Skeleton-based Human Action Recognition using Basis Vectors

Stylianos Asteriadis
Department of Knowledge Engineering
University of Maastricht, Netherlands
stelios.asteriadis@maastrichtuniversity.nl

Petros Daras
Information Technologies Institute
Centre for Research and Tehcnology, Hellas
daras@iti.gr

ABSTRACT

Automatic human action recognition is a research topic that has attracted significant attention lately, mainly due to the advancements in sensing technologies and the improvements in computational systems' power. However, complexity in human movements, input devices' noise and person-specific pattern variability impose a series of challenges that still remain to be overcome. In the proposed work, a novel human action recognition method using Microsoft Kinect depth sensing technology is presented for handling the above mentioned issues. Each action is represented as a basis vector and spectral analysis is performed on an affinity matrix of new action feature vectors. Using simple kernel regressors for computing the affinity matrix, complexity is reduced and robust low-dimensional representations are achieved. The proposed scheme loosens action detection accuracy demands, while it can be extended for accommodating multiple modalities, in a dynamic fashion.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*3D/stereo scene analysis*; I.5.5 [**Pattern Recognition**]: Implementation—*Interactive Systems*

General Terms

Algorithms, Human Factors, Design

Keywords

Action recognition, Gesture Recognition, Kinect data analysis

1. INTRODUCTION AND RELATED WORKS

The Microsoft Kinect depth sensor has attracted a lot of attention, thanks to its ability to capture and release, in real time, 2.5D data with registered RGB information. Moreover, human motion can be easily extracted in the form of moving skeletons [2]. Thanks to the above reasons, various methodologies have been proposed in recent bibliography, in the area of human action recognition [7]. In the proposed work, a novel, skeleton-based human action recognition method, is introduced. The framework approaches the problem by taking into account constraints imposed by spontaneous environments, as well as high amounts of noise and data, usually resulting into high complexity problems. In particular, a low-dimensional representation of large dimensionality feature vectors is utilized, by following a landmark-based spectral analysis scheme. In this way, low-dimensional subspaces, encoding valuable information, are built and new, unknown actions are projected on them. Moreover, the employed features are of global character, modeling qualitative, expressive characteristics and, thus, the ability of utilizing the proposed system for loosening demands in accurate temporal segmentations is also handled.

In recent literature, Dynamic Time Warping (DTW) [12] is one of the most well-known schemes in human action analysis. One of the major advantages of the method is its adjustability to varying time lengths, but it usually requires a very large number of training examples, as it is basically a template matching technique. Models describing statistical dependencies have also been used extensively, mainly in order to encode time-related dependencies. One of the classical approaches, in this vein, are the Hidden Markov Models (HMMs) [6]. Authors in [14], propose a discriminative parameter learning method for hybrid dynamic network in human activity recognition. They showcase results on walking, jogging, running, hand waving and hand clapping activities. The probabilistic behavior of human motion-related features has also been widely used through SVMs, which seek hyperplanes in the feature space for separating data into classes. Authors in [9] use non-linear SVMs for the task of recognizing daily activities of small temporal length (answer the phone, sit down/up, kiss, hug, get out of car). The output of an Artificial Neural Network (ANN) can also be used for modelling the probability $P(y|x)$ of an activity y to occur, given input feature vector x. Typical is the work in [4], where the authors perform indoors action recognition, using wearable and depth sensors. Using ANNs, special attention should be paid to high complexity during training and overfitting. Classical classification schemes, such as k-Nearest Neighbors (k-NNs) and binary trees have also been widely used in the bibliography. The authors in [8] employ Discrete Fourier Transform (DFT) as their representation and feed the corresponding parameters to a k-NN. The main drawback of these systems is that they are quite sensitive to parameter fine tuning and tend to generalize poorly for unknown subjects.

