Structural Laplacian Eigenmaps for modelling sets of multivariate sequences

Michał Lewandowski, Dimitrios Makris, *Member, IEEE*, Sergio A. Velastin, *Member, IEEE*, and Jean-Christophe Nebel, *Senior Member, IEEE*

***Abstract*—A novel embedding-based dimensionality reduction approach, called Structural Laplacian Eigenmaps, is proposed to learn models representing any concept which can be defined by a set of multivariate sequences. This relies on the expression of the intrinsic structure of the multivariate sequences in the form of structural constraints which are imposed on dimensionality reduction process to generate a compact and data-driven manifold in a low dimensional space. This manifold is a mathematical representation of the intrinsic nature of the concept of interest regardless of the stylistic variability found in its instances. In addition, this approach is extended to model jointly several related concepts within a unified representation creating a continuous space between concept manifolds. Since a generated manifold encodes the unique characteristic of the concept of interest, it can be employed for classification of unknown instances of concepts. Exhaustive experimental evaluation on different datasets confirms the superiority of the proposed methodology to other state-of-the-art dimensionality reduction methods. Finally, the practical value of this novel dimensionality reduction method is demonstrated in three challenging computer vision applications, i.e. view-dependent and view-independent action recognition as well as human-human interaction classification.**

***Index Terms*—computer vision, machine learning, multidimensional, pattern analysis, time series analysis, video analysis.**

I. Introduction

In many perceptual tasks, such as speech, text, EEG-signal, gesture and action recognition, modelling of a class intrinsic characteristics is key to successful classification. In such cases, a spoken accent, font type, handwriting style or individual personality can be seen as stylistic variations of the class. Here, we call 'concept' a class which refers to a general phenomenon derived from a set of specific instances or occurrences recorded as multivariate sequences.

In this work, we aim at modelling any concept through a meaningful low dimensional representation. In such context, each observed instance is described by two independent factors termed 'content' and 'style', where 'content' is the invariant factor related to the essence of the concept and 'style' are variations of that concept between instances [1]. The objective is to generate a low dimension manifold which generalises a set of instances by representing mathematically the intrinsic nature of the concept of interest regardless of stylistic variability. For example, in the context of modelling a message containing a set of instructions (e.g. 'drop to the floor, seek cover under a piece of sturdy furniture and hold on tight'), spoken or written words are essential to communication of those instructions, so they are considered as content. On the other hand, accents, font types or handwriting styles are stylistic variations of the content, which should not change the message meaning; thus, they could be marginalised out of the model.

Since classification in high dimensional spaces is very challenging [2], dimensionality reduction methods have been used to address this problem. While they have proved efficient at reducing the complexity of many problems [3]–[8], they generally fail to produce a coherent representation of the class of interest when training data consist of instances varying significantly in terms of style [9]. To overcome this issue, we propose to model the content of a concept using the manifold generated by a dimensionality reduction process where constraints are imposed to reflect the intrinsic structure of multivariate sequences. Furthermore, we suggest an extension of this methodology where several concepts of similar nature are modelled jointly within a unified representation creating a continuous space between concept manifolds. An action that is observed from slightly different views or related activities, such as walking and running, are intuitive examples of similar concepts which share common structural information.

The structure of this paper is organised as follows. After discussion of related work in §2, we explain the fundamental principles of the proposed methodology in §3. Subsequently, in §4, we introduce the Structural Laplacian Eigenmaps (SLE) algorithm and its integration into a general action recognition framework. Afterwards in §5, first, SLE is validated qualitatively and quantitatively, and then performance of our action recognition framework is reported. Finally, discussion and conclusions are presented in §6 and §7 respectively.

II. Related Work

Dimensionality reduction is formally defined as the transformation or/and combination of the original multidimensional features in order to generate more informative, descriptive and practical data representation in a space of fewer dimensions [10]. This process is achieved by eliminating redundancies and irrelevant relationships present in datasets while ensuring maximum preservation of the original information. These techniques have proved an essential step in many machine learning applications in domains such as computer vision [11], computer graphics [12], robotics [13], speech recognition [14], data visualisation [7] and pattern recognition [15].

Although Principal Component Analysis (PCA) is a well known approach for dimensionality reduction [16], it fails to model nonlinear structures embedded in complex data. As a consequence, many dimensionality reduction algorithms able to deal with nonlinearity have been proposed. They

can be classified in two main categories: mapping-based and embedding-based approaches.

Embedding-based approaches either estimate the local [3], [4], [17] or global [5], [6], [18] structure of the underlying manifold by preserving some geometrical relationships between data points [3], [4], [6], [17], [18] or maintaining pair wise similarities between data points [5]. The geometrical constraints of Laplacian Eigenmaps (LE) [4], Locally Linear Embedding (LLE) [3] and Isomap [6] are expressed in the form of local neighbourhoods on a manifold. This idea is further extended by [19] in the context of LLE for a joint representation of multiple datasets with a common underlying manifold. This is achieved by assembling neighbourhoods not only within one manifold but also between different manifolds. As a consequence, some inter manifold correspondences are estimated which allows the embedding of all manifolds in a single space. In turn, Local Tangent Space Alignment [17] first explores the geometric relations between neighbouring data points in a local tangent space and then aligns them into a global coordinate system. A drawback of these geometrically motivated approaches is that they do not provide any mapping between low and high dimensional spaces which is necessary when dealing with unseen data. Moreover, they are very sensitive to the choice of neighbourhood size [10], [20]. Alternatively, kernel PCA [5] expresses the pair wise similarities between data points in the form of a kernel function. However, it is very sensitive to the choice of that function: each kernel generates a specific low dimensional structure whose performance is difficult to predict. When no a priori knowledge is available, the whole space of kernel functions would have to be explored in order to find the most suited to a particular task. To address that, Maximum Variance Unfolding [18] aims at learning kernel matrix from neighbourhood graph restrictions using semi-definite programming. Finally, [21] proposes the patch alignment framework and reveals that all these approaches intrinsically consist of only two steps: the different patch optimisation stage and an almost identical whole alignment stage.

On the other hand, mapping-based approaches, such as Gaussian Process Latent Variable Model (GPLVM) [7] and Generative Topographic Mapping (GTM) [22], use probabilistic nonlinear functions to map the embedded space to the data space. As a result, these methods approximate the underlying distribution of the observed space which, in turn, allows generalising the learned space to unseen data. However, their main limitation is their computational complexity which prevents their usage when dealing with large datasets [7]–[9]. Furthermore, since the GPLVM objective function is underconstrained in the general case [55], it is sensitive to local minima if the initialisation of the model is poor [8].

In addition to the intrinsic limitations of these two classes of dimensionality reduction methods, and despite the considerable amount of work which has been devoted to their development [3]–[7], [10], [19], [22], the specific nature of data is rarely taken into consideration. In particular, the sequential structure of multivariate sequence data should be preserved in their low dimensional representations.

The most well known representative of embedding-based approaches for modelling sequential data structure is a spatio-temporal extention of Isomap which alters empirically the original distance weights in the graph of local neighbours to emphasise similarity between temporal related points [23]. Although this method demonstrates the value of integrating sequential constraints, it is sensitive to the empirical selection of parameters. Some mapping-based approaches also integrate temporal information. Back-Constrained Gaussian Process Latent Variable Model (BC-GPLVM) includes temporal coherence constraints to ensure the smoothness of the mapping between spaces [24]. In addition, the GTM algorithm was extended to capture temporal dynamics of sequential data by incorporating this information as an emission density in a Hidden Markov model [25]. In contrast, Gaussian Process Dynamical Model (GPDM) and its variants integrate time information by associating nonlinear autoregressive dynamic model to the embedded space [26]. The addition of temporal information allows these approaches to produce smoother low dimensional representations, i.e. spaces with trajectories of points without gaps and jumps within a sequence. Nevertheless these methods are even more computationally expensive than their respective standard formulations [9], [24]–[26].

Although the issue of style has so far been of limited interested in the field of dimensionality reduction, a few attempts have been made to model stylistic variability of a concept in a low dimensional space [8], [11], [27]–[29]. However, in classification applications, content is essential to characterise the concept, whereas style should be marginalised out since it is an irrelevant and unhelpful factor. To date, no approach has attempted to tackle the classification problem by suppressing style in the low dimensional representation of the concept of interest. Contrary to the few methods tailored for dimensionality reduction of multivariate sequences [23]–[26], the presented approach takes into account not only the sequential structure of data but also correspondences between different instances of the concept. Consequently, a common underlying manifold can be learnt to represent the unique content of any concept, that is represented as a set of multivariate sequences. Finally, we extend the applicability of our dimension reduction method by proposing a unified representation between concept manifolds so that similar concepts, i.e. which share common structural information, can be modelled jointly within a continuous space.

III. Principle

In addition to low computational cost, spectral dimensionality reduction methods offer a powerful framework based on graph theory which is able to describe constraints between data points in the form of neighbourhood graphs [3], [4], [6]. By taking advantage of that framework, we propose to encode the structure of the concept using two types of constraints expressed by novel structural neighbourhoods for each data point in the multivariate sequences. In contrast with our previous work [9], the strong theoretical framework used to define our new methodology allows its generalisation to any sequentially ordered data. Note that we deal with the lack of mapping between low and high spaces by learning advanced mapping functions separately.

