A COMPARATIVE STUDY OF REPRESENTATIONS FOR FOLK DANCES RECOGNITION IN VIDEO

Eftychia Fotiadou, Nikos Nikolaidis, Anastasios Tefas

Department of Informatics
Aristotle University of Thessaloniki
{eftifot, nikolaid, tefas} @aiia.csd.auth.gr

ABSTRACT

Dance traditions constitute a significant aspect of cultural heritage around the world. The organization, semantic analysis, and retrieval of dance-related multimedia content (i.e., music, video) in databases is, therefore, crucial to their preservation. In this paper we explore the problem of folk dances recognition from video recordings, focusing on Greek folk dances, using different representations for the data. To this end we have employed the well-known Bag of Words model, in combination with dense trajectories, as well as with streaklines descriptors. Furthermore, we have adopted a representation based on Linear Dynamic Systems, including a novel variant that uses dense trajectories descriptors instead of pixel intensities. The performance of the aforementioned representations is evaluated and compared, in a classification scenario involving 13 different dance classes.

Index Terms— dance recognition, Bag of Words model, Linear Dynamic Systems, dense trajectories, streaklines

1. INTRODUCTION

Activity recognition is a well studied problem in the field of computer vision and several approaches have been proposed in the last years [1], [2], using mainly video data. The automatic classification of human activities in distinct classes constitutes a crucial part in various domains, such as human-computer interaction systems, health and elder care applications, video surveillance, and indexing in video databases.

In this paper, we deal with a more complex activity class, namely dance. Dance can be thought of as a composition of elementary motions (i.e. dance steps), the combination of which, in a certain temporal order, gives rise to the choreography. Therefore, the activity of dancing, unlike other types of everyday activities commonly treated in activity recognition research (e.g. running, jumping or waving), is characterized by high complexity and variation, rendering the recognition problem more challenging.

Although a significant amount of research has been conducted in the field of recognition of general activities, the bibliography regarding dance classification is very limited. Moreover, the majority of the existing dance recognition algorithms operate upon motion capture data, or, on 3D skeletal motion representations. A real-time framework for classifying dance gestures using skeleton animation data can be found in [3]. It introduces an angular representation for the skeleton and a cascaded correlation-based classifier. In [4], an alternative SVD-based scheme, called Segmental SVD is introduced, which is suitable for recognizing dance motions obtained through motion capture. A Dynamic Time Warping (DTW)-based method is proposed in [5], for detection of motion patterns in dance movements, in the context of an interactive dance system. In [6], a multimodal approach for classification of Salsa dance movements is presented. The proposed method fuses information from the 3D trajectories of body joints and from the footstep impacts on the floor. In [7], the authors approach the problem of classification of ballet moves recorded with the Kinect sensor. Dance sequences are represented using the joint angle information of the legs, while Nearest Neighbor and Support Vector Machine (SVM) classifiers are used for classification. Another approach for ballet moves classification can be found in [8], where a deformable model is introduced for the representation of the joint trajectories obtained through motion capture.

Certain methods are also concerned with the recognition of folk dances, which can be regarded as more relevant to our problem. Such a method is presented in [9], where dance gesture recognition of Bali traditional dances is performed. The proposed approach employs a skeletal representation similar to the one introduced in [3] for the motion data, while a linguistics motivated method is used for classification. An approach for recognizing gestures in Indian classical dance using 3D coordinate data recorded with the Kinect sensor can be found in [10]. Finally, a video-based method for recognition of different Indian classical dances is proposed in [11]. According to it, each frame in a video sequence is represented

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS-UOA-ERASITECHNIS MIS 375435.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

in a hierarchical way, using a pose descriptor which is based in the histogram of oriented optical flow. Classification of the video data is subsequently performed using a SVM classifier.

In this paper, we study the problem of distinguishing between different folk dances recorded in video. We focus our efforts on Greek folk dances, although the investigated approaches can be used for the recognition of such dances from other countries or regions. Usually, folk dances exhibit a great variability according to the region they originate from and constitute an important part of a country's cultural heritage. In folk dances, old customs, aspects of everyday life, as well as historic events are usually reflected.

