MoCap Data Segmentation and Classification Using Kernel Based Multi-channel Analysis

Sergio García-Vega, Andrés Marino Álvarez-Meza,
and César Germán Castellanos-Domínguez

Universidad Nacional de Colombia, Sede Manizales,
Signal Processing and Recognition Group
km 7 vía al Magdalena, Colombia
{segarciave,amalvarezme,cgcastellanosd}@unal.edu.co
http://portal.manizales.unal.edu.co/gta/signal/

Abstract. A methodology for automatic segmentation and classification of multi-channel data related to motion capture (MoCap) videos of cyclic activities are presented. Regarding this, a kernel approach is employed to obtain a time representation, which captures the cyclic behavior of a given multi-channel data. Moreover, we calculate a mapping based on kernel principal component analysis, in order to obtain a low-dimensional space that encodes the main cyclic behaviors. From such, low-dimensional space the main segments of the studied activity are inferred. Then, a distance based classifier is used to classified each MoCap video segment. A well-known MoCap database is tested which contains different activities performed by humans. Attained results shows how our approach is a simple alternative to obtain a suitable classification performance in comparison to complex methods for MoCap analysis.

Keywords: Multi-channel data, kernel methods, MoCap, human activity recognition.

1 Introduction

Human action recognition from video data are a growing area of study in the computer vision field. For a correct recognizing, it is necessary to develop a system that allows to identify and classify characteristic patterns from the input data [1] [2]. In real life, there are some human activities involving a cyclic behavior along the time, such as: walking, running, swimming, among others. Commonly, it is important to identify the main cyclic behavior that describes each action to find relevant information about the process [3]. For such purpose, it is necessary to develop three main stages: preprocessing, segmentation, and classification. However, the segmentation stage is not always developed in an automatic way, which can lead to unstable results and low classification performances. Moreover, when the data segmentation stage is fixed manually, it could lead in a time demanding process for the user. Then, it is necessary to develop an automatic segmentation stage that allows to obtain a suitable data analysis.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 495–502, 2013.

There are some works in the state-of-the-art related to the analysis of Motion Capture - MoCap data for human activity recognition. In [4], it is used a well-known MoCap database and the dynamics of each action class is modeled by a Bayesian based approach using Hidden Markov Models - HMM. The achieved accuracy classification results are over the 90%, however, the system requires that the multi-channel data is previously segmented, such that each segment contains a whole course of one action. Moreover, a complex classifier is employed to train the data, which requires a high computational load. Other approaches that require a manual MoCap data segmentation can be found in [5].

Here, a methodology for automatic segmentation and classification of multi-channel data is proposed. In this sense, a kernel function is employed to discover the time relationships among multi-channel data. Our aim is to highlight the cyclic behavior of the studied process, which is assumed to be hidden into the input samples. Indeed, an eigen-based decomposition is used to find a low-dimensional space that allows to segment the cyclic segments of the input data. Thus, proposed methodology is able to capture cyclic behaviors hidden into multi-channel data, avoiding the need of a manual segmentation that could lead in biased and unstable results. A well-known MoCap database is tested, which contains different activities executed by humans. Furthermore, two classification alternatives are studied: by considering each MoCap frame as an unique sample, and by considering a set of frames.

The remainder of this work is organized as follow. Section 2 introduces the proposed methodology for automatic segmentation and recognition of multi-channel data using kernel based methods. In Sections 3 and 4, the experimental results are described and discussed, respectively. Finally, in Section 5, the work conclusions are presented.

2 Kernel Based Multi-channel Data Representation

Let $\boldsymbol{X} \in \Re^{N \times P}$ be a multi-channel input matrix, with P channels and N samples, where $\boldsymbol{x}_i \in \Re^{1 \times P}$ is a row vector containing the information of all the provided channels at different time instants, with $i \in \{1, \ldots, N\}$. Our aim is to identify the main relationships that the channels share along the time to highlight hidden cyclic patterns into the studied process. For such purpose, a kernel function is employed to discover such relationships taking into account a nonlinear mapping $\varphi : \Re^{N \times p} \to \mathcal{H}$, where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space - RKHS [6]. Thus, the kernel based representation allows to deal with nonlinear structures that can not be directly estimated by traditional operators, such as, the linear correlation function. Regarding this, the inner product between two samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is computed in RKHS as $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j) \rangle_{\mathcal{H}}$, being $\kappa(\cdot, \cdot)$ a Mercer's kernel [6]. Taking advantage of the so-called kernel trick, the kernel function can be computed directly from $\boldsymbol{X}$. Here, the well-known Gaussian kernel is considered, which can be defined as

