

# A Gesture Recognition System for Detecting Behavioral Patterns of ADHD

Miguel Ángel Bautista, Antonio Hernández-Vela, Sergio Escalera, Laura Igual, Oriol Pujol, Josep Moya, Verónica Violant, and María Teresa Anguera

**Abstract**—We present an application of gesture recognition using an extension of Dynamic Time Warping (DTW) to recognize behavioural patterns of Attention Deficit Hyperactivity Disorder (ADHD). We propose an extension of DTW using one-class classifiers in order to be able to encode the variability of a gesture category, and thus, perform an alignment between a gesture sample and a gesture class. We model the set of gesture samples of a certain gesture category using either GMMs or an approximation of Convex Hulls. Thus, we add a theoretical contribution to classical warping path in DTW by including local modeling of intra-class gesture variability. This methodology is applied in a clinical context, detecting a group of ADHD behavioural patterns defined by experts in psychology/psychiatry, to provide support to clinicians in the diagnose procedure. The proposed methodology is tested on a novel multi-modal dataset (RGB plus Depth) of ADHD children recordings with behavioural patterns. We obtain satisfying results when compared to standard state-of-the-art approaches in the DTW context.

**Index Terms**—Gesture Recognition, ADHD, Gaussian Mixture Models, Convex Hulls, Dynamic Time Warping, Multi-modal RGB-Depth data.

## I. INTRODUCTION

Nowadays, human gesture recognition is one of the most challenging tasks in computer vision. Due to the large number of potential applications involving human gesture recognition in fields like surveillance [8], sign language recognition [24], or clinical assistance [17] among others, there is a large and active research community devoted to deal with this problem. Current methodologies have shown preliminary results on very simple scenarios, but they are still far from human performance.

In the gesture recognition field there exists a wide number of methods based on dynamic programming algorithms for both alignment and clustering of temporal series [25]. Probabilistic methods such as Hidden Markov Models (HMM) or Conditional Random Fields (CRF) are also very usual in

the literature [24]. Nevertheless, one of the most common methods for Human Gesture Recognition is Dynamic Time Warping (DTW) [19], [3]. It offers a simple yet effective temporal alignment between sequences of different lengths. However, the application of such methods to gesture detection in complex scenarios becomes a hard task due to the high variability of the environmental conditions among different domains. Some common problems are: wide range of human pose configurations, influence of background, continuity of human movements, spontaneity of human actions, speed, appearance of unexpected objects, illumination changes, partial occlusions, or different points of view, just to mention a few. These effects can cause dramatic changes in the description of a certain gesture, generating a great intra-class variability. In this sense, since usual DTW is applied between a sequence and a single pattern, it fails to take into account such variability.

In addition, the release of the Microsoft Kinect<sup>TM</sup> sensor in late 2010 has allowed an easy and inexpensive access to synchronized depth imaging with standard video data. This data combines both sources into what is commonly named RGB-D images (RGB plus Depth). This data fusion, very welcomed by the computer vision community, has reduced the burden of the first steps in many pipelines devoted to image or object segmentation and opened new questions such as how this data can be effectively described and fused. This depth information has been particularly exploited for human body segmentation and tracking. Shotton [21] introduced one of the greatest advances in the extraction of the human body pose using RGB-D, which is provided as part of the Kinect<sup>TM</sup> human recognition framework. The method is based on inferring pixel label probabilities through Random Forest from learned offsets of depth features. Girshick and Shotton [7] proposed later a different approach in which they directly regress the positions of the body joints, without the need of an intermediate pixel-wise body limb classification as in [21]. The extraction of body pose information opens the door to develop more accurate gesture recognition methodologies.

In particular, there is a growing interest in the application of gesture recognition methods in the clinical context. Concretely, gesture recognition methods can be even more valuable on psychological or psychiatric scenarios where the diagnostic of a certain disease is based on the interpretation of certain behavioural patterns of the subject. Up to date, video sequences were analysed on a frame-by-frame fashion by experts which were typically trained for several months to achieve a good performance on the analysis. Of course, this situation is not applicable to large amounts of data since it

Miguel Ángel Bautista, Antonio Hernández-Vela, Sergio Escalera, Laura Igual and Oriol Pujol are with the Applied Mathematics and Analysis Department of the Universitat de Barcelona at Gran Via 585, 08007 Barcelona, Spain. They are also with the Computer Vision Center, Campus UAB, Edifici O 08193 Bellaterra, Barcelona, Spain, (email: {mbautista,ahernandez,sescalera,ligual,opujol}@ub.edu)

Josep Moya is with the Parc Taulí Foundation at Parc Taulí, 1, 08208 Sabadell, Barcelona, Spain, (email:jmoya@tauli.cat)

Verónica Violant is with the Didactics and Educational Organization Department at University of Barcelona, P. de la Vall d'Hebron 171, 21 planta, 08035 Barcelona, Spain, (email:vviolant@ub.edu)

