LEVERAGING SKELETON STRUCTURE AND TIME DEPENDENCIES IN THE SCOPE OF ACTION RECOGNITION

Ioannis Tsingalis, Nicholas Vretos, Petros Daras

Centre for Research and Technology Hellas, Information Technologies Institute
6th Km Charilaou-Thermi Road, Thessaloniki, Greece
e-mail:{tsingalis,vretos,daras}@iti.gr

ABSTRACT

In this work, the structure of the moving skeleton, which is a time varying graph, along with the temporal dependencies of human action were leveraged in the scope of skeleton action recognition. The optimisation of the proposed model shares similarities with the optimisation problem of Slow Feature Analysis (SFA) enabling a well defined solution. Moreover, due to the incorporated skeleton structure, the learned slow functions enclose information regarding the geometry of the skeleton movement which is very useful in the action recognition problem. Two skeleton action datasets were used to evaluate our method, the MSR Action 3D and a dataset whose actions were inspired by psychological studies. Both datasets were captured by depth cameras. The proposed method yielded promising results when evaluated on the aforementioned datasets.

Index Terms— Skeleton Tracking, Action Recognition, Slow Feature Analysis, Human Activity.

1. INTRODUCTION

During the last decade, Human Activity Recognition became a very active field in computer vision including applications for motion analysis in sports, robotics, health-care, etc. The first action recognition methods were based on the information of RGB video frames. Even though this approach has attracted a lot of attention, recognition results were not satisfactory due to the irrelevant and noisy background, and the lack of three dimensional information which is useful for the distinction of the same actions performed from different view points.

In order to overcome the difficulties that arise from the noisy background, local descriptors were proposed. Laptev et al. [1] proposed Histogram of Gradients (HoG) and Histogram of optic Flows (HoF) in order to obtain a more robust description of the captured action in 2D domain, i.e., per single video frame. Klaser et al. [2] extended 2D HoG [1] features to 3D domain, i.e., taking into consideration multiple video frames. A drawback in [2] was the sparseness of interest points which may lead to information loss. Thus, Dollar et al. [3] proposed a method where a rich corpus of interest points was computed by applying a series of spatiotemporal filters. Laptev & Lindeberd [4] also proposed spatiotemporal interest points for action representation that demonstrate rotational and translational invariance. Oikonomopoulos et al. [5] proposed spatiotemporal features by computing the variations between neighborhood regions extending the salient point detector in [6]. Schuldt et al. [7] introduced space-time video representations and integrated them with an Support Vector Machine (SVM) for action recognition.

With the advent of depth sensors and the almost real time pose estimation techniques [8], human skeleton action datasets with three dimensional information became publicly available making skeleton action recognition a new trend. However, problems like wrongly estimated postures due to noisy depth maps and different executions of same actions are still open problems. Gowayyed et. al. [9] proposed a novel 2D trajectory descriptor combined with temporal pyramids and classified the input sequences by applying SVMs. Vemulapalli et. al. [10] proposed a skeleton representation that lies into a Lie group which is a curved manifold. Similarly to [9], they applied Fourier temporal pyramids and SVMs to classify the given action. Du et. al. [11] based on a hierarchical decomposition of the human body parts, proposed an application of hierarchical Recurrent Neural Networks (RNNs) in the scope of action recognition. Yanhu et. al. [12] applied SFA [13] to extract skeleton feature representations. Similarly to [9, 10] the applied Temporal Pyramids and SVMs for action recognition.

In our approach, we leverage the information of the human skeleton node speed along with the skeleton structure in order to introduce a graph embedded subspace learning feature extraction. The extracted features are inspired by the Principle of Slowness [13] similarly to [12], but also combine information from the geometry of the moving skeleton over time.

The remainder of this paper is organized as follows. The Section 2 introduces some basic definitions. In addition, in subsection 2.1, a description of standard SFA is presented because of the similar optimization procedure followed by our method. The details of the proposed framework are discussed in subsection 2.2. Parameter selection and experimental results are given in Section 3. Finally, conclusions are drawn in Section 4.