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), \tag{1}$$

being $\sigma \in \Re^{+}$ the kernel band-width. Then, from equation (1) the similarity matrix $\boldsymbol{S} \in \Re^{N \times N}$ can be estimated as $S_{i,j} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. It is important to note that other kind of kernels could be used, e.g. linear, polynomial, Laplacian, tangential, among others. However, due to the smooth nature of the input data and considering the universal approximating capability, the Gaussian kernel is used [7]. Each application task could be adapted or not to each kernel function according to the prior knowledge about the input data (see [6,8]). In this sense, $\boldsymbol{S}$ encodes the temporal dynamics of the multi-channel input data. Analyzing the pair similarities information into $\boldsymbol{S}$, it is possible to cluster (segment) samples that are related to a cyclic behavior of the studied process. Note that, the above mentioned kernel representation assumes that the multi-channel data shares an unique cyclic behavior. In case that the input data is composed by different processes, or when the multi-channel data is non-stationary, an unique kernel function could be not enough to deal with such changes along the time, not mentioning the need to consider the time structure of such kind of processes.

2.1 Automatic Multi-channel Data Segmentation

From the above mentioned kernel based multi-channel representation, and in order to find out the cyclic behavior into $\boldsymbol{X}$, we propose to use an eigen-based decomposition of $\boldsymbol{S}$ to calculate a low-dimensional space $\boldsymbol{Y} \in \Re^{N \times m}$, with $m < P$, which reveals the main components of $\boldsymbol{X}$. Therefore, the well-known Kernel Principal Components Analysis - KPCA algorithm is performed over $\boldsymbol{S}$. KPCA is a nonlinear generalization of PCA in the sense that it performs PCA in $\mathcal{H}$, which can be viewed as a feature space of arbitrarily large dimensionality [6]. Before applying KPCA, a Laplacian based normalization is employed to avoid the effect of outliers, thus, the matrix $\boldsymbol{L_M} \in \Re^{N \times N}$ is computed as $\boldsymbol{L_M} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D} \in \Re^{N \times N}$ is a diagonal matrix with elements $d_{ii} = \sum_{i=1}^{n} S_{ij}$. Afterwards, the low-dimension KPCA mapping is obtained as $\boldsymbol{Y} = \boldsymbol{L_M}\boldsymbol{V}$, where $\boldsymbol{V} \in \Re^{P \times m}$ is a matrix containing the first m eigenvectors of $\boldsymbol{L_M}$, after discarding the first one as trivial solution.

As a result, we obtain a low-dimensional representation that contains the main cyclic components of $\boldsymbol{X}$. Hence, we find the local maxima or *peaks* vector $\boldsymbol{\rho} \in \Re^{B}$, where B indicates the number of found *peaks* into the first coordinate (column vector) $\boldsymbol{y}$ of $\boldsymbol{Y}$. Note that, each column vector of $\boldsymbol{Y}$ could be related to a different cyclic component of $\boldsymbol{X}$. However, for complex dynamics and/or non-stationary environments, such components can mix more than one cyclic behavior. As a first approach, here, we assume that the given multi-channel data encodes an unique cyclic dynamic. Besides, the signal to noise ratio is high enough to ensure stable performances. Then, each element of $\boldsymbol{\rho}$ is estimated as follows. We compare each element y_i against its two nearest neighbors y_{i-1} and y_{i+1}. If y_i value is higher than the value of its neighbors, so, y_i is labeled as a local *peak* and $\rho_b = i$, with $b \in \{1, \ldots, B\}$. After that, we compute the differences between adjacent elements of $\boldsymbol{\rho}$ and finally, take into account the amount of *peaks* found B, we obtain $B - 1$ segments of $\boldsymbol{X}$. Fig. 1 illustrates

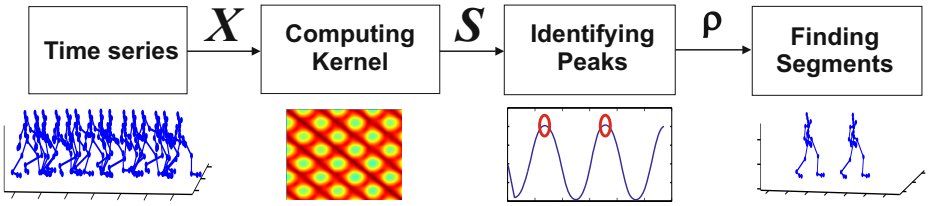


Fig. 1. Proposed methodology for multi-channel data segmentation

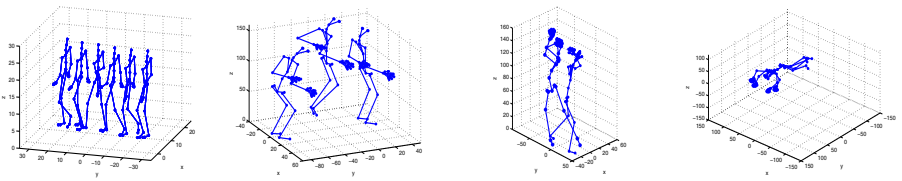
the proposed approach for automatic multi-channel segmentation using a kernel based representation (a motion capture video analysis example is described).