María Teresa Anguera is with the Department of Behavioral Sciences Methodologies at University of Barcelona. Psychology School, Campus de Mundet - Edifici Ponent Passeig de la Vall d'Hebron, 171 08035 Barcelona, Spain. (email:tanguera@ub.edu)

is a very time consuming procedure and its automatization is highly desirable. Specifically, the case of Attention Deficit Hyperactivity Disorder (ADHD) is one of the most notable scenarios, since it is the most commonly studied and diagnosed psychiatric disorder in childhood, globally affecting about 5 percent of children [14]. In this line of research some works can be found in literature [9], [23], which develop tools to assist children with autism-related disorders. Nevertheless, one of the main problems that clinicians experiment when diagnosing ADHD is the huge subjective component of the interpretation of symptoms, because their definition is either ambiguous or inaccurate. In this sense, an objective gesture recognition tool which is able to detect behavioural patterns defined by a set of psychiatric/psychological experts will be of great value in order to help the clinicians with ADHD diagnose. This work pretends to be a study on a concrete set of ADHD patterns, which aims to be extended in future works.

We propose to use an extension of the DTW method, that is able to perform an alignment between a sequence and a set of  $N$  pattern samples from the same gesture category. The variance caused by environmental factors is modelled using either a Gaussian Mixture Model (GMM) [22] or an approximation of a Convex Hull [5]. Consequently, the distance metric used in the DTW framework is redefined in order to provide a probability-based measure. The proposed method is evaluated in a novel ADHD behavioural pattern dataset, in which both subject diagnosed with ADHD and a control group where recorded in a class-room environment, obtaining satisfying results. Our list of contributions is as follows: i) An extension of classical DTW by modelling the intra-class variability of gestures is proposed. ii) GMMs and approximated Convex Hulls are embedded in the DTW by defining novel distances. iii) A novel multi-modal ADHD behavioural pattern dataset is presented. iv) We test our proposal in the novel ADHD behavioural patterns dataset obtaining very satisfying results.

The rest of the paper is structured as follows: Section II presents the Gesture Recognition proposal. Section IV presents a novel ADHD dataset and shows the experimental results on a novel ADHD behavioural pattern dataset. Finally, Section V summarizes the conclusions.

## II. DEFINITION OF ADHD BEHAVIOURAL PATTERNS AND FEATURE EXTRACTION

We split the methodology of the proposal in different stages. First, we define the ADHD behavioural patterns to be learnt. Second, the considered set of multi-modal features for each frame is described, and finally, the novel DTW extension based on GMM and Convex Hull modelling is presented.

### A. Definition of ADHD Behavioural Patterns

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common childhood disorders and can continue through adolescence and adulthood. Symptoms include difficulty staying focused and paying attention, difficulty controlling behaviour, and hyperactivity. ADHD has three subtypes, defined by DSM IV and CIE X [15], [12]:

- 1) Predominantly hyperactive-impulsive

- 2) Predominantly inattentive

- 3) Combined hyperactive-impulsive and inattentive

In addition children who have symptoms of inattention may:

- Be easily distracted, miss details, forget things, and frequently switch from one activity to another.
- Have difficulty focusing on one task.
- Become bored with a task after only a few minutes, unless they are doing something enjoyable.
- Have difficulty focusing attention on organizing and completing a task or learning something new.

Children who have symptoms of hyperactivity may:

- Fidget and squirm in their seats.
- Dash around, touching or playing with anything and everything in sight.
- Have trouble sitting still during dinner, school, and story time.
- Be constantly in motion.

Children who have symptoms of impulsiveness may:

- Have difficulty waiting for things they want or waiting their turns in games.
- Often interrupt conversations or other activities.

In order to develop a system that automatically detects ADHD behavioural patterns, first we have to define a set of ADHD behavioural patterns (gestures to detect) that are both objective and descriptive yet discriminable. In other words, the set of patterns has to be descriptive enough to provide an ADHD profile of the subject, and simple enough in order to be able to automatize the detection.

In order to define the behavioural patterns to be automatically detected, an analysis of the context in which the video sequences take place has to be performed. Taking into account that video sequences were recorded in a school class context, including mathematical exercises and computer gaming, with no disturbing events taking place, the set of defined ADHD behavioural patterns is the following (an example is shown in Figure 1):

- **Head turning behavioural pattern**

The definition of this behavioural pattern takes its reason from the different symptoms in the *inattention* branch. Behaviours like *be easily distracted, miss details, forget things, and frequently switch from one activity to another.* or *have difficulty focusing on one thing* have a close relationship with turning the head from the goal task to other unrelated task. Therefore, this indicator is defined as a head turn to either right or left sides.

