Video analysis based on Multi-Kernel Representation with automatic parameter choice
A.M. Álvarez-Meza a,*, J. Valencia-Aguirre a, G. Daza-Santacoloma c, C.D. Acosta-Medina a,b, G. Castellanos-Domínguez a

a Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales, Colombia
b Scientific Computing and Mathematical Modeling Group, Universidad Nacional de Colombia, Manizales, Colombia
c Faculty of Electronic Engineering, Universidad Antonio Nariño, Bogotá, Colombia

A R T I C L E   I N F O
Available online 7 May 2012
Keywords:
Video analysis
Manifold learning
Multi-Kernel Representation
Automatic parameter selection
Multiple Manifold Learning

A B S T R A C T
In this work, we analyze video data by learning both the spatial and temporal relationships among frames. For this purpose, the nonlinear dimensionality reduction algorithm, Laplacian Eigenmaps, is improved using a multiple kernel learning framework, and it is assumed that the data can be modeled by means of two different graphs: one considering the spatial information (i.e., the pixel intensity similarities) and the other one based on the frame temporal order. In addition, a formulation for automatic tuning of the required free parameters is stated, which is based on a tradeoff between the contribution of each information source (spatial and temporal). Moreover, we proposed a scheme to compute a common representation in a low-dimensional space for data lying in several manifolds, such as multiple videos of similar behaviors. The proposed algorithm is tested on real-world datasets, and the obtained results allow us to confirm visually the quality of the attained embedding. Accordingly, discussed approach is suitable for data representability when considering cyclic movements.

1. Introduction

Essentially, most of the video-based methods employed to learn and discover the real object motion and/or human behavior are subject to two main attributes: the pixel spatial disposition and their variations among time. The common strategy of the nonlinear dimensionality reduction (NLDR) techniques is to presume that the high-dimensional input data (video) lie on a manifold, defined by a smaller set of characteristics [4,9]. Then, computing the pixel intensity similarities among a sequence of images, NLDR aims to reveal the underlying data structure in a low-dimensional space, which allows to infer the latent variables that govern the studied phenomenon.

There are several techniques for implementing NLDR based on manifold learning, such as Isometric Feature Mapping (ISOMAP) [16], Locally Linear Embedding (LLE) [14], Maximum Variance Unfolding (MVU) [20], Laplacian Eigenmaps (LEM) [1], etc. Unlike other iterative NLDR techniques (such as MVU and ISOMAP), LEM has an analytic solution. Thereby, LEM requires lower computational load when dealing with reasonable sample size, and it does not require a regularization process such as LLE. Anyway, all above mentioned techniques determine the assumed manifold by modeling the data topology through local interconnections among samples, just based on the spatial similarities, thus is, by a pixel intensity comparison. Accordingly, they discarded an amount of valuable information related to the soft transitions between adjacent pictures, which can be considered as a time constraint. Furthermore, the temporal information captured on a video should be taken into account to differentiate the cycles on a periodic movement. Furthermore, the temporal information captured on a video should be taken into account to differentiate video data regarding periodic movement cycles (activities that return to its beginning and repeats itself in the same sequence, e.g., rotating, walking, running, etc.), when traditional NLDR are not suitable to identify each repetition in the low-dimensional space.

On the other hand, it had been suggested the possibility of incorporating prior knowledge to the embedding topology, which allows to obtain enhanced low dimensional representations of the phenomenon in hand [18,19]. However, those techniques are based on complex probabilistic models comprising some free heuristic parameters that are far from easy to be tuned by an
inexpert user (not mentioning their huge computational load). Recently, an approach for incorporating temporal information to the embedding process is discussed in [8] that considers adjacent temporal neighbors to find out the structure of repetitive activities. Nonetheless, since the time variable is not reflected in the mapping process, it is not possible to identify different repetitions of a movement, and thus, the cycles are overlapped in the embedding space. To cope with this, a general model for multiview learning called Distributed Spectral Embedding (DSE), which aims to unfold the underlying data structure from different feature spaces is presented in [10]. Then, DSE calculates a common low-dimensional space that is close to each representation as much as possible. Although DSE allows to handle different space representations, the original multiview data are invisible to the final learning process, being inappropriate to explore the complementary nature of different views, and its computational load is too dense [21].