2. PROPOSED METHOD

When working with skeleton data, each action is a sequence of moving skeleton frames. Each skeleton frame consists of a number of N joints-nodes connected with edges. Based on this point of view, we provide the following definitions.

Basic definitions on spatial domain: We define a skeleton action as a time varying graph, $G^{(t)} = (V^{(t)}, E)$, $t \in [0, T]$, where $V^{(t)}$ represents the skeleton nodes at a specific time point t and E the corresponding skeleton edges. Moreover, a skeleton frame is defined by a matrix $\mathbf{\Phi}^{(t)} = \left[\boldsymbol{\phi}_1^{(t)}, \ldots, \boldsymbol{\phi}_N^{(t)}\right] \in \mathbb{R}^{I \times N}$ where each column $\boldsymbol{\phi}_n^{(t)} \in V^{(t)}$ represents a skeleton node. More specifically, each skeleton node is described by $\boldsymbol{\phi}_n^{(t)} \equiv \boldsymbol{\phi}(\mathbf{x}_n^{(t)})$ where $\mathbf{x}_n^{(t)} = [x_n^{(t)}, y_n^{(t)}, z_n^{(t)}]^T$ is a three dimensional vector ($I = 3$) that represents the node position in three dimensional space and $\boldsymbol{\phi}(\cdot)$ a vector valued function that maps the node representation into a higher dimensional feature space. When $\boldsymbol{\phi}(\cdot)$ is the identity func-

tion, it is $\boldsymbol{\phi}_n^{(t)} = \boldsymbol{\phi}(\mathbf{x}_n^{(t)}) = \mathbf{x}_n^{(t)}$.

Given the definition of the time varying graph $G^{(t)}$ in a specific time point t, we also define the corresponding time varying degree, weight and Laplacian matrices as $\mathbf{D}^{(t)}$, $\mathbf{\Gamma}^{(t)}$ and $\mathbf{L}^{(t)} = \mathbf{D}^{(t)} - \mathbf{\Gamma}^{(t)}$, respectively, with $D_{ii}^{(t)} = \sum_j \Gamma_{ij}^{(t)}$. Since the skeleton nodes in a specific time point t are denoted by $\{\boldsymbol{\phi}_n^{(t)}\}_{n=1}^N$, the elements $\Gamma_{ij}^{(t)}$ of the weight matrix $\mathbf{\Gamma}^{(t)}$ are given by Radial Basis Function (RBF) $\Gamma_{ij}^{(t)} = \exp(||\boldsymbol{\phi}_i^{(t)} - \boldsymbol{\phi}_j^{(t)}||^2 / 2\sigma_i^{(t)}\sigma_j^{(t)})$. This means that the final weight matrix contains spatial information about the human skeleton structure. Moreover, all skeleton action frames and the corresponding Laplacian matrices are grouped into the extra bold matrices $\mathbf{\Phi} = \left[\mathbf{\Phi}^{(1)}, \ldots, \mathbf{\Phi}^{(T)}\right] \in \mathbb{R}^{I \times NT}$, and $\mathbf{L} = \left[\mathbf{L}^{(1)}, \ldots, \mathbf{L}^{(T)}\right] \in \mathbb{R}^{N \times NT}$, respectively. Finally, the block diagonal matrix whose block elements are the Laplacian matrices $\mathbf{L}^{(t)}$ in different time points t is defined by:

$$\text{diag}(\mathbf{L}) = \begin{bmatrix} \mathbf{L}^{(1)} & & \\ & \ddots & \\ & & \mathbf{L}^{(T)} \end{bmatrix} \in \mathbb{R}^{NT \times NT} \tag{1}$$

Basic definitions on speed domain: For a skeleton node $\boldsymbol{\phi}_n^{(t)}$ in spatial domain, the speed of this node in speed domain is defined by $\dot{\boldsymbol{\phi}}_n^{(t)} = \boldsymbol{\phi}_n^{(t+1)} - \boldsymbol{\phi}_n^{(t)}$. Similar to the spacial domain, we also define the matrices $\dot{\mathbf{\Phi}}^{(t)}$, $\dot{\mathbf{L}}^{(t)}$, $\dot{\mathbf{\Gamma}}^{(t)}$, $\dot{\mathbf{D}}^{(t)}$, $\dot{\mathbf{\Phi}}$ and $\dot{\mathbf{L}}$ in speed domain by using $\dot{\boldsymbol{\phi}}_n^{(t)}$ instead of $\boldsymbol{\phi}_n^{(t)}$. **Main idea**: Given a set of skeleton nodes, our goal is to find a mapping such that the new node representations will preserve their speed relation which is described by the graph $\dot{G}^{(t)} = (\dot{V}^{(t)}, E)$. In other words, if two nodes are "close" in speed in the input space, they should also be "close" in the new feature-space.

