeled by second-order stationary stochastic processes and CCA is deployed for comparing the models. As noted above, extraction of trajectories is difficult and the model is limited to repetitive motions. An image set is collected either from a video or sparse observations and is represented by a linear subspace (or hyperplane) [31]. CCA measures angles between two subspaces (cosine of the angles are called canonical correlations) for similarity between two image sets. A probabilistic interpretation of CCA [23] yields a model that reveals how well two input variables (i.e., two sets of vectors) are represented by a common source (latent) variable. Computation of canonical correlations has been extended into a nonlinear feature space by a positive definite kernel function [32]. In our earlier work [21], we proposed a CCA-based image-set classification with a discriminative transformation and successfully demonstrated this for various image-set-based object recognition tasks. Allowing data interpolation of image sets in CCA facilitates recognition of high-dimensional imagery data under small sample cases (see Sections 2.1 and 4 for details on CCA). Despite the success of CCA for image set matching (i.e., a collection of images without any temporal coherence), CCA is not sufficient to represent and match action video volumes in which both temporal and spatial information are important.

Fig. 1. **Sample actions bounded in spatiotemporal domain**. The bounding boxes indicate the spatial alignment and the superimposed images of the initial, intermediate, and the last frames of each action show the temporal alignment. The alignment can be automatically done by the proposed detection method.

Proposed study. In this study (conference version [22]), we propose an action recognition method by extending CCA of two sets of vectors into that of two video volume tensors. The method is a pairwise analysis of aligned and holistic action volumes. The proposed method is first applied to classification of aligned actions (see Fig. 1 for examples of actions in spatiotemporal bounding boxes) and then to action detection in input videos. The advantages and disadvantages of the proposed method over existing works are summarized in Table 1.

The proposed method focuses on the view points seen from training examples as in previous studies [1], [7], [5]. In spite of a limited view scope, there remain a number of other variations to consider such as changes in illumination, actors, backgrounds (indoor and outdoor), and clothes, as well as moderate changes in either view or camera movement, as contained in the experimental data sets (see Figs. 6, 10, and 14). Rather than explicitly modeling all of the variations, we take an exemplar-based approach that exhibits reasonable generalization over new data changes. With regard to complex motions that involve nonlinear time warping, these may be tackled in a so-called *divide and conquer* manner by a method that works well with simpler motions. Importantly, many existing works that make strong assumptions on inputs are not readily applicable to real-world problems. Our experiments also do not favor strong assumptions on inputs. The methods of these works are, moreover, based mainly on motion information, ignoring the spatial domain of video data, which provides strong evidence of action.

Harshman [30] has also presented a concept of CCA of multiway data arrays. Although it was carried out independently of our work, it has a common concept that supports the ideas presented in this paper. Our work not only comprises our new Tensor CCA (TCCA) method but also describes new applications of TCCA to action classification and detection.

The remainder of this paper is arranged as follows: CCA and multilinear algebra are briefly reviewed in Section 2. The extension of CCA to video volume tensors and its solution are given in Section 3. We perform action

TABLE 1
Advantages and Disadvantages of the Proposed Method

ADVANTAGES	DISADVANTAGES
Directly operates with video volumes, requiring neither heuristics to set up important parameters like local methods or assumptions on input such as trajectories or silhouettes.	Alignment process in the method requires prior camera motion compensation for moving cameras.
As a so called spatiotemporal appearance based method, makes use of both spatial and temporal information for maximum discrimination of action classes.	Can not handle large view point changes from those of exemplar actions.
Allows data interpolation in matching facilitating recognition of high-dimensional data which typically undergoes significant changes (for further discussion see Section 4).	Suited to simple motions that have approximate linear time-warping. To deal with the linear time-warping, the frames between the defined initial and last posture of actions are uniformly sampled for a fixed number in the temporal alignment.

classification in the Nearest Neighbor (NN) sense with the canonical correlation features, as explained in Section 4. Section 5 is devoted to the action detection method. The experimental results and conclusions are given in Sections 6 and 7, respectively.

2 Background

2.1 Review on Canonical Correlation Analysis

Given two random vectors $\mathbf{x} \in \mathbb{R}^{m_1}$, $\mathbf{y} \in \mathbb{R}^{m_2}$, a pair of transformations $\mathbf{u}, \mathbf{v}$, called canonical transformations, is found to maximize the correlation of $x' = \mathbf{u}^T\mathbf{x}$ and $y' = \mathbf{v}^T\mathbf{y}$ as

$$\rho = \max_{\mathbf{u},\mathbf{v}} \frac{\hat{\mathbb{E}}[x'y']}{\sqrt{\hat{\mathbb{E}}[x'^2]\hat{\mathbb{E}}[y'^2]}} = \frac{\mathbf{u}^T\mathbf{C}_{\mathbf{xy}}\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{C}_{\mathbf{xx}}\mathbf{u}\mathbf{v}^T\mathbf{C}_{\mathbf{yy}}\mathbf{v}}}, \quad (1)$$

where $\hat{\mathbb{E}}[f]$ denotes empirical expectation of function f and ρ is called the canonical correlation. Multiple canonical correlations $\rho_1, \ldots, \rho_d$, where $d \leq min(m_1, m_2)$, are defined by the next pairs of $\mathbf{u}, \mathbf{v}$, which are orthogonal to the previous ones. Canonical correlations are affine-invariant to inputs, i.e., $\mathbf{Ax} + \mathbf{b}$, $\mathbf{Cy} + \mathbf{d}$ for arbitrary (nonsingular) $\mathbf{A} \in \mathbb{R}^{m_1 \times m_1}$, $\mathbf{b} \in \mathbb{R}^{m_1}$, $\mathbf{C} \in \mathbb{R}^{m_2 \times m_2}$, $\mathbf{d} \in \mathbb{R}^{m_2}$. The 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 canonical transformations $\mathbf{u}, \mathbf{v}$.

Given two vector sets as matrices $\mathbf{X} \in \mathbb{R}^{N \times m_1}$ and $\mathbf{Y} \in \mathbb{R}^{N \times m_2}$, Goloub's SVD solution [34] is given as follows: If $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{N \times d}$ denote two eigenvector matrices of $\mathbf{X}, \mathbf{Y}$, respectively, where $N \gg m_1$, $m_2 \geq d$, canonical correlations are obtained as singular values of $(\mathbf{P}_1)^T\mathbf{P}_2$ by

$$(\mathbf{P}_1)^T\mathbf{P}_2 = \mathbf{Q}_1\mathbf{\Lambda}\mathbf{Q}_2^T, \quad \mathbf{\Lambda} = diag(\rho_1, \ldots \rho_d), \quad (2)$$

where $\mathbf{Q}_1$, $\mathbf{Q}_2$ are arbitrary rotating matrices such that $\mathbf{Q}_1\mathbf{Q}_1^T = \mathbf{Q}_2\mathbf{Q}_2^T = \mathbf{I}_d$. As d is typically a small number, the complexity of SVD, $O(d^3)$, is very low.

2.2 Multilinear Algebra and Notations

Following the notations in [24], [28], a video volume is a third-order tensor, which 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 $\mathbf{U} \in \mathbb{R}^{N \times J}$ is

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

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

We generalize the canonical correlation analysis of two vector sets into that of two high-order tensors. Previous studies [31], [32], [21] have made a comparison of vectorized image sets in a standard way of CCA. If a video volume is simply taken as a set of vectorized images for input of CCA, temporal information of action videos would be lost as CCA is invariant to ordering of image vectors. An extension is proposed for considering both spatial and temporal information for action classification.

3.1 Tensor Representation of Standard CCA

Standard CCA is first represented by tensor notations. Given two vector sets as matrices $\mathbf{X} \in \mathbb{R}^{N \times m_1}$, $\mathbf{Y} \in \mathbb{R}^{N \times m_2}$ $(N \gg m_1, m_2)$, CCA is written as

$$\rho = \max_{\mathbf{u},\mathbf{v}} \mathbf{x}'^T\mathbf{y}', \quad \text{where } \mathbf{x}' = \mathbf{X}\mathbf{u}, \mathbf{y}' = \mathbf{Y}\mathbf{v}. \quad (4)$$

Note that the canonical transformations $\mathbf{u}, \mathbf{v}$ are hereinafter defined to be such that $\mathbf{XU} = \mathbf{P}_1\mathbf{Q}_1$, $\mathbf{YV} = \mathbf{P}_2\mathbf{Q}_2$, where $\mathbf{U}, \mathbf{V}$ have $\mathbf{u}, \mathbf{v}$ in their columns, respectively, and $\mathbf{P}, \mathbf{Q}$ are eigenvector and rotating matrices defined in (2), respectively. If we take $\mathbf{X}, \mathbf{Y}$ as second-order tensors $(\mathcal{X})_{ij}$, $(\mathcal{Y})_{ij}$, the standard CCA is then represented as