- **Torso in table behavioural pattern**

The Torso in table behavioural pattern is related to *hyperactive* symptoms such *fidget and squirm in their seats* and *have trouble sitting still during dinner, school, and story time.*

- **Classmate's desk invasion behavioural pattern**

This behavioural pattern takes its root from the *impulsive* symptoms like *often interrupt conversations or others' activities* or *have difficulty waiting for things they want or waiting their turns in games.*

- **Movement with/without a pattern behavioural pattern**

The last pattern aims to provide a detection for those symptoms across all ADHD branches (inattentiveness, hyperactivity and impulsiveness) that involve a high quantity of motion.

This set of behavioural patterns is representative enough of the different symptoms of ADHD and provides a generalization analysis of the feasibility of our approach for supporting diagnosis.

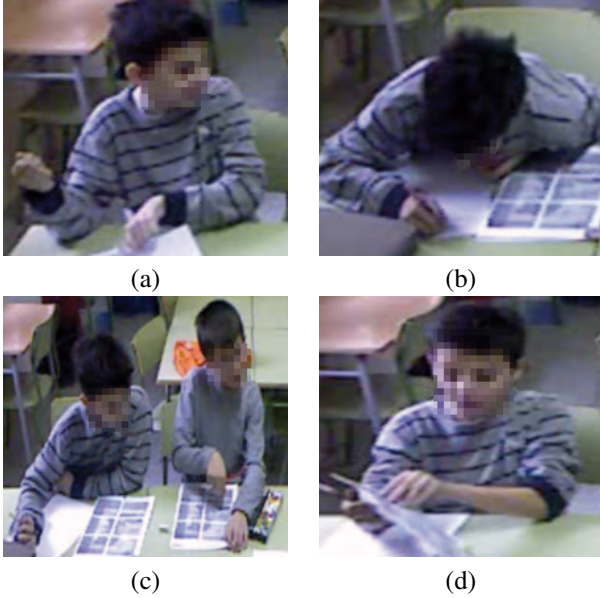


Fig. 1. (a) An example of a the head turning behavioural pattern. (b) Torso in table pattern example, notice how the torso of the subject is completely laid on the table. (c) Sample of a class mate invasion in which the left subject invades the right subject space. (d) Movement behavioural pattern sample.

### B. Image Acquisition, Pre-processing and Feature Extraction

We use the Kinect<sup>®</sup> sensor in order to capture video sequences in which subjects diagnosed with ADHD and subjects not diagnosed with ADHD (control group) were recorded. In this sense, we use the depth information provided by the Kinect<sup>®</sup> sensor to obtain a segmentation of the subjects in the scene obtaining a complete segmentation of their upper-body limbs.

Given a frame  $I^t, t = 1, \dots, T$ , the corresponding segmentation  $S^t$  on the depth map is computed by the OTSU method [16], keeping the biggest convex unconnected components in relation with the number of subjects appearing in the scene. In other words, if three subjects appear on the scene, the three biggest components were kept. Otherwise, if two subjects appear on the scene, the biggest components are kept as the segmentation. Moreover, Random Forest segmentation is applied over the foreground objects [21] in order to segment the regions corresponding to different subjects.

1) *Head Turning Behavioural Pattern Feature*: The features for the head rotation detection are computed for each frame  $t$  as follows. First of all, we obtain the bounding box  $B^t$  containing the head, by means of GrabCut segmentation [20]. As GrabCut is a semi-automatic method, a manual bounding box has to be provided by the user at the first frame. With the

resulting segmentation mask, the bounding box for that frame can be easily computed. Additionally, some morphological operations are applied on the segmentation mask in order to initialize the segmentation of the following frame, as in [10]. Once a bounding box  $B^t$  is detected for one frame, a color-based descriptor  $FHead^t$  is extracted from the pixels inside it. The bounding box is firstly divided in  $\bar{O} \times O$  cells, and each one of them is described with a label  $\gamma \in \{1, \dots, G\}$  corresponding to the most frequent color as follows:

$$FHead_{i,j}^t = \arg \max_{l \in \gamma} \left( \sum_{\mathbf{x} \in B_{i,j}^t} \delta(\text{ColorName}(\mathbf{x}) - l) \right), \quad \forall i \in 1, \dots, \bar{O}, \quad \forall j \in 1, \dots, O, \quad (1)$$

where  $B_{i,j}^t$  is the  $(i, j)$ -th cell of the head bounding box at time  $t$ . In addition,  $\text{ColorName}(\mathbf{x})$  is a function which returns the color name of an RGB pixel  $\mathbf{x}$ , and  $\delta(\cdot)$  is a Dirac delta function. The Color-naming data with  $G = 11$  basic colors (red, orange, brown, yellow, green, blue, purple, pink, white, grey, black) presented in [18] has been used. An example of the feature computation procedure is show in Figure 2.



Fig. 2. Color descriptor for the 'Head turning' behavioural pattern. Image in first column shows a subject turning the head, while in the image at the last column shows a frontal face. Bounding boxes are overlaid in green color. Images in the central column show the respective color naming descriptors. They are composed by  $4 \times 4$  cells, each one of them containing a color name label.