2.1. Slow Feature Analysis (SFA)

Given an I dimensional input signal $\boldsymbol{\phi}^{(t)} = [\phi_1^{(t)}, \phi_2^{(t)}, \ldots, \phi_I^{(t)}]^T \in \mathbb{R}^I$, with $t \in [0, T]$, SFA computes a vector-valued function $\mathbf{g}(\boldsymbol{\phi}^{(t)}) = [g_1(\boldsymbol{\phi}^{(t)}), g_2(\boldsymbol{\phi}^{(t)}), \ldots, g_J(\boldsymbol{\phi}^{(t)})]^T$ in order to obtain the final output signal $\mathbf{y}^{(t)} = [y_1^{(t)}, y_2^{(t)}, \ldots, y_J^{(t)}]^T \in \mathbb{R}^J$ with $y_j^{(t)} = g_j(\mathbf{x}^{(t)})$. In the linear case, the function g_j function is defined by $g_j(\boldsymbol{\phi}^{(t)}) = \mathbf{w}_j^T \boldsymbol{\phi}^{(t)}$. Therefore, for all g_j components, we obtain $\mathbf{y}^{(t)} = \mathbf{g}(\boldsymbol{\phi}^{(t)}) = \mathbf{W}^T \mathbf{x}^{(t)}$ where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J]$ and $\mathbf{W} \in \mathbb{R}^{I \times J}$ [13]. Given the aforementioned definitions and the optimization problem described below, SFA aims to compute the transformation matrix $\mathbf{W}$ that achieves the following map $\boldsymbol{\phi}^{(t)} \xrightarrow{\mathbf{W}} \mathbf{y}^{(t)}$. The SFA optimization problem is given by:

$$\begin{aligned}
&\underset{g_j, \forall j}{\text{minimize}} && \mathbb{E}_t[\dot{g}_j^2(\boldsymbol{\phi}^{(t)})] && (2)\\
&\text{subject to} && \mathbb{E}_t[g_j(\boldsymbol{\phi}^{(t)})] = 0,\ \mathbb{E}_t[g_j^2(\boldsymbol{\phi}^{(t)})] = 1 && (3)\\
& && \mathbb{E}_t[g_j(\boldsymbol{\phi}^{(t)}) g_i(\boldsymbol{\phi}^{(t)})] = 0 && (4)\\
& && \forall j \neq i, i = 1, \ldots, J && (5)
\end{aligned}$$

where $\dot{g}_j$ and $\mathbb{E}_t[\cdot]$ denote the first order time derivative of the output function g_j and the time averaging, respectively. It can be proved that the above optimization problem is equivalent to:

$$\begin{aligned}
&\underset{\mathbf{W}_{I \times J}(\mathbb{R})}{\text{minimize}} && \text{trace}(\mathbf{W}^T \mathbf{C}_{\dot{\mathbf{\Phi}}} \mathbf{W}) && (6)\\
&\text{subject to} && \mathbf{W}^T \mathbf{C}_{\mathbf{\Phi}} \mathbf{W} = \mathbf{I} && (7)
\end{aligned}$$

The solution of the aforementioned optimization problem leads to a generalized eigenvalue problem [14]:

$$\mathbf{C}_{\dot{\mathbf{\Phi}}} \mathbf{W} = \mathbf{C}_{\mathbf{\Phi}} \mathbf{W} \mathbf{D} \tag{8}$$

where $\mathbf{C}_{\dot{\mathbf{\Phi}}}$ and $\mathbf{C}_{\mathbf{\Phi}}$ are the covariance matrices of the time differentiated and the original input signal $\boldsymbol{\phi}^{(t)}$, respectively, with dimension $I \times I$.