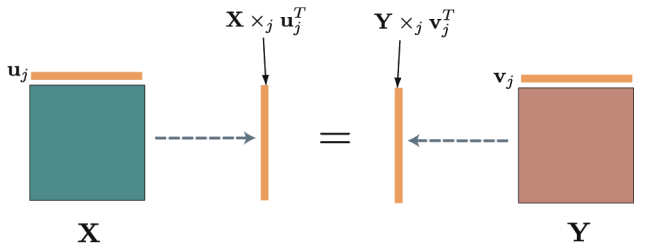

Fig. 2. **Tensor representation of standard CCA**. A pair of canonical transformations, $\mathbf{u}$ and $\mathbf{v}$, are applied to the two data matrices $\mathcal{X}$ and $\mathcal{Y}$ to yield maximally correlated vectors (called canonical vectors).

$$\rho = \max_{\mathbf{u},\mathbf{v}} \langle \mathcal{X} \times_j \mathbf{u}^T, \mathcal{Y} \times_j \mathbf{v}^T \rangle. \tag{5}$$

CCA has one shared mode (index i) and mode products by canonical transformations (index j), which is illustrated in Fig. 2. The two data matrices, for which $\mathbf{P}_1, \mathbf{P}_2$ are computed, can be written with respect to the j-mode vector matrices such that $\mathbf{X} = \mathbf{X}_{(j)}^T$, $\mathbf{Y} = \mathbf{Y}_{(j)}^T$. The j-mode products $\mathcal{X} \times_j \mathbf{U}^T$, $\mathcal{Y} \times_j \mathbf{V}^T$ in terms of j-mode vector matrices are $\mathbf{U}^T\mathbf{X}_{(j)} = (\mathbf{P}_1\mathbf{Q}_1)^T$, $\mathbf{V}^T\mathbf{Y}_{(j)} = (\mathbf{P}_2\mathbf{Q}_2)^T$, respectively. The canonical transformations are obtained by $\mathbf{U} = (\mathbf{X}_{(j)}\mathbf{X}_{(j)}^T)^{-1}\mathbf{X}_{(j)}\mathbf{P}_1\mathbf{Q}_1$, $\mathbf{V} = (\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^T)^{-1}\mathbf{Y}_{(j)}\mathbf{P}_2\mathbf{Q}_2$. Note that there is no loss of generality in applying formulation (5) to high-order tensors.

3.2 Joint/Single-Shared-Mode TCCA

A single channel video volume is represented as a third-order tensor denoted by $(\mathcal{A})_{ijk}$ that has the three axes of space (X and Y) and time (T). We assume that actions are spatiotemporally bounded as shown in Fig. 1 and every bounded video volume is uniformly resized to be $\mathbb{R}^{I\times J\times K}$ (Note that this preserves unique spatiotemporal patterns of video volumes). Tensor data, therefore, have all three indices (i, j, k) in common. Two different architectures of TCCA are proposed according to the number of shared modes.

Joint-shared-mode TCCA shares any two axes (i.e., a plane) and applies canonical transformations to the remaining single axis of tensor data. It involves three pairs of canonical transformations for given two tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I\times J\times K}$ as

$$\rho = \max_{\Phi} \langle \mathcal{X}', \mathcal{Y}' \rangle, \tag{6}$$

where

$$(\mathcal{X}')_{ijk} = \left(\mathcal{X} \times_i \mathbf{u}_i^T\right)_{jk} \cdot \left(\mathcal{X} \times_j \mathbf{u}_j^T\right)_{ik} \cdot \left(\mathcal{X} \times_k \mathbf{u}_k^T\right)_{ij},$$
$$(\mathcal{Y}')_{ijk} = \left(\mathcal{Y} \times_i \mathbf{v}_i^T\right)_{jk} \cdot \left(\mathcal{Y} \times_j \mathbf{v}_j^T\right)_{ik} \cdot \left(\mathcal{Y} \times_k \mathbf{v}_k^T\right)_{ij},$$

and $\Phi = \{(\mathbf{u}_i, \mathbf{v}_i), (\mathbf{u}_j, \mathbf{v}_j), (\mathbf{u}_k, \mathbf{v}_k)\}$. The resulting two tensors $\mathcal{X}'$, $\mathcal{Y}'$ are called canonical tensors. TCCA is seen as an aggregation of three different canonical correlation analyses, each of which is conceptually for two sets of vectorized IJ planes (involving k-mode product) and two sets of IK (j-mode product) or JK planes (i-mode product, see Fig. 3a). Note that the CCA in previous studies [31], [32], [21] is equivalent to that for two sets of vectorized IJ planes (i.e., images).

Single-shared-mode TCCA takes any single axis in common (i.e., a scan line) and applies canonical transformations to the remaining two axes of tensor data as

$$\rho = \max_{\Phi} \langle \mathcal{X}', \mathcal{Y}' \rangle, \tag{7}$$

where

$$(\mathcal{X}')_{ijk} = \left(\mathcal{X} \times_i \mathbf{u}_i^T \times_j \mathbf{u}_j^T\right)_k \cdot \left(\mathcal{X} \times_i \mathbf{u}_i^T \times_k \mathbf{u}_k^T\right)_j \cdot \left(\mathcal{X} \times_j \mathbf{u}_j^T \times_k \mathbf{u}_k^T\right)_i,$$
$$(\mathcal{Y}')_{ijk} = \left(\mathcal{Y} \times_i \mathbf{v}_i^T \times_j \mathbf{v}_j^T\right)_k \cdot \left(\mathcal{Y} \times_i \mathbf{v}_i^T \times_k \mathbf{v}_k^T\right)_j \cdot \left(\mathcal{Y} \times_j \mathbf{v}_j^T \times_k \mathbf{v}_k^T\right)_i,$$

and $\Phi = \{(\mathbf{u}_i, \mathbf{v}_i), (\mathbf{u}_j, \mathbf{v}_j), (\mathbf{u}_k, \mathbf{v}_k)\}$. Note that the canonical tensors are given by the outer products of the three vectors. Similarly, it is an aggregation of three different canonical correlation analyses, each of which can be conceptually for sets of I (involving j, k-mode product), J (i, k-mode product), or K (i, j-mode product) scan lines (see Fig. 3b).

Multiple canonical correlations $\rho_1, \ldots, \rho_d$ are defined for both joint-shared-mode and single-shared-mode TCCA, analogously to standard CCA. Compared with the previous study [30], Harshman only considered a single-shared mode, while we have proposed a general concept of multiple-shared modes.

3.3 Alternating Solution

Intuitively, the proposed TCCA process in (6) and (7) involves three subanalyses, each of which explains canonical correlations in different data domains. We, therefore, propose a solution that performs a subanalysis independently of the others. Each independent process is associated with the respective canonical transformations and yields canonical correlations as inner products of the respective canonical tensors. This section is devoted to explaining the solution for the I single-shared mode for example. This involves two sets of canonical transformations $\{(\mathbf{U}_j, \mathbf{V}_j), (\mathbf{U}_k, \mathbf{V}_k)\}$, 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 $(\rho_1, \ldots \rho_d)$, where $d \leq \min(K, J)$ for given two data tensors, $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I\times J\times K}$ as

$$\max_{\mathbf{U}_j,\mathbf{V}_j,\mathbf{U}_k,\mathbf{V}_k} \left\langle \mathcal{X} \times_j \mathbf{U}_j^T \times_k \mathbf{U}_k^T, \mathcal{Y} \times_j \mathbf{V}_j^T \times_k \mathbf{V}_k^T \right\rangle. \tag{8}$$

That is, canonical correlations are defined by the inner product of two resulting canonical tensors. The solution is obtained by performing the SVD method (see (5)) alternatively until convergence, as detailed in Table 2.

The J and K single-shared-mode TCCAs are performed in the same alternating fashion while the IJ, IK, JK joint-shared-mode TCCAs (e.g., IJ joint-shared-mode TCCA corresponds to the process involving k-mode product in (6)) by performing the SVD method (5) a single time without iterations.