The first constraint which is introduced in the dimensionality reduction method focuses on preserving the sequential structure of a multivariate sequence in the low dimensional space. This constraint is expressed by a **sequential neighbourhood**, where data points are connected according to the order in the sequence. The second constraint aims at minimising stylistic variations displayed by different sequences or even within a sequence, if it contains repetitions. This is achieved by finding correspondences between and within sections of the multivariate sequences that exhibit high similarity. Then, within these regions, each data point is associated to corresponding points within the same sequence (intra) or in other sequences (inter). In this work, we will refer to these points expressing the second constraint as **intra and inter-sequence neighbours** respectively.

These two sets of constraints associated to each data point are then assembled into adjacent structural graphs to encapsulate effectively all sequential as well as mutual intra and inter dependencies of instances of multivariate sequences defining a single concept. Those graphs are then processed using an extended spectral dimensionality reduction scheme to generate a single low dimensional manifold which should model the concept of interest independently from style.

Since structural relationships may also exist between different concepts of similar nature, the inter-sequence neighbours can also be identified between such concepts. As a result, when the manifold for each concept has been generated independently, such correspondences can be employed to align these concept spaces into a single coherent representation of a meta-concept. Such meta-concept space, when associated with advanced mapping functions, allows extrapolating a model to instances of unknown concepts.

IV. Proposed Methodology

This paper presents Structural Laplacian Eigenmaps (SLE) which is a novel and efficient unsupervised nonlinear method for dimensionality reduction which learns data-driven manifolds designed for any concept which can be represented by a set of multivariate sequences. In addition, an extension of this schema is introduced to model jointly similar concepts in a continuous manner.

A concept is mathematically modelled by a set of P multivariate sequences $Y = \{\mathbf{y}_\zeta : 1 \leq \zeta \leq P, \zeta \in \mathbb{N}, P \in \mathbb{N}\}$ distributed on a manifold in some high dimensional space of dimension D, where a sequence is defined as $\mathbf{y}_\zeta = \{y_\zeta[t] : 1 \leq t \leq T_\zeta, t \in \mathbb{N}, T_\zeta \in \mathbb{N}, y_\zeta[t] \in \mathbb{R}^D\}$. Here t denotes a discrete sequential index, e.g. time, while ζ corresponds to different instances of the concept with various lengths T_ζ. In turn, N is the total number of samples in the dataset $N = \sum_{r=1}^{\zeta} T_r$. Given a set Y of such data sequences, SLE learns their low dimensional representation $X = \{\mathbf{x}_\zeta : 1 \leq \zeta \leq P\}$ of dimension d, where $\mathbf{x}_\zeta = \{x_\zeta[t] : 1 \leq t \leq T_\zeta, x_\zeta[t] \in \mathbb{R}^d, d \ll D\}$. This is achieved by modelling the intrinsic structure of the data sequence manifold instead of its local geometry as is the case with the standard Laplacian Eigenmaps.

Ideally, the reduced dimensionality d should correspond to the intrinsic dimensionality of the data so that the reduced space represents the observed properties of the data without information loss. The intrinsic dimensionality can be understood as the minimum number of independent variables needed to explain satisfactory a concept of interest [10]. Formally, from a geometrical point of view, a low dimensional representation of a concept is expected to be d-dimensional if the dataset elements lie entirely within a d-dimensional subspace of $\mathbb{R}^D$ [30]. The determination of d has been an active field of research where many approaches have been proposed (see survey [30]), and is beyond the scope of this paper.

The proposed methodology first constructs local patches from neighbourhoods around each data point based on structural sequences (§IV-A). Then, these individual local constraints are assembled into two sparse graphs to represent complementary relationships between all data points in the sequences (§IV-B). The first graph expresses the constraint of sequential consistency within each data sequence, whereas the second graph encodes mutual intra and inter region dependencies between sequences. These local constraints are appropriately modelled by taking advantage of a unique property of the standard LE framework: the preservation of relative distances between neighbourhood points in the low dimensional space. SLE extends this framework by introducing structural graphs which impose proximity relations between multivariate sequences. As a consequence, these graphs, when employed simultaneously to constrain the extended dimensionality reduction process (§IV-C), allow representing the intrinsic nature of the concept regardless of style. The whole pipeline is depicted in Fig. 1. Note that in the context of the patch alignment framework [21], SLE proposes a new patch optimisation stage; whereas the global alignment step is consistent with other embedded based approaches.

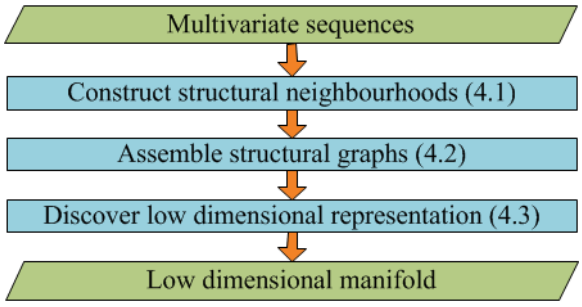

Fig. 1. Framework of Structural Laplacian Eigenmaps.

A. Construction of Structural Neighbourhoods

The content of a concept is extracted implicitly during the dimensionality reduction process which relies on preserving its structure using a set of constraints. They have the form of two sets of neighbourhoods for each data point. They express relationships within as well as between multivariate sequences. Let's define any point in dataset Y as $u_i = y_\zeta[t]$ where $u_i \in \mathbb{R}^D$ and:

$$i = \begin{cases} t \text{ for } \zeta = 1 \\ \sum_{r=1}^{\zeta-1} T_r + t \text{ for } \zeta > 1 \end{cases} \quad (1)$$

Then, the sets are defined as follows:

- **Sequential neighbours** (S): assuming a first-order Markov dependency between consecutive points, the sequential neighbours, S_i, of $y_\zeta[t]$ are the previous and subsequent closest point in the sequential order of Markovian process associated with the concept (green dots in Fig. 2):

$$S_i = \{y_\zeta[t-1], y_\zeta[t+1]\} \quad (2)$$

- **Intra and inter-sequence neighbours** (R): let's associate to each point $2s$ sequential neighbours which define a sequence fragment F_i. The intra and inter-sequence neighbours, R_i, of $y_\zeta[t]$ are the centres of the q_i sequence fragments, $F_{i,k}$, which are similar to F_i according to some mathematical criterion (magenta dots in Fig. 2):

$$R_i = \{F_{i,1}(C), ..., F_{i,q_i}(C)\} \quad (3)$$

where $F_{i,k}(C)$ returns the centre point of $F_{i,k}$, in the instance ξ of the concept. The intra neighbours are extracted from different fragments within the current sequence $\xi = \zeta$, whereas inter neighbours correspond to centres of fragments from different sequences $\xi \neq \zeta$.

The size of the intra and inter-sequence neighbourhood q_i corresponds to the number of times a local fragment is repeated, i.e. how many times the same subset of the content is available within the set of instances describing the concept of interest. The optimal intra and inter-sequence neighbourhood size as well as a selection of these neighbours are automatically determined. First, each data point, $y_\zeta[t]$, is associated to $2s$ sequential points to create the local fragment F_i:

$$F_i = \{y_\zeta[t-s], ..., y_\zeta[t-1], y_\zeta[t], y_\zeta[t+1], ..., y_\zeta[t+s]\} \quad (4)$$

In order to calculate similarity between local fragments F_i, a similarity function f, e.g. dynamic time warping (DTW) [31], is selected. Comparisons are then performed against all fragments created by sliding a warping window through the entire training set (j is defined according to eq. (1)):

$$M = f(F_i, F_j) \quad (5)$$

Afterwards, the obtained similarity matrix M is diagonally windowed by applying a moving average filter on distances between fragments using a history window of size $2s$:

$$m'_{i,j} = \frac{1}{2s} \sum_{b=0}^{2s-1} m_{i-b,j-b} \quad (6)$$

Finally, intra and inter-sequence neighbours R_i are identified by extracting the centres of similar frag-ments, $F_{i,\xi}$. These centres correspond to local minima in each row of neighbourhood similarity matrix $M' = \{m'_{i,j}\}$. More formally, the similarity is defined here as **b**, typically 1.5, standard deviations σ_i from the mean μ_i in each row i:

$$R_i = \{F_{i,j}(C) : m'_{i,j} < \mu_i - \mathbf{b}\sigma_i\} \quad (7)$$

Although, the process of sliding the warping window is time-consuming for large training datasets, it can be significantly reduced by using constraints during DTW computation such as Sakoe-Chiba band [32].

