# View and Style-Independent Action Manifolds for Human Activity Recognition

Michał Lewandowski, Dimitrios Makris, and Jean-Christophe Nebel

Digital Imaging Research Centre, Kingston University, London, United Kingdom
{m.lewandowski,d.makris,j.nebel}@kingston.ac.uk
http://dircweb.king.ac.uk/

**Abstract.** We introduce a novel approach to automatically learn intuitive and compact descriptors of human body motions for activity recognition. Each action descriptor is produced, first, by applying Temporal Laplacian Eigenmaps to view-dependent videos in order to produce a stylistic invariant embedded manifold for each view separately. Then, all view-dependent manifolds are automatically combined to discover a unified representation which model in a single three dimensional space an action independently from style and viewpoint. In addition, a bidirectional nonlinear mapping function is incorporated to allow projecting actions between original and embedded spaces. The proposed framework is evaluated on a real and challenging dataset (IXMAS), which is composed of a variety of actions seen from arbitrary viewpoints. Experimental results demonstrate robustness against style and view variation and match the most accurate action recognition method.

**Keywords:** action manifold, activity recognition.

## 1 Introduction

Since video recording devices have become ubiquitous, the automated analysis of human activity from a single video is now an essential area of research in computer vision. Applications for such technology include video surveillance, indexing of film archives, sports video analysis and human-computer interactions.

Variability in human shape, appearance, posture and individual style in performing some motion makes the unified description of a given action difficult. In addition, camera view, perspective and scene environment have a critical impact on the aspect of recorded data. Consequently, the task of action recognition from a single video is extremely challenging. In this paper, we propose a solution which deals with this complexity within a single powerful framework. It allows accurate action recognition from a single uncalibrated camera in a fully automatic approach which exhibits high robustness to action style and view variation.

Previous work in this field falls into two categories: view-dependent and view-independent approaches. View-dependent methods assume that all actions are recorded from a fixed viewpoint [3,7,9,1]. The standard approach uses temporal templates to represent an action by encoding the history of silhouette deformation over time [3]. Actions were also described in the space-time domain. Local

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 547–560, 2010.

space-time features were extracted from the volumetric space-time action shape derived from sequence silhouettes by solving the Poisson equation [7]. Alternatively, the structure of local 3D patches was analysed by extending interest points into the spatio-temporal domain [9]. Moreover, by taking into account dynamics, action descriptors were defined in terms of chaotic invariant features from joint tracking [1]. Although these approaches have proved very accurate, the fact they rely on videos captured from a specific view limits their practicality in real world scenarios.

As a consequence, many researchers focused on multiple camera systems to achieve view-invariant action recognition. For instance, 2D temporal templates were extended into 3D motion history volumes [27]. If point correspondences between actions are assumed to be known, then either epipolar geometry [29] or projective invariants of coplanar landmark points can be exploited [19]. The main drawback of these methods is that, since they all require multiple cameras setups, they can only be applied in a controlled environment.

More recently, research has tackled the task of action recognition from an arbitrary view, i.e. from a single video, where multi camera data are used for training. Typically, a database of exemplars from different views is created to recognise actions based on the best matching score. Although silhouettes can be used to represent an action, their intrinsic ambiguity leads to high density sampling of the view space to obtain accurate results [18]. In contrast, richer action descriptors based on 3D exemplars represented by visual hulls and hidden Markov model allow reducing significantly the size of action templates [25]. Consequently, matching between observation and exemplars has to be performed in 2D by projecting visual hulls. Since such projection from high dimensional space to low dimensional is multimodal, it impacts on the quality of the recognition rate [25]. Junejo et al. [8] proposed to represent image sequences using self-similarity based descriptors which are fairly stable under view variation and characterises well the dynamics of the scene. However, this approach relies on the rough localisation and tracking of people in the video [8]. In [28], a video is represented by a combination of 3D visual hulls with spatio-temporal volumes to build 4-dimensional action feature models. Alternatively, a video can be described as a bag of spatio-temporal features called video-words (BOW) by quantising extracted 3D points of interest [16]. Initially, a SVM was trained on BOW to recognise actions [16], but this feature was also extended with a bag of spin-images [15]. Although these schemes perform accurate action recognition, the absence of continuous action model limits their applicability.

The methods most closely related to our approach model activities by reducing dimensionality of each sequence to obtain view-invariant manifold representations [21,6,5]. [21] used R-transform as a descriptor and Isomap [23] for dimensionality reduction, whereas [5,6] chose implicit distance function representation and locally linear embedding [22]. In these approaches [21,5], generative view-independent functions are designed to interpolate between intermediate views. This generative function was also extended to handle stylistic variation of data [6,5]. However, due to the limitations of the chosen dimensionality reduction

methods, none of these approaches managed to produce consistent style invariant representations, i.e. representations which are valid for a variety of individuals. Consequently, the accuracy of their systems was limited. This problem was addressed be applying non-rigid transformation [17] to artificially unify manifold representations of different people [21,6]. However, since such transformation affects manifold geometry, they may no longer reflect relationships between points in the high dimensional space. Alternatively, in [5] the topological structure of a torus was artificially constrained on the manifold to explicitly deal with stylistic variation instead of being learned from the data.

