A CORRESPONDENCE BASED METHOD FOR ACTIVITY RECOGNITION IN HUMAN SKELETON MOTION SEQUENCES

Eftychia Fotiadou, Nikos Nikolaidis

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

ABSTRACT

In this paper we present an algorithm for efficient activity recognition operating upon human skeleton motion sequences, derived through motion capture systems or by analyzing the output of RGB-D sensors. Our approach is driven from the assumption that, if two such sequences describe similar activities, then, consecutive frames (poses) of one sequence are expected to be similar to consecutive frames of the other. The proposed method adopts a quaternion based distance metric to calculate the similarity between poses and an intuitive method for estimating a similarity score between two skeleton motion sequences, based on the structure of a pose correspondence matrix. Our method achieved 99.5% correct activity recognition, when applied on motion capture data, in a classification task consisting of 18 classes of activities.

Index Terms— activity recognition, classification

1. INTRODUCTION

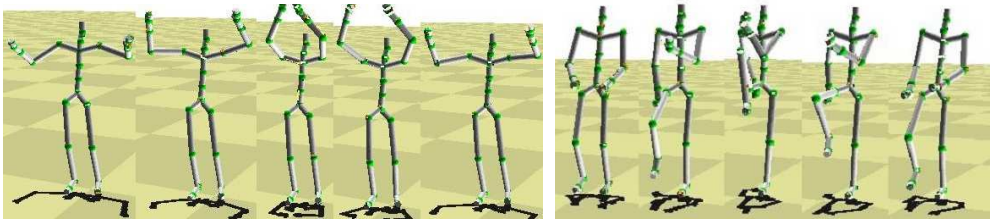
Activity recognition, i.e. the identification of the activity performed by a human subject, can be a crucial part in many applications, such as video surveillance, semantic annotation and labeling of multimedia data for summarization, indexing and retrieval in databases, or human-machine interaction. As a result, activity recognition constitutes an important research field of computer vision and various approaches have been proposed [1]. Activities can be simple, everyday actions such as walking, jumping or waving, or more complex ones, such as playing basketball or dancing. Although the majority of the activity recognition algorithms operate upon video data and rely on features calculated from these data to describe human motion, several methods adopt a 3D representation of the human skeleton. Such representations, namely sequences of 3D skeletal poses over time, can be derived from motion capture devices (infrared, ultrasonic, magnetic etc.). Furthermore, skeleton motion sequences can be obtained from the analysis of video data, or through the processing of RGB-D (RGB + depth) data generated from the Microsoft Kinect or similar sensors. Some frames from a skeleton motion sequence describing the activities "clap above head" and "elbow to knee" are illustrated in Fig. 1.

An activity recognition method utilizing skeleton motion data is described in [2], as part of an interactive dance game framework. The system receives an input stream of motion data from the player and performs movement recognition, based on a standard set of template moves. A block matching approach is adopted, where segments of the incoming motion stream are continuously compared to motion templates. In [3], human motion recognition is performed using Support Vector Machines (SVMs). In addition, the importance of certain skeleton points to the recognition task is explored. A different approach can be found in [4], where a distance function for motion capture sequences suitable for activity classification, clustering and anomaly detection is introduced. This method, which is based on the kinetic energy of each joint, allows for fast computations, as its complexity depends only on the number of the joints used as features, rather than the length of each sequence.

An algorithm for sequence alignment and activity recognition, called IsoCCA, is described in [5]. IsoCCA extends the Canonical Correlation Analysis (CCA) algorithm, by means of introducing a number of alternative monotonicity constraints. The activity classification task performed in this paper is based on a 1-Nearest Neighbor (1-NN) classifier, that uses the alignment cost between sequences as distance metric, and yields improved classification rates in comparison to other alignment algorithms, such as Canonical Time Warping (CTW), Dynamic Time Warping (DTW), Hungarian and CCA.

In [6], a method for activity segmentation and classification of motion data is proposed, based on the derivation of a set of simple movements from the data, called primitives. Another method for segmentation and recognition of motion capture sequences is described in [7], where Singular Value

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.



(a) (b)

Fig. 1. Sample frames from skeleton motion sequences of the classes "clap above head" (b) and "elbow to knee" (c), from the HDM05 database.

Decomposition (SVD) and a multi-class SVMs are combined.

In [8], a new representation of skeleton motion data, suitable for activity recognition, called Sequence of the Most Informative Joints (SMIJ), is proposed. This representation relies on the selection of the skeletal joints carrying the most information relevant to the motion, according to the values of measures such as the mean or the variance of the joint rotations. Classification is then performed using an 1-NN classifier as well as SVMs. The activity recognition method proposed in [9] can be applied to motion capture data as well as to data acquired by RGB cameras with depth sensors, such as the Microsoft Kinect sensor. As far as features are concerned, several alternatives are explored, based on the coordinates of the skeleton joints and on spatial and temporal differences between them. The classification task is performed using the Extreme Learning Machine (ELM) algorithm. Specifically, each frame of a test sequence is classified separately by the ELM and subsequently, the whole sequence is assigned the label of the class that yielded the most votes.