2) *Torso on Desk Behavioural Pattern Feature*: The torso on desk feature computes the relative distance of the subject's torso to the desk, in order to provide a measure of how close the torso is in relation to the desk. In this sense, this distance is computed as the Euclidean distance of the top pixel of the head to the closest desk pixel. This distance can be easily computed by finding the uppermost pixel  $x^{top} = \{x_i | (x_i, y_i) \in S^t, (x_j, y_j) \in S^t, y_i \leq y_j, \forall i \neq j\}$  in the segmentation mask  $S^t$  of the subject, and its corresponding lowermost pixel in vertical direction  $x^{bot} = \{x_i | (x_i, y_i) \in S^t, (x_j, y_j) \in S^t, y_i \geq y_j, \forall i \neq j\}$ :

$$FTorso^t = \|x^{top} - x^{bot}\|_2. \quad (2)$$

An example of the feature calculation is shown in Figure 3.

3) *Classmate's Desk invasion feature*: In order to compute the Classmate's Desk Invasion feature, we also use the segmentation mask  $S^t$ . For a given subject, the feature is basically defined as the minimum distance between the pixels in the subject's unconnected components of the mask  $S^t$ , and the pixels in the neighbour classmate's components  $S_{ne}^t$  ( $ne = 1, 2$  in our case):

$$FInv^t = \min_{ne \in N} \left( \min_{\mathbf{x}_n \in S_n^t} \left( \min_{\mathbf{x}_s \in S^t} \|\mathbf{x}_s - \mathbf{x}_n\|_2 \right) \right). \quad (3)$$

An example of this computation is shown in Figure 3.

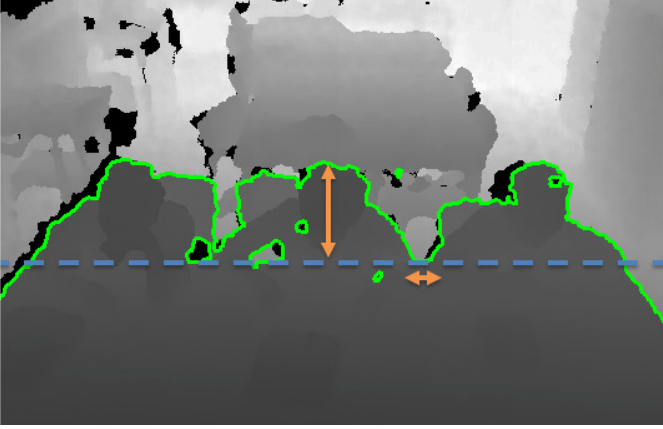


Fig. 3. Distances computed using the segmentation of the depth image. Green contour indicates the boundary of the segmentation mask. Blue dashed line shows the table limit. Vertical and horizontal orange arrows show the distances computed for the "Torso in table" and "Classmates desk invasion" behavioural patterns, respectively.

4) *Movement with/without a pattern feature*: In the movement with/without a pattern feature we want to describe general movements of the subject, so we first compute the optical flow [13] between current and next frames. Then, we compute the average optical flow magnitude over the pixels belonging to the segmentation mask of the subject:

$$FMov^t = \frac{1}{|S^t|} \sum_{\mathbf{x} \in S^t} \sqrt{u_{\mathbf{x}}^2 + v_{\mathbf{x}}^2}, \quad (4)$$

where  $u_{\mathbf{x}}$  and  $v_{\mathbf{x}}$  are the components of the flow vector between current frame  $I^t$  and  $I^{t+1}$ , and  $|\cdot|$  computes the number of elements of the set.

### III. DYNAMIC TIME WARPING BASED ON ONE-CLASS CLASSIFIERS

The original DTW algorithm was defined to match temporal distortions between two models, finding an alignment/warping path between the two time series  $Q = \{q_1, \dots, q_n\}$  and  $C = \{c_1, \dots, c_m\}$ . 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 alignment cost between  $c_i$  and  $q_j$ . Then, a warping path of length  $\tau$  is defined as a set of contiguous matrix elements, defining a mapping between  $C$  and  $Q$ :  $W = \{w_1, \dots, w_\tau\}$ , where  $w_i$  indexes a position in the cost matrix. This warping path is typically subjected to several constraints:

*Boundary conditions*:  $w_1 = (1, 1)$  and  $w_\tau = (m, n)$ .

*Continuity and monotonicity*: Given  $w_{\tau'-1} = (a', b')$ , then  $w_{\tau'} = (a, b)$ ,  $a - a' \leq 1$  and  $b - b' \leq 1$ . This condition forces the points in  $W$  to be monotonically spaced in time.

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

$$DTW(M) = \min_W \left\{ \frac{M(w_\tau)}{\tau} \right\}, \quad (5)$$

where  $\tau$  compensates the different lengths of the warping paths. This path can be found very efficiently using dynamic programming. The cost at a certain position  $M(i, j)$  can be found as the composition of the Euclidean distance  $d(i, j)$  between the feature vectors of the sequences  $c_i$  and  $q_j$  and the minimum cost of the adjacent elements of the cost matrix up to that point, i.e.:  $M(i, j) = d(i, j) + \min\{M(i-1, j-1), M(i-1, j), M(i, j-1)\}$ .