Several approaches that deal with multiple kernels within the machine learning context (classification and regression) are also presented in [7,12,3]. Their main goal is to employ different sources of information to identify the similarities among samples, and then, a combination of these similarities is calculated by means of statistical kernel learning. In this regard, a convenient approach is to consider that the calculated multiple kernel is actually a convex combination of a basis kernel [3,12]. In [21], a similar approach is used based on the mathematical framework of LEM for obtaining a Multiview Spectral Embedding (MSE) of the input data. MSE approach takes advantage of different views (feature space) to find out a low-dimensional space wherein each view is sufficiently smooth. Particularly, MSE is tested in image retrieval, video annotation, and document clustering problems, mainly combining low-level visual features. Due to there is no closed-form solution for MSE, they drive an alternating-optimization to obtain the embeddings.

In this work we propose a methodology for analyzing videos based on a Multi-Kernel Representation (MKR) of the input data, improving the LEM technique to compute and learn both spatial and temporal relationships among frames. The spatial relationships refer to the change of the pixel intensity among samples. The temporal information is related to the sequence order of data, more precisely, the order of appearance of the frames. When the spatial and temporal information are considered, the low-dimensional representation reveals the real motion of the objects. In addition, a formulation for automatic tuning of the free parameters required is presented. This formulation is based on a tradeoff between the contribution of the spatial and temporal information, to minimize, as well as possible, each error representation in the low-dimensional space based on a L-curve criteria [6]. Our work is inspired by the multiple kernel learning framework for machine learning [12,17], which is adapted in a NLDR scheme for cyclic motion analysis from video data. The presented approach is tested for revealing the spatial and temporal dynamics of several real-world videos related to cyclic motions. The main goal is to understand the true motion behavior of different scenes, making easier their interpretation. Particularly, the experiments are conducted on video sequences of rotating objects, walking humans, handwaving, and head movements. Obtained results exhibit a better performance of our method for visualizing the videos in a low-dimensional representation than traditional NLDR embeddings.

On the other hand, most of the NLDR algorithms are constrained to deal with a single manifold, attaining inappropriate low-dimensional representations when input data lie on multiple manifolds, moving apart each manifold from the others, regardless of whether the behavior among them is similar. For dealing multiple manifolds, a novel methodology was proposed in [17], which learns a joint representation from data lying on multiple manifolds. They seek to preserve the local structure on each manifold, and in the same time, collapse the different manifolds into one manifold in the embedding space, preserving the implicit correspondences between the points across different datasets. Still, this technique performs a pixel comparison, that is, the analysis is carried out in the high-dimensional space, which implies that the methodology is limited to analyze frames of video sharing a similar appearance. Therefore, the approach presented in [17] is not suitable to visually compare similar process when the appearance between the objects/subjects is different.

In this sense, we also proposed a Multiple Manifold Learning – (MML) scheme to compute a common representation in a low-dimensional space for multiple videos of similar behaviors, without making a pixel comparison between data points from different datasets. Moreover, we combine MKR with MML for revealing the spatial and temporal dynamics of a given cyclic motion, learning it from several videos. Attained results show that our proposal outperforms the method presented in [17] when the mapping to a low-dimensional space deals with multiple videos at the same time.

This work is organized as follow. In Section 2, the proposed MKR for NLDR based on the LEM algorithm is presented. Section 3 introduces the proposed method for automatic parameter selection in MKR. Section 4 shows how our approach computes a common representation in a low-dimensional space for multiple videos of similar behaviors. In Section 5 the experimental conditions and results are described. Finally, in Sections 6 and 7, we discuss and conclude about the attained results.