4 Tensor CCA for Action Classification

Multiple canonical correlations computed in all subprocesses yield a total number of $2 \times 3 \times d$ canonical correlation features. (Each joint-shared mode or single-shared mode has three different CCA processes and each CCA

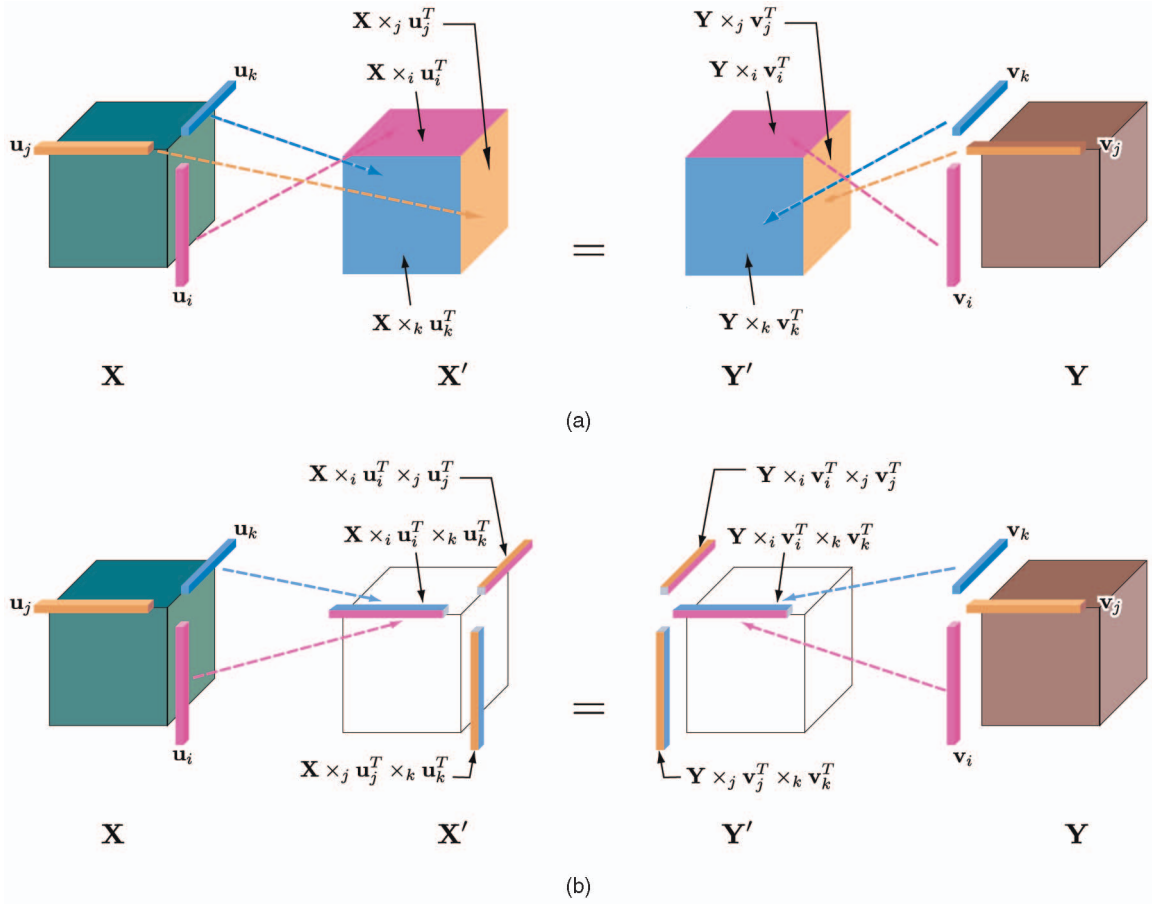

Fig. 3. **Representation of Tensor CCA**. (a) Joint-shared mode. Each canonical transformation, $\mathbf{u}_i$, $\mathbf{u}_j$, or $\mathbf{u}_k$, applied to the tensor data $\mathcal{X}$ yields a canonical plane. The three canonical planes make up the canonical tensor $\mathcal{X}'$. Likewise, canonical transformations $\mathbf{v}_i$, $\mathbf{v}_j$, $\mathbf{v}_k$ are applied to $\mathcal{Y}$ for the canonical tensor $\mathcal{Y}'$. (b) Single-shared mode. Any two canonical transformations (e.g., $\mathbf{u}_i$, $\mathbf{u}_k$) applied to the tensor $\mathcal{X}$ yields a canonical vector. Other two canonical vectors are similarly obtained and the canonical tensor $\mathcal{X}'$ is obtained by the outer products of the three canonical vectors. The same process is done for $\mathcal{Y}$.

TABLE 2
Proposed Alternating Algorithm for Tensor Canonical Correlations

Algorithm 1. Alternating solution for I single-shared-mode TCCA

Input: Two data tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K}$ **Output:** canonical correlations $\rho_1, ...\rho_d$

1.Given a random guess for $\mathbf{U}_j, \mathbf{V}_j$, $\widetilde{\mathcal{X}} \leftarrow \mathcal{X} \times_j \mathbf{U}_j^T, \widetilde{\mathcal{Y}} \leftarrow \mathcal{Y} \times_j \mathbf{V}_j^T$.
2. Do iterate the following until convergence:
3.Find $\mathbf{U}_k^*, \mathbf{V}_k^*$ that maximize $\langle \widetilde{\mathcal{X}} \times_k \mathbf{U}_k^T, \widetilde{\mathcal{Y}} \times_k \mathbf{V}_k^T \rangle$ by the SVD method (5).
Let $\widetilde{\mathcal{X}} \leftarrow \widetilde{\mathcal{X}} \times_k \mathbf{U}_k^{*T}$, $\widetilde{\mathcal{Y}} \leftarrow \widetilde{\mathcal{Y}} \times_k \mathbf{V}_k^{*T}$,
4.Find $\mathbf{U}_j^*, \mathbf{V}_j^*$ that maximize $\langle \widetilde{\mathcal{X}} \times_j \mathbf{U}_j^T, \widetilde{\mathcal{Y}} \times_j \mathbf{V}_j^T \rangle$ by the SVD method (5).
Let $\widetilde{\mathcal{X}} \leftarrow \widetilde{\mathcal{X}} \times_j \mathbf{U}_j^{*T}$, $\widetilde{\mathcal{Y}} \leftarrow \widetilde{\mathcal{Y}} \times_j \mathbf{V}_j^{*T}$,
5.End
6.Obtain $\rho_1, ..., \rho_d$ ($d \leq min(K, J)$) from the latest SVD solution.

process yields d features.) In general, each feature carries a different amount of discriminative information for action classification. We propose the discriminative feature selection method and NN classification, where the sum of selected canonical correlations serves as a similarity measure between action video volumes.

4.1 TCCA Features

4.1.1 *Explaining Data Similarity in Different Domains*

Intuitively, canonical correlation features explain data similarity in different data subspaces and dimensions. In Fig. 4a, we have visualized the first few canonical tensors

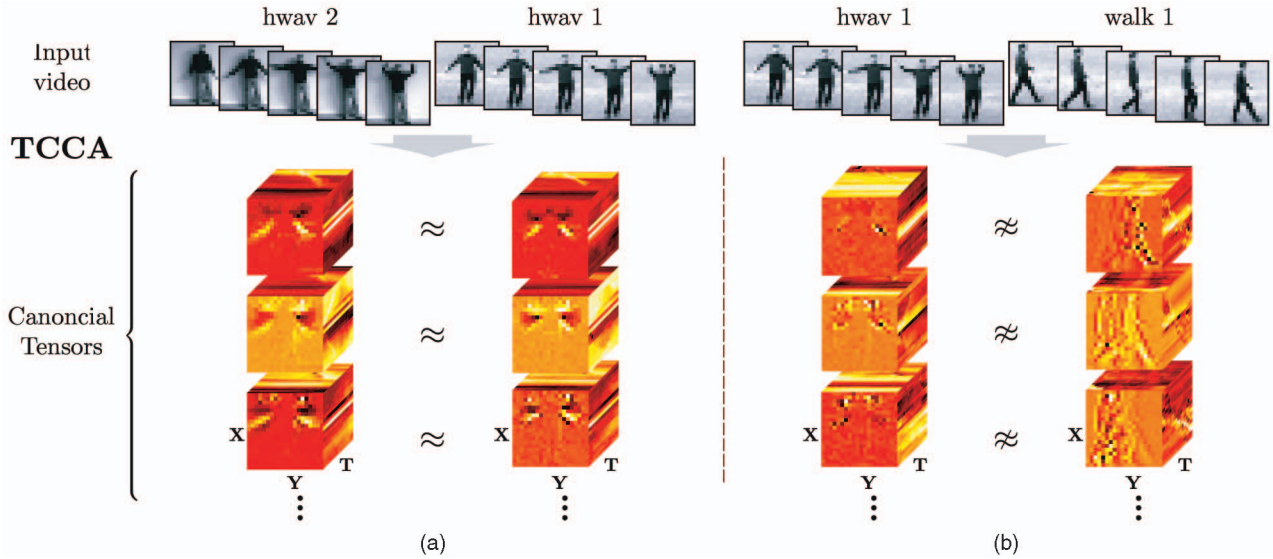

Fig. 4. **Examples of pairwise canonical tensors**. This visualizes the first few canonical tensors computed for the pair of input sequences of (a) the same action class and (b) the two different action classes. Canonical tensors of IJ, IK, JK joint-shared modes are the XY, XT, YT planes of the cubes, respectively. (a) Note that the canonical tensors in each pair are very much alike, although the two hand-waving sequences are captured under different environmental conditions and posed by different people wearing different clothes. (b) On the other hand, the canonical tensors are much dissimilar despite the sequences of the same person in the same environment.

computed by the joint-shared-mode TCCA from the two hand-waving sequences. Canonical tensors of IJ, IK, JK joint-shared modes are the XY, XT, and YT planes of the cubes, respectively. The canonical tensors (XY planes) of the IJ joint-shared mode show the common spatial components of the two hand-waving videos. Note that the canonical transformations applied to the K-axis (temporal axis) in the IJ joint-shared mode make the mode independent of temporal information, i.e., temporal ordering of video frames, whereas all other modes remain dependent. Similarly, the canonical tensors of the IK, JK joint-shared mode reveal the common components of the two videos in the joint space-time domain. The two modes are independent of J and I-axes, respectively. Likewise, the single-shared mode yields canonical correlations of other data domains.