2.2. Speed Relation Preserving Slow Feature Analysis (srpSFA)

Similar to standard SFA, our goal is to obtain the transformation matrix $\mathbf{W} \in \mathbb{R}^{I \times J}$ in order to acquire the new skeleton node representations $\mathbf{y}_n^{(t)} = \mathbf{W}^T \boldsymbol{\phi}_n^{(t)}$. In this section, we will define the loss function of our approach and the imposed constraints that align our method with standard SFA optimisation approach. **Loss function:** In order to fulfill the preservation of speed between skeleton nodes in the new feature space, as described in Section 2, the objective function:

$$\sum_{ij} \mathbb{E}_t\left[(\dot{\mathbf{y}}_i^{(t)} - \dot{\mathbf{y}}_j^{(t)})^2 \dot{\Gamma}_{ij}^{(t)}\right] \tag{9}$$

needs to be minimized. Similar to standard SFA constraints, the extracted features of the new mapped skeleton nodes must be uncorrelated and have zero mean and unit variance. The weight factor $\dot{\Gamma}_{ij}^{(t)}$ in (9) penalises the distance between the new skeleton node representations $\dot{\mathbf{y}}_i^{(t)}$ and $\dot{\mathbf{y}}_j^{(t)}$. By definition, a high value of $\dot{\Gamma}_{ij}^{(t)}$ describes a close speed relation between $\dot{\boldsymbol{\phi}}_i^{(t)}$ and $\dot{\boldsymbol{\phi}}_j^{(t)}$ which is retained between $\dot{\mathbf{y}}_i^{(t)}$ and $\dot{\mathbf{y}}_j^{(t)}$ through the loss function (9). Provided the new skeleton node representation of the n-th node $\mathbf{y}_n^{(t)} = \mathbf{W}^T \boldsymbol{\phi}_n^{(t)}$, the matrix notation of the loss function in (9) is given by:

$$\text{tr}\left(\mathbf{W}^T \dot{\mathbf{\Phi}} \text{diag}(\dot{\mathbf{L}}) \dot{\mathbf{\Phi}}^T \mathbf{W}\right) \tag{10}$$

Zero mean features: Given that the input data have zero mean, i.e., $\mathbb{E}\left[\boldsymbol{\phi}_i^{(t)}\right] = 0$, the zero mean constraint is also automatically fulfilled in the new feature space:

$$\mathbb{E}_t\left[\mathbf{y}_i^{(t)}\right] = \mathbb{E}_t\left[\mathbf{W}^T \boldsymbol{\phi}_i^{(t)}\right] = \mathbf{W}^T \mathbb{E}_t\left[\boldsymbol{\phi}_i^{(t)}\right] = \mathbf{0}_I \tag{11}$$

where $\mathbf{0}_I$ is a column vector of zeros in $\mathbb{R}^I$. □

Unit variance and uncorrelated features: Given the mapped representation $\mathbf{y}_n^{(t)}$ of the n-th skeleton node $\boldsymbol{\phi}_n^{(t)}$, the unit variance constraint is imposed by:

$$\begin{aligned}
\mathbb{E}_t\left[y_{i,n}^{(t)} y_{j,n}^{(t)}\right] &= \mathbb{E}_t\left[\mathbf{w}_i^T \boldsymbol{\phi}_n^{(t)} \left(\boldsymbol{\phi}_n^{(t)}\right)^T \mathbf{w}_j\right] && (12)\\
&= \mathbf{w}_i^T \mathbb{E}_t\left[\boldsymbol{\phi}_n^{(t)} \left(\boldsymbol{\phi}_n^{(t)}\right)^T\right] \mathbf{w}_j && (13)\\
&= \begin{cases} 1, & \text{for } i = j \\ 0, & \text{for } i \neq j \end{cases} && (14)
\end{aligned}$$

□

The matrix notation of the aforementioned constraints imposed for all N skeleton nodes is given by:

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{n=1}^N \mathbf{y}_n^{(t)} \left(\mathbf{y}_n^{(t)}\right)^T\right] = \mathbf{I}_{J \times J} \iff && (15)\\
&\mathbf{W}^T \mathbb{E}_t\left[\mathbf{\Phi}^{(t)} \left(\mathbf{\Phi}^{(t)}\right)^T\right] \mathbf{W} = \mathbf{I}_{J \times J} \iff && (16)\\
&\mathbf{W}^T \mathbf{\Phi} \mathbf{\Phi}^T \mathbf{W} = \mathbf{I}_{J \times J} && (17)
\end{aligned}$$

Finally, provided the matrix notation of the loss function in (10) and the constraints in (11) and (17), we have the following optimisation problem:

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} \quad & \text{tr}\Big(\mathbf{W}^T\dot{\mathbf{\Phi}}\text{diag}(\dot{\mathbf{L}})\dot{\mathbf{\Phi}}^T\mathbf{W}\Big) \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{\Phi}\mathbf{\Phi}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{18}$$

3. EXPERIMENTS

3.1. Evaluation protocol

Data mapping: Given an input skeleton action sequence $\mathcal{V} \in \mathbb{R}^{I\times N\times T}$, and the learned mapping matrix $\mathbf{W} \in \mathbb{R}^{I\times J}$ provided by the optimisation problem 18, the new skeleton action representation is $\tilde{\mathcal{V}} = \mathcal{V} \times_1 \mathbf{W}^T \in \mathbb{R}^{J\times N\times T}$ by applying n-mode multiplication [15]. Each mapped frontal slice $\tilde{\mathbf{V}}_k$ [15] represents the new mapped posture representation which will be used later in key-posture dictionary learning.

Data preprocessing: The input skeleton nodes $\mathbf{x}_n^{(t)}$ were transformed into the polynomial feature space by applying the polynomial feature expansion function $\boldsymbol{\phi}(\mathbf{x}_n^{(t)}) = [x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2, x_1, x_2, x_3]$. Next, in order to eliminate the effects of different camera setups in the accuracy of the model, the new skeleton representations were transformed into a unified coordinate system [11]. Moreover, in order to eliminate the intra-class variance imposed by the same action execution from different actors, each video was separately normalised to have zero mean and unit standard deviation [12].

Parameter selection: The feature extraction process demands the specification of three parameters, namely, the structure of skeleton graph $G^{(t)}(\cdot)$, the choice of the scale parameters $\sigma_i^{(t)}$ and $\sigma_j^{(t)}$ of the weight computation $\Gamma_{ij}^{(t)}$ and the output dimension J of the mapped data. The skeleton graph $G^{(t)}(\cdot)$ is defined according to human body skeleton structure, but also additional edge connections where added between peripheral skeleton nodes and a node close to the centre of the skeleton. These additional edges were added in order to capture the relation of the body limbs with the centre of the body and are presented with red colour in the pictorial representation of Figure 2.

Similar to [16], the scale values where defined by $\sigma_i = \text{d}^2(\boldsymbol{\phi}_i^{(t)}, \boldsymbol{\phi}_K^{(t)})$ where $\boldsymbol{\phi}_K^{(t)}$ is the K-th neighbour skeleton node of the i-th node $\boldsymbol{\phi}_i^{(t)}$, with $K = 3$. The neighbor is defined by traversing the skeleton edges depicted in Figure 2. Finally, the parameter J was optimised according to the accuracy results and was set to $J = 8$.

Video representation and classification: Given the mapped postures $\tilde{\mathbf{V}}_k$, a dictionary of key-postures was constructed by applying k-means clustering algorithm. The optimal number of key-postures (number of clusters) C was chosen after evaluation. These key-postures were used to extract a histogram representation of a given video. In order to capture the time dependency during the given action, a temporal pyramid of histograms was applied [12]. A similar approach was followed in [10]. The obtained $7C$ dimensional skeleton action feature vectors were fed to an SVM. The optimal SVM parameters were chosen through a grid-search procedure, using $\mathcal{X}^2$-kernel.