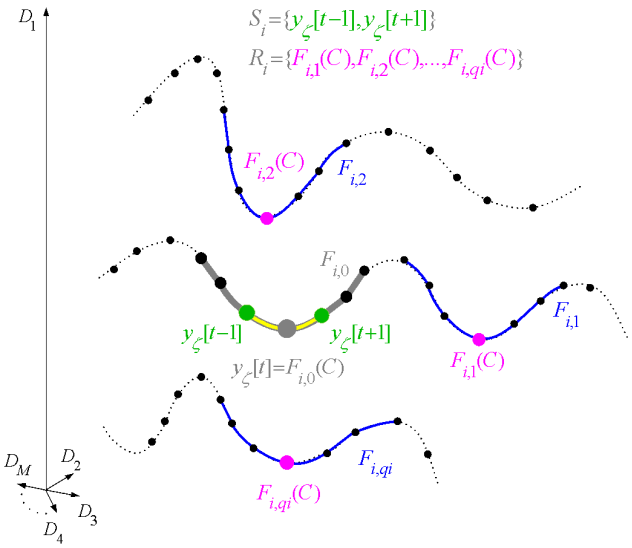

Fig. 2. Sequential (green dots) and intra and inter-sequence (magenta dots) neighbours of a given data point, $y_\zeta[t]$ (grey dot), where $s = 3$ (grey and blue sequences).

B. Assembling of Structural Graphs

The obtained structural neighbour relations are used for assembling two structural graphs $G = \{S, R\}$, where any two vertices in these graphs are connected only when a neighbourhood relation exists between these points. Weights W are assigned to the edges of each graph separately using the standard LE formulation:

$$W_G^{i,j} = \begin{cases} \exp(-\|u_i - u_j\|^2/\alpha) & \text{if } u_i, u_j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where α is a global coefficient, i.e. width of Gaussian kernel. In contrast to the standard LE graph obtained by the naive K-nearest neighbours procedure [1], SLE's graphs are derived from structural relationships. The conceptual difference between them is illustrated in Sup. 1. Neighbourhood connections defined in the Laplacian graphs impose point closeness in the embedded space. Consequently, the sequential neighbours allow modelling the sequential nature of successive data points into the resulting embedding. In turn, intra and inter-sequence neighbourhoods discard style variability. This is achieved by aligning the sequences in the embedded space, so that the intrinsic pattern of the concept is implicitly extracted.

C. Manifold Generation

Let's define any low dimensional point in dataset X as $v_i = x_\zeta[t]$ where $v_i \in \mathbb{R}^d$ and i,j are given by eq. (1). Following the standard LE formulation, we introduce an extended cost function to combine information from both structural graphs:

$$\begin{aligned} \varepsilon &= \frac{1}{2}\sum_{i,j} \|v_i - v_j\|^2 W_S^{i,j} + \frac{1}{2}\sum_{i,j} \|v_i - v_j\|^2 W_R^{i,j} \\ &= X^T L_S X + X^T L_R X \end{aligned} \quad (9)$$

where $L_G = Z_G - W_G$ is the Laplacian matrix derived from the corresponding graph $G = \{S, R\}$, while $Z_G = diag(Z_G^{1,1}, Z_G^{2,2}, ..., Z_G^{N,N})$ is a diagonal matrix whose entries are: $Z_G^{i,i} = \sum_{j=1}^{N} W_G^{i,j}$.

The objective of the dimensionality reduction process is to minimise eq. (9) with respect to the embedded coordinates X subject to constraints:

$$\text{argmin}_X \quad tr(X^T L_S X + X^T L_R X) \tag{10}$$

$$\text{subject to} \quad X^T Z_S X + X^T Z_R X = I \tag{11}$$

This formulation doesn't constrain manifold to be derived from closed topology, which allows modelling cyclic, quasi cyclic and noncyclic multivariate sequences. Since L_G is the positive semi-definite Hermitian matrix, the minimum of the objective function can be found analytically by applying Lagrange multipliers to eq. (10) subject to the constraint expressed by eq. (11):

$$(L_S + L_R)X = \lambda(Z_S + Z_R)X \tag{12}$$

where the corresponding eigenvectors x_i $(i = 1..D)$ ordered according to their eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_D$ satisfy:

$$\begin{aligned} (L_S + L_R)x_1 &= \lambda_1(Z_S + Z_R)x_1 \\ (L_S + L_R)x_2 &= \lambda_2(Z_S + Z_R)x_2 \\ &... \\ (L_S + L_R)x_D &= \lambda_D(Z_S + Z_R)x_D \end{aligned} \tag{13}$$

The final embedded space X is spanned by the eigenvectors x_i which correspond to the d smallest nonzero eigenvalues λ_i obtained by the solution of the sparse generalised eigenvalue problem (eq. (12)) [33] according to the generalisation of the Rayleigh-Ritz theorem [34]. Note the bottom $d+1$ eigenvectors of $L_S + L_R$ can be determined without performing a full matrix decomposition [35]. Moreover, the combined matrix is extremely sparse, which results in substantial computational savings. As a consequence, the complexity of the optimisation process is $O(pN^2)$, where N denotes the number of points in a dataset and p is the ratio of nonzero elements in a sparse matrix $L_S + L_R$ to the total number of elements N. Since the analytical optimisation process of SLE has the same complexity as the standard LE, the proposed methodology is computationally efficient especially in comparison to mapping based approaches with complexity of $O(I * N^3)$, where I denotes the number of iterations in an optimisation process.

According to the patch alignment framework [21], the proposed objective function eq. (10) can be decomposed over $m = 1..N$ patches within the patch optimisation step:

$$\text{argmin}_X \sum_{m=1}^{N} tr(X_m^T L_{S,m} X_m + X_m^T L_{R,m} X_m) \tag{14}$$

, whereas the second stage, i.e. whole alignment, follows the standard formulation:

$$\text{argmin}_X tr(X^T L X) \tag{15}$$

where $X_m = X S_m$, the S_m denotes the selection matrix [21], and $L = \sum_{m=1}^{N} S_m L_m S_m^T$ allows forming a global coordinate system using the alignment trick [17]. Finally, the L_m for each patch is defined by:

$$L_m = \begin{bmatrix} \sum_{j=1}^{q_m+2}(W_S^{m,m_j} + W_R^{m,m_j}) & -\overrightarrow{W_m}^T \\ \overrightarrow{W_m} & diag(\overrightarrow{W_m}) \end{bmatrix} \tag{16}$$

so that $\overrightarrow{W_m} = [W_S^{m,m_j} + W_R^{m,m_j}]_{j=1}^{q_m+2}$ is a vector weighted by 2 sequential and q_m intra and inter sequence neighbours in the local patch around sequence point m according to eq. (8).

D. Joint modelling of similar concepts

When several concepts share a similar content, the proposed schema can be extended to model jointly these concepts in a unified representation. Such model allows representing a meta-concept by approximating the continuity of the content space in a meaningful manner and consequently provides generalisation abilities to unknown instances of related concepts.

This is achieved by first reducing independently dimensionality of each concept space using SLE to d-dimensions. Then correspondences between concepts are estimated to align and join all individual concept spaces into a single coherent representation of meta-concept. Finally, the continuity of the obtained model is approximated by learning mapping functions between low and high dimensional spaces, for instance using a variant of the RBF network [11], [36], [37].

If correspondences between multivariate sequences are unknown, e.g. in walking and running activities, they can be inferred using our DTW-based procedure which estimates inter-sequence neighbours. Based on estimated relations, for any given manifold point, a single correspondence neighbour is chosen on each of other concept manifolds. Alternatively the required correspondences between similar concepts may be known apriori. For example in computer vision, when a specific action, i.e. concept, is observed from slightly different camera views, there is a unique correspondence between postures in each view-dependent manifold. The obtained correspondences allows aligning spaces in a single representation using any transformation which preserves the internal structure of the manifold, e.g. Procrustes analysis [38].

E. Integration into an Activity Recognition Framework

In order to demonstrate a practical application of the proposed methodology, it is integrated into an activity recognition framework able to handle challenging scenarios where data are captured from uncalibrated cameras. In the context of these scenarios, any specific action recorded from a given camera angle can be interpreted as a concept to model. In turn, an action observed from several cameras corresponds to a view-independent meta-concept.

A classic action recognition framework consist of three steps: extraction of features, which are often spatio-temporal [39]–[44], generation of action models and classification of unseen action instances. A variety of approaches has been proposed to produce action models. They include Hidden Markov Models [45], [46], Conditional Random Fields [47], Action Nets [48],Random Forest [42], Bag of Words [40], [41], [41], [49]–[52] and low dimensional models [53]. Finally,

a classifier is trained on these models to perform the final annotation of a new action. The most popular approaches are nearest neighbour classification [39], [43], [50], [52]–[54], probabilistic classification [45]–[48] or Support Vector Machine [40], [44], [49], [51].

Since different instances of a given action reside only in a subspace of the entire high dimensional feature space, we develop the idea of an action manifold [55], which is an intuitive and compact descriptor of human body motion embedded in a low dimensional space. In the context of this work, the action manifold corresponds to the concept manifold. Here, style corresponds to morphological and biomechanical differences between people induced by body size, body shape, gender, mood, etc. as well as motion execution variability or speed [56]. Following [57], [58], the action content is represented as a 1-dimensional manifold embedded in 2-dimensional space. As a low level image feature, the space-time cube is adopted [43], whereas classification is performed according the nearest neighbour procedure.

1) Action Recognition Framework: The proposed action recognition system is composed of two pipelines performing training (Fig. 3a) and classification (Fig. 3b). On the training side (Fig. 3a), binary silhouettes are extracted from the image sequences and used to produce a spatio-temporal shape descriptor for each motion instance. Then, for each action of interest, the dimensionality of the motion space is reduced using SLE to create an action manifold (§IV-E3). Subsequently, a mapping function is learned in order to provide a bidirectional projection mechanism between the original space and the action manifold. Note that SLE is used to generate an action manifold for each concept defined as an action observed from a given angle. The production of a single model for the meta-concept of an action observed from any angle can be achieved by integrating a set of manifolds modelling actions seen from slightly different angles. This new model, which is called view-independent action manifold, can then be used to perform view-independent action recognition.