3 Experimental Set-up and Results

We test our automatic segmentation and classification methodology for multi-channel data analysis, using a well-known Motion Capture Database - MoCap database, with the purpose to find the main cyclic patterns of human motion activities. In this sense, the CMU MoCap is used[1]. Such data were recorded in a MoCap lab at Carnegie Mellon University, which contains 12 Vicon infrared MX-40 cameras, each of which is capable of recording 120 Hz with images of 4 mega pixel resolution. The cameras are placed around a rectangular area, of approximately $3m \times 8m$, in the center of the room. Subjects wear a black jump suit and have 57 markers taped on, and the Vicon cameras see the markers in infra-red. The images that the various cameras pick up are triangulated to get 3D data representation. The subjects are asked to perform several human motions activities, which are captured by the MoCap system. Then, a video in BVH format for each motion activity by a given subject is recorded. Thus, 146 videos of 31 different subjects are considered for 11 different activities: *jump*, *walk*, *run*, *marching*, *salsa dance*, *golf*, *boxing*, *swimming*, *yoga*, *monkey (human subject)* and *chicken (human subject)*. For each video, an input multi-channel matrix $\boldsymbol{X} \in \Re^{N \times P}$ is obtained, where $P = 57 \times 3 = 171$ corresponds to the 57 joints in 3D coordinates, and N represents the number of frames in the BVH file. As seen from Fig. 2, it is possible to notice some examples from the database used on this work.

In order to avoid the bias effect due to the subject translation along the 3-D space when performing a human activity, e.g. walking and running, a preprocessing stage is carried out, where each input frame is normalized with respect to the Hips joint 3-D coordinates. Thus is, this joint will be always centered at the $(0, 0, 0)$ position for every time instant. After that, we compute the kernel matrix as shown in equation 1, where σ is computed according to the empirical estimation of the Gaussian kernel band-width by the *Sylverman's rule* [9]. From such kernel based representation, we estimate the different segments for

[1] `http://mocap.cs.cmu.edu/`



(a) `walk` (b) `jump` (c) `golf` (d) `swimming`

Fig. 2. Some MoCap human activities representative frames

each video as described in subsection 2.1, fixing $m = 1$. Table 1 describes the amount of segments found automatically for each activity with our approach. Such segments contain the main cycles of the dynamics for the different considered activities. As result we found 910 different segments that represent 112045 frames. The main stages for the proposed automatic segmentation approach are presented in Fig. 3 for two MoCap videos examples.

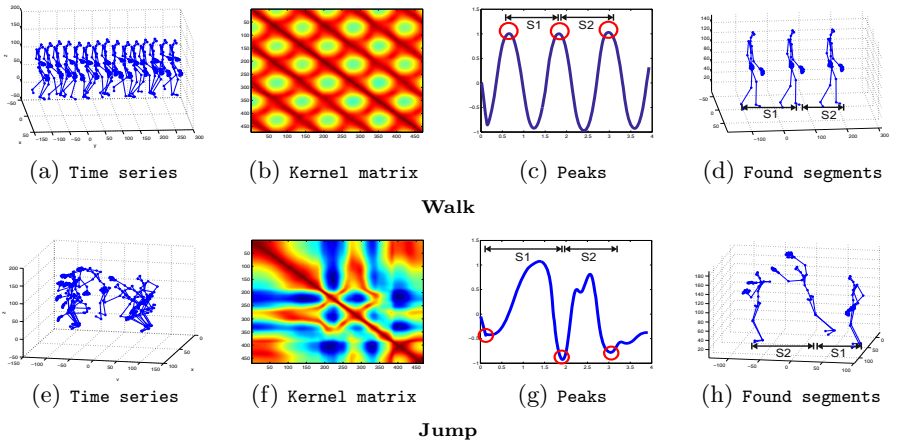
Furthermore, given the computed segments, the generalization abilities for the provided experimental conditions are tested by using a 10-fold cross-validation scheme. Regarding this, a k-nearest neighbors (KNN) classifier is used to recognize automatically different activities. The number of neighbors for this classifier is optimized with respect to the leave-one-out error of the training set. In this case, two kind of experiments are provided. First, each frame is employed as an unique sample. Second, for a given video segment, its class membership is estimated as the mode of the labels of the frames within the segment. In Tables 2 and 3 the mean confusion matrices for the above mentioned classification conditions are presented. Finally, at the bottom of the Table 3, the performance of the proposed methodology is compared against the results obtained in [4].