Recognition of folk dances is a very challenging problem in comparison to recognition of general activities, due to the complexity and the particularities of this specific motion category. This is particularly true for Greek folk dances. Often, different folk dances consist of similar steps or have the same rhythm. Furthermore, large intra-class variabilities are observed. The tempo may change from slow to fast within the same song across various performances of the same dance, which strongly affects the execution style. In addition, the dancers' costumes may differ across performances. Finally, further obstacles in the recognition task are introduced by occlusions in dance recordings. These occlusions might occur due to the fact that folk dances are often performed by a group of dancers in circle (as is the case of Greek folk dances), or due to the long skirts sometimes worn by female dancers, hiding the feet movements.

The work we present here extends our previous work on the folk dances recognition problem [12], by exploring alternative representations for the dance data and by considering a more complex classification scenario, involving 13 different classes instead of 5. In contrast to the majority of the aforementioned methods, which are based on motion capture data, our approach is based on video recordings. In order to classify (Greek) folk dances, we explore a number of different approaches. First, we apply the method proposed in [13] for activity recognition, where a dense trajectory representation is combined with a Bag of Words (BoW) model and SVMs for classification. This method was employed in our previous work [12], and here it serves as a baseline. Second, we adopt a representation for the video data based on streaklines. Streaklines are borrowed from the fluid mechanics field, where they have been used for modeling the flow vector field. In [14], they were employed in order to model the flow in crowded scenes, with application to crowd segmentation and abnormal behavior detection. Following [14], we apply the streakline representation to an activity recognition problem, namely folk dances recognition, and integrate it into a BoW+SVMs classification scheme. Finally, we adopt a Linear Dynamic Systems (LDS)-based representation along with the Bag of Systems (BoS) model proposed in [15]. This representation is further extended in a novel way, by combining a dense trajectory-based feature descriptor. In the following sections, the aforementioned methodologies for folk dances recognition are discussed in more detail.

2. METHODS DESCRIPTION

2.1. Dense Trajectories and Bag of Words model

Dense trajectories ([16],[13]) have been proved very efficient feature descriptors in human activity classification tasks. Their calculation is performed by first densely sampling points of the video at different spatial scales. The sampled points located in video regions where motion occurs are tracked on each spatial scale separately, using the information from the dense optical flow field. A trajectory is formed by points in successive frames: $(\mathbf{P}_t, \mathbf{P}_{t+1}, \mathbf{P}_{t+2}, ..., \mathbf{P}_{t+L})$, where L is the length of the trajectory. Consequently, each trajectory is represented by a descriptor given by the following formula:

$$\mathbf{T} = \frac{(\Delta\mathbf{P}_t, \Delta\mathbf{P}_{t+1}, ..., \Delta\mathbf{P}_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta\mathbf{P}_j\|}, \tag{1}$$

where $\Delta\mathbf{P}_t = (\mathbf{P}_{t+1} - \mathbf{P}_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$ is the displacement vector between two successive frames. Apart from the aforementioned trajectories, the following additional descriptors are calculated in the spatio-temporal volumes surrounding each trajectory: Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), as well as Motion Boundary Histograms in horizontal (x) and vertical (y) directions (MBHx and MBHy respectively).

The process described above results in the extraction of five different descriptor types for each trajectory. Subsequently, a BoW model is constructed for each descriptor type separately. The first step of the BoW construction is the creation of a codebook: by applying k-means clustering on the features calculated for the training data, a set of K codewords is obtained. Subsequently, a histogram of occurrences of each codeword is calculated for a training video, by assigning each feature vector of the video to the closest codeword. Using the same procedure, a histogram for a test sequence can be obtained. Once the training and testing histograms have been calculated, classification is performed by means of a SVM classifier. For the SVM classifier, a χ^2 kernel is selected, given by the following formula:

$$K(\mathbf{s}_i, \mathbf{s}_j) = \exp(-\frac{1}{A} D(\mathbf{s}_i, \mathbf{s}_j)) \tag{2}$$

where $D(\mathbf{s}_i, \mathbf{s}_j)$ is the χ^2 distance between histograms $\mathbf{s}_i = \{s_{i1}, s_{i2}, ..., s_{iK}\}$ and $\mathbf{s}_j = \{s_{j1}, s_{j2}, ..., s_{jK}\}$:

$$D(\mathbf{s}_i, \mathbf{s}_j) = \frac{1}{2} \sum_{k=1}^{K} \frac{(s_{ik} - s_{jk})^2}{s_{ik} + s_{jk}}, \tag{3}$$

and A is a scaling factor, calculated as the mean of χ^2 distances between all training samples.