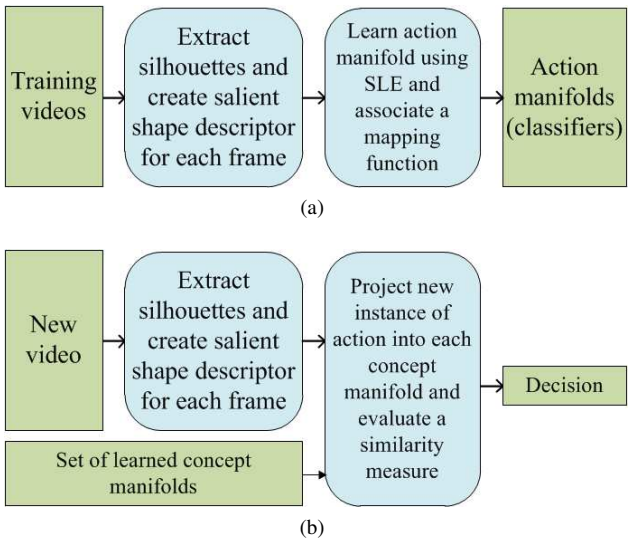

Fig. 3. Action recognition framework: learning process (a) and classification process of a new video (b).

On the classification side (Fig. 3b), the shape descriptor is generated from the video of interest. It is then projected using the learned mapping function into each action manifold. Finally, the process of video labeling is performed by nearest neighbour classification.

2) Neighbourhood Selection Procedure: First, temporally segmented action videos captured from a particular angle are processed to extract a sequence of binary silhouettes which are then represented as local space-time saliency features [43]. This representation assigns highest gradient values within fast moving limbs which are usually much more informative for identifying actions (Sup. 11 and Sup. 10, 1st row). Since the shape descriptor is derived from a unique region of interest, or action bounding box, it includes several interacting subjects in the case of actions involving interactions while it only contains a single actor when the action is performed by an individual.

Successful dimensionality reduction using SLE depends on the appropriate identification of intra and inter-sequence neighbours for each frame. This is achieved automatically by adopting the similarity metric proposed by [43] and extending the local space-time saliency shape descriptor with 6 local space-time orientation features, which correspond to temporal (blue), horizontal (red), and vertical (green) directions of local 'plates' (Sup. 10, 2nd row) and 'sticks' (Sup. 10, 3rd row) within a human shape. All these features are then combined to form a space-time cube for each frame [43] by sliding a warping window in time of size s (see §IV-A). This cube is a compact and temporally constrained representation of the sequence fragment which encapsulates local shape and orientations features within a given window. Since, each sequence fragment is now expressed by a single feature vector; the similarity between them can be computed effectively using the standard Euclidean norm without the need of computationally expensive temporal alignment of points sequences. This procedure computes a neighbourhood similarity matrix M (Eq. (5)) of Euclidean distances between all space-times cubes among all sequences. An example of neighbourhood similarity matrix produced by this process for SLE is depicted in Sup. 11.

3) View-independent Action Recognition : An action manifold represents the temporal structure of the action observed from a given view. On the other hand, the same action can be observed from slightly different views, for example, in terms of camera azimuth angle. We propose to take advantage of the sequential structure along azimuth angle by modelling jointly the action manifolds in a meaningful representation associated with the meta-concept of view-independent action. Here, we demonstrate that the extension for joint modelling of similar concepts which was proposed in §IV-D can be successfully implemented to achieve view-independent action recognition. The learning procedure of a view-independent action manifold is summarised in Fig. 4: it is based on the framework we presented in [55].

First, SLE extracts the action manifold, i.e. the style invariant content of the action of interest, for each view. As a result, a set of style-invariant but view-dependent 1-dimensional action manifolds embedded in 2-dimensional space is obtained. Secondly, these models are combined to produce a compact and view-independent action manifold model of the consid-

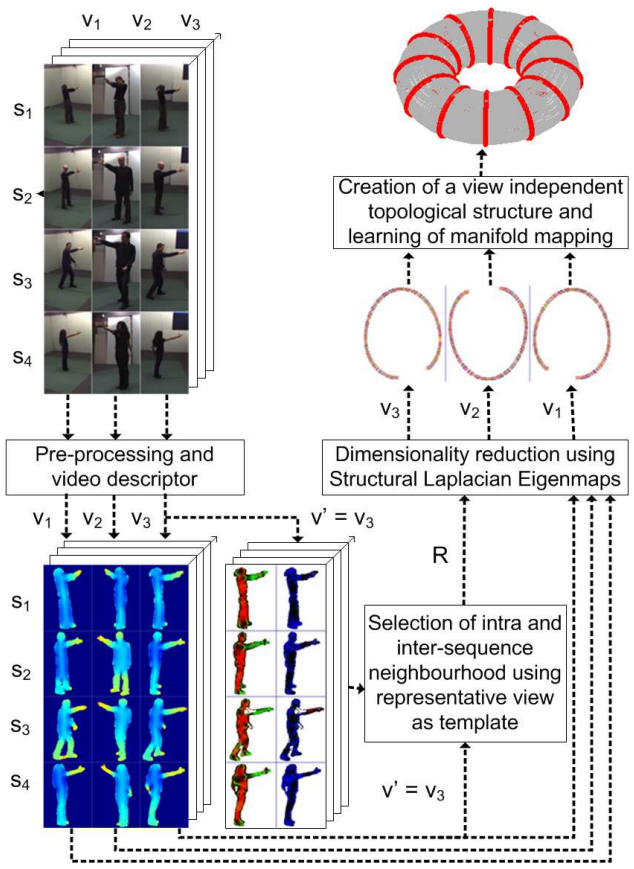

Fig. 4. Description of the view-independent action recognition framework for the 'point' action.

ered action. The 1-dimensional action manifold represents the structure of an action along the time dimension captured from a given view. Here, we assume that similar concepts are evenly sampled every Z degrees of azimuth view angle in the range $\{i : i \in [0°, 360°), i = i + Z\}$ and point correspondence between them have been established. In this context, every view-dependent action manifold M_i has two neighbouring manifolds M_{i-Z} and M_{i+Z}. Connections of corresponding points create closed 1-dimensional manifolds (topologically equivalent to circles) which are the view-independent embedded spaces of postures present in the action. Therefore, we define the unified representation of an action as the combined space of the two sets of continuous 1-dimensional manifolds, i.e. content and view, which are placed orthogonally to each other and embedded nonlinearly in a 3-dimensional space. In order to extrapolate that space to unseen instances of actions, a mapping function is required. We propose to achieve this using a geometrically based process. This requires that view-dependent representations are first aligned with respect to a reference manifold. Here Procrustes analysis [38] is employed to preserve the internal structure of each manifold. Then, all embedded representations are assembled into a three-dimensional space to encode the sequential structure along the view dimension. The outcome of this procedure reveals a torus-like structure which encapsulates both a unique content and view variation (Sup. 12). We call this structure a view-independent action manifold. This result is in line with previous work [11], where the usage of a torus was justified as an ideal representation for modelling both the viewpoint and the body configuration of different actions. However, while in that work the topological correspondence between data points and an ideal torus was artificially enforced, in our work this torus-like representation is data-driven and reflects the intrinsic content of the view-dependent data. Therefore, in our approach any type of motion, even non-periodic ones can be handled using a single framework (Sup. 12). Finally, the continuity of the descriptor is approximated by learning a generative decomposable model [36] which has previously been used to interpolate style and view factors of unknown instances of actions. As a consequence, a flexible mapping function is derived which enables projecting between the low dimensional view-independent action space and high dimensional observed space.

V. SLE Experimental Evaluation

The proposed methodology is validated qualitatively and quantitatively in a range of perception tasks to examine its properties and demonstrate its key characteristics. In particular, a comparative analysis of performance is performed between the proposed SLE and current state-of-the-art approaches for dimensionality reduction.

First, we evaluate the proposed approach qualitatively using datasets for which the underlying structure is known so that the quality of the embedded space can be judged visually (§V-B). Secondly, a quantitative comparison of SLE against state-of-the-art approaches is presented in a 3D pose recovery application (§V-C). Finally, we validate our extension for joint modelling of similar concepts by representing walking and running activities in a coherent continuous space (§V-D).

Unlike SLE, all other embedding-based methods require manual parameter tuning, which is very sensitive to the dataset of interest. Therefore, in such cases, extensive testing was conducted to determine the optimal settings for each experiment independently. In addition, the number of nontrivial neighbours required for ST-Isomap [23] was calculated using the SLE neighbourhood estimation procedure from §IV-A. Regarding mapping-based approaches, we used the default parameters provided with the Matlab implementations of BC-GPLVM [24] and GPDM [26].