The rest of the paper is structured as follows: Section 2 provides the technical details of the proposed low-dimensional embedding method, section 3 outlines the feature extraction strategy, while section 4 presents experimental results. Section 5 concludes the paper and describes future directions.

2. LANDMARK-BASED ACTION RECOGNITION

Based on the idea that similar, person-independent activities lay close to each other on a manifold space, a feature vector $\mathbf{x}_i \in \mathbb{R}^m$ representing a certain action can be approximated by the linear combination of representation vectors $\mathbf{z}_i \in \mathbb{R}^k$ ($k << m$) with a set of basis vectors $\mathbf{l}_j \in \mathbb{R}^m$. A natural assumption is that basis vectors $\mathbf{l}_j$ correspond to action-specific descriptors, in an action recognition problem. Thus, the problem becomes an optimization problem of minimizing $||X - LZ||$, with $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ being a set of n instances, $L = [\mathbf{l}_1, ..., \mathbf{l}_k] \in \mathbb{R}^{m \times k}$ the table of feature vectors of landmark-activities and $Z = [\mathbf{z}_1, ..., \mathbf{z}_n] \in \mathbb{R}^{k \times n}$ the low-dimensional representation of X.

A common approach of finding low-dimensional representations of data points $\mathbf{x}_i \in \mathbb{R}^m$ in a manifold space, is to apply classical spectral clustering [13]. According to this method, all n data vectors are compared to each other, using a distance metric, leading to the construction of the adjacency matrix $W = (\mathrm{w}_{i,j})_{i,j=1}^n$. From W, the degree matrix D is built, which is a diagonal matrix whose elements are the column (or row) sums of W. Subtracting W from D gives the graph Laplacian matrix L, and the eigenvectors corresponding to its k smallest eigenvalues are the low (k)-dimensional representation of the initial dataset. However, large datasets lead to time consuming construction and eigen-decomposition of the Laplacian. Moreover, real-time action classification, using a spectral clustering scheme, requires a per-frame unfolding of local submanifolds, as well as the use of a pre-defined number of closest feature points in it. We hereby make use of the idea introduced in [3] for solving the optimization problem of finding low-dimensional representations, taking advantage of basis vectors $\mathbf{l}_j$. In [3], the authors introduce the idea of Large Scale Spectral Clustering with Landmark-based representation (LSC). Instead of finding point-to-point distances for constructing the adjacency matrix, they make use of a small number of feature (basis) vectors and the adjacency matrix is constructed by them. According to this method, the n data points $\mathbf{x}_i \in \mathbb{R}^m$ can be represented by linear combinations of k ($k \ll n$) representative landmarks (basis vectors). This representation can be used in the spectral embedding. The new representations are k-dimensional vectors $\mathbf{b}_i \in \mathbb{R}^k$ while the landmarks are the result of random selection or a k-means algorithm.

In the proposed work, it is straightforward to extract landmark basis vectors, as feature vectors representing whole actions. Each of these k classes of a training dataset can constitute a basis for building the landmark matrix $L \in \mathbb{R}^{m \times k}$. Here, we consider each action-specific landmark as the average of the corresponding m-dimensional feature vectors. The original data matrix $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ can be approximated by the product of L and the representation matrix $Z \in \mathbb{R}^{k \times n}$ as $X \approx LZ$. Each element z_{ji} of the representation matrix Z can be found as the output of a kernel function $k_h(\cdot)$ (here, we use the Laplacian Kernel) of feature vector $\mathbf{x}_i$ and landmark $\mathbf{l}_j$ normalized with the sum of the corresponding values for all landmark vectors:

$$z_{ji} = \frac{e^{\frac{-\|\mathbf{x}_i - \mathbf{l}_j\|}{\sigma}}}{\sum_j e^{\frac{-\|\mathbf{x}_i - \mathbf{l}_j\|}{\sigma}}} \tag{1}$$

with $\| \cdot \|$ being a vector distance metric, while σ is the width of the kernel. Z represents the similarity values between data vectors and actions' representative landmarks and defines an undirected graph $G = (V, E)$ with graph matrix $W = \hat{Z}^T \hat{Z}$, where:

$$\hat{Z} = D^{-1/2} Z \tag{2}$$

with D being a diagonal matrix whose elements are the row sums of Z. Since each column of the representation matrix sums up to 1, it is straightforward to check that the degree matrix of W is the identity matrix. Consequently [10], the eigenvectors of W are the same as those of the corresponding Laplacian matrix.