The main contribution of this paper is a new continuous view and style invariant action descriptor in a form of an Action Manifold. The proposed descriptor overcomes above limitations, since, not only, it is obtained automatically from labelled training data, but it encapsulates both style and view in a coherent torus-like two-dimensional manifold. The novel procedure used for generating torus-like descriptors takes advantage of several advanced techniques which have never been used in a view independent action recognition. They include Temporal Laplacian Eigenmaps [14] (TLE), Decomposable Generative Model [12] and Poisson Equation [7]. In addition, the method used for determining repetition neighbourhood in the TLE algorithm has been refined to handle for complex and dynamic videos of human actions. Finally, our descriptors are validated in a challenging real-life scenario of a view independent action recognition.

The structure of this paper is organised as follows. First, we describe our framework. This includes the processes of view-dependent discovery, view-independent manifold construction and mapping and a brief description of the dimensionality reduction algorithm. Secondly, the framework is validated quantitatively on a real dataset of human actions. Finally, conclusions and future work are presented.

## 2 View and Style-Independent Action Manifold

An action can be implicitly defined by a set of videos of a variety of people performing similar movements seen from different cameras. In our work, we aim to produce a single compact and informative model, i.e. action manifold, which represents an action independently from camera views and individuals' styles.

In our framework, the set of videos defining an action includes a variety of individuals, each of them captured on their own by a set of calibrated and synchronised cameras. Moreover, for each action, a video is labelled as a good representative; usually it is captured from a side view. We do not impose restrictions regarding video length variability for a given action and an individual may perform an action several times.

Let $Y$ denote the set of $N$ videos defining an action performed by different people and captured from different views. For a given view, action repetitions and variability of people define action style. Therefore, $Y$ can be defined as $Y = \{Y^{sv}\}_{(s=1..N_s, v=1..N_v)}$, where $v$ denotes the view class index and $s$ is the style index. Each frame $y$ of video is represented by $D$ pixels: $Y^{sv} = \{y_i^{sv}\}_{(i=1..T^{sv})}$,