2. Multi-Kernel Representation based on LEM

Laplacian Eigenmaps (LEM) is a nonlinear dimensionality reduction technique based on preserving the intrinsic geometric structure of the manifold. Let \( \mathbf{X} \in \mathbb{R}^{n \times p} \) the input data matrix with sample objects \( \mathbf{x}_i \), \( i = 1, ..., n \). The goal is to provide a mapping to a low-dimensional Euclidean space \( \mathbf{Y} \in \mathbb{R}^{n \times m} \), with sample vectors \( \mathbf{y}_i \), being \( m < p \). The LEM algorithm has three main steps. First, an undirected weighted graph \( G(V, E) \) is built; where \( V \) are the vertices and \( E \) are the edges. In this case, there are \( n \) vertices, one for each \( \mathbf{x}_i \). Nodes \( i \) and \( j \) are connected by the \( E_{ij} = 1 \), if \( i \) is one of the \( k \) nearest neighbors of \( j \) (or viceversa), being measured by means of the Euclidean distance [2]. The second step is to construct a weight matrix \( \mathbf{W} \in \mathbb{R}^{n \times n} \). For this purpose, two variants can be implemented: heat kernel or simple minded. In the heat kernel variant, if nodes \( i \) and \( j \) are connected, then \( W_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \), being \( k(\cdot, \cdot) \) a kernel function, otherwise, \( W_{ij} = 0 \). For the simple minded option, \( W_{ij} = 1 \) if vertices \( i \) and \( j \) are connected by an edge, otherwise, \( W_{ij} = 0 \).

Then, the \( \mathbf{L} \in \mathbb{R}^{n \times n} \) graph Laplacian is given by \( \mathbf{L} = \mathbf{D} - \mathbf{W} \), where \( \mathbf{D} \in \mathbb{R}^{n \times n} \) is a diagonal matrix with elements \( D_{ii} = \sum_j W_{ij} \). A reasonable criterion for choosing a good map is to minimize the following objective function:

\[
\sum_{q} (\mathbf{y}_q - \mathbf{y}_{\text{ref}})^{T} \mathbf{W}_{ij} (\mathbf{y}_q - \mathbf{y}_{\text{ref}})
\]

subject to \( \mathbf{Y}^{T} \mathbf{Y} = \mathbf{I}_{m \times m} \). Minimizing (1) implies a penalty if neighboring points \( \mathbf{x}_i \) and \( \mathbf{x}_j \) are mapped far apart. Finally, the LEM problem can be accomplished solving the generalized eigenvalue problem \( \mathbf{LY} = \lambda D \mathbf{Y} \), where \( \lambda_i \) is the eigenvalue corresponding to the \( Y_i \) eigenvector, with \( i = 1, ..., n \). First eigenvector is the unit vector with all equal components, while the remaining \( m \) eigenvectors form the embedded space found by LEM.

Recently, machine learning approaches have shown that using multiple kernels instead of just one, can be useful to improve the data interpretability [3,12,7]. Given \( Z \) feature representations for each observation \( \mathbf{x}_i = [\mathbf{x}_i^T : z = 1, ..., Z] \), the Multi-Kernel Learning
(MKL) methods aim to infer the function:

\[ k_\tau(x_i, x_j) = g_{\tau'\tau}(\langle k(x_i, x_j) \rangle^{\tau' \tau}_i) \]

subject to \( \xi, \zeta \geq 0 \), and \( \sum \xi = 1 \) (\( k_{\tau\tau} \) are kernels). Taking into account the MKL formulation presented in Eq. (2), we propose to analyze high-dimensional data considering different topological representations, to unfold the underlying data structure of a video sequence. Therefore, we assume that a given video can be visualized as a manifold, which is modeled by different graph representations (topologies) \( G_{\tau\tau} \).

We employ the MKL formulation (3) to enhance the input data representability. The goal is to compute a function that allows to map the input data \( X \) to an embedded space \( Y \), properly combining different information sources that are contained in each \( G_{\tau\tau} \).