Given the streaming nature of our problem, the input vector  $Q$  has no definite length and may contain several occurrences a gesture class, namely  $C$ . At that point the system considers that there is correspondence between the current block  $k$  in  $Q$  and a gesture if satisfying the following condition,  $M(m, k) < \beta$ ,  $k \in [1, \dots, \infty]$  for a given cost threshold  $\beta$ .

This threshold is estimated in advance using leave-one-out cross-validation strategy on the training set. 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.

Once the threshold is defined and a possible end of pattern of gesture 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  $d(i, j)$  is the cost function which measures the difference among our descriptors  $c_i$  and  $q_j$ .

An example of a begin-end gesture recognition together with the warping path estimation is shown in Figure 5.

#### A. Handling temporal deformation in sequences

Consider a training set of  $N$  sequences  $\{S_1, S_2, \dots, S_N\}$ , where all sequences belong to a certain gesture class. Then, each sequence  $S_g$  is composed by a set of feature vectors at each time  $t$ ,  $S_g = \{s_1^g, \dots, s_{L_g}^g\}$ , where  $L_g$  is the length in frames of sequence  $S_g$ . Let us assume that sequences are ordered according to their length, so that  $L_{g-1} \leq L_g \leq L_{g+1}, \forall g \in [2, \dots, N-1]$ , and the median length sequence is  $\tilde{S} = S_{\lceil \frac{N}{2} \rceil}$ . This sequence is used as a reference, and the rest of the sequences are aligned with respect to it using the classical Dynamic Time Warping with Euclidean distance, in order to avoid the temporal deformations of different samples from the same gesture category. Therefore, after the alignment process, all sequences have length  $L_{\lceil \frac{N}{2} \rceil}$ . We define the set of warped sequences as  $\{\tilde{S}_1, \tilde{S}_2, \dots, \tilde{S}_N\}$ .



Once all samples are aligned, the feature vectors corresponding to a certain time  $t$  among all sequences  $\tilde{s}_t^g \forall g \in [1, \dots, N]$  are modelled by means of one-class classifiers (i.e GMMs) in order to encode intra-class variability. An example of the process using GMMs is shown in Figure 4.

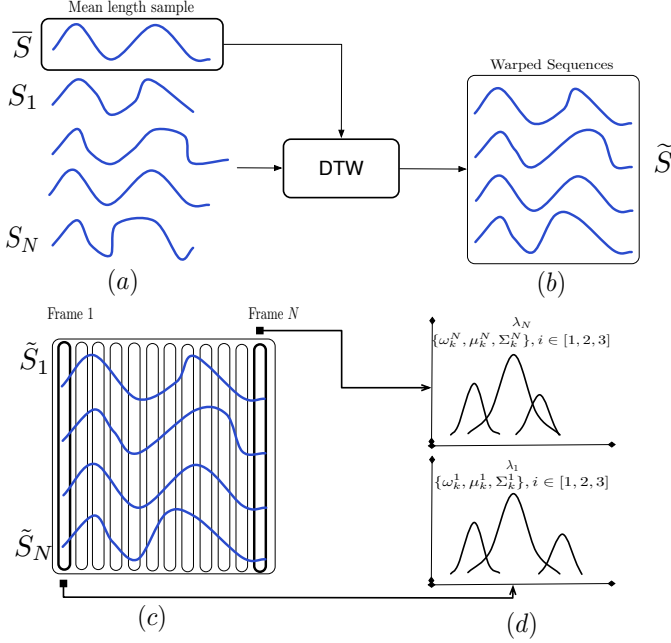


Fig. 4. (a) Different sample sequences of a certain gesture category and the mean length sample. (b) Alignment of all samples with the mean length sample by means of Euclidean DTW. (c) Warped sequences set  $\tilde{S}$  from which each set of  $t$ -th elements among all sequences are modelled. (d) Gaussian Mixture Model learning with 3 components.

### B. Embedding One-Class Classifiers in DTW

In the classical DTW, a pattern and a sequence are aligned using a distance metric, such as the Euclidean distance. Since our pattern is modelled by means of one-class models, if we want to use the principles of DTW, the distance needs to be redefined. Next, we propose two cost distances, one based on GMM and the other on approximated Convex Hull.

1) *Gaussian Mixture Models*: We propose to use Gaussian Mixture Models (GMM) to learn the features among all sequence samples (of a gesture category) at a certain time  $t$ ,  $\tilde{s}_t^g \forall g \in [1, \dots, N]$ . Since after the alignment step all the sequences have the same length,  $L_{\lceil \frac{N}{2} \rceil}$ , we learn  $L_{\lceil \frac{N}{2} \rceil}$  GMMs, one per each component.