Fusion of the five histograms, corresponding to the different descriptor types, is performed using two alternatives: first we calculate separate kernels for each descriptor type and subsequently add them. Second, the five different histograms are concatenated in a feature vector and a single kernel is calculated.

2.2. Streaklines and Bag of Words model

Streaklines are commonly used in fluid mechanics and fluid visualization in order to describe the flow fields. A streakline is defined by the locations of all the fluid particles at a given time that passed through a particular point [14]. Let us denote with $(x_i^{\mathbf{p}}(t), y_i^{\mathbf{p}}(t))$ the position of a particle at time step t, which was initialized at point $\mathbf{p}$ at time step (frame) i, where $i = 1, 2, ..., t$. We suppose that, at each time step, a new particle is initialized at point $\mathbf{p}$, therefore, $(x_i^{\mathbf{p}}(i), y_i^{\mathbf{p}}(i)) = (x_0^{\mathbf{p}}(0), y_0^{\mathbf{p}}(0))$. The particle motion is calculated according to:

$$\begin{aligned} x_i^{\mathbf{p}}(t+1) &= x_i^{\mathbf{p}}(t) + u(x_i^{\mathbf{p}}(t), y_i^{\mathbf{p}}(t), t) \\ y_i^{\mathbf{p}}(t+1) &= y_i^{\mathbf{p}}(t) + v(x_i^{\mathbf{p}}(t), y_i^{\mathbf{p}}(t), t), \end{aligned} \tag{4}$$

where u and v are the (optical) flow components in the horizontal and vertical direction, respectively. At each time step, the particles are advected following the above equation, while a new particle is initialized at each pixel. A streakline is, therefore, formed by the positions at a specific time, of all the particles that were initialized at the same pixel. More detailed information on the streaklines calculation can be found in [14].

If a video frame consists of N pixels, N streaklines will be calculated at each frame. Since using the information from all the frame's streaklines results in increased computing time and storage demands, and given that only certain regions in the frame contain information associated to the depicted motion, only the streaklines in these regions were selected. Specifically, we select the streaklines at regions of size 10×10 surrounding the critical points of the optical flow field. As critical points we considered the extrema of the irrotational component of the optical flow field.

Using the aforementioned procedure, each video is represented by a set of streaklines and a BoW model is calculated, as described in section 2.1. In more detail, k-means clustering is applied on the streaklines from the training videos and a codebook of K codewords is obtained. Subsequently, a histogram of frequencies of occurrence of each codeword is calculated for the training and test videos. Finally, the histograms are classified using an SVM with a χ^2 kernel.

2.3. Linear Dynamic Systems and Bag of Systems model

Linear Dynamic Systems (LDS) have been successfully used in order to model dynamic textures [17], [15], as well as for activity recognition [18], [19], [20]. In order to model the dance video data using LDSs, we followed the methodology proposed in [15]. The main concept is to divide each video into spatio-temporal volumes (cuboids) and model each one by a LDS. Let us denote a spatio-temporal volume consisting of F frames with $\{\mathbf{I}(t) \in \mathbb{R}^p\}_{t=1}^F$, where p is the number of pixels in each frame. The pixels intensity $\mathbf{I}(t)$ of a frame t can be modeled as a LDS:

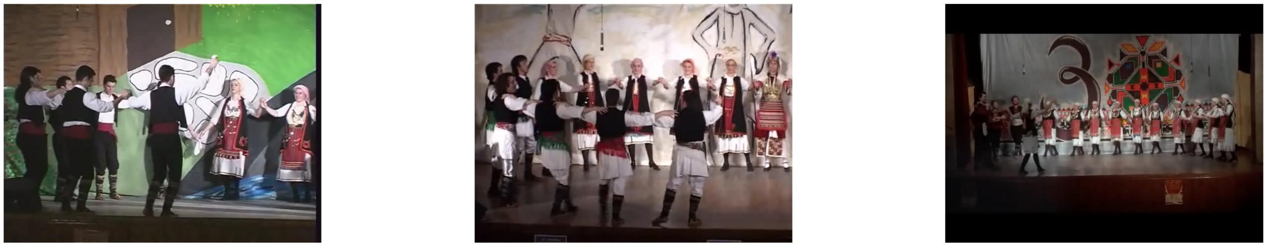
$$\begin{aligned} \mathbf{z}(t+1) &= \mathbf{A}\mathbf{z}(t) + \mathbf{B}\mathbf{v}(t), \\ \mathbf{I}(t) &= \mathbf{C}^0 + \mathbf{C}\mathbf{z}(t) + \mathbf{w}(t), \end{aligned} \tag{5}$$

where $\mathbf{z} \in \mathbb{R}^n$ denotes the hidden state at t, matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ models the dynamics of the state vector, matrix $\mathbf{C} \in \mathbb{R}^{p\times n}$ transforms the state vector to the system's output, $\mathbf{C}^0 \in \mathbb{R}^p$ is the mean of the spatio-temporal volume, while $\mathbf{w}(t) \sim \mathcal{N}(0, \mathbf{R})$ and $\mathbf{B}\mathbf{v}(t) \sim \mathcal{N}(0, \mathbf{Q})$ are the measurement and the process noise, respectively. In the aforementioned equations, n denotes the order of the system. In order to identify the parameters of the dynamical model for a specific spatio-temporal video volume, the method proposed in [21] is employed, which uses a Principal Component Analysis (PCA)-based approach. The LDS is finally represented by a tuple $M = (\mathbf{A}, \mathbf{C})$.

Subsequently, a variate of the BoW model, called Bag of Systems (BoS) model, is constructed. Similar to the BoW model, the BoS represents a sample (video), using a histogram of occurrences of codewords from a codebook. Since the LDS descriptors do not lie in a Euclidean space, the application of common clustering techniques, such as the k-means, is not straightforward. Therefore, in order to construct a codebook from a training set of LDSs, a suitable distance metric needs to be determined. To this end, the Martin distance between two systems M_1, M_2 is adopted, which is based on the subspace angles between the two systems.

Let us denote with $\mathbf{D} \in \mathbb{R}^{T\times T}$ the matrix of Martin distances between the LDSs , where T is the total number of LDSs extracted from a training videos set. By applying Multidimensional Scaling (MDS) on matrix $\mathbf{D}$, a mapping of the data to a space of lower dimension, $\mathbf{M}_i^e \in \mathbb{R}^{d_e}, i = 1, ..., T$ can be obtained, where d_e denotes the dimensionality of the low dimensional space. This mapping produces a set of Euclidean low dimensional points, which approximately preserve the distances in the original, non-Euclidean space, i.e. the Martin distances between LDSs. Therefore, since the low-dimensional space is Euclidean, k-means clustering can be applied on the transformed data, and a set of K codewords $\mathbf{w}_i^e, i = 1, ..., K$ is determined. However, these codewords do not correspond to any of the original LDSs and, due to the application of MDS, there is not a mapping from the low dimensional to the original space. In order to obtain LDS codewords, such that the assignment of LDSs to codewords is possible, training LDSs are chosen whose mapped points lie closest to the cluster centers $\mathbf{w}_i^e$.

After the LDS codewords have been determined, each video is represented by histogram of occurrences of the code-



(a) (b) (c)

Fig. 1: Sample frames from different dances in our dataset: (a) Ramna, (b) Stankena, (c) Syrtos.

words $\mathbf{h} = \{h_1, ..., h_K\}$. In more detail, each LDS of a video is assigned to the closest codeword and, subsequently a histogram is calculated by counting the number of occurrences of the codewords. Following the aforementioned process, we calculate histograms for the training as well as for the testing videos. For the classification of the histograms we employed a multi-class SVM classifier with a χ^2 kernel.