The LEM optimization problem presented in (1) is directly related with the weight matrix \( W \), which can be analyzed as a kernel learning problem, whence, it is possible to employ different topological representations (Multi-Kernel Representation – MKR) in the LEM formulation by means of MKL. When we are dealing with video sequences related with periodic motions (e.g., rotation, gait, running, etc.), it is necessary to take into account the temporal dynamics of the observations. Otherwise, to analyze a couple of points \( x_i \) and \( x_j \) representing the same activity state but belonging to different repetitions, the edge \( E_{ij} \) can be set to one, if only considering the spatial similarity (as traditional LEM). Hence, the weight \( W_{ij} \) will incur in a penalty if both \( x_i \) and \( x_j \) are mapped far apart according to Eq. (1), causing overlapping between samples of different time periods.

To cope with this drawback, we propose to model the assumed manifold \( M \) by two topologies: one based on the spatial local geometry similarity, and the other to reveal the time order sequence similarity among samples. As a result, we have two neighborhood sets for each observation. The first set \( \eta_{ij} \) is composed by the \( k \) nearest neighbors of \( x_i \) in \( X \) via the Euclidean distance. For the second set, we use the time vector \( t_{ij} \), with \( t_i \) representing the time index order of \( x_i \) in \( X \). So, we compute the set \( \theta_{ij} \) by the \( k \) nearest neighbors of \( x_i \) in \( X \), taking into account the temporal distance among observations. Given the pair of sets \( \eta_{ij} \) and \( \theta_{ij} \), we approximate the structure of \( M \) constructing the undirected weighted graphs \( G_{\tau\tau}(V, E_{\eta}) \) and \( G_{\tau\tau}(V, E_{\theta}) \), with edges:

\[ E_{\eta} = \left\{ \begin{array}{ll} 1 & x_i \in \eta_{ij}, \ 0 & x_i \notin \eta_{ij} \end{array} \right. \]

\[ E_{\theta} = \left\{ \begin{array}{ll} 1 & x_i \in \theta_{ij}, \ 0 & x_i \notin \theta_{ij} \end{array} \right. \]

\[ \text{therefore, two different weight matrices } W_{\eta} \text{ and } W_{\theta} \text{ can be computed as} \]

\[ W_{\eta} = \left\{ \begin{array}{ll} k(x_i, x_j) & E_{\eta} = 1, \ 0 & E_{\eta} = 0 \end{array} \right. \]

\[ W_{\theta} = \left\{ \begin{array}{ll} k(x_i, x_j) & E_{\theta} = 1, \ 0 & E_{\theta} = 0 \end{array} \right. \]

Inspired by MKL, it is possible to find out a weight matrix \( W \) that model the total data structure as

\[ W = \zeta W_{\eta} + \xi W_{\theta} \]

subject to \( \zeta + \xi = 1 \). Given \( W \), the combined Laplacian matrix is calculated as \( L_T = D_T - W \), where \( D_T \) is a \( n \times n \) diagonal matrix with elements \( D_T = \sum W_{ij} \). Therefore, the new LEM objective function that must be minimized can be written as

\[ \sum_{y_i, y_j} (y_i - y_j)^2 W_{ij} \]

subject to \( Y^T Y = I \).

In Fig. 1, it is possible to find out a weight matrix \( W \) that model the total data structure as

\[ W = \zeta W_{\eta} + \xi W_{\theta} \]

subject to \( \zeta + \xi = 1 \). Given \( W \), the combined Laplacian matrix is calculated as \( L_T = D_T - W \), where \( D_T \) is a \( n \times n \) diagonal matrix with elements \( D_T = \sum W_{ij} \). Therefore, the new LEM objective function that must be minimized can be written as

\[ \sum_{y_i, y_j} (y_i - y_j)^2 W_{ij} \]

subject to \( Y^T Y = I \).