In this sense, a  $G$ -component Gaussian Mixture Model, is defined as,  $\lambda_t = \{\alpha_k^t, \mu_k^t, \Sigma_k^t\}$ ,  $k = 1, \dots, G$ , where  $\alpha$  is the mixing value and  $\mu$  and  $\Sigma$  are the parameters of each of the  $G$  Gaussian models in the mixture. As a result, each one of the GMMs that model each set of  $t$ -th components  $\tilde{s}_t$ , among all warped sequence samples, is defined as follows:

$$p(\tilde{s}_t) = \sum_{k=1}^G \alpha_k \cdot e^{-\frac{1}{2}(q-\mu_k)^T \cdot \Sigma_k^{-1} \cdot (q-\mu_k)}. \quad (6)$$

The resulting model is composed by a set of  $L_{\lceil \frac{N}{2} \rceil}$  GMMs

corresponding to the modelling of each one of the component elements of the warped sequence  $\tilde{s}_t$  for each gesture pattern.

In this paper we consider a soft-distance based on the probability of a point belonging to each one of the  $G$  components in the GMM, i.e., the posterior probability of  $q \in Q$  is obtained according to Equation 6. In addition, since  $\sum_k \alpha_k = 1$ , we can compute the probability of  $x$  belonging to the whole GMM  $\lambda$  as the following:

$$P_{\text{GMM}}(q, \lambda) = \sum_{k=1}^M \alpha_k \cdot P(q)_k, \quad (7)$$

$$P(x)_k = e^{-\frac{1}{2}(x-\mu_k)^T \cdot \Sigma_k^{-1} \cdot (x-\mu_k)}, \quad (8)$$

which is the sum of the weighted posterior probability of each component. However, an additional step is required since the standard DTW algorithm is conceived for distances instead of similarity measures. In this sense, we use a soft-distance based measure of the probability, which is defined as:

$$D(x, \lambda) = e^{-P_{\text{GMM}}(x, \lambda)}. \quad (9)$$

An example of the use of GMMs framework to detect a given gesture is shown in Figure 5.

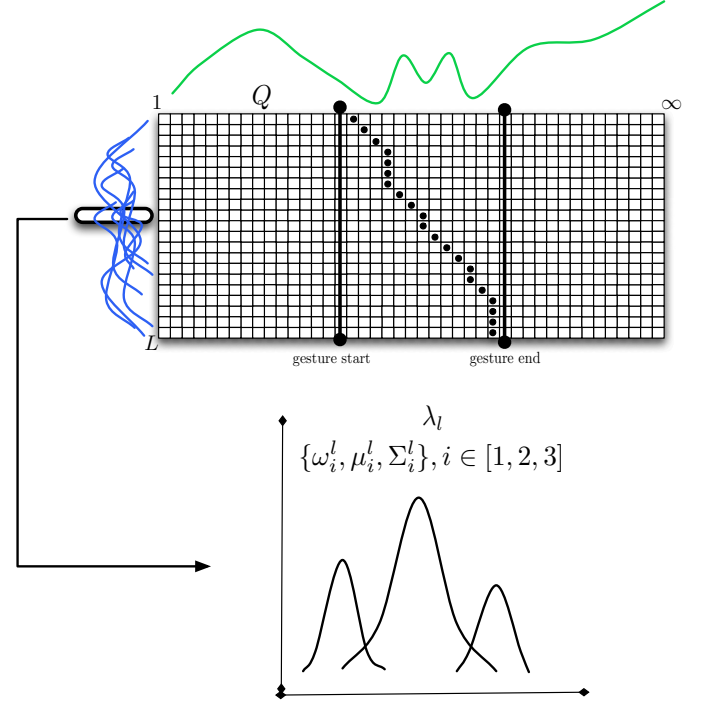


Fig. 5. Begin-end of gesture recognition of a gesture pattern in an infinite sequence  $Q$  using the probability-based DTW. Note that different samples of the same gesture category are modelled with a GMM and this model is used to provide a probability-based distance. In this sense, each cell of  $M$  will contain the accumulative  $D$  distance.

2) *Convex Hulls and Approximate Convex Polytope decision Ensemble*: In addition to the use of GMMs as One-class classifiers, we also propose to use Convex Hulls to model the set of features,  $\tilde{s}_t^g \forall g \in [1, \dots, N]$ . The underlying idea of Convex Hulls is to model the boundary of the set of

points defining the problem. If the boundary encloses a convex area, then the convex hull, defined as the minimal convex set containing all the training points, provides a good general tool for modelling the target class, which in our case will be the set of features of all sequence samples at a certain time.