In [10], activity recognition is based on modeling the spatio-temporal relationships between joints, which are represented by Sparse Granger Causality Graph Models (SGCGM). Each motion capture sequence is transformed to a causality graph and classified by a sparse regression classifier. In [11], a representation for motion capture data, useful for activity recognition is proposed. Specifically, each frame of a motion sequence is represented by a matrix containing the distances between the skeleton joints. Principal Component Analysis (PCA) is applied, in order for the dimensionality of the data and the noise associated with it to be reduced. Finally, classification is performed using action graphs in conjunction with a probabilistic model.

The method proposed in this paper requires no preprocessing on the skeleton motion data, such as alignment, feature extraction or temporal segmentation of an activity sequence into elementary movements such as steps. The proposed algorithm is based on the hypothesis, that similar skeleton motion sequences exhibit strong similarity between successive frames in one or more segments within them, which is expressed through specific patterns in a pose correspondence matrix. In order to classify skeleton motion sequences to distinct classes, we developed a scheme for similarity estimation between such sequences. In the following sections, we discuss the details of our method as well as its experimental evaluation, when applied on motion capture data.

2. PROPOSED METHOD

The proposed activity recognition method is based on the similarity between two skeleton motion sequences and comprises of two distinct steps: first, a correspondence matrix, that describes which frame in the second sequence is the most similar to each frame in the first sequence, is calculated. Subsequently, a similarity score between the two sequences is calculated, based on the correspondence matrix. In the following subsections, the aforementioned steps are described in detail.

2.1. Correspondence Matrix Construction

Let us consider two skeleton motion sequences denoted with $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M\}$ and $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N\}$, consisting of M and N frames respectively. Each frame in a sequence describes the pose of the subject, i.e. the configuration of the human body parts, at a certain time instance and consists of the rotation angles of each joint. In order to construct the correspondence matrix, the distances from each pose (frame) $\mathbf{Y}_i$ of the sequence $\mathbf{Y}$ to every pose $\mathbf{X}_i$ of sequence $\mathbf{X}$ are calculated.

The distance between two poses can be calculated using a quaternion-based pose distance, as described in [12]. Let a pose configuration $\mathbf{X}_i$ be expressed in the following form:

$$\mathbf{X}_i = (\mathbf{t}_r, \mathbf{R}_r, (\mathbf{R}_b)_{b \in B}), \tag{1}$$

where $\mathbf{t}_r$ denotes the position of the root of the skeletal hierarchy, $\mathbf{R}_r$ the absolute root rotation and $(\mathbf{R}_b)_{b \in B}$ the relative joint rotation of the bone b, that consists element of the set of bones B, with respect to its parent in the skeletal hierarchy. Additionally, let $\mathbf{q}_b$ denote the unit quaternion describing the relative joint rotation $\mathbf{R}_b$ of the bone b.

Taking the aforementioned notation into consideration, the distance between two pose configurations (frames) $\mathbf{X}_i$

(a) (b)

Fig. 2. Examples of correspondence matrices for two movements of the same class "walk"(a) and different classes "walk" and "punch left side" (b).

and $\mathbf{Y}_j$ can be expressed as follows:

$$d^{Quat}(\mathbf{X}_i, \mathbf{Y}_j) = \sum_{b \in B} w_b \cdot \frac{2}{\pi} \cdot \arccos |\langle \mathbf{q}_b | \mathbf{q'}_b \rangle|, \quad (2)$$

where $\langle \cdot | \cdot \rangle$ denotes the inner product in $\mathbb{R}^4$ and $(w_b)_{b \in B}$ are weights corresponding to the rotation of each joint, with $\sum_{b \in B} w_b = 1$. The assignment of weights to the joint rotations reflects the fact, that certain joint rotations, specifically those attached closer to the torso, have a greater effect on the pose than others. This can be easily perceived by an example: if we consider the movement of an arm that is being raised, the movement of the upper arm is more important to the overall pose than the movement of the hand. This assumption is, of course, application-dependent and may not hold for particular activity vocabularies. However, for everyday activities, such as the ones being of interest to this work, the assumption is in general valid. The above distance function takes values in the interval $[0, 1]$. Note, that, the information for the root translation ($\mathbf{t}_r$) and absolute root rotation ($\mathbf{R}_r$) is not used in the distance function.