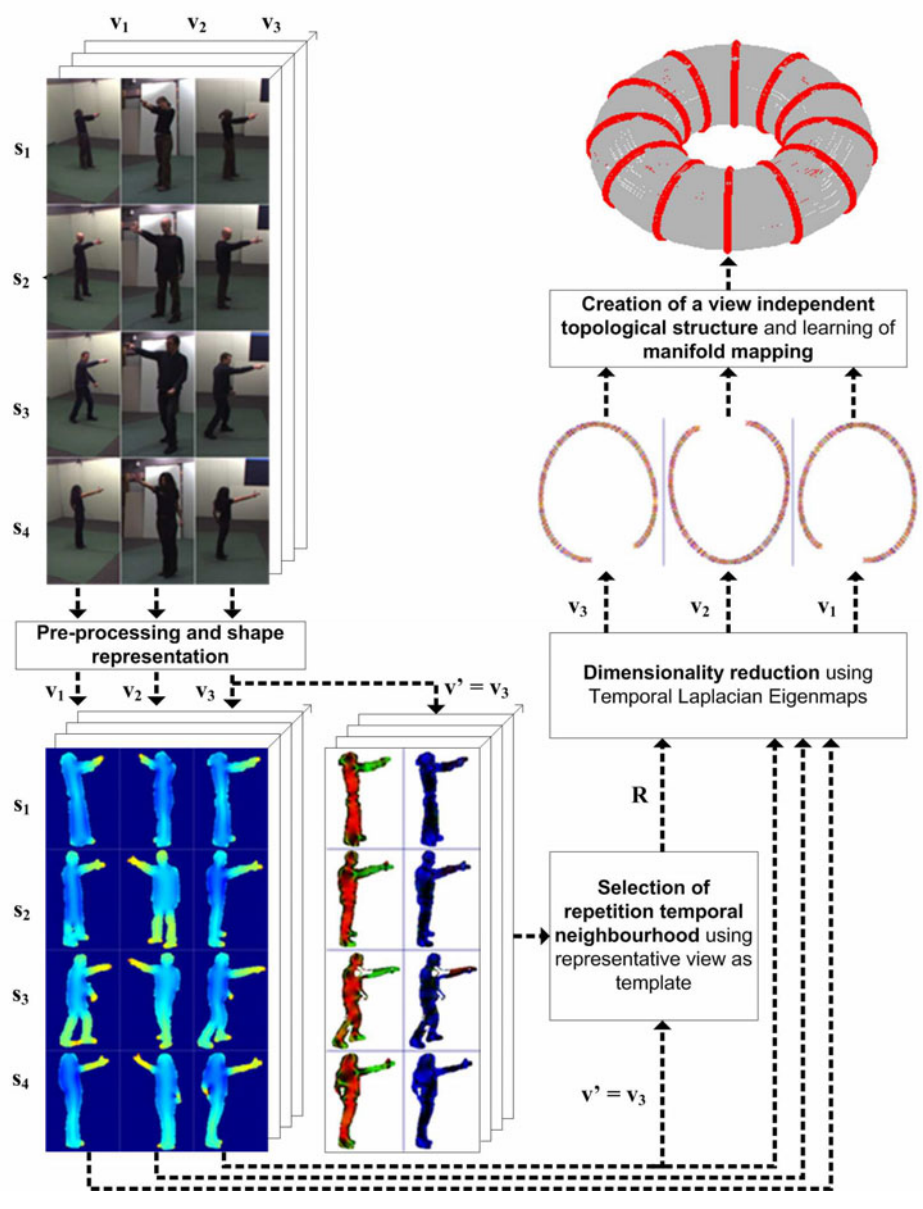


**Fig. 1.** Description of the action recognition framework for the "point" action

$y_i^{sv} \in R^D$, where $T^{sv}$ is the number of frames in the sequence. Fig. 1 summarises the processing pipeline used to produce a unified and compact action model, $X$, of dimension $d \ll D$, defined by $X = \{X^{sv}\}_{(s=1..N_s, v=1..N_v)}$, where $X^{sv} = \{x_i^{sv}\}_{(i=1..T^{sv})}$ and $x_i^{sv} \in R^d$.

Our algorithm is divided into two parts. First, view-dependent analysis of action data generates a style invariant action model for each view. This is performed using Temporal Laplacian Eigenmaps, a dimension reduction algorithm with excellent generalisation properties [14]. Then, these models are combined to learn a single compact and view invariant generative model of the action using generative decomposable model [12]. Fig. 1 provides an overview of our method.

### 2.1 View-Dependent Manifold

**Pre-processing and Shape Representation.** A frame $y_i^{sv}$ is generally defined by grey scale or colour pixel values. This very high dimensional description makes the process of learning an activity model from a frame sequence costly and inaccurate. However, many studies [18,25,5,6,12] have revealed that a binary representation of moving objects, i.e. silhouettes, are sufficient to capture the activity described by a frame sequence. Consequently, we adopt this approach in our framework.

We extract binary silhouettes $y_i^{sv}$ from each video by a standard background subtraction technique which models each pixel as a Gaussian in RGB space [27]. When videos consist of multiple instances of a given motion, temporal segmentation is required to extract elementary motion segments $Y^{sv}$ [26,4].

All silhouettes are normalised to deal with translation and scale variations by using the largest silhouette square bounding box available within the entire action dataset. In order to improve the quality of the normalised silhouettes, two morphological operations, i.e. bridge and open, and a median filter are applied. Lengths of all sequences $Y^{sv}$ are also normalised to match the length of the shortest sequence $T$' in the set $Y$ using the standard bicubic spline interpolation technique.

A sequence of binary silhouettes can be considered as a space-time shape surrounded by a closed surface [7]. This allows representing each silhouette by a local space-time saliency feature extracted from the solution of the Poisson equation of the corresponding volumetric surface, which takes into account the time domain [7]. This representation assigns highest gradient values within fast moving limbs which are much more informative for identifying actions, whereas torso has relatively smaller values inside (Fig. 1). As a consequence, such descriptor is significantly more powerful than binary representation [7] and essential, as it will be shown later, in the procedure allowing the selection of the TLE repetition neighbourhoods.

**Dimensionality Reduction.** Even with the generation of the previously described shape descriptor, the high dimension of $Y$ remains unsuitable for analysis. Consequently, we propose to produce an informative and unified model of the action using a nonlinear dimensionality reduction method. However, most of these techniques [23,22,2,11] cannot handle large variations within a dataset such as an action performed by different people. As a result, they tend to capture the intrinsic structure of each manifold separately without generalisation. Consequently, the common embedded space shows separate and highly distorted

manifolds. To deal with this fundamental issue, in this work we use the TLE algorithm which shows excellent generalisation properties [14].

TLE is an unsupervised nonlinear method for dimensionality reduction designated for time series data. It aims to preserve the temporal structure of data manifolds by introducing the concept of simultaneous exploitation of two types of neighbourhood graphs, which express implicitly temporal dependencies between data points. In our framework both graphs are constructed for the view $Y^{v\prime}$ which was labelled as a good representative. Each graph is based on a different definition of neighbour:

a. Adjacent temporal neighbours ($A$): the next and previous closest points in the sequential order of input.
b. Repetition temporal neighbours ($R$): the points similar to input but extracted from the different repetitions of activity which may vary in style. The number of $R$ neighbours should match the number of styles $N_s$ contained in the training set $Y^{v\prime}$.