4 Discussion

At the top of the Fig. 3, it is possible to see the main segmentation results by using the proposed approach to analyze a walk MoCap video. Particularly, Fig. 3 *(b)* shows the computed kernel matrix, which properly identifies the cyclic similarities (green circles) into the video. Now, Fig. 3 *(c)* describes how our method

Table 1. Number of identified segments per each human activity

Activity	Found Segments	Activity	Found Segments
jump	68	golf	23
walk	127	boxing	36
run	78	swimming	41
marching	81	yoga	86
salsa dance	114	monkey(HS)	135
chicken(HS)	121		
Total: 910 Segments			



(a) Time series (b) Kernel matrix (c) Peaks (d) Found segments

Walk

(e) Time series (f) Kernel matrix (g) Peaks (h) Found segments

Jump

Fig. 3. Some automatic segmentation results - Main stages

Table 2. Mean confusion matrix - frames classification results

	Jump	Walk	Run	Marching	Salsa dance	Chicken (HS)	Golf	Boxing	Swimming	Yoga	Monkey (HS)
Jump	92,94	0,16	1,75	0,52	0,96	0,00	0,60	0,00	0,33	1,97	0,46
Walk	1,06	98,22	6,00	0,84	0,52	0,00	0,70	0,00	0,08	0,10	0,00
Run	0,70	1,06	79,84	3,25	1,24	0,00	0,58	0,00	1,93	0,63	0,00
Marching	0,37	0,42	5,16	94,34	0,95	0,00	0,21	0,00	1,47	0,84	0,00
Salsa dance	0,69	0,05	3,38	0,89	93,13	0,00	1,50	0,00	0,39	2,86	0,00
Chicken (HS)	0,00	0,00	0,00	0,00	0,08	100,00	0,00	0,00	0,00	0,14	0,00
Golf	0,09	0,00	0,62	0,00	0,97	0,00	92,28	1,47	0,00	0,80	0,00
Boxing	0,22	0,00	0,00	0,00	0,21	0,00	2,75	97,73	0,00	0,98	0,00
Swimming	0,00	0,09	2,40	0,00	0,12	0,00	0,39	0,10	95,80	0,00	0,00
Yoga	3,83	0,00	0,34	0,15	1,38	0,00	0,77	0,44	0,00	90,02	0,62
Monkey (HS)	0,10	0,00	0,52	0,00	0,45	0,00	0,23	0,27	0,00	1,64	98,92

Table 3. Mean confusion matrix - video segments classification results

	Jump	Walk	Run	Marching	Salsa dance	Chicken (HS)	Golf	Boxing	Swimming	Yoga	Monkey (HS)
Jump	98,57	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,83
Walk	0,00	100,00	2,50	0,00	0,91	0,00	0,00	0,00	0,00	0,00	0,00
Run	0,00	0,00	93,75	1,25	0,00	0,00	0,00	0,00	2,36	0,00	0,00
Marching	0,00	0,00	0,00	98,75	0,00	0,00	0,00	0,00	2,50	0,71	0,00
Salsa dance	0,00	0,00	2,50	0,00	99,09	0,00	0,00	0,00	0,00	1,48	0,00
Chicken (HS)	0,00	0,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00	0,00
Golf	0,00	0,00	0,00	0,00	0,00	0,00	97,50	2,00	0,00	0,00	0,00
Boxing	0,00	0,00	0,00	0,00	0,00	0,00	2,50	98,00	0,00	0,00	0,00
Swimming	0,00	0,00	1,25	0,00	0,00	0,00	0,00	0,00	95,14	0,00	0,00
Yoga	1,43	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	96,32	0,00
Monkey (HS)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,48	99,17

	Simple Activities		Complex Activities	
	Benchmark [4]	Kernel Multi-Channel	Benchmark [4]	Kernel Multi-Channel
Mean accuracy	91.96%	97.77%	92.14%	97.88%

models, in one-dimensional coordinate, the main relationships among frames. In this case, due to walking is a slow motion with smooth changes between adjacent frames, the KPCA mapping can be related to a *sin* function. Now, from Fig. 3

(d) 3 peaks are calculated (red circles), which properly identify the 2 gait cycles performed by the subject. Analogously, at the bottom of the Fig. 3, we have the main segmentation results for a jump MoCap video. Note that, even when jumping is a cyclic activity with a stronger dynamic of change than walking, our approach is able to infer such behaviors. As seen in Fig. 3 *(f)* the computed kernel matrix highlights 2 set of frames that share a strong similarity into them. Such segments can be identified in the first KPCA coordinate as presented in Fig. 3 *(g)*. Again, note that how our approach is able to track the activity cyclic behavior, even when *S1* is smoother and longer than *S2*.