Then, the eigenvectors $A = [\mathbf{a}_1 ... \mathbf{a}_k] \in \mathbb{R}^{k \times k}$ and eigenvalues σ_j^2 of $\hat{Z}\hat{Z}^T$ are calculated. It is obvious that σ_j are the singular values of $\hat{Z}$ and A consists of the left singular vectors of $\hat{Z}$, found through singular value decomposition (3), while $B = [\mathbf{b}_1 ... \mathbf{b}_k] \in \mathbb{R}^{n \times k}$ are the eigenvectors of matrix $W = \hat{Z}^T \hat{Z}$. Each row of B is a low-dimensional representation of the original, high-dimensional feature vectors.

$$\hat{Z} = A \Sigma B^T \tag{3}$$

Consequently, and since $A^T = A^{-1}$, B can be computed directly from (3), as:

$$B = (\Sigma^{-1} A^T \hat{Z})^T \tag{4}$$

Σ is a diagonal with elements σ_j, in decreasing order, and $A = [\mathbf{a}_1 ... \mathbf{a}_k] \in \mathbb{R}^{k \times k}$ are the eigenvectors of $\hat{Z}\hat{Z}^T$.

2.1 Classification of new instances

For classifying a new data instance $\mathbf{x}'$ to an activity, the elements z'_j of the representation vector $\mathbf{z}' \in \mathbb{R}^k$ defined by the similarities between $\mathbf{x}'$ and $L = [\mathbf{l}_1 ... \mathbf{l}_k]$ is found as:

$$z'_j = \frac{e^{\frac{-\|\mathbf{x}' - \mathbf{l}_j\|}{\sigma}}}{\sum_j e^{\frac{-\|\mathbf{x}' - \mathbf{l}_j\|}{\sigma}}} \tag{5}$$

The representation $\mathbf{b}'$ of the new feature vector in the low dimensional domain is given by:

$$\mathbf{b}' = \Sigma^{-1} A^T D^{-1/2} \mathbf{z}' \tag{6}$$

Classification result is given as the label C of the action with low-dimensional representation matrix B_a (as calculated in training) that minimizes a distance metric $d(\cdot)$ from $\mathbf{b}'$:

$$C = argmin_a d(\mathbf{b}', B_a) \tag{7}$$

Thus, for new data vectors, no local sub-manifold unfolding is necessary and, for inference, simple matrix operations are needed. This is of great significance, since it allows for real-time action recognition. Consequently, the proposed method allows for online evaluation of whether the projection of extracted expressivity features over the course of an action is close to the subspace classes of a trained model.

3. FEATURE EXTRACTION

Tracked skeletal joints used in this work refer to the head, neck, shoulders, elbows, hands, torso, hips, knees, feet. Their x, y, z positions are dependent on sensor position and using them directly would yield unreliable results. One intuitive feature representation would be to consider a coordinate system with origin on a body joint and re-calculate all joints' positions with reference to this. However, this would impose a demand to a classifier that joints belonging to different individuals follow the same path for reproducing the same gestures. Thus, it is desired to consider features that describe salient qualities of the actions, expected to be uniform among different individuals or different action reproductions. Moreover, the skeletal representation considered should be invariant to sensor's position or subject's orientation, so that different actions can be recognized independently of extrinsic parameters.