4.1.2 Linear Data Interpolation

Canonical correlations are of linear combinations (by canonical transformations) of data vectors of two respective data sets. That is, CCA does interpolation of vectors to find maximum correlations and additional data generated by the interpolation facilitates generalization on new data and recognition of high-dimensional imagery data that typically undergoes significant variations. The invariance afforded by the interpolation is equivalent to the mathematical affine invariance of CCA in Section 2.1.

In Fig. 4a, we can see that the canonical tensors in each pair are very much alike. The two input sequences belong to the same action class, hand waving, but have different backgrounds, lighting conditions. They are also posed by different people wearing different clothes. Despite all of the differences, the canonical tensors, however, capture mutual information of the two inputs yielding high correlations. The first pair of canonical tensors corresponds to the most similar direction of variation of the two data sets and the next pairs represent other directions of similar variations. The canonical tensors corresponding to XY planes emphasize the movements of arms, which define the hand-waving class, as a common source of information. All other canonical tensors (XT, YT planes) are also pairwise similar. On the other hand, the canonical tensors are significantly different from the paired ones in Fig. 4b, where the two input sequences are from two different action classes (one is hand waving and the other walking). Although these sequences were captured under the same environment and posed by the same person, TCCA returns least correlations.

4.1.3 CCA as Subspace Angles

The proposed method embodies CCA. The geometrical interpretation of CCA, which is equivalent to the standard formulation (1), gives another intuitive explanation. Canonical correlations, which are cosines of principal angles $0 \leq \theta_1 \leq \cdots \leq \theta_d \leq (\pi/2)$ between any two d-dimensional linear subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, are uniquely defined as

$$\rho_i = \cos\theta_i = \max_{\mathbf{a}_i \in \mathcal{L}_1} \max_{\mathbf{b}_i \in \mathcal{L}_2} \mathbf{a}_i^T \mathbf{b}_i \tag{9}$$

subject to $\mathbf{a}_i^T \mathbf{a}_i = \mathbf{b}_i^T \mathbf{b}_i = 1$, $\mathbf{a}_i^T \mathbf{a}_j = \mathbf{b}_i^T \mathbf{b}_j = 0$, $i \neq j$. CCA as subspace-based matching (measuring angles between two subspaces) effectively places uniform prior on subspaces and yields invariance to pattern variations subject to the subspaces. The subspace angle is intuitively a natural extension of prior subspace-based recognition methods. When a single vector is given as an input, there is a standard way to classify it by subspaces: We measure the distances of the vector to the subspaces and pick the nearest one. As we now need to classify a subspace instead of a single vector, the distance is no longer valid, but angles between subspaces become a reasonable measurement.

TABLE 3
Accuracy Comparison of the Joint-Shared-Mode TCCA and Dual-Mode TCCA (Using Both Joint and Single-Shared Modes)

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

4.2 Joint versus Single-Shared Mode

Generally, the single-shared mode is more flexible and preserves less original data structures in matching than the joint-shared mode. The single-shared mode involves two pairs of canonical transformations, whereas the joint-shared mode has a single pair. Any ideal feature for classification should balance the flexibility (for minimizing intraclass variation) against the data-preserving power (for maximizing interclass variation). We have observed from our experiments that the joint-shared-mode TCCA delivers more discriminative features than the single-shared-mode TCCA. Note again that the CCA of image sets [21] is identical to the IJ joint-shared-mode TCCA method. The proposed single-shared-mode TCCA is, however, important: It consolidates the unified TCCA method and improves the accuracy of the joint-shared mode. Superiority of one type to the other may be application dependent.

4.3 Feature Selection

A discriminative boosting method is proposed to select useful tensor canonical correlation features. First, the intraclass and interclass feature sets (i.e., canonical correlations ρ_i, $i = 1, \ldots, 6 \times d$, computed from any pair of videos) are generated from the training data comprising of several class examples. We use each TCCA feature to build a simple weak classifier $\mathcal{M}(\rho_i) = sign[\rho_i - C]$ and aggregate the weak learners using the AdaBoost algorithm [35] (C is a classifier threshold and optimized in the AdaBoost). In an iterative update scheme, classifier performance is optimized on the training data to yield the final strong classifier by

$$\mathcal{M}(\rho) = sign\left[\sum_{i=1}^{M} w_{L(i)}\mathcal{M}\big(\rho_{L(i)}\big) - \frac{1}{2}\sum_{i=1}^{M} w_{L(i)}\right], \qquad (10)$$

where w contains the weights and L is the list of selected features. NN classification by sum of selected canonical correlations is performed to categorize a new test video.

5 Action Detection

The proposed TCCA is time efficient provided that actions are aligned in space-time domain. However, searching nonaligned actions by TCCA in 3D (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 iterations of the solutions and delivers sufficient discriminative power (see Table 3), time-efficient action detection is proposed by applying joint-shared-mode TCCA, which may be followed by the TCCA method using both joint and single-shared modes. For example, 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 with the joint mode 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.

5.1 Incremental PCA

An efficient update scheme of eigensubspaces has been developed when a new set of vectors is added to an existing data set [36], [37]. Given two data sets (an existing and a new set) represented by eigenspace models $\{\boldsymbol{\mu}_i, M_i, \mathbf{P}_i, \boldsymbol{\Lambda}_i\}_{i=1,2}$, where $\boldsymbol{\mu}_i$ is the mean, M_i is the number of samples, $\mathbf{P}_i$ is the matrix of eigenvectors, and $\boldsymbol{\Lambda}_i$ is the eigenvalue matrix of the ith data set, the combined eigenspace model $\{\boldsymbol{\mu}_3, M_3, \mathbf{P}_3, \boldsymbol{\Lambda}_3\}$ is efficiently computed. The eigenvector matrix $\mathbf{P}_3$ can be represented by $\mathbf{P}_3 = \boldsymbol{\Phi}\mathbf{R} = h([\mathbf{P}_1, \mathbf{P}_2, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2])\mathbf{R}$, where $\boldsymbol{\Phi}$ is the orthonormal column matrix spanning the entire combined data space, $\mathbf{R}$ is a rotation matrix, and h is a vector orthonormalization function. Using this representation, an original eigenproblem for $\mathbf{P}_3$, $\boldsymbol{\Lambda}_3$ is converted into a smaller eigenproblem as

$$\mathbf{S}_{T,3} = \mathbf{P}_3\boldsymbol{\Lambda}_3\mathbf{P}_3^T \Rightarrow \boldsymbol{\Phi}^T\mathbf{S}_{T,3}\boldsymbol{\Phi} = \mathbf{R}\boldsymbol{\Lambda}_3\mathbf{R}, \qquad (11)$$

where $\mathbf{S}_{T,3}$ is the scatter matrix of the combined data. Note that the matrix $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{S}_{T,3}\boldsymbol{\Phi}$ has the reduced size $d_{T,1} + d_{T,2} + 1$, where $d_{T,1}$, $d_{T,2}$ are the number of eigenvectors in $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. Thus, the eigenanalysis here only takes $O((d_{T,1} + d_{T,2} + 1)^3)$ computations, whereas the eigenanalysis in the left-hand side of (11) requires $O(\min(N, M_3)^3)$, where N is the input data dimension and M_3 is the total number of data points. Usually, $N, M_3 \gg d_{T,1} + d_{T,2} + 1$.