In Fig. 1, it is possible to find out a weight matrix \( W \) that model the total data structure as

\[ W = \zeta W_{\eta} + \xi W_{\theta} \]

subject to \( \zeta + \xi = 1 \). Given \( W \), the combined Laplacian matrix is calculated as \( L_T = D_T - W \), where \( D_T \) is a \( n \times n \) diagonal matrix with elements \( D_T = \sum W_{ij} \). Therefore, the new LEM objective function that must be minimized can be written as

\[ \sum_{y_i, y_j} (y_i - y_j)^2 W_{ij} \]

subject to \( Y^T Y = I \).

In Fig. 1, it is possible to find out a weight matrix \( W \) that model the total data structure as

\[ W = \zeta W_{\eta} + \xi W_{\theta} \]

subject to \( \zeta + \xi = 1 \). Given \( W \), the combined Laplacian matrix is calculated as \( L_T = D_T - W \), where \( D_T \) is a \( n \times n \) diagonal matrix with elements \( D_T = \sum W_{ij} \). Therefore, the new LEM objective function that must be minimized can be written as

\[ \sum_{y_i, y_j} (y_i - y_j)^2 W_{ij} \]

subject to \( Y^T Y = I \).
matrix between $X^c$ and $X^b$. In [17], a methodology called Learning a Joint Manifold Representation (LJM) is proposed, which can be related with the matrix $A$ of (9), but computing each $M^{cb}$ by solving a matrix $P$ that permutes the rows of $U^{cb} \in \mathbb{R}^{n_c \times n_b}$, with $U^{cb}_q = k(x_q, x_r), x_q \in X^c$, and $x_r \in X^b (q = 1, \ldots, n_c; r = 1, \ldots, n_b)$. However, the application of this technique is limited to analyze frames of video sharing a similar geometry, due to $U^{cb}$ is inferred in the high-dimensional space (pixels frame comparison).

We propose to identify the correspondence between data points from different manifolds without making a pixel comparison. In other words, the similarity among frames of different videos is not calculated directly in each pair $X^c$ and $X^b$. In this sense, we compute each soft correspondence matrix $M^{cb}$ in (9) as

$$M^{cb}_{qr} = \frac{\langle w^c_q, w^b_r \rangle}{\|w^c_q\|\|w^b_r\|}, \quad (10)$$

where $w^c_q \in \mathbb{R}^{1 \times n_c}$ and $w^b_r \in \mathbb{R}^{1 \times n_b}$ are row vectors of $W^c$ and $W^b$, respectively. It is important to note that Eq. (10) is not well defined when $n_c \neq n_b$, thereby, a conventional interpolation method can be used for oversampling the lowest size vector to properly compute the inner product between $w^c_q$ and $w^b_r$. Our approach for Multiple Manifold Learning (MML) aims to calculate the relationships among samples of different manifolds, comparing the intra manifold similarities contained in each $W^c$.

Furthermore, it is possible to couple MKR with a multiple manifold learning approach (MML or LJM) by using $W^c$ (Eq. (6)) to compute $A$ in (9). Thence, a common low-dimensional space for visualizing the spatial and temporal dynamics of a given activity recorded by multiple videos, could be inferred.

5. Experimental set-up and results

5.1. Single video analysis

We test the MKR methodology (Section 2) for finding a low-dimensional space that allows to visually identify the spatial and temporal dynamics of a single real-world video. In this sense, four real-world datasets are studied: COIL-100 [11], CMU MoBo [13], Action [15], and Head. The first database is the Columbia Object Image Library, which contains 72 RGB-color images for several objects in PNG format. Pictures are taken while each object is rotated every 5 degrees from 0 to 360. We create a video of three repetitions of rotation for the Duck (object 15, Fig. 2 (e)). The size of the images is $128 \times 128$, and they are transformed to gray scale. Hence, an input space with $n = 216$ and $p = 16,384$ is obtained.