The process of dimensionality reduction can be summarised briefly by the following steps. First, view-dependent weights $W^v$ are assigned to the edges of graph $G' \in \{A, R\}$ to construct graphs for all views $G^v$ using the standard LE formulation [2]. Then for each view the extended cost function is defined to combine information from both graphs:

$$argmin_{X^v}((X^v)^T L_A^v X^v + (X^v)^T L_R^v X^v) \tag{1}$$

$$subject\ to\ (X^v)^T D_A^v X^v + (X^v)^T D_R^v X^v = I \tag{2}$$

where $D^{v,G} = diag\{D_{11}^{v,G}, D_{22}^{v,G}, , D_{T^vT^v}^{v,G}\}$ is a diagonal matrix with entries $D_{ii}^{v,G} = \sum_{j=1}^{T^v} W_{ij}^{v,G}$, and $L_G^v = D^{v,G} - W^{v,G}$ is the Laplacian matrix. The minimum of the objective function can be found by applying Lagrange multipliers to Eq. 1 subject to the constraint expressed by Eq. 2 and solving the generalised eigenvalue problem:

$$(L_A^v + L_R^v)X^v = \lambda(D_A^v + D_R^v)X^v \tag{3}$$

The embedded space $X^v$ is spanned by the eigenvectors given by the $d$ smallest nonzero eigenvalues $\lambda$ ($d = 2$). The output of this stage is a view-dependent and style-independent one-dimensional action manifold $X^v$ (Fig. 1 and 2b).

**Selection of Repetition Temporal Neighbourhood.** The size of the repetition neighbourhood corresponds to the number of times an activity is repeated in the training set. Although video lengths were normalised for each action, it cannot be assumed that these videos are synchronous for two reasons. Firstly, they may start on different posture and, secondly, due to style variations, there may not be frame to frame correspondences between two action instances. Consequently, the estimation of the size and location of the repetition neighbourhood is essential. We automatically determine the optimal repetition neighbourhood by adopting the action detection procedure proposed in [7]. This schema is used

to find similar motion patterns in each sequence of the training set from which $R$ neighbours can be extracted (see lower part of Fig. 1).

First, the local space-time saliency shape descriptor defined in section 2.1 is extended with a local space-time saliency feature which is composed of 6 local space-time orientation attributes [7]. This allows indentifying regions with vertical, horizontal, and temporal "plates" and "sticks" within body and define orientation local features. Fig. 1 illustrates an example of "plate" and "stick" local features for a good representative view. Blue, red, and green colour regions correspond to temporal, horizontal, and vertical directions of local "plates" and "sticks" [7].

In the next step, a space-time cube is associated to each frame $y_i^{v\prime}$ in a sequence $Y^{v\prime}$ by sliding a warping window in time. The cube, i.e. the global space-time descriptor, combines local shape and orientations features using weighted moments of the form [7]:

$$m_{oqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(p_x, p_y, t)g(p_x, p_y, t)p_x^o p_y^q t^r dp_x dp_y dt \tag{4}$$

where $p_x$,$p_y$ are pixels coordinates, $g(p_x, p_y, t)$ denotes the characteristic function of the space-time shape, $w(p_x, p_y, t)$ is one of the seven possible weighting functions which corresponds to local features. As suggested in [7], spatial and time moments are considered up to order $o + q \leq 2$ and $r \leq 2$ respectively. Each space-time cube is centred around its space-time centroid and uniformaly scaled to preserve spatial aspect ratio.

Secondly, we calculate the matrix $M$ $(N_v \times N_v)$ of Euclidean distances between all space-times cubes among all sequences for a particular view. To emphasise continuity and temporal coherence of the underlying action between sequentially adjacent points in time, we perform temporal windowing of matrix $M$ by averaging distances through time within boundaries of each sequence. This implicitly leads to introducing a temporal history into each data point.

Finally, for each cube we look for the most similar motion pattern in each different repetition of activity based on $M$. The centre point of each most similar space-time cube becomes a repetition neighbour.

Because of possible substantial differences in speed and imperfect segmentation of action, the repetition neighbours may still not align coherently along time what may result in distortions in the embedded space. To address this problem, we incorporated a neighbourhood refinement procedure. In principle, we accept only these $R$ neighbours for given point $P$ which are within specific range from a corresponding point in each other sequence:

$$R' = \{P_{(i-1)*T+1} - T' \leq R_j \leq P_{iT} + T'\}, i = 2..N_s, j = 1..N_s \tag{5}$$

where $T'$ is defined as 10% of the normalised sequence length $T$. As it was mentioned earlier, the entire procedure is performed only once per action for the most discriminative view, because the temporal structure of an action is not view-dependent.