The convex hull of a set  $\mathcal{C} \subseteq \mathbb{R}^d$ , denoted  $\text{conv } \mathcal{C}$ , is defined as the smallest convex set that contains  $\mathcal{C}$  and is defined as the set of all convex combinations of points in  $\mathcal{C}$ :

$$\text{conv } \mathcal{C} = \{\theta_1 x_1 + \dots + \theta_m x_m \mid x_i \in \mathcal{C}, \theta_i \geq 0, \forall i; \sum_i \theta_i = 1\} \quad (10)$$

In this scenario, the one-class classification task is reduced to the problem of knowing if test data lie inside or outside the hull. Although the convex hull provides a compact representation of the data, a small amount of outliers may lead to very different shapes of the convex polytope. Thus, a decision using these structures is prone to over-fitting. In [5], the authors show that it is useful to define a parametrized set of convex polytopes associated with the original convex hull of the training data. This set of polytopes are shrunk/enlarged versions of the original convex hull governed by a parameter  $\varphi$ . The goal of this family of polytopes is to define the degree of robustness to outliers. The parameter  $\varphi$  defines a constant shrinking ( $-\|\varphi - \varsigma\| \leq \varphi \leq 0$ ) or enlargement ( $\alpha \geq 0$ ) of the convex structure with respect to the center  $c$ . If  $\varphi = 0$  then  $\wp_0 = \text{conv } \mathcal{C}$ .

However, the creation of high-dimensional convex hulls is computationally intensive. In general, the cost for computing a  $d$ -dimensional convex hull on  $N$  data examples is  $\mathcal{O}(N^{\lfloor d/2 \rfloor + 1})$ . This cost is prohibitive in time and memory and, for the classification task, only checking if a point lies inside the multidimensional structure is needed. Instead, we propose to use the Approximate convex Polytope decision Ensemble (APE) of [5]. This method consists in approximating the decision made using the extended convex polytope in the original  $d$ -dimensional space by aggregating a set of  $F$  decisions made on low-dimensional random projections of the data.

Since the projection matrix is created at random, the resulting space does not preserve the norm of the original space. Hence, a constant value of the parameter  $\varphi$  in the original space corresponds to a set of values  $\gamma_i$  in the projected one. As a result, the low-dimensional approximation of the expanded polytope is defined by the set of vertices as follows:

$$\bar{\wp}^\varphi : \{\bar{\wp}_i + \omega_i \frac{(\bar{\wp}_i - \bar{\varsigma})}{\|\bar{\wp}_i - \bar{\varsigma}\|}\}, i = 1, \dots, N, \quad (11)$$

where  $\bar{\varsigma} = \rho \varsigma$  represents the projected center,  $\bar{\wp}_i$  is the set of vertices belonging to the convex hull of the projected data and  $\gamma_i$  is defined as follows:

$$\omega_i = \frac{(\wp_i - \varsigma)^T \rho^T \rho (\wp_i - \varsigma)}{\|\wp_i - \varsigma\|} \alpha, \quad (12)$$

where  $\rho$  is the random projection matrix,  $\varsigma$  is the center and  $\wp_i$  is the  $i$ th vertex of the convex hull in the original space. Note that there exist a different expansion factor for

each vertex  $\wp_i$  belonging to the projected convex hull. Thus, we defined an APE model as:

$$\psi = \{\bar{\wp}_f^\varphi\}, \quad (13)$$

where  $f \in [1, \dots, F]$ , and  $F$  is the number of total random projections used to approximate the original convex hull. In this sense, to obtain the probability of a point belonging to the extended/shrunk convex polytope ensemble we compute the proportion of low-dimensional random projections in which the testing point  $q$  lies inside the extended convex polytope. In this sense, we get an approximate measure of how probable is the point to be inside the original Convex Hull. The calculation of the proportion is as follows:

$$P_{APE}(q, \psi) = \frac{\sum_{f=1}^F \mathbb{1}_{q \in \text{conv } \bar{\wp}_f^\varphi}}{F}. \quad (14)$$

Following the same scheme used with GMM, we compute a soft distance based on the proportion of random projections in which the testing point  $q$  lies inside the extended convex polytope. This soft-distance is defined as follows,

$$D(q, \psi) = e^{-P_\psi(q, \psi)}. \quad (15)$$

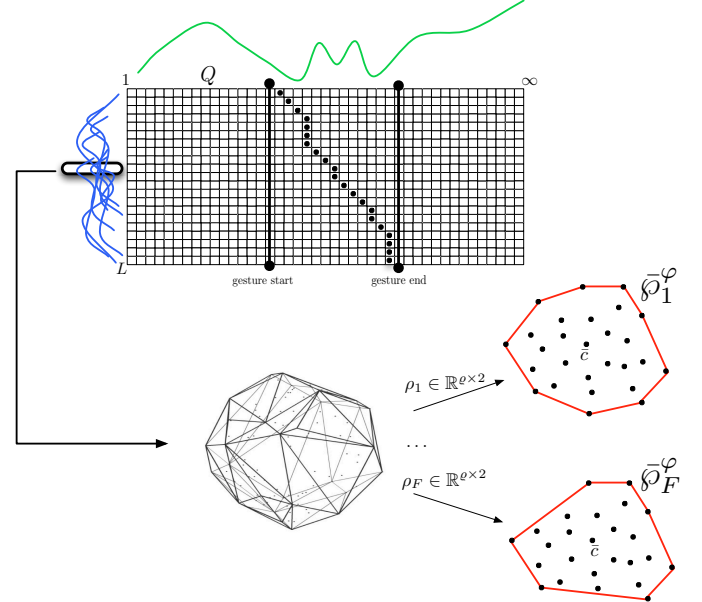


Fig. 6. Begin-end of gesture recognition of a gesture pattern in an infinite sequence  $Q$  using the probability-based DTW. In this example, APEs are used to model each set of  $i$ -th frames.

