ADHD indicators modelling based on Dynamic Time Warping from RGBD data: A feasibility study

Antonio Hernández-Vela $^{*+}$, Miguel Reyes^{*+}, Laura Igual^{*+},
Josep Moya†, Verónica Violant‡, and Sergio Escalera^{*+}

* *Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra, Spain*
E-mail:{ahernandez,mreyes,sescalera}@cvc.uab.cat
$^{+}$ *MAIA Dept., University of Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain*
E-mail:{sergio,ligual}@maia.ub.es
† *Mental Health Observatory of Catalunya, Parc Taulí 1, 08208 Sabadell, Spain*
E-mail:jmoya@tauli.cat
‡ *Dept. of Didactics and Educational Organisation, University of Barcelona,*
Campus Mundet, Edifici Llevant 2nd, Passeig de la Vall d'Hebron 171, 08035 Barcelona, Spain
E-mail:vviolant@ub.edu

Abstract

In this paper, we present a feasibility study for the automatic modelling of visual communicative indicators in video sequences of children with attention deficit hyperactivity disorder (ADHD). Our methodology is based on RGBD sequences (RGB + Depth), recorded with the recent Microsoft's Kinect sensor. More specifically, a feature vector is extracted from a previously fitted human skeleton model in the RGBD sequence, and compared to a training set using Dynamic Time Warping (DTW). Finally, some qualitative results are presented, concluding that the presented methodology is feasible for the modelling of ADHD visual indicators.

Keywords: ADHD, Human pose, Motion analysis, Dynamic Time Warping, Depth maps.

1 Introduction

Attention deficit disorder –with or without hyperactivity– is one of the main reasons of consultation in mental health centers for children and adolescents. The basic characteristics of ADHD are excessive and harmful levels of activity, inattention, and impulsiveness. Currently, the diagnosis is made following the criteria of the DSM IV-TR (American Psychiatric Association) and / or CIE-10 (World Health Organization). These criteria include mechanisms to validate three different blocks: attention deficit, hyperactivity, and impulsivity. More specifically, these validation mechanisms are based on visual communicative indicators, which are defined by some specific gestures performed by the patient. This requires observation of patients for long periods of time, and it is often not feasible in practise; hence, we propose an automation in order to help doctors diagnose the disorder.

From the point of view of data acquisition, many methodologies treat images captured by visible-light cameras. Computer Vision is then used to detect, describe, and learn visual features of the human body [1, 2]. The main difficulties of visual descriptors on RGB data is the discrimination of shapes, textures, background objects, changes in lighting conditions and viewpoint. On the other hand, depth information is invariant to color, texture and lighting, making it easier to distinguish between the background and the foreground object. Nowadays, several works have been published related to this topic because of the emergence of inexpensive structured light technology, which is reliable and robust to capture depth information along with the corresponding synchronized RGB image. This technology has been developed by the PrimeSense [4] company and marketed by Microsoft XBox under the name of Kinect. Using this sensor, Shotton et al. [6] present one of the greatest advances in the extraction of the human body pose from depth images, representing the body as a skeletal form comprised by a set of joints. Moreover, some works have started a general behaviour study based on RGBD data [5].

Our proposal is to perform a temporal analysis of the skeletal data given by [6], applied to some captured RGBD sequences of children with ADHD. More specifically, this temporal analysis is performed using DTW in order to segment and model visual communicative indicators which may be potentially useful for the diagnosis of ADHD.

The rest of the paper is organized as follows: Section 2 presents our methodology, section 3 shows some preliminary results we obtained, and finally section 4 concludes the paper.

2 Methodology

In this section, we first describe the feature vector extraction step based on the skeletal model returned by the method of [6]. Secondly, we briefly introduce the DTW framework used for the temporal analysis, and finally, we apply the system to perform begin-end of gesture detection in large video data.

2.1 Feature Vector Extraction

The articulated human model is defined by the set of 15 reference points shown in Figure 1.