### 2.2 View-Independent Manifold

**Generation of a View-Independent Topological Structure.** Discovery of a compact representation of any human activity requires modelling both the view and body configuration jointly in a single space. Here we assume that human motion is observed from different viewpoints along a view circle at fixed camera height. Although such cylindrical setting appears limited, its robustness to view elevation variations, up to 45 degrees as shown in experimental section, makes it appropriate for many real life applications such as visual surveillance and sport analysis [5]. it is important to note that this configuration is not critical to our framework since it can easily be extended to a full view sphere-like model using training videos captured from different camera heights.

In section 2.1 style invariant body configuration manifolds were discovered for each view. Since the embedded spaces share the same topology regardless of the view, see Fig. 1 and 2b, for a given posture there is a unique correspondence on each of these manifolds. Consequently, the connection of those corresponding points in the order of view angle values creates a closed one dimensional manifold (topologically equivalent to a circle) which is the view-independent embedded space of the posture. Therefore, we define the unified representation of an activity as the combined space of the two sets of continuous one dimensional manifolds, i.e. posture and view, which are placed orthogonally to each other.

The process of producing the unified manifold comprises two steps. First, the view-dependent representations are combined: the embedded spaces $X^v$ are aligned with respect to a good representative $X^{v'}$ using Procrustes analysis [24]. Since this is a rigid transformation of the spaces, the internal structure of each manifold is not changed. Secondly, each embedded representation $X^v$ is aligned into a three-dimensional structure according to the view angle parameter $\mu^v \in [0, 2\pi]$. The outcome of this procedure reveals a torus-like structure which encapsulates both style and view (Fig. 1 and 2c). We called this structure a view and style-independent action manifold. This result is in line with previous work [5], where the usage of a torus is justified as an ideal representation for modelling

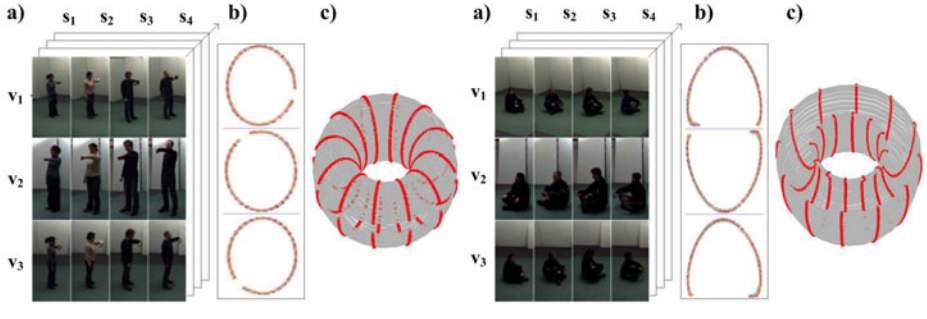


**Fig. 2.** Training results for quasi periodic action "check watch" (left) and non periodic action "sit down" (right): a) training videos; b) style-independent low dimensional representation for each view; c) style and view-independent manifolds

both the viewpoint and the body configuration of different activities. However, while, in that work, the topological correspondence between data points Y and an ideal torus is artificially enforced, in our approach, the torus-like representation reflects the temporal structure of the view-dependent data. Therefore, in our approach all types of motions, i.e. periodic, quasi-periodic and non-periodic, see Fig. 2, can be handled using the same framework.

### 2.3 Manifold Mapping

**Mapping Function.** In the previous section, view descriptors have been combined to form a unique view-independent action manifold. Since TLE is a spectral dimensionality reduction method, there is no mapping function between initial and embedded spaces. However, the ability to project data points from one space to the other is required for classification.

In order to provide a single projection function which allows dealing not only with stylistic variations, but also view changes, a decomposable generative model is learned [12]. This model aims at separating the intrinsic action configuration from other factors such as the motion style and view. Following [12] approach, the generative mapping function is modelled using three factors:

- Content $C$: a representation of the intrinsic body configuration which characterises motion as a function of time. It is invariant to either person or view.
- Style $S$: a time-invariant person parameter which describes the person appearance, shape and motion style.
- View point $V$: a time-invariant view parameter which characterises the view point from which the performed action is captured.

In our framework, content is represented by a continuous manifold while style and view are represented by the discrete classes present in the training data. For the last two factors, intermediate states can be interpolated. As a result, we are able to approximate view and style continuity. In addition, we assume that both style and view factors are time-invariant, i.e. both parameters remain constant during any instance of an action.

The procedure of fitting the decomposable generative model to the data consists of two steps. First, a set of style and view-dependent functions is trained. Then, all functions are combined into a single style and view-independent projection function.