The second database, the CMU motion of body (Mobo) holds 25 individuals walking on a treadmill. Each person perform four different walk patterns: slow walk, fast walk, incline walk, and walking with a ball. The subjects are captured using six high resolution color cameras distributed around the treadmill. Each sequence is recorded at 30 frames per second. In this case, we used a one person sequence for slow walk captured from a side view (subject 02, camera vr03_7). The images are resized to $80 \times 61$ and transformed to gray scale. Further, we take the first 105 frames in JPG format, obtaining an input space with $n = 105$. 

![Fig. 2. Duck three cycles embeddings (n=216, p=16,384, m=3). (a) Spatial representation, (b) temporal representation, (c) spatio-temporal representation using MKR, (d) L-curve for parameter selection in MKR and (e) video sequence example.](image-url)
Fig. 3. Gait three cycles embeddings ($n=105, p=4880, m=3$). (a) Spatial representation, (b) temporal representation, (c) spatio-temporal representation using MKR, (d) L-curve for parameter selection in MKR and (e) video sequence example.

Fig. 4. Hand-waving two cycles embeddings ($n=80, p=19,200, m=3$). (a) Spatial representation, (b) temporal representation, (c) spatio-temporal representation using MKR, (d) L-curve for parameter selection in MKR and (e) video sequence example.
and \( p = 4880 \) (Fig. 3(e)). The third database, called Action, contains six types of human actions (walking, jogging, running, boxing, hand-waving and hand clapping) performed several times by 25 subjects. All sequences were down-sampled to the spatial resolution of 160 \( \times \) 120 pixels and have a length of four seconds in average. For concrete testing, we use one subject Hand-waving sequence (outdoor condition), transforming the images to gray scale. Thereby, we generate an input space with \( n = 80 \) and \( p = 19,200 \) (Fig. 4(e)). Finally, the fourth dataset is a rotating head video recorded by ourselves using a webcam with a resolution of 640 \( \times \) 480 pixels. Each frame is transformed to gray scale and resized to 60 \( \times \) 60 pixels (Fig. 5(e)), resulting an input space with \( n = 200 \) and \( p = 3600 \).

For all the experiments, the dimensionality of the embedded space is set to \( m = 3 \), and the number of nearest neighbors is fixed as \( k = 10 \) for the COIL-100 and Head, and \( k = 5 \) for Mobo and Action. A linear kernel to compute \( W_s \) and \( W_t \) in (5) is used. Moreover, we employ the proposed methodology for automatic parameter choice (Section 3) to select an optimum value for \( \xi_1 \) and \( \xi_2 \) in the mapping process, in order to obtain a suitable MKR of the studied video. The embedding results obtained by MKR with automatic parameter selection (our approach) are compared against the embedding only considering the spatial relationships (traditional NLDR methods) or the temporal similarities among frames. Figs. 2-5 show the results for single video analysis.

### 5.2. Multiple video analysis

The second kind of experiments aims to calculate a common low-dimensional space that allows to visualize the spatial and temporal dynamics of a given activity recorded by multiple videos. More precisely, our goal is to find out the visual representation of a given activity learning it from different subjects or objects. In this sense, we use MKR with automatic parameter selection to identify the relationships among frames of a single video, and then we use MML (Section 4) to compute the relationships among different manifolds.

Three real-world datasets are analyzed: COIL-100, CMU Mobo, and Action, which are described in Section 5.1. Here, more than one video of each database is considered. From the COIL-100 the Maneki-Neko and the Cat are tested (objects 14 and 17, respectively, Fig. 6(e)). For each object, a video sequence of 216 frames containing three rotation cycles is analyzed. So, the input space has the following characteristics: \( C = 2, n_1 = n_2 = 216, p_1 = p_2 = 16,384 \). On the other hand, the slow walk sequences of three persons from the Mobo database are used (subjects 02, 06, and 15; camera vr03_7; Fig. 7(e)). The first 150 frames are considered for each video, therefore, an input space with \( C = 3, n_1 = n_2 = n_3 = 105, p_1 = p_2 = p_3 = 4880 \) is obtained. Lastly, the Hand-waving sequences of two subjects are employed (subjects 02 and 16, Fig. 8(e)). So, an input space with \( C = 2, n_1 = n_2 = n_3 = 80, p_1 = p_2 = p_3 = 19,200 \) is obtained.