The distances between all pairs of poses in the two sequences $\mathbf{X}$,$\mathbf{Y}$ are calculated by utilizing (2) and subsequently used to construct a correspondence matrix of dimensionality MxN, denoted by $\mathbf{C}$. The rows / columns of $\mathbf{C}$ correspond to poses of sequence $\mathbf{X}$ / $\mathbf{Y}$ respectively. For each pose $\mathbf{X}_i$ of $\mathbf{X}$, the nearest pose $\mathbf{Y}_j$ of $\mathbf{Y}$ is found and the element (i, j) of $\mathbf{C}$ is set to one, whereas all other elements $(i, k), k \neq j$ of the i-th row are set to zero.

The result of this process is, that $\mathbf{C}$ exhibits distinct patterns depending on the similarity between the two sequences under examination. When the two sequences compared describe movements of the same class (e.g. two walking sequences), the correspondence matrix contains diagonal segments of ones, of various lengths, either continuous or interrupted, since successive poses from one sequence are in general most similar to successive poses from the other. Specifically, these diagonal segments extend from the upper left to the bottom right of the matrix. When two sequences describe movements from different classes, the similarity matrix exhibits different structures, with two general characteristics: First, there may exist long vertical lines, implying that many poses in sequence $\mathbf{X}$ are matched to the same pose in $\mathbf{Y}$. This is often the case, when the two movements described in the sequences are different, but share one or more similar poses. Second, the correspondence matrix may exhibit diagonal segments of limited length or the ones (units) may be arranged with no particular structure, a fact indicating, that the two sequences describe completely different movements. Examples of correspondence matrices are shown in Fig. 2, where the unit entries are represented by white pixels.

2.2. Similarity Score Evaluation

The similarity between two skeleton motion sequences $\mathbf{X}$,$\mathbf{Y}$, is determined by means of a score S, calculated over their correspondence matrix $\mathbf{C}$. The calculation of S is based on the existence and the structure of diagonal segments in the correspondence matrix, and the higher its value, the more similar the two sequences are.

For each row of matrix $\mathbf{C}$ (which corresponds to a pose of sequence $\mathbf{X}$), the position of the unit entry (column index) is retrieved, in order to determine the relative position of the unit entries in subsequent poses and to identify possible diagonal segments. The relative position of a unit in the next row with respect to the unit in the current row defines whether the next unit lies in a "legal" position or not, according to a number of rules. These rules try to take into account the fact that, although units (matching poses) should ideally form a diagonal (45° slope) segment consisting of connected elements (i.e. units should be arranged in matrix cells (i, j), $(i + 1, j + 1)$, $(i + 2, j + 2)$ and so on), deviations from this ideal situation (i.e. gaps of limited extend) should be allowed. More specifically, a maximum of three consecutive points within a diagonal segment are allowed to lie in a vertical placement, such as (i, j),$(i+1, j)$,$(i+2, j)$. This means, that up to three consecutive poses of sequence $\mathbf{X}$ are allowed to be matched

Table 1. Legal (marked with a tick) and illegal (marked with an x) positions for a unit entry in the correspondence matrix.

1	0	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
x	$\surd$	$\surd$	$\surd$	$\surd$	x

to the same pose of sequence $\mathbf{Y}$. Furthermore, if the current unit is at cell (i, j), then the unit in the next row can be either in the "ideal" position $(i+1, j+1)$, or in positions $(i+1, j)$ (vertical placement), $(i+1, j+2)$, $(i+1, j+3)$. In other words, gaps of length 1 and 2 are allowed. Finally, the length of each diagonal segment, that is the number of units lying on it, should be larger than a threshold T_l, in order for it to contribute to the final score. The optimal value for T_l is determined by testing various values and selecting the one that achieves the highest recognition rate. It was observed that, a minimum length of 5 points produced the best results. An example of legal and illegal positions for a unit entry in the 4-th row of a correspondence matrix, given the arrangement of units in the previous 3 rows, is shown in Table 1.

After the aforementioned process has been performed for every pose of sequence $\mathbf{X}_s$, all the units lying in "legal" positions, have been assigned to a diagonal segment. Each unit entry to the correspondence matrix lying in a diagonal segment, is assigned a weight $\alpha_{i,j}$. A weight equal 1 is assigned if consecutive poses of sequence $\mathbf{X}$ are matched to exactly consecutive poses in sequence $\mathbf{Y}$, (i.e. units are in an arrangement (i, j), $(i+1, j+1)$) and a weight equal to 0.8 otherwise, so as to penalise "imperfect" diagonal segments.

At the end of the procedure, the segments containing a number of points below threshold T_l are discarded as invalid, while the valid segments contribute to the calculation of the total score. The total score S is estimated by summing the weights of the units lying in valid diagonal segments:

$$S = \sum_{(i,j) \in V} \alpha_{i,j}, \tag{3}$$

where V is the set of all units lying in valid diagonal segments.