Since mapping between the embedded manifold and the original space is highly nonlinear, generalised Radial Basis Function network [12] is applied to provide the nonlinear view-dependent mapping. It is expressed by $N_s$ style-dependent mapping functions:

$$y^{sv} = B^{sv} * \Psi(x^{sv}) \tag{6}$$

where $B$ is a $D \times E$ matrix of mapping coefficients. The kernel vector $\psi(\mathbf{.})$ is defined by:

$$\Psi(x^{sv}) = [\Phi(\parallel x^{sv} - z_1 \parallel)..\Phi(\parallel x^{sv} - z_E \parallel)\ 1\ x^{sv}]^T \tag{7}$$

where $Z = \{z_i\}_{(i=1..E)}$ is a set of distinctive representative points in each embedded space and $\phi(\centerdot)$ is a radial basis function; here we use a thin plate spline. $B^{sv}$ is calculated by applying the Moore-Penrose pseudo-inverse on matrix $\psi(X^{sv})$ and solving a linear system of equations: $B^{sv} = Y^{sv} * \psi(X^{sv})^{+}$ like in [12]. The set $Z$ is obtained by calculating a mean style and view manifold, which is then transformed by a non-rigid point registration procedure, called Coherent Point Drift [17], to better fit the data.

The final view-independent decomposable generative model is obtained by multi-linear tensor analysis in the space of nonlinear mapping coefficients [12]. Each coefficient matrix $B^{sv}$ is represented as the coefficient vector $b^{sv}$ of dimensionality $N_e = D * E$ by column wise stacking (columns of the matrix are concatenated to form a vector). Afterwards, all coefficient vectors $b^{sv}$ are arranged in an order three coefficient tensor $B$ whose dimensionality is $N_s \times N_v \times N_e$. The view and style orthogonal factors are decomposed from the assembled coefficient tensor $B$ using higher order singular value decomposition [10]:

$$B = C \times_1 S \times_2 V \times_3 F = G \times_1 S \times_2 V \tag{8}$$

where $S$ $(N_s \times N_s)$ is the mode-1 basis of $B$, which represents the orthogonal basis for the style space. Similarly, $V$ $(N_v \times N_v)$ is the mode-2 basis matrix which spans the space of viewpoint parameters and $F$ $(N_e \times N_s * N_v)$ represents the mode-3 basis for the mapping coefficient space. $C$ is a core tensor $(N_s \times N_v \times N_e)$ which governs the interactions between orthogonal factors represented in mode basis matrices. Coefficient eigenmodes $G$ is a new core tensor formed by $G = C \times_3 F$ whose dimensionality is $N_s \times N_v \times N_e$. Mode-i is a tensor product as defined in [10]. As the result, view-independent and style-independent projection function is expressed by equation $y = B * \Psi(x)$.

**Action Recognition.** The task is performed by projecting a motion sequence into each action descriptor using the generative decomposable model presented in the previous section. Then, the dynamic time warping distance [20] is calculated to measure similarity between actions.

Given a new instance of action $\tilde{Y}^{sv}$, its length is first normalised as described in section 2.1. Then the embedded coordinates $\tilde{X}^{sv}$ of the new action are obtained by least square solution of the following nonlinear system:

$$argmin_{B\Psi} \parallel \tilde{Y}^{sv} - \tilde{B}^{sv}\Psi(\tilde{X}^{sv}) \parallel \tag{9}$$

It's minimum solution can be found by determining and optimising coefficient matrix $\tilde{B}^{sv}$ given a learned model and then projecting data by solving a linear system of equations using the Moore-Penrose pseudo-inverse :

$$\Psi(\tilde{X}^{sv}) = (\tilde{B}^{sv})^{+} * \tilde{Y}^{sv} \tag{10}$$

Coordinates of $\tilde{X}^{sv}$ are provided by the last $d$ rows of the matrix $\Psi(\tilde{X}^{sv})$. In order to determine the optimal coefficient matrix $\tilde{B}^{sv}$, we adopt an iterative procedure [12]. First, we calculate a mean view manifold $Z$ over all aligned

mean styles manifolds $Z^v$ to obtain a homeomorphic manifold [12]. Then, the coefficient matrix is initialised by solving the following equation:

$$\tilde{B}^{sv} = \tilde{Y}^{sv} * \Psi(Z)^{+} \tag{11}$$

Let's $\tilde{b}^{sv}$ denote a vector obtained by column wise stacking of matrix $\tilde{B}^{sv}$. Then given a mapping model as described in the previous section and any style vector, $\tilde{s}$, and any view vector $\tilde{v}$, we can define a coefficient vector $\tilde{b}^{sv}$ by the tensor product $b^{\tilde{s}\tilde{v}} = G \times_1 \tilde{s} \times_2 \tilde{v}$.