5.2 Dynamic Subspace Learning for TCCA

The computational complexity of the joint-shared-mode TCCA in (6) depends on the computation of eigenvector matrices $\mathbf{P}_1, \mathbf{P}_2$ and the Singular Value Decomposition (SVD) of $(\mathbf{P}_1)^T\mathbf{P}_2$ (see (5) and (2)). The total complexity trebles this computation for the IJ, IK, and JK joint-shared modes. If $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{N\times d}$, where d is the number of the first few eigenvectors corresponding to the most data energy (usually a small number), the complexity of the SVD of $(\mathbf{P}_1)^T\mathbf{P}_2$ taking $O(d^3)$ is negligible. Time-efficient detection is achieved by incrementally learning the three sets of eigenvectors, corresponding to the mode vector matrices $\mathbf{X}_{(k)}^T$, $\mathbf{X}_{(j)}^T$, $\mathbf{X}_{(i)}^T$, of every possible volume $\mathcal{X}$ (*cuboid*) of an input video for the IJ, IK, JK joint-shared modes, respectively. See Fig. 5 for the concept. There are three separate steps that are carried out in the same fashion, each of which is to compute one of three eigenvector matrices of every possible volume of an input video. First, the subspaces of every cuboid of the initial slices of the

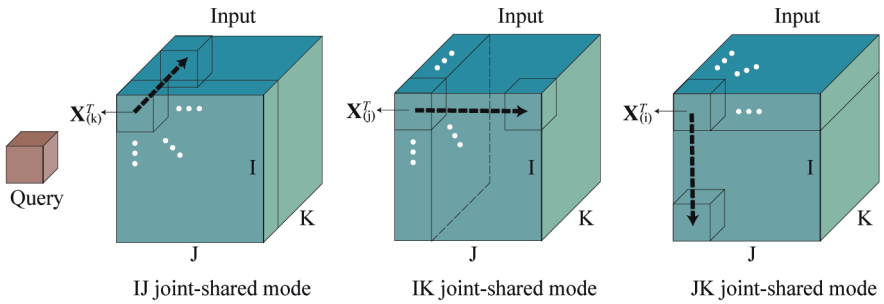

Fig. 5. **Detection scheme**. A query video is searched in a large volume input video. TCCA between the query and every possible volume (*cuboids*) of the input video can be speeded up by dynamically learning the three subspaces of cuboids for the IJ, IK, JK joint-shared-mode TCCAs. While moving the initial slices along one axis, subspaces of cuboids are dynamically computed from those of the initial slices. See Section 5.2 for further explanation.

input video are learned, then the subspaces of all remaining cuboids are incrementally computed while moving the slices along one of the axes. For the IJ joint-shared-mode TCCA, as an example, the subspace $\mathbf{P}$ of every possible cuboid, represented by the transposed k-mode vector matrix $\mathbf{X}_{(k)}^T$, in the initial IJ slice of the input video is computed. Then, the subspaces of all next cuboids are dynamically computed, while pushing the cuboids in the initial slice along the K-axis to the end as follows (for simplicity, the size of a query video and input video is set to be $\mathbb{R}^{m\times m\times m}$, $\mathbb{R}^{M\times M\times M}$, where $M \gg m$):

Any cuboid at z on the K-axis, $\mathcal{X}^z$ is represented by the k-mode vector matrix $\mathbf{X}_{(k)}^T = \{\mathbf{x}^z, \ldots, \mathbf{x}^{z+m-1}\}$. The scatter matrix $\mathbf{S}^z = (\mathbf{X}_{(k)}^T)(\mathbf{X}_{(k)}^T)^T$ is written with respect to the scatter matrix of the previous cuboid at $z-1$ as $\mathbf{S}^z = \mathbf{S}^{z-1} + (\mathbf{x}^{z+m-1})(\mathbf{x}^{z+m-1})^T - (\mathbf{x}^{z-1})(\mathbf{x}^{z-1})^T$. This involves both incremental and decremental learning. A new vector $\mathbf{x}^{z+m-1}$ is added and an existing vector $\mathbf{x}^{z-1}$ is removed from the $(z-1)$th cuboid. The sufficient spanning set[1] of the current scatter matrix can be $\Upsilon = h([\mathbf{P}^{z-1}, \mathbf{x}^{z+m-1}])$, where h is a vector orthogonalization function and $\mathbf{P}^{z-1}$ is the eigenvector matrix of the previous cuboid. The current eigenvector matrix can be the product of the sufficient spanning set by an arbitrary rotation matrix $\mathbf{R}$ as $\mathbf{P}^z = \Upsilon\mathbf{R}$. Therefore, the original eigenproblem to solve is reduced to a much smaller eigenproblem as

$$\mathbf{S}^z = \mathbf{P}^z\mathbf{\Lambda}^z(\mathbf{P}^z)^T \Rightarrow \Upsilon^T\mathbf{S}^z\Upsilon = \mathbf{R}\mathbf{\Lambda}^z\mathbf{R}. \tag{12}$$

The matrices $\mathbf{\Lambda}^z$, $\mathbf{R}$ are computed as the eigenvalue and eigenvector matrix of $\Upsilon^T\mathbf{S}^z\Upsilon$. The final eigenvectors are obtained as $\mathbf{P}^z = \Upsilon\mathbf{R}$ after removing the components in $\mathbf{R}$ corresponding to the least eigenvalues in $\mathbf{\Lambda}^z$, keeping the dimension of $\mathbf{P}^z$ as $\mathbf{R}^{m^2\times d}$.

5.2.1 Computational Cost

Similarly, the subspaces for $\mathbf{X}_{(j)}^T$, $\mathbf{X}_{(i)}^T$ for the IK, JK joint-shared-mode TCCAs are computed by moving the all cuboids of the slices along the I, J-axes, respectively. In this way, the total complexity of learning the three kinds of the subspaces of every cuboid is significantly reduced such that

$$O(M^3 \times m^3) \longrightarrow O(M^2 \times m^3 + M^3 \times d^3), \tag{13}$$

as $M \gg m \gg d$. $O(m^3)$, $O(d^3)$ are the complexity for solving eigenproblems in a batch (i.e., the left-hand side of (12)) and the proposed way (the right-hand side of (12)). Efficient multiscale search, as a future work, may be performed by merging two or more subspaces of smaller cuboids by the incremental learning.

6 Experimental Results

6.1 Hand Gesture Recognition

We acquired the *Cambridge-Gesture database*[2] consisting of 900 image sequences of nine hand gesture classes, which are defined by three primitive hand shapes and three primitive motions (see Fig. 6). Each class contains 100 image sequences (5 illuminations $\times$ 10 arbitrary motions of two subjects). Each sequence was recorded in front of a fixed camera having roughly isolated gestures in space and time. All training was performed on the data acquired in the single plain illumination setting (the leftmost part in Fig. 6b), while testing was done on the data acquired in the remaining settings. The 20 sequences per class in the training set were randomly partitioned into 10 sequences for training and the other 10 sequences for validation.

All video sequences were uniformly resized into 20 $\times$ 20 $\times$ 20 in our method. The proposed alternating solution in Section 3.3 was performed to obtain the TCCA features of every pairwise training sequence. The iterative method stably converged, as shown in Fig. 7a. Feature selection was performed for the TCCA features based on the weights and the feature list learned from the AdaBoost method in Section 4. NN classification was performed for a new test sequence by the sum of the selected TCCA features. In Fig. 7b, it is shown that about the first 60 features contained most of the discriminatory information. Of the first 60 features, the number of features is shown for the different TCCA modes in Fig. 7c. 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 3, 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 Fig. 7, where the weights of more than 20 features are

1. The sufficient spanning set is an economical set of bases which spans most data energy. This helps obtain a small eigenproblem to solve [36], [37].

2. The database is publicly available at http://mi.eng.cam.ac.uk/~tkk22.

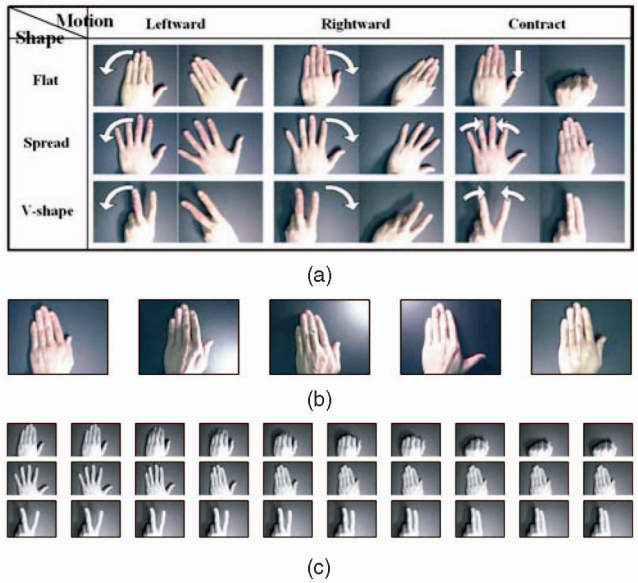

Fig. 6. **Hand gesture database**. (a) Nine gestures generated by three primitive shapes and motions. (b) Five illumination conditions in the database. (c) Three sample sequences of the contraction motion.