If it is not stated otherwise, the DTW distance is used in SLE to measure similarity of sequence fragments during determination of the intra and inter-sequence neighbours. In turn, the length of the sequence fragment s was set empirically to a value 10 in all our experiments (see eq. (4)). Due to the transitivity of neighbourhood connections, the choice of this parameter is not critical as shown in Sup. 2. Finally, standard LE coefficient α was set to 1 (eq. (8)).

A. Datasets

The proposed methodology is evaluated using three different datasets: one is composed of images, whereas the other two contain 3D motion capture (MoCap) data.

Using the Columbia Object Image Library [59], we selected 6 sets of colour images which show similar objects in terms of global shape, i.e. rectangular cuboid, but displaying significantly different appearance: 27.car, 31.box_1, 39.container, 46.cigarette_packet, 55.jar, 79.box_2 (see Sup. 3). In the dataset, each object was placed on a turntable in front of a uniform black background. Then, 72 views, i.e. every 5 degrees, were captured by a fixed camera. Pictures were normalised to a size of 128x128 pixels. In such context, the angular position of an object with respect to the camera view represents the concept to model, whereas the type of object corresponds to its style. After conversion to grey level scale, all images represent points in a 16384-dimensional space within multivariate sequences corresponding to the sequential change of the object appearance when rotated on the turntable. Although the appearance of objects differs significantly, the global shape of these objects is similar when represented as 2D contours. Such contours change smoothly when the turntable is rotated. As a result, the image sequence captures the visual deformation of the global 3D object geometry across different views. Contours are extracted by thresholding binary masks from images and then tracing their boundaries. Subsequently they are normalised so that they display the same height to width ratio.

The second dataset is a subset ot the HumanEva dataset [60], i.e. MoCap data of walking and jogging actions performed by three different subjects. In turn, the third dataset, called "walking2running" [27], consists of walking, walking fast and jogging sequences and relevant transitions performed on a treadmill by one subject. In both MoCap datasets, the 13 joints skeleton-based MoCap data are first normalised and then parameterised by a quaternion representation [11] to form temporal sequences of 52-dimensional feature vectors. The walking, walking fast and jogging actions were chosen as concepts, since their intrinsic dimensionality as well as their underlying manifold structure is well known [12], [61], [62]. These actions are cyclic, since the same intrinsic joint configuration of the human body reoccurs every two steps. Intuitively, any two steps correspond to a continuous curve in a human motion space, since there is only one degree of freedom, i.e. the innate state/configuration of the motion over time.

B. Qualitative Evaluation

Qualitative evaluation is performed in two experiments using the image and MoCap datasets, where style, i.e. object nature and person style should be discarded, respectively, to model view angle and innate pose configuration, respectively. In both cases, their intrinsic dimension is 1 and the hidden structure of their content is embedded into a 2-dimensional space to take into account their cyclic nature ($d = 2$). Whereas the low dimensional representation of human motion as a 2D oval shape has already been shown in many studies [12], [37], [61], [62], object orientation is the only dimension related to the content of the image dataset. Examples of neighbourhood similarity matrices generated during dimensionality reduction using SLE are depicted in Fig. 5 for the human MoCap data and in Sup. 3 for image dataset.

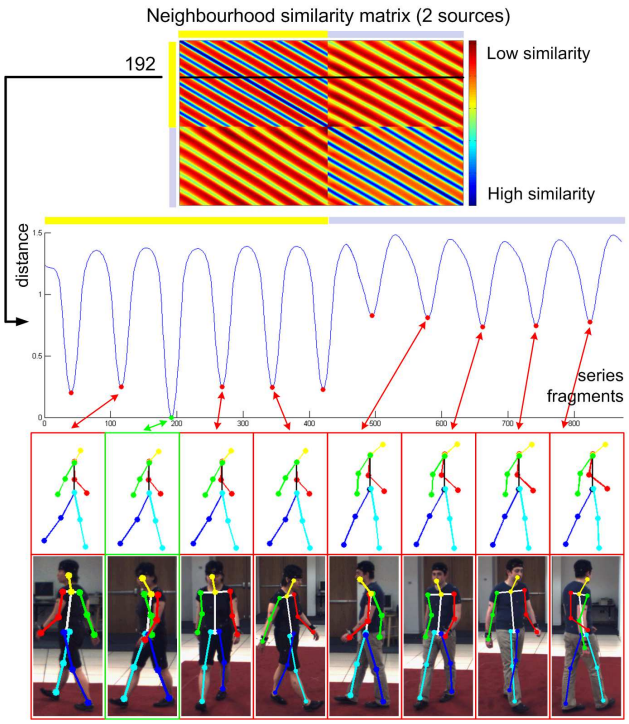

Fig. 5. Neighbourhood similarity matrix created by the SLE using two subjects. Each local minima (red) corresponds to the most similar intra and inter-sequence neighbours in relation to the reference object (green).

Using various nonlinear dimensionality reduction methods, i.e. LE, Isomap, BC-GPLVM, GPDM, ST-Isomap and SLE, we generated the 1-dimensional manifolds of the 6 image objects (Sup. 5) and two human actions (walking in Fig. 6 and jogging in Sup. 4). In addition, Sup. 6 shows projections of image views on the produced manifolds.

The geometrically motivated LE and Isomap approaches as well as the probabilistic BC-GPLVM and GPDM fail to model the expected single ellipse structure of the content. The representations of all these embeddings appear to be dominated by stylistic variations (Fig. 6a,6b,6c,6e, Sup. 4a,4b,4c,4e and 5a,5b,5c,5e). In all cases, these methods fail to generalise between the different styles, either 2 subjects or 6 objects, which leads to the generation of manifolds displaying a set of disjointed circular shapes. Since BC-GPLVM and GPDM use only a constraint of temporal continuity, these results suggest it is insufficient to model the content of action regardless of style. In contrast, the incorporation of sequence-based constraints using either ST-Isomap or SLE, allows, at least, some integrations of the different styles in a single space (Fig. 6d,6f, Sup. 4d,4f,5d,5f). However, not only the spaces obtained by ST-Isomap are distorted, but, in the case of the image dataset, although some global pattern of concentric ellipses emerges (Sup. 5d), this is not satisfied by the sequences of 2 objects, i.e. 27.car and 39.container (black and red curves respectively) (Sup. 5d, 6c).

In contrast, SLE produces an single ellipse-like representation which is in line with the expected structure of the dataset content. Fig. 7 clearly shows that all the image objects are arranged according to the view point in this representation,

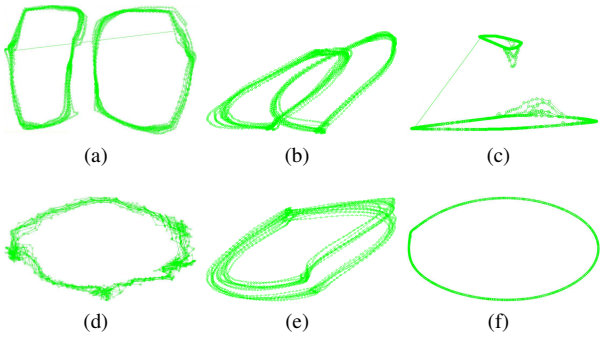

Fig. 6. Embedded spaces for walking (2 subjects) using a) Isomap, b) BC-GPLVM, c) LE, d) ST-Isomap, e) GPDM and f) SLE.

Fig. 7. The 1-dimensional joint view manifold embedded in the 2-dimensional space obtained by SLE with visualisation of corresponding objects.

which is invariant to object appearance. Similarly Sup. 7 and corresponding Sup. 8 confirm the subject invariant representation in the case of human motion. These experiments demonstrate that SLE is able to embed the common intrinsic dimension of non linear data by discarding style variability between different sequences.

C. Quantitative Evaluation

A comparative analysis of quantitative performance between the proposed SLE and other state-of-the-art approaches is performed using a 3D pose refinement framework [9] which takes advantage of the embedded spaces of the human motions generated in the previous section.

This consists of two modules: training and pose refinement. During the training stage, the space of each human motion is reduced to its intrinsic dimensionality ($d = 2$). Following [28], a Radial Basis Function network is learned to provide a bidirectional projecting mechanism between the high and low dimensional spaces. The second module of the framework deals with the actual problem of 3D pose refinement. Given a sequence of (inaccurate) 3D pose estimates, the framework projects each 3D skeleton into the embedded space using the corresponding mapping function. Then, this projection is associated to its nearest low dimensional training neighbour according to the Euclidean distance. Finally, the selected neighbour is projected back to the human motion space as the refined 3D pose estimate.

In this experiment, two actions are considered, i.e. walking and jogging. Although test pose estimates could be obtained from any 3D pose recovery framework, in order to provide a comprehensive evaluation platform for quantitative comparison, a large testing dataset of 6000 pose estimates was simulated by introducing Gaussian noise to ground truth poses with an average error per joint of 80mm, i.e. the average error of recently proposed approaches [60], [63]. To measure performances, experiments are conducted using cross-validation taking either one or two subjects for training leaving respectively two or one subjects for testing and averaging over 5 test sequences. The visual representations of the generated low dimensional spaces are provided in the previous section to highlight the content extraction abilities of the methods under study.