The authors in [11] propose a set of features, structured in a hierarchical manner, by considering three separate sets of body positions: Torso, first and second order joints. Based on this scheme, they describe dance movements by a series of 19-dimensional vectors containing Tait-Bryan angular data. This representation is appropriate for our proposed methodology since it fulfills our criteria while, at the same time, it guarantees signal continuity and stability (e.g. feature representation does not suffer from the gimbal lock effect). In particular, joints belonging to the torso (neck, shoulders, heaps and torso) can be used for the calculation of the overall body orientation. For this, Principal Components Analysis (PCA) is applied on the matrix composed of the corresponding joints' x, y, z positions. The first principal component $\mathbf{u}$ has the same directionality with the longest dimension of the torso, while, directionality $\mathbf{r}$ is directly calculated by the shoulders' position, and $\mathbf{t}$ is found as the cross product of $\mathbf{u}$ and $\mathbf{t}$. The above basis fully describes torso orientation. Here, the average first derivative of the corresponding angles is also considered, throughout the course of an action; in this way, the feature representation is view-independent, while qualitative measurements, related to body directionality and speed of an action are captured.

The calculation of hierarchical features for the first-order (elbow, knees) joints is made as follows: A spherical coordinate system is defined at each parental joint (torso) with $\mathbf{u}$ and $\mathbf{r}$ being the zenith and the azimuth axis, respectively. The position of the child joint is described by radius R, inclination θ (the angle between $\mathbf{u}$ and the vector connecting the two joints) and its azimuth ϕ, which is the angle between $\mathbf{r}$ and the projection of the child joint on the plane whose normal is the $\mathbf{u}$ basis vector. In the second-order case (hands, feet), the zenith axis becomes the vector $\mathbf{b}$ connecting the parental joint (elbow, knee) with its adjacent torso joint. Azimuth ϕ is calculated as the angle between the projection of $\mathbf{r}$ onto the plane S whose normal is $\mathbf{b}$, $\mathbf{r}_p$ and the vector defined by the parental joint and the projection of the second-order joint onto plane S. Inclination θ is the angle between $\mathbf{b}$ and the vector defined by the second and the first-order joints. As in [11], we ignored R, since it is fixed for all joints. In our experiments, the relative differences between successive values of the hierarchical features were considered as feature representation, similar to the case of the torso orientation, and time segments corresponding to actions are described by the corresponding average values.

Figure 1: Example from the Huawei/3DLife Dataset 1 [1].

The average speed $\overline{\mathbf{v}^j}=(\overline{v^j}_x, \overline{v^j}_y, \overline{v^j}_z)$ for all joints j, its standard deviation, as well as the differences of speeds between the first and the second half of the duration of each action were also employed as features modelling expressivity parameters [5]. The above feature representation leads to 154-dimensional feature vectors for each action.

4. EXPERIMENTAL EVALUATION

In order to have its accuracy validated, the proposed methodology has been tested on the publicly available Huawei/3DLife Dataset, Session 1 [1], in which 17 subjects participated, each performing a set of 16 repetitive actions. These actions are either sports-related activities, or involve some standard movements (e.g. knocking on the door), as shown in Fig. 1. Each action was performed at least 5 times by each subject (apart from one person who performed 15 out of the 16 repetitive actions). Subjects' motion was captured using a series of depth sensors (Microsoft Kinect), while wearable inertia sensors information is also available. In order not to bias the parameters of the subspace vectors, the same amount of actions (i.e., 5) was used for all users, during the extraction of the training data.

Prior to training, all data were normalized between -1 and 1 and a leave-one-subject-out cross validation protocol was followed. More specifically, for each user, a subset S_{trn} of 11 different users was employed for training subspace parameters for different kernel widths (1). The resulted parameters were applied on validation data S_{val} of 5 users and the kernel for which the highest accuracy for S_{val} was achieved, was used for the user. Throughout all experiments, the Mahalanobis distance was employed for inferring the correct classification labels in (7). Recognition accuracy, on all 16 repetitive actions of the dataset reaches a total of 83.6%.