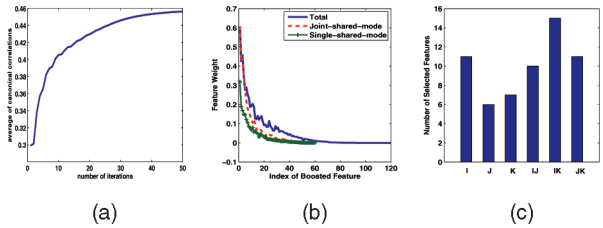

Fig. 7. **Feature selection**. (a) Convergence graph of the alternating method. (b) The weights of TCCA features learned by boosting. (c) The number of features chosen for the different TCCA modes.

nonsignificant. Note that the accuracy monotonically increased delivering the best accuracy at 60 even without feature selection. The single-shared mode alone gave relatively poor accuracy, which is yet meaningful compared with those of other methods in Table 4. The dual-mode TCCA (using both joint and single-shared modes) improved the accuracy of the joint-shared mode by 5 percent. Fig. 8 shows the example of canonical tensors computed from the two lighting sequences of the same hand gesture class. Only one of each pair of canonical tensors is shown here as the other looks similar.

Table 4 shows the recognition rates of the proposed TCCA method (exploiting both joint and single-shared-mode features), the simple CCA method [21], Niebles et al.'s [5] method (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 [5]), Wong et al.'s method (SVM or Relevance Vector Machine (RVM) with the Motion Gradient Orientation (MGO) images [18]), NN classifier in the sense of Euclidean Distance (NN-ED) and Normalized Correlation (NN-NC) of video vectors (all pixels in a video are concatenated into a column vector), and SVM of the video vectors. 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 codebook) were used in the evaluation for the previous works. The two methods of SVM or RVM on the MGO images turned far worse. As observed in [18], using RVM improved the accuracy of SVM by about 10 percent. However, both methods often failed to discriminate the gestures, which have the same motion of the different shapes as the methods are mainly based on motion information of gestures. The two methods, NN-ED and NN-NC, which exploit vector distance and vector correlation, respectively, as a similarity between two gesture videos, were also far poorer than the proposed method. The SVM applied to the vector representation enhanced the accuracy of the NN-ED/NC methods, but is again much worse than the proposed method. Although the vector representation of videos encodes space-time shape information, its high dimension interrupts obtaining good generalization on novel data under small sample size. The unsupervised learning method pLSA with the space-time interest points and the simple CCA method achieved the second-rank accuracy by either a flexible representation or matching: The pLSA method is based on distribution of local patterns and CCA provides the affine invariance in matching. Note, however, that the accuracy of the pLSA method is highly compromised with good parameter setting (of the space-time descriptors), which is difficult in practice. Both methods do not make use of full video information: pLSA does not encode global shape information, while CCA does not consider temporal information. The proposed method, TCCA, significantly outperformed all compared methods. The proposed method improved the simple CCA method by around 17 percent. By matching both spatial and temporal information with the affine invariance, the proposed method is far better in correct identifications of the sequences of distinct shapes subject to similar motion as well as the similar shape sequences having different motions. See Fig. 9 for the confusion matrix of our method.

TABLE 4
Hand Gesture Recognition Accuracy (in Percent) of the Four Illumination Sets

Methods	set1	set2	set3	set4	total	Methods	total
TCCA	81	81	78	86	82±3.5	MGO/SVM [18]	30
CCA [21]	63	61	65	69	65±3.2	NN-ED	29.44
pLSA [5]	70	57	68	71	66±6.1	NN-NC	29.03
MGO/RVM [18]	-	-	-	-	44	SVM	41.25

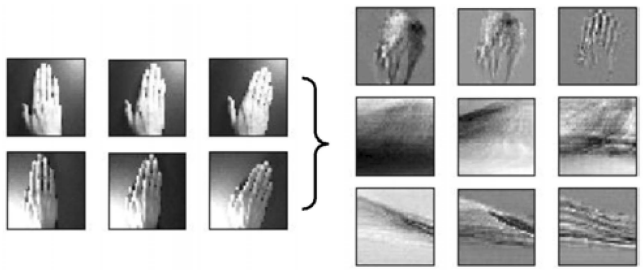

Fig. 8. **Example of canonical tensors**. Given two lighting sequences of the same hand gesture class (the left two rows), the first three canonical tensors of the IJ, IK, JK joint-shared modes are shown in the top, middle, bottom rows, respectively.

6.2 Action Categorization on KTH Data Set

We followed the experimental protocol of Niebles et al.'s [5] work on the KTH action data set, which is the largest public action database [16]. The data set contains six types (boxing, hand clapping, hand waving, jogging, running, and walking) of human actions performed by 25 subjects in four different scenarios. The original input videos contain actions that are not space-time aligned and are repeated several times. Leave-one-out cross-validation was performed to test the methods, 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 Fig. 10 with the indication of the action alignment (or cropping). This space-time alignment of actions was manually done for accuracy comparison, but can also be automatically achieved by the proposed detection scheme as shown below. The defined aligned actions contain unit atomic motions without repetitions. Most competing methods are based on the histogram representations with SVM (ST/SVM) [15], [16] or pLSA [5]. Ke et al. applied the spatiotemporal volumetric features [17]. Note that all of these methods do not require action alignment in nature because they do not consider global space-time shape information. These methods were, therefore, applied to the original input videos. For comparison, we quoted the accuracy of the methods reported in [5] and further performed the simple CCA method, the pLSA method [5],

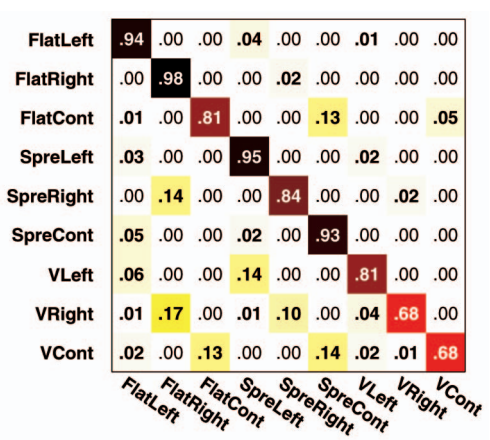

Fig. 9. Confusion matrix of the TCCA method for hand gesture recognition.

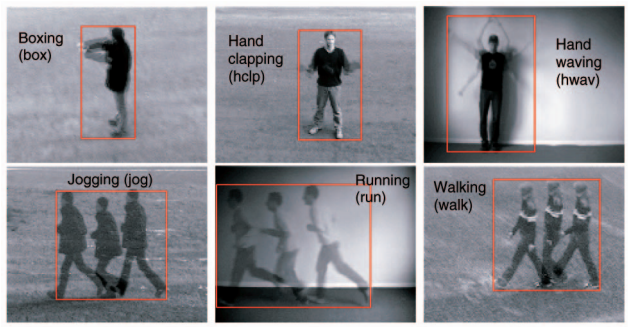

Fig. 10. **Example action videos in the KTH data set**. The bounding boxes indicate the spatial alignment and the superimposed images of the initial, intermediate, and the last frames of each action show the temporal segmentation of action classes.

and the proposed TCCA method (exploiting both joint and single-shared-mode features) on the aligned videos. In the TCCA method, the aligned video sequences were uniformly resized to $20 \times 20 \times 20$ by NN interpolation (see Table 7 for the original volume size). See Table 5 for accuracy comparison of several methods and Fig. 11 for the confusion matrices of the TCCA method and CCA method. The pLSA method on the cropped videos dropped the accuracy of the same method on the original input videos by about 10 percent, maybe due to insufficient amount of interest points detected in the cropped videos. Note that the original sequences contain several repetitions of the actions giving fluent interest points. The SVM applied to the same histogram representation as that of the pLSA method [15] delivered the similar accuracy. While most of the histogram-based methods showed the accuracy around 60 percent to 80 percent, the proposed TCCA method and the CCA method achieved impressive accuracy at 95 percent and 89 percent, respectively. From the good accuracy of the CCA method that does not consider temporal information, we infer that the six action classes of the KTH data set discriminate well in spatial domain. The histogram-based methods lost important information in the global space-time shapes of actions resulting in ambiguity for spatial variations of the different action classes. The TCCA method improved the CCA method by using joint spatial-temporal information, being particularly better in discrimination between the jogging and running actions, which is shown in Fig. 11.

There have been recent attempts to incorporate the global space-time shape information based on the histogram

TABLE 5
Recognition Accuracy (in Percent) on the KTH Action Data Set

Methods	(%)	Methods	(%)
TCCA	95.33	ST/SVM [15]	81.17
CCA [21]	89.50	ST/SVM [16]	71.72
pLSA [5]	81.50	Ke et al. [17]	62.96
pLSA* [5]	68.53		
pLSA-ISM [19]	83.92	Savarese et al. [40]	86.83

pLSA denotes the pLSA method applied to the cropped videos.*

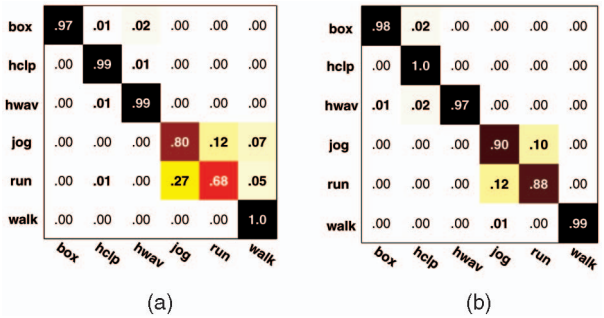

Fig. 11. **Confusion matrix of (a) CCA and (b) TCCA methods for the KTH data set**. TCCA improved CCA especially by better discriminating the jogging from the running actions.

representation [19], [40]. As shown in the last row of Table 5, they achieved reasonable improvements over the previous histogram methods but were still inferior to the method proposed.