The qualitative evaluation of human motion from the previous section is supported here by a quantitative comparison of the obtained accuracy (Fig. 8 and Sup. 9). First, performance analysis confirms the generalisation abilities of the methods integrating temporal constraints since data from a second subject improves their accuracy. Conversely, the inability of Isomap and LE to generate a coherent manifold from data comprising several individuals leads to significant degradation of pose refinement performance. Among the temporal approaches, BC-GPLVM and SLE benefit the most from additional training samples (accuracy $+12\%$). On the other hand, GPDM's dynamic model seems to be able to optimise most of its parameters from a single subject. Note that the

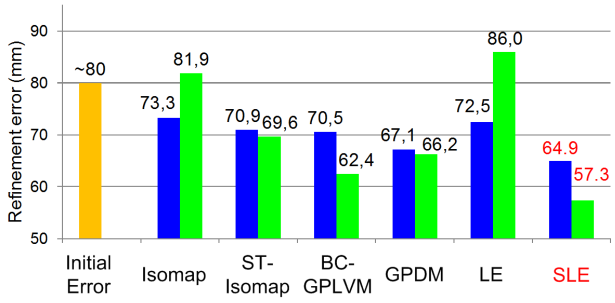

Fig. 8. Average refinement root mean square error for walking sequences using either one (blue) or two (green) subjects for training.

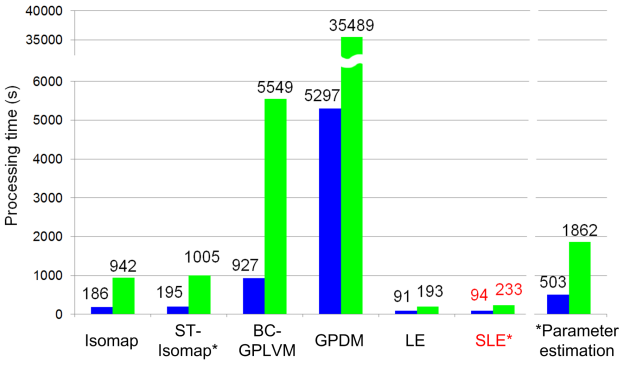

Fig. 9. Training times based on either 1 (blue) or 2-subject (green) walking sequences (parameter estimation is manual for all embedding-based methods).

presented quantitative results are influenced by the quality of both the embedded space and the mapping function between spaces. For that reason, although ST-Isomap produces more convincing visual representation than BC-GPLVM (Fig. 6 and Sup. 4), it performs worse because of less advanced mapping, i.e. RBF-based mapping. We can therefore conclude that SLE and BC-GPLVM are the most successful approaches. SLEs superiority was predictable theoretically. First, since structural relationships between sequences, e.g. temporality, are local properties of data, the local constraints of SLE allows encoding this crucial information and thus discovering more accurate data driven models than ST-Isomap which relies on global constraints. Second, the lower complexity of the optimisation process of SLE should lead to better convergence in comparison to the Gaussian based approaches. As consequence, SLE not only displays the best accuracy and produces more meaningful embedded spaces (Fig. 6 and Sup. 4), but it is also significantly faster by an order of magnitude, even when the cost of the neighbourhood selection procedure is added (Fig. 9, last column). This is very important because this shows that, unlike BC-GPLVM, SLE has the ability to learn models from much larger training sets which should conduce to even better results.

D. Validation of joint modelling of similar concepts

Walking and running are human bipedal motions which are based on cyclical movements of the hind limbs [64]. One complete cycle is called a stride and consists of two phases. During the stance phase, each limb spends a part of the stride in contact with the ground. Then it is followed by the swing phase where the foot leaves the ground and is brought forward for the next stance phase. The right and left legs alter between phases, so when the right leg is in the middle of its stance phase, the left leg is in the middle of its swing phase. Thus walking and running can be considered as similar concepts with a common content structure which differs, in particular, in the relative duration of the stance and swing phases of the stride. In walking, each foot spends more than half of the stride in stance, while in running it is shorter thus creating overlapping swing phases with both feet off the ground. Fig. 10 presents a representation of those activities in a single coherent space. Using the "walking2running" dataset, this space was obtained by reducing dimensionality of each action independently using SLE and then combining both concept spaces in a unified model of the meta-concept. Correspondences between concepts were determined using our DTW-based procedure which estimates intra and inter-sequence neighbours. For each point, the most similar neighbour in the other concept was chosen. These correspondences allowed aligning spaces in a single representation using Procrustes analysis [38]. Finally, the continuity of this meta-concept space was approximated by learning RBF mapping functions [37].

Such meta-concept space encodes not only walking and running actions, which form the boundaries of the space, but also intermediate actions between these concepts, e.g. fast walking. In order to demonstrate that the interpolation between the two spaces is meaningful, we conducted an experiment where the sequence defining a fast walking model generated from the continuous meta-concept space was projected in a motion capture space containing examples of walking, running and fast walking sequences.

First, we extracted a model expected to represent fast walking by selecting a sequence of points located at mid distance between the walking and running manifolds. (Fig. 10, green circle). Then, the low dimensional models of walking, walking fast and running were projected to the high dimensional space. Subsequently, we looked for the best fit for the generated sequences of 3D skeletons within a dataset of actual motion capture data of an individual performing those three actions. This was achieved using DTW by comparing the sequences of interest with sequence fragments generated by sliding a warping window through the dataset. Finally, the average of mean absolute angle error was computed between the 3D skeletons of each generated sequence and their best match in the dataset.

Table I shows that, as expected, each generated sequence matches the most a MoCap sequence with the same label, even the fast walking one which was created from manifold interpolation. In addition, the generated walking and running sequences are closer to fast walking data than running, respectively walking, data. This experiment confirms that the generated fast walking model is a useful approximation, which illustrates the value of meta concept spaces. Note that the lower accuracy displayed here by the fast walking model can be explained by the fact it does not rely on actual training data, but interpolation.

TABLE I
AVERAGE OF MEAN ABSOLUTE ANGLE ERROR OF PROJECTED ACTION MANIFOLDS WITH ACTUAL MOCAP DATA

Low dimensional representation	**Walking**	**Running**	**Fast walking**
Walking	$\mathbf{2.1}^\circ$	7.4°	2.6°
Running	7.4°	$\mathbf{3.7}^\circ$	3.9°
Fast walking	7.2°	5.9°	$\mathbf{5.5}^\circ$

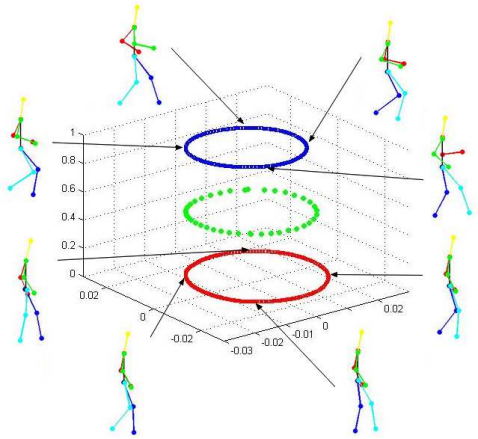

Fig. 10. Joint representation of learned walking (red) and running actions (blue) together with fast walking action approximated by an interpolated model (green).

VI. Application to Action Recognition

Evaluation of classification performance is achieved using the standard approach used in the action recognition community [40], [41], [44], [45], [48], [50], [51], [54], [65], [66]. As a consequence, recognition rates are computed by the leave-one-actor-out strategy, i.e. at each run, one subject is selected for testing, whereas all remaining actors are used for models learning. Then, all actions performed for that actor are evaluated independently and a final error is estimated by the average error rate over all experiments. Two scenarios are evaluated, i.e. view-dependent and view-independent action recognition using action-based and subject-based bounding boxes respectively.

A. Datasets

View-dependent action recognition is performed on the Weizmann [43] and UT-Interaction [67] datasets. The first dataset consists of 9 different subjects repeating individually several times 10 actions such as wave, run, jump, bend, in an outdoor environment with a static background. The second dataset is currently the most complete in terms of available training material and actions involving interactions. All videos are captured from a single view and show interactions between two characters seen sideways. It is composed of two sets of 10 videos including 6 different actions such as kicking, shaking hands and hugging. These two sets (D1 & D2) differ in terms of character's resolution (260 against 220 pixels) and background complexity (D1's is more uniform).

View-independent action recognition is evaluated on the publicly available multi-view IXMAS dataset [45]. It is comprised of 13 actions, such as sitting, waving, pointing and checking watch, performed by 12 different actors. Each activity instance was recorded simultaneously by 5 calibrated cameras, and a reconstructed 3D visual hull is provided. In line with other experiments using IXMAS dataset [39]–[41], [50], [65], the poorly discriminative top view is discarded from evaluation. Since no specific instruction was given to actors regarding their position and orientation, action viewpoints are arbitrary and unknown.

B. View Dependent Action Recognition

1) Single-subject Action Recognition: Similarly to other state-of-the-art approaches, our framework reports a perfect recognition rate for view-dependent action recognition using single-subject bounding box in Table II.

TABLE II
Performances obtained on Weizmann dataset.

%	Average accuracy
SLE [55]	100.0
Blank [43]	100.0
Wang [53]	100.0
Weinland [44]	100.0
Junejo [51]	95.3
Liu [39]	90.4
Zhang [47]	89.3
Vezzani [46]	86.7

2) Multiple-subjects Action Recognition: In turn, Table III reports performances using action-based bounding boxes and interacting subjects. Examples of trained models are depicted in Sup. 13. Our results are compared with the Random Forest (RF) framework [42], which is considered to be the current state-of-the-art approach on this dataset. In addition, results of two popular bag of words (BoW) frameworks using nearest neighbour classification are presented as a baseline [67].