In real life problems, different individuals perform the same action at different durations or, many times, action detection algorithms fail to accurately detect the exact time boundaries of an action. To this aim, uncertainty was introduced into the time an action is expected to be completed, by shortening and extending it by a certain amount of time. Representation features were considered for each time segment T and classification followed, as previously.

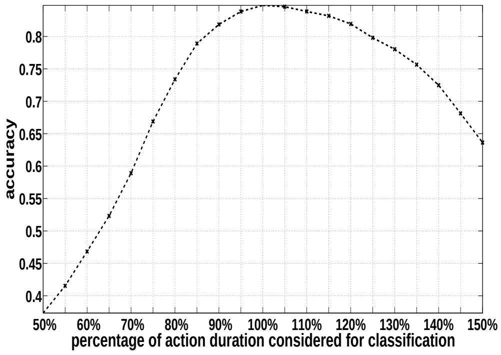

Figure 2: Average accuracy achieved on the Huawei/3DLife Dataset 1, for different considerations of action duration.

In particular, for every representation $\mathbf{b}'_t$ corresponding to time t, its distances from every cluster of the trained model were considered. The cluster and time that correspond to the smallest distance (8) are considered as the final estimate and duration of the action.

$$\{C, t\} = argmin_{a,t \in T} d(\mathbf{b}'_t, B_a) \tag{8}$$

Figure 2 shows the system's accuracy at estimating correct actions when inserting uncertainty with respect to the time an action is expected to be completed (as a percentage of the real duration).

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we used action-dependent basis vectors for projecting large-dimensionality feature vectors to low - dimensional spaces. An affinity matrix between feature vectors and basis vector was constructed, instead of the full adjacency matrix. Initial results showed that the method is promising, even at discriminating between similar actions in the dataset (e.g. knocking on the door versus hand waving or throwing an object). Using features describing global expressivity showed the robustness of the system to varying temporal boundaries of an action's duration, achieving high accuracy results when there exists fuzziness in the estimate of the time an action is completed. Future work will cater for different action styles among different individuals (or within the same person), as sub-classes of an activity.Moreover, multiple modalities, dynamically and adaptively weighted, over the duration of an action, will be considered in the representation matrix.

6. ACKNOWLEDGMENTS

The research leading to this work has been supported by the University of Maastricht and the EU Horizon 2020 Framework Programme under grant agreement no. 644204 (ProsocialLearn project).

7. REFERENCES

[1] Huawei/3dlife acm multimedia grand challenge for 2013, http://mmv.eecs.qmul.ac.uk/mmgc2013/.

[2] S. Asteriadis, A. Chatzitofis, D. Zarpalas, D. S. Alexiadis, and P. Daras. Estimating human motion from multiple kinect sensors. In *MIRAGE*, 2013.

[3] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *AAAI*, 2011.

[4] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. F. S. Mejia. Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors. In *Advances in Computational Intelligence*, pages 216–223. Springer, 2013.

[5] B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Gesture Workshop*, pages 188–199, 2005.

[6] E. Kim, S. Helal, and D. Cook. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, 9(1):48–53, Jan. 2010.

[7] H. S. Koppula, R. Gupta, and A. Saxena. Human activity learning using object affordances from rgb-d videos. *CoRR*, abs/1208.0967, 2012.

[8] S. Kumari and S. K. Mitra. Human action recognition using dft. *Computer Vision, Pattern Recognition, Image Processing and Graphics, National Conference on*, 0:239–242, 2011.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, jun 2008.

[10] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[11] M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 147–156, NY, USA, 2011.

[12] A. Veeraraghavan, S. Member, and A. K. Roy-chowdhury. Matching shape sequences in video with applications in human movement analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1896–1909, 2005.

[13] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17, 2007.

[14] X. Wang and Q. Ji. Learning dynamic bayesian network discriminatively for human activity recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 3553–3556, 2012.