![Fig. 5. Juliana rotating head three cycles embeddings \((n = 200, p = 3600, m = 3)\). (a) Spatial representation, (b) temporal representation, (c) spatio-temporal representation using MKR, (d) L-curve for parameter selection in MKR and (e) video sequence example.](image-url)
The dimensionality of the embedded space is set to $m = 3$ for all the experiments, and the number of nearest neighbors is fixed as $k = 10$ for the COIL-100 and $k = 5$ for the Mobo dataset. A linear kernel to compute $W_s$ and $W_t$ in (5) is used. Besides, the performance of the proposed methodology is compared against the state of the art LJM [17]. Figs. 6–8 show the results for multiple video analysis.

6. Discussion

According to the single video analysis results for original LEM (spatial representation), which are shown in Figs. 2(a), 3(a), 4(a), and 5(a), it is possible to notice that traditional NLD formulations that only consider the spatial relationships among observations, lead in low-dimensional representations highlighting the underlying motion structure of the studied video. However, it does not allow to reveal the temporal dynamics of the data. Thus, it is not possible to separate each cycle of the movement. In addition, when just the temporal relationships among samples is considered to analyze a single video, the projections found by the NLDR algorithm lose the global structure of the phenomenon, it is shown in Figs. 2(b), 3(b), 4(b), and 5(b). Even when it is possible to visualize the temporal sequence without overlaps, the motion structure of the studied video is not found. So, the algorithm is not able to recognize the similarity among observations representing the same state of the movement for different repetitions of the activity.

Otherwise, the proposed methodology MKR, which is devised to comprise the spatial and temporal relationships of the input data, generates enhanced embeddings reflecting topological similarity. As a result, the final transformations display the real behavior of the data, where the correct geometry of the movement is preserved without overlapping different analyzed periods (Figs. 2(c), 3(c), 4(c), and 5(c)). This improvement can be explained because of the trade-off between the temporal and spatial weights. Particularly, the proposed methodology for automatic parameter selection is able to mix both topologies fixing suitable values for $\xi_1$ and $\xi_2$. Our method calculates an appropriated combination for both source of information, minimizing as well as possible the spatial and temporal error representation using an L-curve. This can be corroborated in Figs. 2(d), 3(d), 4(d), and 5(d).

On the other hand, the presented results for multiple video analysis using LJM (Figs. 6(a,b), 7(a,b), and 8(a,b)), let us to conclude that it cannot properly learn the relationships among objects performing the same activity. Even if sometimes it tries to identify the soft correspondence between similar frames, the embedded space does not expose the spatial and temporal dynamics of the phenomenon. Finally, the embedding results obtained with the proposed methodology MML for dealing with multiple videos, show that the computed low-dimensional space
exhibits the whole dynamic of a given activity. Figs. 6(c,d), 7(c,d), and 8(c,d) shown how this method learns the spatial and temporal relationships among frames of videos related to a similar activity, conserving a soft correspondence among different subjects.