SLE performs better than BoW approaches as illustrated by performances obtained in the more complex and dynamic background of D2. Although a RF-based framework outperforms SLE, instead of using action-based bounding boxes it requires the extraction of a bounding box per subject, which is not a trivial problem when people interact. In their implementation, this is achieved using an advanced tracking framework. Such a scheme could be integrated in our approach which should also allow better performance.

TABLE III
Performances obtained on UT-interaction dataset

%	**BoW**		**SLE**	**RF**
	Action bounding box			Single-subject bounding box
	[52], [67]	[49], [67]	ours	[42]
D1	57	63	75	80%
D2	50	62	67	*NA*

C. View Independent Action Recognition

For view-independent action recognition, in addition to a single view classification, we also report results of using multiple views for recognition by applying a simple majority voting rule [39]–[41], [45], [50], [65]. Note that testing in view-independent scenario is performed with views which are not included in the training data. Therefore, not only azimuth view angles are unknown but also elevation angle varies within a 45-degree range. As a consequence, testing is performed using examples of unknown action primitives performed by unknown people captured from an arbitrary and unknown view. Similarly to [68], our framework requires a dense set of action videos regarding viewpoints for the training of proposed action manifolds. To generate them we follow the same approaches [68] where the animated visual hulls are projected onto 12 evenly spaced virtual cameras located around the vertical axis of the subject. Since synchronisation among views is provided, processing time can be reduced in the generation of the meta-concept models. This is achieved by estimating only once intra and inter-sequence neighbours. Then, these neighbourhoods are used in producing each individual action manifold using SLE. Examples of trained models are depicted in Sup. 12.

In addition to SLE results, Table IV reports those published for other state-of-the-art algorithms. [68] was not included, because they use a completely different evaluation framework where testing is performed using artificially generate

observations from the visual hulls. To allow fair comparison with [44], [45], [51], [65], results using 11 actions are also reported, where the 'point' and 'throw' actions are discarded. Unfortunately, since some authors use less challenging evaluation frameworks, it is very difficult to draw any definitive conclusion based only on this table. However, SLE shows highest accuracy when compared to methods which have been evaluated using the same stringent framework. This shows that our descriptors are robust not only to subject style variability and view variations in terms of azimuth, but also to variation in elevation angles, which vary within a range of 45 degrees in the IXMAS dataset.

Although [44] and [54] seem to obtain better results, both frameworks are actually trained and tested using the same camera views, whereas our evaluation is based on completely unknown testing views. Thus, it is unclear how performance of these two algorithms [44], [54] would extrapolate in the more complex scenario of action recognition in unfamiliar views. Similarly, performance of [48] is reported for a single sequence (out of three) per actor which was selected to achieve best accuracy. In such settings, our framework produces similar performance (82.4%) using all available subjects. One should also note that evaluations of [48] and others, i.e. [44], [45], [51], are conducted only on subsets of available subjects which makes them less comprehensive. Finally, performances of [40] and [50] approaches are very similar to ours. However, since they both rely on codebooks, they are likely to be less scalable than ours to a higher number of actions.

TABLE IV
AVERAGE RECOGNITION ACCURACY OVER ALL CAMERAS USING EITHER SINGLE OR MULTIPLE VIEWS FOR TESTING.

%	Subjects \Actions	Average accuracy	
		Single view	All views
SLE [55]	12 / 13	73.2	83.3
Lv [48]	10 / 14	82.9	-
Tran [54]	12 / 13	80.2	-
Liu [40]	12 / 13	73.7	82.8
Kaanische [50]	12 / 13	71.7	90.6
Liu [39]	12 / 13	71.7	78.5
Reddy [41]	12 / 13	66.5	72.6
SLE [55]	12 / 11	74.7	83.1
Weinland [44]	10 / 11	86.9	-
Junejo [51]	10 / 11	73.7	-
Yan [65]	12 / 11	64.0	78.0
Weinland [45]	10 / 11	63.9	81.3

VII. DISCUSSION

Exhaustive validation demonstrates the value of the proposed methodology to model concepts defined by a set of multivariate sequences showing stylistic variability. In comparison to current state-of-the-art approaches, SLE was able to discover intrinsic nature of angle change in the image object dataset as well as innate body configuration of 3D MoCap data. In addition, SLE improved its performance in the 3D pose recovery task by introduction of additional training samples. This suggests that SLE could benefit from even larger training datasets. Although the cost of its neighbourhood selection procedure adds extra computational complexity compared to standard embedding-based approaches, SLE remains significantly faster, i.e. by an order of magnitude, than Gaussian process methods as expected from our theoretical analysis of complexity (see §IV-C). Finally, SLE does not require tuning of any parameter to perform well. Setting a unique value for the length of the sequence fragment s in all experiments and the more detailed analysis shown in Sup. 2 demonstrate that the method is not sensitive to that parameter since wide range of values is acceptable. Investigation of the impact of the low dimension space dimensionality is beyond the scope of this paper. However, according to the statistical learning theory [2], our framework is subject to 'peaking phenomenon' [69], i.e. there is an optimal number of dimensions for a given training dataset that allows best performance. As a consequence, the increase of dimensionality - $d = 2$ has been used in all experiments - may further improve results when enough training data is available. This may be an interesting direction for future research.

Since a generated manifold by SLE encodes the unique characteristic of the concept of interest, it is suitable for classification of unknown instances of concepts in different real life action recognition scenarios. Although the discussed methods cannot be compared purely on the reported performances, all experiments confirm the versatility of proposed methodology producing very competitive results while overcoming drawbacks of state-of-the-art methods. In particular, the action manifold demonstrates its superiority in generalisation over variations of style, view and speed within one class while accurately distinguishing between actions of different classes. However, when a concept is given a broad definition, e.g. kicking which can be represented by instances of stand kick, jump kick, turn around kick, features describing it lack consistency makes the process of finding intra and inter-sequence neighbours very challenging .This may lead to the generation of a noisy model and poor classification performance. Fig. 11 highlights this issue by showing the action model associated to the broad kicking concept with a few not perfectly aligned instances of the action.

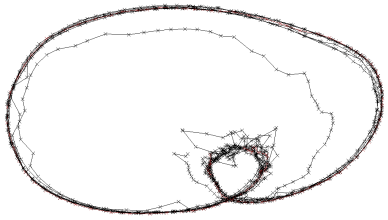

Fig. 11. Heterogenous nature of the kicking action manifold.

VIII. CONCLUSION

This paper presented a novel embedding-based dimensionality reduction approach, Structural Laplacian Eigenmaps, which learns data-driven manifolds designed for any concept which

can be represented by a set of multivariate sequences. This is achieved by encoding the intrinsic structure of multivariate sequences in the form of two structural neighbourhoods, which are then incorporated into the extended LE-based dimensionality reduction scheme. Then, this methodology is further developed to model jointly several concepts of similar nature within unified representation creating continuous space between concept manifolds. The conducted experiments on various datasets prove that the proposed methodology is able to generate a low dimension manifold which summarises a set of instances. The obtained manifold represents mathematically the intrinsic nature of the concept of interest regardless of stylistic variations, which is essential for classification tasks. Based on SLE, a flexible and intuitive action recognition framework was developed. It is competitive to current state-of-the-art methodologies. Moreover, it is the only framework which is able to report first-class performances in both view-independent action recognition and human interaction classification. This confirms practical value of the proposed methodology.

REFERENCES

[1] J. Tenenbaum and W. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, 2000.

[2] V. Vapnik, *Statistical Learning Theory*, 1998.

[3] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, 2000.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, vol. 14, pp. 585–591, 2002.

[5] B. Schölkopf, A. Smola, and K. Müller, "Kernel principal component analysis," *Artificial Neural Networks*, 1997.

[6] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, 2000.

[7] N. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[8] R. Urtasun, D. Fleet, A. Geiger, J. Popović, T. Darrell, and N. Lawrence, "Topologically-constrained latent variable models," *Proc. of Int'l Conf. on Machine Learning*, vol. 307, 2008.

[9] M. Lewandowski, J. Martinez-del Rincon, D. Makris, and J.-C. Nebel, "Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series," *Proc. of Int'l Conf. on Pattern Recognition*, 2010.

[10] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Tilburg University Technical Report, TiCC-TR-2009-005, 2009.

[11] A. Elgammal and C.-S. Lee, "Tracking people on a torus," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, 2009.

[12] K. Grochow, S. Martin, A. Hertzmann, and Z. Popović, "Style-based inverse kinematics," *ACM Trans. on Graphics*, vol. 23, no. 3, pp. 522–531, 2004.

[13] A. Shon, K. Grochow, A. Hertzmann, and R. Rao, "Learning shared latent structure for image synthesis and robotic imitation," *Advances in Neural Information Processing Systems*, vol. 18, p. 1233, 2006.

[14] A. Jafari and F. Almasganj, "Using laplacian eigenmaps latent variable model and manifold learning to improve speech recognition accuracy," *Speech Communication*, 2010.