6.2.1 Discussions

We have tried two different regularization methods. Each image in the videos is Gaussian-smoothed with histogram equalization or is just Gaussian-smoothed. We achieved 92.00 percent and 95.33 percent recognition accuracies by Gaussian smoothing with or without histogram equalization, respectively.

The volume size of $20 \times 20 \times 20$ gave a good compromise between the recognition accuracy and computational resource. We set the volume size as $10 \times 10 \times 10$, $20 \times 20 \times 20$, and $40 \times 40 \times 20$, obtaining 90.67 percent, 95.33 percent, 96.00 percent recognition accuracies, respectively.

To check the sensitivity of the proposed method on temporal misalignment, we added Gaussian noise ϵ to both start and end times of actions, such that $t' = t + \epsilon$. The Gaussian noise had zero mean and 10 percent of the average volume size in T as standard deviation. For example, the standard deviation is set to be 3.2 for the boxing videos that have 32-pixel temporal duration on average (see Table 7). The TCCA method exhibited reasonable degradation in performance for temporal misalignment, showing 90 percent accuracy for the noisy data.

We have also performed an experiment for background change. We used only outdoor samples in training and indoor samples in testing. Despite the quite different backgrounds in the indoor and outdoor videos (see Fig. 10), the TCCA method obtained the same accuracy (95 percent) as that reported in Table 5. Segmentation or any better representation method (rather than raw pixels) may further improve the TCCA method for significant background changes and clutters.

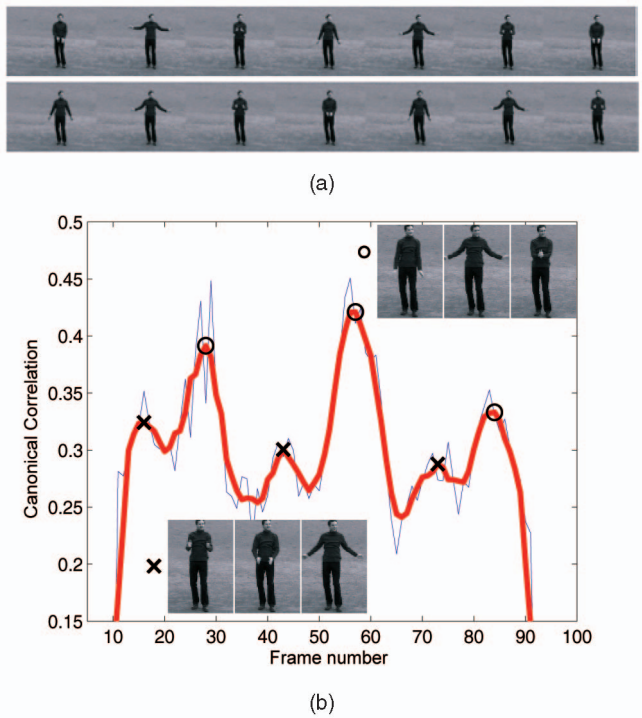

Fig. 12. **Action detection result**. (a) A sample input video sequence of continuous hand-clapping actions. (b) The detection result: All three correct hand-clapping actions are detected at the highest three peaks, with the three intermediate actions at the three lower peaks. The thin line (joint-shared-mode TCCA) was smoothed using a five-point moving average to yield the bold line.

6.3 Action Detection on KTH Data Set

The action detection was performed by the exemplar set consisting of the sequences of five people that are not contained in the testing sequences. Every possible volume (for both fixed-scale and multiscale search) in an input video is scanned and is matched with the sample sequences by TCCA (the joint-shared mode).

For the fixed-scale search, detection results are shown in Fig. 12 for the continuous hand-clapping video, which is comprised of the three correct unit clapping actions. The maximum canonical correlation is shown along time. All three correct hand-clapping actions are detected at the three highest peaks, with the three intermediate actions at the three lower peaks. The three highest peaks correspond to the video volumes that are synchronized to the query video in both spatial and temporal domains. When it goes far

TABLE 6
Action Detection Time (in Seconds) for Fixed-Scale Search by a Single Query Sequence

action class	box	hclp	hwav	jog	run	walk
dynamic subspace learning	43.01	35.42	19.27	12.60	5.16	10.70
or batch subspace learning	240.26	245.45	47.62	64.10	19.72	40.34
+ TCCA	9.96	8.43	2.26	3.09	1.14	2.21

The detection speed differs for the size of input volume with respect to the size of query volume.

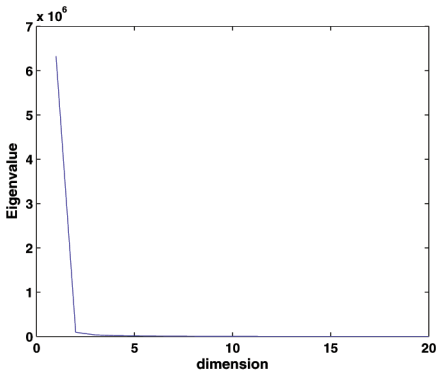

Fig. 13. **Eigenvalue plot**. Averaged eigenvalue plot of the three kinds of subspaces of action videos.

from the peaks, a video volume is less synchronized to the query, having lower correlations in both spatial and temporal aspects. However, at a certain point, it starts recovering correlations in spatial domain by containing most but permuted frames of the query video, exhibiting local maxima between any two correct hand-clapping actions. Note that the IJ joint-shared-mode TCCA is invariant to permutation of frames.

The detection time of the proposed method (using the joint-shared-mode TCCA) is reported in Table 6 on a Pentium 3 GHz PC using nonoptimized Matlab codes. The proposed incremental subspace learning reduced the detection time of the batch computation. The detection time differs for the size of input volume with respect to the size of query volume. For example, the input and query volume sizes of 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, and 1 (for the respective three dimensions). In the reduced dimension, the size of the query video, m in (13), was 20. The dimensions of the subspaces, d in (13), were set to be 5 as the number that reflects most data energy from the eigenvalue plot (see Fig. 13). When the search area M and the size of the query video m are larger, the computational saving by the proposed method over the batch method would be greater. The obtained speed seems to be comparable to that of the state of the art [1]. Video processing techniques such as moving area segmentation may be conveniently incorporated into the proposed method for further speedup.

Fig. 14 shows the sample action detection results with scale variations, which are obtained by three steps in each axis. We set the three steps as the mean and mean plus/minus the standard deviation of the scales of video volumes (see Table 7). The detection results show the best response space-time region in each input sequence. Despite the small training samples (of only five people, as mentioned before) and the coarse three-step scale search, the alignments look close to the manual settings shown in Fig. 10. Efficient multiscale search would help obtain more accurate and yet time-efficient action detection.

7 Conclusions

We have proposed a novel method called TCCA, which extracts pairwise flexible and yet descriptive correlation features of videos in joint space-time domain. The proposed features combined with NN classifier significantly improved

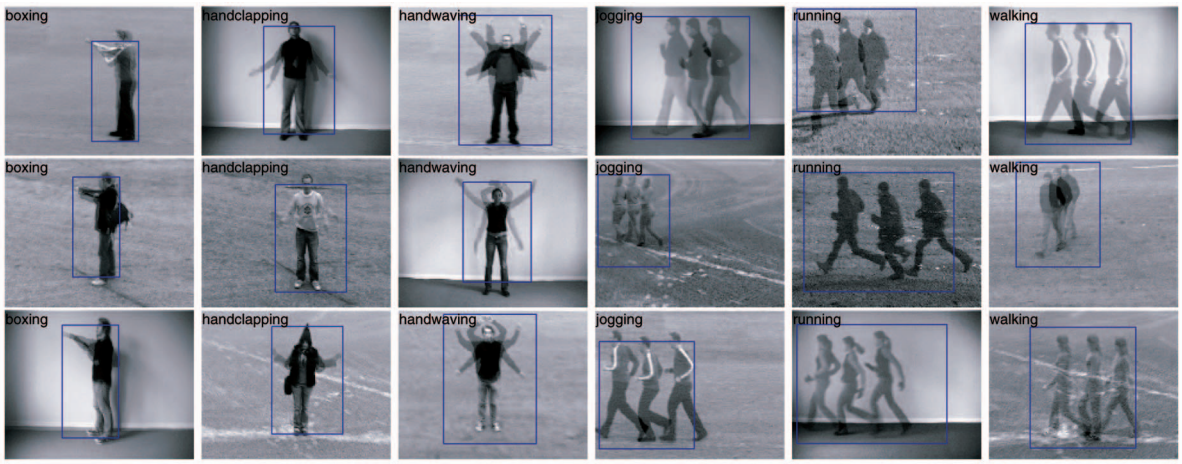

Fig. 14. Automatic multiscale action detection result.