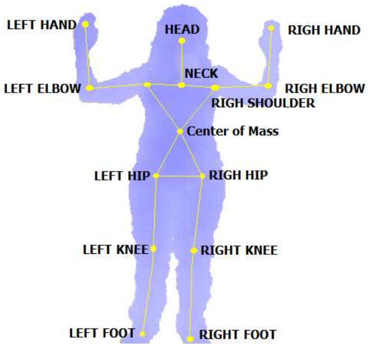

Figure 1: The 3D articulated human model consisting of 15 distinctive points.

This model has the advantage of being highly deformable, and thus, able to fit to complex human poses. In order to subsequently make comparisons and analyze the different extracted skeletal models, we need to normalize them. In this sense, we use the neck joint of the skeletal model as the origin of coordinates (OC). Then, the neck is not used in the frame descriptor, and the remaining 14 joints are using in the frame descriptor computing their 3D coordinates with respect to the OC. This transformation allows us to relate pose models that are at different depths, being invariant to translation, scale, and tolerant to corporal differences of subjects. Thus, the final feature vector $\mathbf{V_j}$ at frame j that defines the human pose is described by 42 elements (14 joints × three spatial coordinates),

$$V_j = \{\{v^1_{j,x}, v^1_{j,y}, v^1_{j,z}\}, ..., \{v^{14}_{j,x}, v^{14}_{j,y}, v^{14}_{j,z}\}\}$$

2.2 Dynamic Time Warping

The DTW algorithm [3] was defined to match temporal distortions between two models, finding an alignment warping path between the two time series $Q = \{q_1, .., q_n\}$ and $C = \{c_1, .., c_m\}$. In order to align these two sequences, a $M_{m\times n}$ matrix is designed, where the position (i, j) of the matrix contains the distance between c_i and q_j. The Euclidean distance is the most frequently applied.

Then, a warping path,

$$W = \{w_1, .., w_T\}, \max(m, n) \leq T < m + n + 1$$

is defined as a set of ”contiguous” matrix elements that defines a mapping between C and Q. This warping path is typically subjected to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_T = (m, n)$.

Continuity: Given $w_{t-1} = (a', b')$, then $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$.

Monotonicity: Given $w_{t-1} = (a', b')$, $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this forces the points in W to be monotically spaced in time.

We are generally interested in the final warping path that satisfying these conditions minimizes the warping cost,

$$DTW(Q, C) = \min \left\{ \frac{1}{T} \sqrt{\sum_{t=1}^{T} w_t} \right\}, \qquad (1)$$

where T compensates the different lengths of the warping paths. This path can be found very efficiently using dynamic programming to evaluate the following recurrence, which defines the cumulative distance $\gamma(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distance of the adjacent elements,

$$\gamma(i, j) = d(i, j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}. \quad (2)$$

Given the nature of our system to work in uncontrolled environments, we continuously review the stage for possible actions or gestures. In this case, our input feature vector Q is of ”infinite” length, and may contain segments related to gesture C at any part.

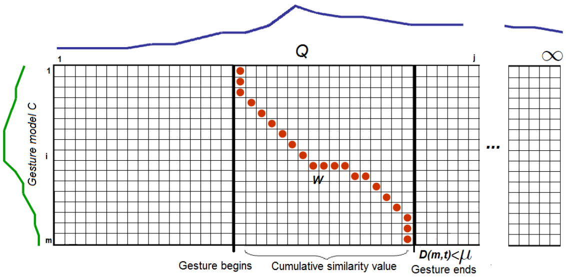

Figure 2: Begin-end of gesture recognition of a model C in an infinite sequence Q.

2.3 Begin-end of gesture detection

In order to detect a begin-end of gesture $C = \{c_1, .., c_m\}$ in a possible infinite sequence $Q = \{q_1, .., q_\infty\}$, a $M_{m \times \infty}$ matrix is designed, where the position (i, j) of the matrix contains the distance between c_i and q_j, quantifying its value by the Euclidean distance, as commented before. Finally, our warping path is defined by $W = \{w_1, .., w_\infty\}$ as in the standard DTW approach. Our aim is focused on finding segments of Q sufficiently similar to the sequence C. The system considers that there is correspondence between the current block k in Q and a gesture if satisfying the following condition,

$$M(m, k) < \mu, k \in [1, .., \infty]$$

for a given cost threshold μ. This threshold value is estimated in advance for each of the categories of actions or gestures using leave-one-out cross-validation strategy. This involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. At each iteration, we evaluate the similarity value between the candidate and the rest of the training set. Finally we choose the threshold value which is associated with the largest number of hits within a category.