In the method described in [15], the video information used is in the form of pixel intensity values. Apart from this description, we also tested a novel variant that employed the dense trajectories descriptors. In this case, each frame of a spatio-temporal volume is represented by a set of dense trajectories. Since the number of trajectories varies from frame to frame, and in order to obtain a feature vector of constant length for all the frames, we first construct a BoW model for the dense trajectories, using the methodology described in 2.1. In this way, each frame of a spatio-temporal volume is described by the histogram of occurrences of the -trajectory based- codewords. Subsequently, the LDSs are calculated based on these histograms, using exactly the same procedure as before. By modeling a video using the dense trajectory descriptors instead of using all the pixel values of a frame, we automatically select regions of interest, i.e., regions in the video where motion is observed.

3. DATASET AND EXPERIMENTAL EVALUATION

The dataset we used in our experiments consists of 39 video recordings (3 for each dance) from the following 13 different Greek folk dances: Aloniotikos, Kori Eleni, Mpaintouskino, Mpoufiko, Poustseno, Ramna, Sarakina, Stankena, Syrtos, Tikfeskino, Tsourapia, Zavlitsena and Zaramo. Their duration varies between 1-2 min. From each dance, two videos were used for training and one for testing. Therefore, the training set consisted of 26 videos, while the testing of 13 videos. Several of the dances included in the dataset are very similar to each other with regards to the rhythm and steps they comprise of. Furthermore, they are performed in a circle, which results in frequent occlusions. Finally, the recordings for each different dance (class) are obtained from different performances, meaning that the recorded conditions, i.e. background, lighting, camera motion, vary significantly in the training and testing sets. These facts render the recognition task particularly challenging, but also more realistic, at the same time. Sample frames of the videos used in the experiments are shown in Fig. 1, where the aforementioned challenges of the recognition problem are obvious.

As far as the dense trajectories features are concerned, the default length (equal to 15) produced optimal results. For the length of the streaklines representation, we also experimented with different values, and optimal results were obtained for streaklines of length 20. In both dense trajectories and streaklines representations, the BoW models were based on codebooks of 4000 words. For the implementation of the LDS-based method we used the toolbox provided in [22]. In the case the pixel values were used for the video description, spatio-temporal volumes of size $60 \times 60 \times 50$ were selected. In the case of dense trajectory-based description, the five different descriptors were concatenated in a single vector (of 426 elements) and a codebook of 4000 codewords was selected for the BoW model. Furthermore, spatio-temporal volumes of size $120 \times 120 \times 50$ were used. The aforementioned sizes for the spatio-temporal volumes were determined by trying different sizes and selecting the ones that yielded the best classification performance. During the construction of the Bag of Systems (BoS) model, different values in the range $[10, 300]$ were tested for the number of codewords K, and an optimal performance was observed for 70 codewords. For the multi-class SVM classifiers used by all the methods, a 1-vs-all approach was adopted.

In Table 1, the classification rates for the different methodologies are presented. As can be observed, the streaklines and the LDS-based methods achieved better performance than the baseline dense trajectories method proposed in [12]. In the case of LDS, the novel dense trajectory-based representation of the data yielded an improved classification rate over the standard pixel intensity-based approach.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have studied the problem of recognizing different folk dances, with application in Greek dances. To this

Table 1: Recognition rates for the proposed methods

Method	**Rate**
Dense trajectories + BoW (kernel addition) [12]	15.38%
Dense trajectories + BoW (hist. concatenation) [12]	23.08%
Streaklines	**38.46%**
LDS	30.77%
LDS + dense trajectories	**38.46%**

end, we have employed different methodologies previously used in generic activity recognition tasks, and performed experiments in order to compare their performance. In a classification scenario consisting of 13 classes, the streaklines and the LDS representations exhibited significantly better classification rates than the dense trajectories, which had been previously used for the Greek folk dance recognition problem [12]. Although the obtained recognition rates are rather low, one has to interpret them having in mind the increased challenges and complexity of folk dances recognition compared to generic single-person activity recognition. Moreover, the results show that even state of the art approaches coming from the field of human activity recognition are not very suitable for (folk) dances recognition. Future work will target experimentation on a larger dataset, design and experimental testing of novel approaches tailored to this specific challenging problem, as well as incorporation of audio (music) information.

5. REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[2] P.V.K. Borges, N. Conci, and A. Cavallaro, "Video-Based Human Behavior Understanding: A Survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1993–2008, 2013.

[3] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proc. 2011 ACM SIGGRAPH/Eurographics Symposium Computer Animation*, 2011, pp. 147–156.

[4] L. Deng, H. Leung, N. Gu, and Y. Yang, "Recognizing Dance Motions with Segmental SVD," in *Proc. 20th Int. Conf. Pattern Recognition*, 2010, pp. 1537–1540.

[5] H. Pohl and A. Hadjakos, "Dance Pattern Recognition using Dynamic Time Warping," in *Proc. 7th Sound and Music Computing Conf.*, 2010, pp. 183–190.

[6] A. Masurelle, S. Essid, and G. Richard, "Multimodal classification of dance movements using body joint trajectories and step sounds," in *Proc. 14th Int. Workshop Image Analysis for Multimedia Interactive Services*, 2013, pp. 1–4.

[7] J. Dancs, R. Sivalingam, G. Somasundaram, V. Morellas, and N. Papanikolopoulos, "Recognition of ballet micro-movements for use in choreography," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2013, pp. 1162–1167.

[8] S. Boukir and F. Chenevière, "Compression and recognition of dance gestures using a deformable model," *Pattern Anal. Appl.*, vol. 7, no. 3, pp. 308–316, 2004.

[9] Y. Heryadi, M.I. Fanany, and A.M. Arymurthy, "Grammar of dance gesture from bali traditional dance," *Int. J. Computer Science Issues*, vol. 9, pp. 144–149, 2012.

[10] S. Saha, S. Ghosh, A. Konar, and A.K. Nagar, "Gesture Recognition from Indian Classical Dance Using Kinect Sensor," in *Proc. 5th Int. Conf. Computational Intelligence, Communication Systems and Networks*, 2013, pp. 3–8.

[11] S. Samanta, P. Purkait, and B. Chanda, "Indian classical dance classification by learning dance pose bases," in *Proc. IEEE Workshop Applications of Computer Vision*, 2012, pp. 265–270.

[12] I. Kapsouras, S. Karanikolos, N. Nikolaidis, and A. Tefas, "Feature comparison and feature fusion for traditional dances recognition," in *Proc. 14th Int. Conf. Engineering Applications of Neural Networks*, 2013, pp. 172–181.

[13] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 3551–3558.

[14] R. Mehran, B. E. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. 11th European Conf. Computer Vision*, 2010, pp. 439–452.

[15] A. Ravichandran, R. Chaudhry, and R. Vidal, "Categorizing dynamic textures using a bag of dynamical systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 342–353, 2013.

[16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[17] P. Saisan, G. Doretto, Y. Wu, and S. Soatto, "Dynamic texture recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, vol. 2, pp. 58–63.

[18] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos, "Recognizing activities via bag of words for attribute dynamics," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2587–2594.

[19] H. Wang, C. Yuan, G. Luo, W. Hu, and C. Sun, "Action recognition using linear dynamic systems," *Pattern Recognition*, vol. 46, no. 6, pp. 1710 – 1718, 2013.

[20] W. Wang, H. Liu, L. Yu, and F. Sun, "Human activity recognition using smart phone embedded sensors: A Linear Dynamical Systems method," in *Proc. Int. Joint Conf. Neural Networks*, 2014, pp. 1185–1190.

[21] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int. J. Comput. Vision*, vol. 51, no. 2, pp. 91–109, Feb. 2003.

[22] A. Ravichandran, R. Chaudhry, and R. Vidal, "Dynamic texture toolbox," 2011.