Mapping coefficients $\tilde{b}^{sv}$ can be optimised to reflect style and view of a new instance action $\tilde{Y}^{sv}$ by minimising the following error:

$$argmin_{\tilde{s}\tilde{v}} \parallel b^{\tilde{s}\tilde{v}} - G \times_1 \tilde{s} \times_2 \tilde{v} \parallel \tag{12}$$

where $G$ is derived from learning (equation 8). Since tensor $G$ represents the intrinsic body configuration 'content' of the considered action and manages interactions between all factors, an accurate solution for style and view can only be reach for the same action.

If the style vector, $\tilde{s}$ is known we can obtain a closed form solution for $\tilde{v}$ and vice versa. This leads to an iterative procedure for estimating $\tilde{s}$ and $\tilde{v}$ simultaneously until equation 12 converges [12]. In practice, we follow Lee's approach where $\tilde{s}$ is initialised with a mean style estimate. Since the view classes are discrete, we identify the closest view class and use it to estimate $\tilde{s}$. Finally, vector $\tilde{b}^{sv}$ is unstacked to create matrix $\tilde{B}^{sv}$; then the action $\tilde{Y}^{sv}$ is embedded into the low dimensional space using equation 10.

## 3 Experimental Results

### 3.1 Experimental Setup

The proposed framework was validated on the publicly available multi-view IXMAS dataset [27,25], which is considered as the benchmark for action recognition methods. Since the 'throw action' is not performed by all subjects, we excluded it from our experiments. As a result, the chosen dataset is comprised of 12 actions, performed 3 times by 12 different actors. Each of these 432 activity instances was recorded simultaneously by 5 calibrated cameras, and a reconstructed 3D visual hull is provided. In this dataset, actors' positions and orientations are arbitrary since no specific instruction was given during acquisition. As a consequence, the action viewpoints are arbitrary and unknown.

To obtain a dense set of action descriptors regarding viewpoints for training, we followed [21] approach where the animated visual hulls are projected onto 12 evenly spaced virtual cameras located around the vertical axis of the subject. In line with other experiments made on this dataset [16,15,28], the top view was discarded for testing.

Experiments are conducted using the leave-one-out strategy followed by [28,8,25,21]. In each run, we select one actor for testing and all remaining subjects for training. Two testing schemes were used: recognition using single view,

and recognition using multiple views. In the recognition from multiple views, a simple majority voting rule was applied [16,15]. Finally, performances were compared to the other state of art methods. Unfortunately, results could not be compared with [21], because, instead of evaluating their method with original video data, they did it by using projections of the visual hulls.

### 3.2 Performances

Although different approaches may use slightly different experimental settings, table 1 shows that our framework produces state of art performances. Accuracy rates obtained for an experiment aiming at only 11 actions, i.e. the 'point' action was not considered, reveals that we outperform all methods targeting this task [28,8,25] even if they considered a smaller set of subjects [8,25].

When all actions completed by all subjects are considered, i.e. 12, our framework displays results which are significantly better than Liu [15] and match those obtained by Liu [16]. Although performance alone cannot discriminate between Liu's and our method, we believe that our action models are superior. Indeed, unlike Liu's descriptors which are based on codebooks, ours consists of single integrated continuous models. Consequently, our action manifolds can be applied to many applications beyond action recognition such as synthetic action sequence generation, style recognition and camera view estimation.

Fig. 3 depicts the confusion matrix of recognition for the 'all-view' experiment. It reveals that our framework performed better when dealing with motions involving the whole body, i.e. "walk", "sit down", "get up", "turn around" and "pick up". Since temporal information is essential when dealing with highly dynamic motions and TLE aims at preserving temporal structure in each view, action manifolds of those activities are more representative. The best recognition rates 74.8%, 80.3% are obtained for camera 2 and 4 respectively. This was expected, since both views are the most similar among those used for training. Moreover, when dealing with either different, i.e. camera 1, or even significantly different views, i.e. camera 3, our framework still achieves reasonable recognition, i.e. 71.7% and 65.9% respectively. Details about average accuracy per camera can be found in supplementary material [13].

**Table 1.** Average recognition accuracy over all cameras (top view excluded) using either single or multiple views for testing

| % | Subjects | Actions | Average Accuracy | |
|---|---|---|---|---|
| | | | Single view | All views |
| Weinland [25] | 10 | 11 | 63.9 | 81.3 |
| Yan [28] | 12 | 11 | 64.0 | 78.0 |
| Junejo [8] | 10 | 11 | 74.1 | - |
| **Our** | 12 | 11 | **75.0** | **83.1** |
| Liu [15] | 12 | 13 | 71.7 | 78.5 |
| Liu [16] | 12 | 13 | 73.7 | 82.8 |
| **Our** | 12 | 12 | **73.2** | **83.1** |