3.2. Datasets

MSR-Action3D database [17]: A kinect-like depth sensor was used to obtain the recorded skeletons. It consists of 20 different recorded actions performed by 10 different subjects/actors. In addition, each recorded action was repeated two or three times by each subject. For each skeleton, the 3D joint locations through time were provided. Also, the connections of the nodes that define the recorded skeleton were given. Each recording was captured in 15fps. Noisy videos were removed using the list given by Wang et. al. [18].

Emotional Context Dataset [19]: In this dataset, five different emotional actions were included, namely *anger*, *happiness*, *fear*, *sadness* and *surprise*. The categorisation of these emotional actions was based on social psychology research [20]. For each emotional action, two different action patterns were collected, each having a duration of 4 seconds. A number of 14 subjects (5 women and 9 men) participated in the recording session.

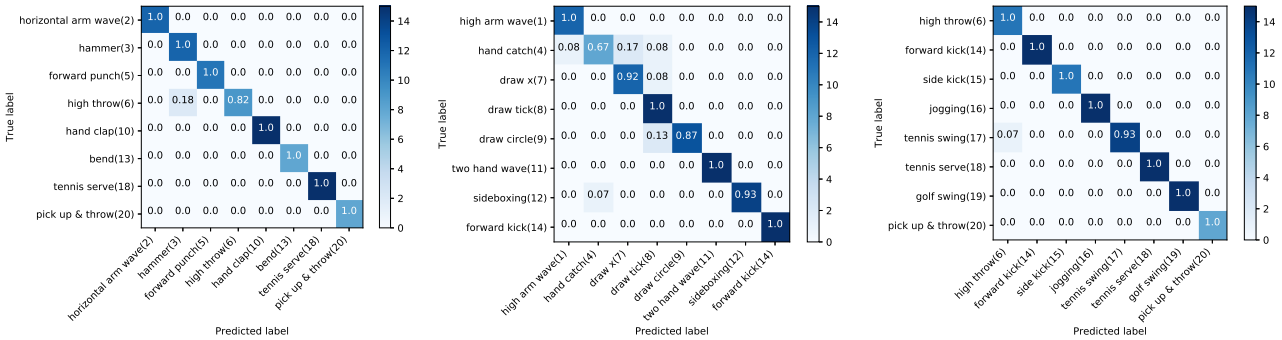
3.3. Experimental results

For MSR Action 3D Dataset, the experimental setup discussed in [17] was followed. More specifically, the dataset was split into three subsets, namely *AS1*, *AS2* and *AS3*. Figure 1 contains the confusion matrix of each subset along with the names of the included actions. Finally, the action samples with odd index were used for training, while those with even index were used for testing. In Table 1, the proposed method was compared to other action recognition methods which use either hand-crafted [21, 9, 12, 10] or automatically learned features, i.e, based on deep learning [11]. From the results presented in Table 1, it is noteworthy that the proposed method demonstrated competitive performance even when compared to robust deep learning based algorithms [11]. The proposed method was also compared to another SFA based action recognition algorithm [12] and yielded better results both in each subset separately and in average. Interestingly, the proposed method also outperformed all hand crafted action recognition algorithms in Table 1.

In order to examine whether the proposed method can group different actions describing the same emotion, the following experiment has been conducted. In [19] a series of features were applied that were capable to capture the context of a specific sentiment and group the action patterns that belong to the same emotion. We have conducted experiments to evaluate our method in the same scope following a leave-one-subject-out experimental setup. The intuition behind this is that the slow functions are capable to capture the content of two different actions that refer to the same emotion. The last row of Table 2 contains the comparative results between the proposed method and the method presented in [19]. We may observe that our method yields better results proving that the learned slow functions can capture the context of a sentiment that is performed by different action patterns. Moreover, separate experiments on the two action patterns were conducted. For Action Pattern 1, we have reach accuracy results of 93.68% while for Action Pattern 2 the accuracy results are 84.54%.