[15] R. Urtasun and T. Darrell, "Discriminative gaussian process latent variable model for classification," *Proc. of Int'l Conf. on Machine Learning*, vol. 227, pp. 927–934, 2007.

[16] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Education Psychology*, vol. 24, pp. 417–441, 1933.

[17] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2005.

[18] K. Q. Weinberger and L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[19] M. Torki, A. Elgammal, and C. Lee, "Learning a joint manifold representation from multiple data sets," *Proc. of Int'l Conf. on Pattern Recognition*, pp. 1068–1071, 2010.

[20] M. Lewandowski, D. Makris, and J.-C. Nebel, "Automatic configuration of spectral dimensionality reduction methods," *Pattern Recognition Letters*, vol. 31, 2010.

[21] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1299 –1313, 2009.

[22] C. Bishop, M. Svensén, and C. Williams, "Gtm: The generative topographic mapping," *Neural computation*, vol. 10, 1998.

[23] O. Jenkins and M. Matarić, "A spatio-temporal extension to isomap nonlinear dimension reduction," *Proc. of Int'l Conf. on Machine Learning*, pp. 441–448, 2004.

[24] N. Lawrence and J. Quinonero-Candela, "Local distance preservation in the gp-lvm through back constraints," *Proc. of Int'l Conf. on Machine Learning*, vol. 148, pp. 513–520, 2006.

[25] C. Bishop, "Gtm through time," *Proc. of Int'l Conf. on Artificial Neural Networks*, pp. 111–116, 1997.

[26] J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1441–1448, 2006.

[27] J. Martinez-del Rincon, J.-C. Nebel, and D. Makris, "Graph-based particle filter for human tracking with stylistic variations," *British Machine Vision Conference*, 2011.

[28] A. Elgammal and C. Lee, "Separating style and content on a nonlinear manifold," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2004.

[29] J. Wang, D. Fleet, and A. Hertzmann, "Multifactor gaussian process models for style-content separation," *Proc. of Int'l Conf. on Machine learning*, pp. 975–982, 2007.

[30] F. Camastra, "Data dimensionality estimation methods: A survey," *Pattern Recognition*, vol. 36, no. 12, 2003.

[31] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[32] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Readings in Speech Recognition*, pp. 159–165, 1990.

[33] W. Arnoldi, "The principle of minimized iterations in the solution of the matrix eigenvalue problem," *Quarterly of Applied Mathematics*, vol. 9, pp. 17–25, 1951.

[34] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[35] Z. Bai, J. Demmel, J. Dongarra, R. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Society for Industrial and Applied Mathematics, 2000.

[36] C. Lee and A. Elgammal, "Homeomorphic manifold analysis: Learning decomposable generative models for human motion analysis," *Workshop on Dynamical Vision at European Conference on Computer Vision*, pp. 100–114, 2006.

[37] A. Elgammal and C. Lee, "Inferring 3d body pose from silhouettes using activity manifold learning," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, vol. 3, 2004.

[38] C. Wang and S. Mahadevan, "Manifold alignment using procrustes analysis," *Proc. of Int'l Conf. on Machine Learning*, 2008.

[39] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, 2008.

[40] J. Liu and M. Shah, "Learning human actions via information maximization," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, 2008.

[41] K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," *Proc. of Int'l Conf. on Computer Vision*, 2010.

[42] D. Waltisberg, A. Yao, J. Gall, and L. Van Gool, "Variations of a hough-voting action recognition system," *Recognizing Patterns in Signals, Speech, Images and Videos*, pp. 306–312, 2010.

[43] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, p. 2247, 2007.

[44] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," *Proc. of European Conference on Computer Vision*, pp. 635–648, 2010.

[45] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," *Proc. of Int'l Conf. on Computer Vision*, vol. 5, no. 7, p. 8, 2007.

[46] R. Vezzani, D. Baltieri, and R. Cucchiara, "Hmm based action recognition with projection histogram features," *Proc. of Int'l Conf. on Pattern Recognition: Contest on Semantic Description of Human Activities*, 2010.
[47] J. Zhang and S. Gong, "Action categorization with modified hidden conditional random field," *Pattern Recognition*, vol. 43, no. 1, pp. 197–203, 2010.
[48] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, 2007.
[49] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," *Proc. of Int'l Conf. on Pattern Recognition*, vol. 3, pp. 32–36, 2004.
[50] M. Kaâniche and F. Brémond, "Gesture recognition by learning local motion signatures," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, pp. 2745–2752, 2010.
[51] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "Cross-view action recognition from temporal self-similarities," *Proc. of European Conference on Computer Vision*, vol. 12, 2008.
[52] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *Proc. of Int'l Conf. on Computer Communications and Networks*, 2005.
[53] L. Wang and D. Suter, "Visual learning and recognition of sequential data manifolds with applications to human movement analysis," *Journal on Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 153–172, 2008.
[54] D. Tran and A. Sorokin, "Human activity recognition with metric learning," *Proc. of European Conference on Computer Vision*, pp. 548–561, 2008.
[55] M. Lewandowski, D. Makris, and J.-C. Nebel, "View and style-independent action manifolds for human activity recognition," *Proc. of European Conference on Computer Vision*, vol. 6316, 2010.
[56] R. Easterby, K. Kroemer, and D. Chaffin, *Anthropometry and Biomechanics: Theory and Application*. Plenum Press, 1982.
[57] K. Jia and D. Yeung, "Human action recognition using local spatio-temporal discriminant embedding," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, 2008.
[58] T. Chin, L. Wang, K. Schindler, and D. Suter, "Extrapolating learned manifolds for human activity recognition," *Proc. of Int'l Conf. on Image Processing*, vol. 1, 2007.
[59] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (coil-100)," Columbia University Technical Report, CUCS-006-96, 1996.
[60] L. Sigal, A. Balan, and M. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int'l Journal of Computer Vision*, vol. 87, no. 1, pp. 4–27, 2010.
[61] J. Darby, B. Li, and N. Costen, "Tracking human pose with multiple activity models," *Pattern Recognition*, vol. 43, 2010.
[62] R. Urtasun, D. Fleet, and P. Fua, "3d people tracking with gaussian process dynamical models," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2006.
[63] P. Kuo, T. Ammar, M. Lewandowski, D. Makris, and J.-C. Nebel, "Exploiting human bipedal motion constraints for 3d pose recovery from a single uncalibrated camera," *Proc. of Int'l Conf. on Computer Vision Theory and Applications*, vol. 1, 2009.
[64] J. Hutchinson and S. Gatesy, "Bipedalism," *Encyclopedia Of Life Sciences*, 2001.
[65] P. Yan, S. Khan, and M. Shah, "Learning 4d action feature models for arbitrary view action recognition," *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, vol. 12, 2008.
[66] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. Velastin, "Recognizing human actions using silhouette-based hmm," *Proc. of the 6th Int'l Conf. on Advanced Video and Signal Based Surveillance*, pp. 43–48, 2009.
[67] M. Ryoo, C. Chen, J. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (sdha)," *Recognizing Patterns in Signals, Speech, Images and Videos*, pp. 270–285, 2010.
[68] S. Richard and P. Kyle, "Viewpoint manifolds for action recognition," *Journal on Image and Video Processing*, 2009.
[69] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice-Hall International, 1982.

Michał Lewandowski received the MSc(Eng) degree in Computer Science in 2006 from the Silesian University of Technology, Gliwice, Poland. He completed in 2010 the PhD degree in Computer Vision/Machine Learning from the Kingston University, London, United Kingdom. He is currently a research associate in the Faculty of Science, Engineering and Computing at Kingston University, London. His research interests include computer vision and machine learning with application to visual surveillance.

Dimitrios Makris (M03) received the diploma in Electrical and Computer Engineering from Aristotle University of Thessaloniki, Greece in 1999 and the PhD in Computer Vision from City University, London in 2004. He is currently a Reader (Associate Professor) in the School of Computing and Information Systems, Kingston University, London. His research interests are in the area of image processing, computer vision and machine learning and particularly in motion analysis. Dr Makris is also a member of the IEEE and the British Machine Vision Association.

Sergio A. Velastin (M90, SM'12) received the B.Sc. degree in Electronics, M.Sc. (Research) degree in Digital Image Processing and the Ph.D. from the University of Manchester in the UK, in 1978, 1979 and 1982 respectively. Currently he is a Research Professor at the Department of Informatic Engineering, Universidad de Santiago de Chile. His research interests include computer vision for pedestrian and traffic monitoring as well as distributed visual surveillance systems. Prof Velastin is also a Fellow of the IET and a member of the British Machine Vision Association.

Jean-Christophe Nebel (SM'08) received the MSc(Eng) degree in Electronics and Signal Processing in 1992 from the Institute of Chemistry and Industrial Physics, Lyon, France. He completed in 1997 the PhD degree in Parallel Programming from the University of St Etienne, France. He is currently a reader/associate professor in the Faculty of Science, Engineering and Computing at Kingston University, London. His research interests include computer vision and bioinformatics. He was awarded in 2004 with co-authors the A. H. Reeve Premium by the Council of the Institute of Electrical and Electronics Engineers for a journal paper describing his pioneer work in developing a 3D Dynamic Whole Body Measurement System.