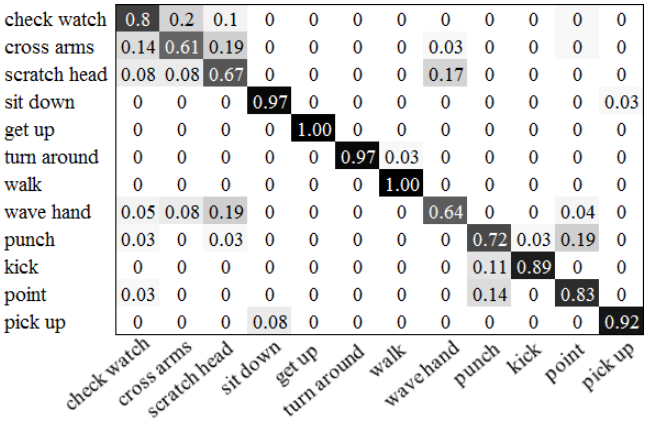


**Fig. 3.** Class-confusion matrix using multiple views. The average performance is 83.1%.

## 4 Conclusion

This paper introduces a novel human action recognition framework for arbitrary individuals and views. Its main contribution is a procedure for learning discriminative and unified action descriptors, which reside in a low dimensional space. These descriptors are constructed automatically by taking advantage of the TLE algorithm and a generative decomposable model. Performance of the proposed methodology has been evaluated using the IXMAS dataset and competitive results have been demonstrated. In addition, since our procedure to produce manifold based descriptor is general, it can be applied to many applications beyond action recognition such as visual surveillance or sport analysis. We plan to investigate some of these directions in future work.

**Acknowledgments.** The authors would like to thank Lena Gorelick from University of Western Ontario and Richard Souvenir from University of North Carolina at Charlotte for sharing their codes.

## References

1. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: ICCV, pp. 1–8 (2007)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. NIPS 14, 585–591 (2001)
3. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. TPAMI 23(3), 257–267 (2001)
4. Cutler, R., Davis, L.: Robust real-time periodic motion detection, analysis, and applications. TPAMI 22(8), 781–796 (2000)
5. Elgammal, A., Lee, C.S.: Tracking people on a torus. TPAMI 31(3), 520–538 (2009)
6. Elgammal, A., Lee, C.S.: Separating style and content on a nonlinear manifold. In: CVPR, vol. 1 (2004)

7. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. TPAMI 29(12), 2247 (2007)
8. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: Cross-view action recognition from temporal self-similarities. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 293–306. Springer, Heidelberg (2008)
9. Laptev, I.: On space-time interest points. IJCV 64(2), 107–123 (2005)
10. Lathauwer, L., Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21(4), 1253–1278 (2000)
11. Lawrence, N.: Gaussian process latent variable models for visualisation of high dimensional data. In: NIPS, vol. 16 (2004)
12. Lee, C., Elgammal, A.: Homeomorphic manifold analysis: Learning decomposable generative models for human motion analysis. In: Vidal, R., Heyden, A., Ma, Y. (eds.) WDV 2005/2006. LNCS, vol. 4358, pp. 100–114. Springer, Heidelberg (2007)
13. Lewandowski, M., Makris, D., Nebel, J.C.: Average recognition rates using single views. (2010); supplied as additional material, `avgrecrates.tif`
14. Lewandowski, M., Martinez-del-Rincon, J., Makris, D., Nebel, J.-C.: Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series. In: Proc. ICPR (2010)
15. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: CVPR (2008)
16. Liu, J., Shah, M.: Learning human actions via information maximization. In: CVPR (2008)
17. Myronenko, A., Song, X., Carreira-Perpinán, M.: Non-rigid point set registration: Coherent Point Drift. In: NIPS, vol. 19, p. 1009 (2007)
18. Ogale, A., Karapurkar, A., Aloimonos, Y.: View-invariant modeling and recognition of human actions using grammars. In: W. on Dyn. Vis. at ICCV, vol. 5 (2005)
19. Parameswaran, V., Chellappa, R.: View invariance for human action recognition. IJCV 66(1), 83–101 (2006)
20. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition (1993)
21. Richard, S., Kyle, P.: Viewpoint Manifolds for Action Recognition. EURASIP J. on Img. and Vid. Proc. 2009 (2009)
22. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
23. Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
24. Wang, C., Mahadevan, S.: Manifold alignment using Procrustes analysis. In: ICML, pp. 1120–1127 (2008)
25. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: ICCV, vol. 5(7), p. 8 (2007)
26. Weinland, D., Ronfard, R., Boyer, E.: Automatic discovery of action taxonomies from multiple views. In: CVPR, vol. 2, pp. 1639–1645 (2006)
27. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding 104(2-3), 249–257 (2006)
28. Yan, P., Khan, S., Shah, M.: Learning 4D action feature models for arbitrary view action recognition. In: CVPR, vol. 12 (2008)
29. Yilmaz, A., Shah, M.: Recognizing human actions in videos acquired by uncalibrated moving cameras. In: ICCV, vol. 1, pp. 150–157 (2005)