TABLE 7
Average Volume Size of Action Classes

(pixels)	box	hclp	hwav	jog	run	walk
X	48±8	60±11	68±10	80±20	101±26	71±18
Y	91±10	87±10	92±12	86±12	82±13	84±13
T	32±8	22±6	19±4	11±2	9±1	15±1

The mean and the standard deviation along each axis.

the accuracy of state-of-the-art action recognition methods. The proposed method is also practically appealing as it does not require any significant tuning parameters. Additionally, the proposed detection method for TCCA could yield time-efficient action detection in large-volume input videos.

In spite of the proposed detection method, the method may require further time efficiency for the scenarios that have a much larger search space and require multiscale search in real time. One may try a hierarchical approach that applies simpler but less accurate methods to filter out a majority of candidates and then to apply our method, which has the benefit of high accuracy. Efficient multiscale search by merging the space-time subspaces of TCCA would constitute useful future work. For further enhancement in accuracy, the proposed method as a general meta-algorithm may be combined with other task-specific representations or segmentation methods. As an example, the raw pixel representation in the TCCA method has been replaced with the Scale-Invariant-Feature-Transform (SIFT) vectors in [42]. Although we have exploited a naive NN classifier for the purpose of demonstrating the power of new features and matching, the use of a more modern classifier remains as future work.

ACKNOWLEDGMENTS

The work of Tae-Kyun Kim was supported by a research fellowship from Sidney Sussex College, University of Cambridge and by a Toshiba Cambridge Studentship. The authors are grateful to the anonymous reviewers for their constructive comments, Kil-Rye Lee for her help in the database annotation, Navid Mavaddat and Björn Stenger for their help in proofreading this work, and Shu-Fai Wong for his effort in the conference version.

REFERENCES

[1] E. Shechtman and M. Irani, "Space-Time Behavior Based Correlation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 405-412, 2005.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1395-1402, 2005.

[3] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of Human Gaits," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 401-417, 2001.

[4] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Matching Shape Sequences in Video with Applications in Human Motion Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1896-1909, Dec. 2005.

[5] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Proc. British Machine Vision Conf.,* 2006.

[6] M.J. Black, "Explaining Optical Flow Events with Parameterized Spatio-Temporal Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1326-1332, 1999.

[7] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* 2003.

[8] D. Ramanan and D.A. Forsyth, "Automatic Annotation of Everyday Movements," *Proc. Advances in Neural Information Processing Systems,* 2004.

[9] C. Rao, A. Yilmaz, and M. Shah, "View-Invariant Representation and Recognition of Actions," *Int'l J. Computer Vision,* vol. 50, no. 2, pp. 203-226, 2002.

[10] V. Parameswaran and R. Chellappa, "Human Action-Recognition Using Mutual Invariants," *Computer Vision and Image Understanding,* vol. 98, no. 2, pp. 294-324, 2005.

[11] A. Yilmaz and M. Shah, "Matching Actions in Presence of Camera Motion," *Computer Vision and Image Understanding,* vol. 104, no. 2, pp. 221-231, 2006.

[12] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "The Function Space of an Activity," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[13] A. Bobick and J. Davis, "The Recognition of Human Movements Using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, Mar. 2001.

[14] I. Laptev and T. Lindeberg, "Space-Time Interest Points," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* pp. 432-439, 2003.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," *Proc. Second Joint IEEE Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* pp. 65-72, 2005.

[16] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. 17th Int'l Conf. Pattern Recognition,* pp. 32-36, 2004.

[17] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. 10th IEEE Int'l Conf. Computer Vision,* pp. 166-173, 2005.

[18] S.-F. Wong and R. Cipolla, "Real-Time Interpretation of Hand Motions Using a Sparse Bayesian Classifier on Motion Gradient Orientation Images," *Proc. British Machine Vision Conf.,* pp. 379-388, 2005.

[19] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning Motion Categories Using Both Semantic and Structural Information," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[20] J.C. Niebles and L. Fei-Fei, "A Hierarchical Model of Shape and Appearance for Human Action Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[21] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1005-1018, June 2007.

[22] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor Canonical Correlation Analysis for Action Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.

[23] F.R. Bach and M.I. Jordan, "A Probabilistic Interpretation of Canonical Correlation Analysis," TR 688, Dept. of Statistics, Univ. of California, Berkeley, 2005.

[24] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear Analysis of Image Ensembles: TensorFaces," *Proc. Seventh European Conf. Computer Vision,* 2002.

[25] C. Bauckhage, T. Kaster, and J.K. Tsotsos, "Applying Ensembles of Multilinear Classifiers in the Frequency Domain," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[26] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Discriminant Analysis with Tensor Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.

[27] H. Wang and N. Ahuja, "Rank-R Approximation of Tensors Using Image-as-Matrix Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[28] C.D.M. Martin, *Proc. Tensor Decompositions Workshop,* 2004.

[29] D. Hardoon, S. Szedmak, and J.S. Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation,* vol. 16, no. 12, pp. 639-2664, 2004.

[30] R. Harshman, "Generalization of Canonical Correlation to *N*-Way Arrays," *Poster at the 34th Ann. Meeting of the Statistical Soc. Canada,* May 2006.

[31] O. Yamaguchi, K. Fukui, and K. Maeda, "Face Recognition Using Temporal Image Sequence," *Proc. Third IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 318-323, 1998.

[32] L. Wolf and A. Shashua, "Kernel Principal Angles for Classification Machines with Applications to Image Sequence Interpretation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.

[33] H. Hotelling, "Relations between Two Sets of Variates," *Biometrika,* vol. 28, no. 34, pp. 321-372, 1936.

[34] Å. Björck and G.H. Golub, "Numerical Methods for Computing Angles between Linear Subspaces," *Math. Computation,* vol. 27, no. 123, pp. 579-594, 1973.

[35] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Proc. Second European Conf. Computational Learning Theory,* pp. 23-37, 1995.
[36] P. Hall, D. Marshall, and R. Martin, "Merging and Splitting Eigenspace Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 9, pp. 1042-1049, Sept. 2000.
[37] D. Skocaj and A. Leonardis, "Weighted and Robust Incremental Method for Subspace Learning," *Proc. Ninth IEEE Int'l Conf. Computer Vision,* pp. 1494-1501, 2003.
[38] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[39] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden Conditional Random Fields for Gesture Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.
[40] S. Savarese, "2D and 3D Spatial Reasoning for Object Categorisation," *Tutorial in Int'l Computer Vision Summer School,* July 2007.
[41] M.A.O. Vasilescu, "Human Motion Signatures: Analysis, Synthesis, Recognition," *Proc. Int'l Conf. Pattern Recognition,* pp. 456-460, 2002.
[42] T.-K. Kim and R. Cipolla, "Gesture Recognition under Small Sample Size," *Proc. Eighth Asian Conf. Computer Vision,* pp. 335-344, 2007.
[43] L. Wolf, H. Jhuang, and T. Hazan, "Modeling Appearances with Low-Rank SVM," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[44] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa, "Shape Activity: A Continuous-State HMM for Moving/Deforming Shapes with Application to Abnormal Activity Detection," *IEEE Trans. Image Processing,* vol. 14, no. 10, pp. 1603-1616, 2005.
[45] P. Saisan, G. Doretto, Y.N. Wu, and S. Soatto, "Dynamic Texture Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 58-63, 2001.

Tae-Kyun Kim received the BSc and MSc degrees from the Korea Advanced Institute of Science and Technology in 1998 and 2000, respectively, and the PhD degree from the University of Cambridge in 2007. He is a research fellow in the Sidney Sussex College at the University of Cambridge. He was a research staff member at the Samsung Advanced Institute of Technology during 2000-2004. His research interests include computer vision, statistical pattern classification, and machine learning. The joint proposal of Samsung and NEC for face image descriptor, for which he developed main algorithms, is the international standard of ISO/IEC JTC1/SC29/WG11.

Roberto Cipolla received the BA degree in engineering from the University of Cambridge in 1984, the MSE degree in electrical engineering from the University of Pennsylvania in 1985, and the DPhil degree in computer vision from the University of Oxford in 1991. His research interests are in computer vision and robotics and include the recovery of motion and 3D shape of visible surfaces from image sequences, visual tracking and navigation, robot hand-eye coordination, algebraic and geometric invariants for object recognition and perceptual grouping, novel man-machine interfaces using visual gestures, and visual inspection. He has authored three books, edited six volumes, and coauthored more than 200 papers. He is a member of the IEEE and the IEEE Computer Society.