Once a possible end of pattern of gesture or action is detected, the working path W can be found through backtracking of the minimum path from $M(m, k)$ to $M(0, z)$, being z the instant of time in Q where the gesture begins. Note that, in this case, $d(i, j)$ is the cost function which measures the difference among our descriptors V_i and V_j. An example of a begin-end gesture recognition for a model and infinite sequence together with the working path estimation is shown in Figure 2.

3 Preliminary Results

In order to validate our proposal, we applied the method on five video sequences of one hour each one at 24 FPS, recorded with the Kinect device. Two different scenarios have been considered. In both of them there are three children between 8-11 years –half of them with ADHD diagnosis– in a classroom, but in the first one they are doing math exercises while in the second one they are playing videogames (see Figure 3). Given this scenario, we defined some specific gestures and trained the system with them. Figure 4 shows a sequence where our method successfully detects the beginning and end for the gesture "lower head", which is one of the visual communicative indicators of ADHD related to the focus of attention and subject agitation.

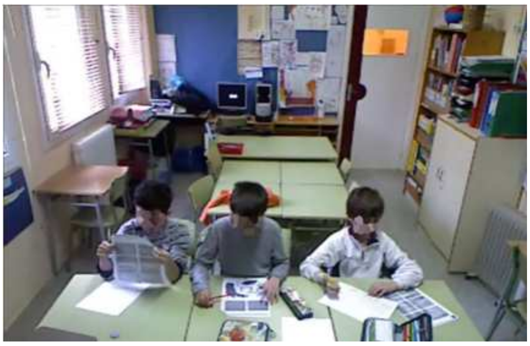

Figure 3: Sample frame from a sequence showing the experiment scenario.

4 Conclusion

In this paper, we have presented a methodology for the detection of the beginning and end of gestures based human skeleton points described using RGBD representation from Kinect device. First results indicate that the presented methodology can successfully recognize a set of defined gestures related to ADHD indicators, showing the viability of the system to be considered for diagnostic support.

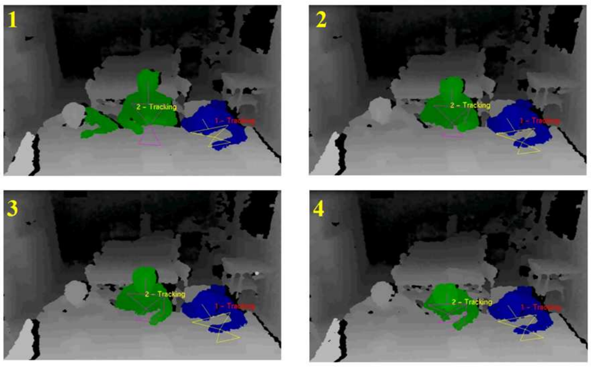

Figure 4: Sequence where the gesture "lower head" is found. Frames 1 and 4 show the beginning and end of the gesture, respectively. Frames 2 and 3 are intermediate frames.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 1:886–893, 2005.

[2] E. N. Mortensen, Hongli Deng, and L. Shapiro. A sift descriptor with global context. *CVPR*, 1:184–190 vol. 1, 2005.

[3] M. Parizeau and R. Plamondon. A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification.

[4] PrimeSense Inc. *Prime Sensor NITE 1.3 Algorithms notes*, 2010. Last viewed 14-07-2011 13:19.

[5] M. Reyes, G. Domnguez, and S. Escalera. Feature weighting in dynamic time warping for gesture recognition in depth data. *HICV workshop, ICCV*, 2011.

[6] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.