Regarding to the classification results, as can be seen in Table 2, the mean confusion matrix for the frame based classification scheme demonstrates how our approach obtains a suitable recognition accuracy. Overall, performances over the 90% are attained for all the provided classes. The worst result is obtained for *run*, where the system is confused with *walk* and *march* classes. Above drawback is expected considering that *run* is the class with lowest number of segmented sequences (see Table 1). Moreover, a frame based classification could not be the best alternative to differentiate between activities that share many MoCap poses, e.g., run and walk. Thus is, such video segments are conformed by some frames where the spatial position of the human body joints are similar for both activities. It is important to note that our method, in most of the cases, obtains a better frame based classification performance in comparison to a closed work presented in [4]. Moreover, our approach is a simple solution that includes both, data segmentation and classification.

Now, taking into account the segment based classification scheme results presented in Table 3, it is possible to see how such alternative is more stable than the frame based classification. Attained results describe an average accuracy over the 95%. Particularly, the worst frame based classification performance (*run*) is improved from $79,84\%$ to $93,75\%$. Above system behavior can be explained by the fact that a segment classification decision considers the mode of the frame labels as the segment membership. So, the mode function can be viewed as a filter that is robust against wrong decisions due to pose mistakes (human body joint similarities). Finally, at the bottom of the Table 3, the performance of the proposed methodology is compared against the results obtained [4]. The classification success of our method lies in the automatic segmentation approach, which suitable identifies the main dynamic cycles of the process.

5 Conclusions

A methodology for automatic segmentation and classification of multi-channel data was presented. In this sense, a kernel based representation is employed to find out the time relationships among channels. Then, a KPCA mapping is calculated to highlight the main dynamics of the studied process in a low-dimensional space. From such low-dimensional space, a local minimum based method is used to cluster different time segments that share a common behavior. Therefore, our approach is able to capture cyclic behaviors hidden into multi-channel data.

A well-known MoCap database was tested, which contains different activities executed by humans. For concrete testing, proposed approach is used to segment automatically the video data. Such segments are employed to train a *k-nearest* neighbors (KNN) classifier for recognizing automatically different activities. Besides, two kind of classify experiments are carried out: by considering each frame as an unique sample, and by considering a set of frames (video segment). The attained results showed that our approach is a simple but efficient alternative to obtain a suitable classification performance in comparison to other complex state of the art methods related to MoCap data classification. Besides, state art methods employs, in most of the cases, a manually video segmentation, which can lead to subjectively results and inefficient real-world implementations. As future work, we are interested in test our methodology in other kind of human activities that involve different cyclic patterns and non-stationary environments by coupling the proposed method with an online based adaptive filter scheme.

Acknowledgements. Research carried out under grants provided by Jóvenes Investigadores e Innovadores - 2012, and a Ph.D. scholarship funded by Colciencias.

References

1. Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., Shotton, J.: Real-time human pose recognition in parts from single depth images. In: CVPR, pp. 1297–1304 (2011)
2. Gan, Y.W.C., Wang, X.: Human motion segmentation by rpca with augmented lagrange multiplier. In: ICALIP, pp. 379–383 (2012)
3. Murugappan, M., Basah, S.N.B., Yaacob, S.B., Ismail, K.N.S.B.K.: Human postures modeling using motion analysis: A review. In: International Conference on Biomedical Engineering (ICoBE), pp. 280–285 (2012)
4. Lv, F., Nevatia, R.: Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 359–372. Springer, Heidelberg (2006)
5. Lan, Z.-Z., De la Torre, F., Hoai, M.: Joint segmentation and classification of human actions in video. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3265–3272 (2011)
6. Scholkopf, B., Smola, A.J.: Learning with Kernels. The MIT Press, Cambridge (2002)
7. Liu, W., Príncipe, J.C., Haykin, S.: Kernel Adaptive Filtering: A Comprehensive Introduction. John Wiley & Sons, Inc. (2010)
8. Genton, M.G., Cristianini, N., Shawe-taylor, J., Williamson, R.: Classes of kernels for machine learning: a statistics perspective. Journal of Machine Learning Research 2, 299–312 (2001)
9. Sheather, S.J.: Density estimation. Statistical Sci. 19, 588–597 (2004)