Table 1: Experimental results in percentage scale on the MSR Action3D Dataset

Method	AS1	AS2	AS3	Ave.
Li et. al. 2010 [17]	72.9	71.9	79.2	74.7
Chen et. al. 2013 [21]	96.2	83.2	92.0	90.47
Gowayyed et. al. 2013 [9]	92.39	90.18	91.43	91.26
Yanhu et. al. 2014 [12]	92.47	82.14	97.17	90.59
Vemulapalli et. al. 2014 [10]	95.29	83.87	98.2	92.46
Du et. al. 2015 [11]	93.33	**94.64**	95.50	94.49
srpSFA	**97.83**	92.86	**99.05**	**96.58**



(a) AS1 (b) AS2 (c) AS3

Fig. 1: Confusion matrices for each action subset in MSR Action3D. Each confusion matrix is labeled with the action categories included in each action subset *AS1*, *AS2* and *AS3* along with the accuracy results.

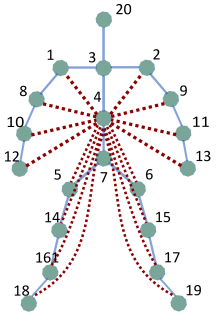


Fig. 2: Anthropocentric structure of the temporal graph $G^{(t)}(\cdot)$. These edges define which edge connections to take into account when computing the weight graph $\Gamma^{(t)}$

.

Table 2: Experimental results in percentage scale on the Emotional context dataset Action3D Dataset

Our method		Method in [19]	
Action Patterns	Acc	Action Patterns	Acc
Action Pattern 1	93.68	Action Pattern 1	-
Action Pattern 2	84.54	Action Pattern 2	-
Both action patterns	**86.42**	Both action patterns	84.78

4. CONCLUSIONS

A method for skeleton action recognition was introduced. According to this approach, concepts like subspace learning with graph structures and SFA were combined. Experiments were conducted to evaluate the proposed method both in the scope of distinct action pattern recognition, but also in a more conceptual way where the context of the emotions that are represented by different action patterns needs to be recognised.

In future work, our efforts will focus on increasing the performance of the algorithm and applying experiments on additional datasets. More specifically, an extension of the unsupervised feature extraction into a supervised one in order to obtain more discriminant features based on the class information will be investigated.

Acknowledgement

The work presented in this document is a result of MaTHiSiS project. This research has received funding from the European Union's Horizon 2020 Programme (H2020-ICT-2015) under Grant Agreement No. 687772

5. REFERENCES

[1] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008, pp. 1–8.

[2] A. Klser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d- gradients," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2008, pp. 275–1.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *International Conference on Computer Communications and Networks*, 2005, pp. 65–72.

[4] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2003, pp. 432–439.

[5] A. Oikonomopoulos, I. Patras, and M. Pantic, "Human action recognition with spatiotemporal salient points," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, pp. 710–719, May 2006.

[6] T. Kadir and M. Brady, "Scale saliency: a novel approach to salient feature and scale selection," in *Proceedings of the IEEE International Conference on Visual Information Engineering (VIE)*, July 2003, pp. 25–28.

[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2004, pp. 32–36.

[8] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[9] Mohammad Abdelaziz Gowayyed, Marwan Torki, Mohammed Elsayed Hussein, and Motaz El-Saban, "Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[10] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 588–595.

[11] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[12] Y. Shan, Z. Zhang, and K. Huang, *Learning Skeleton Stream Patterns with Slow Feature Analysis for Action Recognition*, pp. 111–121, Springer International Publishing, 2015.

[13] Laurenz Wiskott and Terrence J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computations*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[14] E. Kokiopoulou, J. Chen, and Y. Saad, "Trace optimization and eigenproblems in dimension reduction methods," *Numerical Linear Algebra with Applications*, vol. 18, no. 3, pp. 565–602, 2011.

[15] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[16] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems (NIPS)*. 2004, pp. 1601–1608, MIT Press.

[17] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proceedings of IEEE International Conference in Computer Vision and Pattern Recognition-Workshop (CVPRW)*, 2010, pp. 9–14.

[18] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1290–1297.

[19] K. Kaza, A. Psaltis, K. Stefanidis, K. C. Apostolakis, S. Thermos, K. Dimitropoulos, and P. Daras, *Body Motion Analysis for Emotion Recognition in Serious Games*, pp. 33–42, Springer International Publishing, 2016.

[20] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting wiht Computers*, vol. 18, no. 6, pp. 1371–1389, december 2006.

[21] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-Time Image Processing (JRTIP)*, pp. 1–9, 2013.