7. Conclusions

In this work we learn both the spatial and temporal relationships among frames of cyclic motions from video. Therefore, we presented a nonlinear dimensionality reduction methodology based on Laplacian Eigenmaps and multiple kernel learning. Particularly, we showed that considering both spatial and temporal relationships among frames in video analysis enhances the data representability, revealing the underlying structure behind the samples in a low-dimensional space. Moreover, we proposed a method for automatic parameter selection that allows to combine properly different topologies, incorporating a tradeoff for weighting each representation. Thus, our method calculates the best combination for both source of information (spatial and temporal), minimizing as well as possible the error of each representation in the embedding space. According to the achieved results, we showed that if only the spatial topology (traditional NLDR approaches) is used to learn the data structure, the calculated low-dimensional representation can be just related to the data intrinsic geometry, but it is not able to differentiate repetitions or cyclic movements exhibiting a similar behavior. On the other hand, considering just the temporal information, the accomplished mapping may not reflect the global structure of the phenomenon. The spatial–temporal methodology learns the geometrical behavior and the time influence, which is suitable and required, particularly, for analyzing cyclic movements. Furthermore, we proposed a methodology to calculate a common low-dimensional space for multiple videos (multiple manifolds), which allows to identify visually the spatial and temporal dynamics of a given activity recorded from different subjects. According to the obtained results, our approach (MKR) seems to be a good alternative for analyzing cyclic human activities such as gait, rotating, and handwaving, improving traditional NLDR representations. Besides, the attained results for multiple video analysis using MKR with MML (our proposal) outperforms a previous similar work [17]. As future work, we are interested in apply these methodologies to support human motion classification and identification of impairments as well as for computer animation of living beings. Furthermore, based on the attained results of our work, and considering other multiple kernel approaches such as [21,12], we...
are interested in incorporating different similarity measures and feature representations in a NLDR framework, which could be useful in more complex video analysis.

Acknowledgements

This research was carried out under grants provided by the Research Center of Excellence in TIC – ARTICA, the projects 20201006599, 20201006574, and a M.Sc. scholarship funded by Universidad Nacional de Colombia, and the project 20201006594 found by Universidad de Caldas and Universidad Nacional de Colombia.

References


Andrés Álvarez-Meza received his undergraduate degree in electronic engineering from the Universidad Nacional de Colombia in 2009. Actually, he is pursuing a M.Sc. at the same university. His research interests are nonlinear dimensionality reduction and kernel methods for motion analysis and signal processing.

Juliana Valencia-Aguirre received her undergraduate degree in electronic engineering from the Universidad Nacional de Colombia in 2009. Actually, she is pursuing a M.Sc. at the same university. Her research interests are feature extraction for training pattern recognition systems, nonlinear dimensionality reduction for motion analysis and image processing.

Carlos Daniel Acosta-Medina received a B.S. degree in Mathematics from Universidad de Sucre in Colombia in 1996. In 2000 received a M.Sc. Mathematics degree and in 2008 a Ph.D. in Mathematics, both of them from Universidad Nacional de Colombia – Sede Medellin. Currently he is Associated Professor at Universidad Nacional de Colombia – Sede Manizales. His research interests are Regularization, Conservation Laws and Discrete Mollification.

Genaro Daza-Santacoloma received the B.S. degree in electronic engineering (2005), the M.Sc. degree in engineering-industrial automation with honors (2007), and the Ph.D. degree in engineering-automatics with honors (2010), from the Universidad Nacional de Colombia. Currently, he is Assistant Professor at Universidad Antonio Nariño, Bogotá where he is researching about human motion and subspace learning in collaboration with Signal Processing and Recognition Group from Universidad Nacional de Colombia. His research interests are feature extraction/selection for training pattern recognition systems, artificial vision, computer animation, and machine learning.

Carlos Daniel Acosta-Medina received a B.S. degree in Mathematics from Universidad de Sucre in Colombia in 1996. In 2000 received a M.Sc. Mathematics degree and in 2008 a Ph.D. in Mathematics, both of them from Universidad Nacional de Colombia – Sede Medellin. Currently he is Associated Professor at Universidad Nacional de Colombia – Sede Manizales. His research interests are Regularization, Conservation Laws and Discrete Mollification.

Germán Castellanos-Domínguez received his undergraduate degree in radiotechnical systems and his Ph.D. in processing devices and systems from the Moscow Technical University of Communications and Informatics, in 1985 and 1990, respectively. Currently, he is a professor in the Department of Electrical, Electronic and Computer Engineering at the Universidad Nacional de Colombia at Manizales. In addition, he is Chairman of the GCPDS at the same university. His teaching and research interests include information and signal theory, digital signal processing and bioengineering.