The classification of a test sequence is performed using an 1-Nearest Neighbor classifier: the test sequence is tested against all training sequences and is labeled with the activity label of the training sequence that yielded the highest similarity score.

3. EXPERIMENTAL RESULTS

In our experiments we used data from the HDM05 motion capture database [13], which consists of files in ASF/AMC format, for various types of activities, performed by five actors. Our activity recognition task included 18 classes of activities: cartwheel, clap, clap above head, elbow to knee, hop both legs, hop left, hop right, kick right front, kick right side, kick left front, kick left side, punch right front, punch right side, punch left front, punch left side, run on place, sneak and walk. Data from all five actors were used, 1013 motion capture files in total. From the motion capture clips, we selected the angles information for a subset of 13 joints, namely lower back, upper back, thorax, right humerus, right radius, left humerus, left radius, right femur, right foot, left femur, left tibia, left foot and right tibia, since these joints were observed to be the most informative, and therefore more discriminant for our recognition task. The weights w_b used in the quaternion-based distance function were calculated empirically and assigned to each joint according to its position in the skeletal hierarchy. Upper back, thorax, humerus and femur were each assigned a weight of 0.1, radius, tibia and lower back a weight of 0.065, while feet were assigned a weight of 0.0375. A 1-NN classifier was adopted and the classification experiment was performed in a leave-one-out setting, commonly used in the experimental evaluation of activity recognition methods. In more detail, each skeleton motion sequence was tested against all other sequences in the database and was labeled according to the label of the sequence that yielded the biggest similarity score. This procedure was repeated for all sequences of the dataset and the overall correct recognition rate was calculated.

The aforementioned classification experiment resulted in a correct recognition rate of 99.5%, which outperforms the IsoCCA method described in [5]. From the total of 18 classes, 13 yielded 100% recognition rate, while 5 classes yielded rates between 96.43% and 97.77%. For our dataset, the IsoCCA algorithm achieved a recognition rate of 94.97%.

4. CONCLUSIONS AND FUTURE WORK

A method for activity recognition on skeleton motion data has been presented in this paper. Our algorithm is based on the structure of a correspondence matrix between motion sequences and achieves a 99.5% recognition rate for a dataset including 18 different classes from the HDM05 database. Regarding future work, additional features, such as the velocity and acceleration of joints could be explored. Other research directions that will be investigated include testing the algorithm to datasets other than the HDM05, incorporation of our method in a continuous activity recognition framework, and extending the method towards recognizing dance actions depicted in skeleton motion sequences.

5. REFERENCES

[1] D. Weinland, R. Ronfard, and E. Boyer, "A Survey of Vision-based Methods for Action Representation, Segmentation and Recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, Feb 2011.

[2] J. K. T. Tang, J. C. P. Chan, and H. Leung, "Interactive Dancing Game with Real-Time Recognition of Continuous Dance Moves from 3D Human Motion Capture," in *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*. 2011, pp. 50:1–50:9, ACM.

[3] J.-Y. Wang and H.-M. Lee, "Recognition of Human Actions Using Motion Capture Data and Support Vector Machine," in *WRI World Congress on Software Engineering, 2009. WCSE '09.*, May, vol. 1, pp. 234–238.

[4] K. Onuma, C.Faloutsos, and J. K. Hodgins, "FMDistance: A fast and effective distance function for motion capture data," in *Short Papers Proceedings of EUROGRAPHICS*, 2008.

[5] S. Shariat and V. Pavlovic, "Isotonic CCA for sequence alignment and activity recognition," in *2011 IEEE International Conference on Computer Vision (ICCV)*, Nov., pp. 2572–2578.

[6] A. Fod, M. J. Matarić, and O. C. Jenkins, "Automated Derivation of Primitives for Movement Classification," *Auton. Robots*, vol. 12, no. 1, pp. 39–54, Jan. 2002.

[7] C. Li, P. R. Kulkarni, and B. Prabhakaran, "Segmentation and Recognition of Motion Capture Data Stream by Classification," *Multimedia Tools and Applications*, vol. 35, no. 1, pp. 55–77, 2007.

[8] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 8–13.

[9] X. Chen and M. Koskela, "Classification of RGB-D and Motion Capture Sequences Using Extreme Learning Machine," in *Image Analysis*, vol. 7944, pp. 640–651. Springer Berlin Heidelberg, 2013.

[10] S. Yi and V. Pavlovic, "Sparse granger causality graphs for human action classification.," in *ICPR*. pp. 3374–3377, IEEE.

[11] A.W. Vieira, T. Lewiner, W.R. Schwartz, and M. Campos, "Distance matrices as invariant features for classifying mocap data," in *2012 21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 2934–2937.

[12] M. Müller, *Information Retrieval for Music and Motion*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[13] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation Mocap Database HDM05," Tech. Rep. CG-2007-2, Universität Bonn, June 2007.