Finally, Algorithm I shows the proposed DTW algorithm for begin-end gesture detection, where the compute distance  $D$  is computed from APE models.

#### IV. EXPERIMENTAL RESULTS

In order to present the experimental results, first, we introduce the data, methods, and evaluation measurements of the experiments.

TABLE I  
PROBABILITY-BASED DTW APPLIED TO BEGIN-END OF GESTURE  
RECOGNITION, USING APES AS BASE CLASSIFIERS.

**Input:** A gesture model composed by a set of APE models  $C = \{\psi_1, \dots, \psi_m\}$ , a threshold value  $\beta$ , and the testing sequence  $Q = \{q_1, \dots, q_\infty\}$ . Cost matrix  $M_{m \times \infty}$  is defined, where  $\mathcal{N}(x), x = (i, j)$  is the set of three upper-left neighbor locations of  $x$  in  $M$ .  
**Output:** Working path  $W$  of the detected gesture, if any.  
// Initialization  
**for**  $i = 1 : m$  **do**  
    **for**  $j = 1 : \infty$  **do**  
         $M(i, j) = \infty$  **item**      **end**  
**end**  
**for**  $j = 1 : \infty$  **do**  
     $M(0, j) = 0$   
**end**  
**for**  $j = 0 : \infty$  **do**  
    **for**  $i = 1 : m$  **do**  
         $x = (i, j)$   
         $M(x) = \{D(x, \psi) + \min_{x' \in \mathcal{N}(x)} M(x')\}$   
        **end**  
        **if**  $M(m, t) < \mu$  **then**  
             $W = \{\text{argmin}_{x' \in \mathcal{N}(x)} M(x')\}$   
            **return**  
        **end**  
    **end**  
**end**

#### A. The ADHD Behavioural Patterns Dataset

In this section we introduce the novel dataset in which the experiments are performed. The *ADHD behavioural patterns dataset* is composed of 18 video sequences in which both, a group of three subjects diagnosed with ADHD and three control subjects are recorded in a scholar context, performing recreational and mathematical tasks. These video sequences were recorded using the Kinect© sensor, which is able to obtain RGB and depth information. The features of the dataset are the following:

- There is an equal proportion of video sequences of ADHD subjects and the control group.
- There is an equal proportion of video sequences in which the subjects were performing recreational tasks and mathematical tasks.
- The mean length of the video sequences was approximately 5 minutes each.
- Outlier events taking place during the recording sessions were manually filtered from the sequences.

For each one of the video sequences a manual labelling process was performed, in which two independent observers labelled the start and ending points of each one of the four behavioural patterns defined in Section II-A (head turn, torso in table, classmate desk invasion and movement with/without pattern). The agreement of the labelling of the independent observers was measured with the well-known Cohen's kappa coefficient for inter-annotator agreement [4]. In order to obtain this measure we used the GSEQ software presented in [2]. Finally, the mean Cohen's Kappa statistic of the labelling procedure was 0.93, which follows in the interval defined as *almost perfect agreement* in [11], and thus, this labelling is used as the ground truth for evaluating the performance of the proposed methodologies. Table II shows a summary of

the number of samples per subject and behavioural pattern. In addition, in Figure 7 some samples of the ADHD behavioural pattern dataset are shown. The dataset is composed of approximately 50.000 frames.

TABLE II  
NUMBER OF SAMPLES PER SUBJECT AND BEHAVIOURAL PATTERNS.

	Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5
<b>Head Turn</b>	14	17	24	2	3
<b>Torso in Table</b>	4	3	5	0	0
<b>Class. Inv.</b>	7	8	7	0	0
<b>Movement</b>	110	98	130	9	6
<b>ADHD</b>	Yes	Yes	Yes	No	No

#### B. Methods

We compare the following methods, which have been proposed in the paper:

- **DTW random**, aligning the streaming sequence  $Q$  with a sample selected randomly from the training set of gesture samples for a certain behavioural pattern, using the standard Euclidean distance.
- **DTW mean**, aligning the streaming sequence  $Q$  with the mean of the set of warped samples  $\tilde{S}$ , using also the Euclidean distance.
- **DTW GMM**, where the sequence  $Q$  is aligned to a whole gesture category by taking into account the probability of a element in  $Q$  on the whole GMM, proposed in Section III-B1.
- **DTW APE**, where the sequence  $Q$  is aligned to a certain gesture category by modelling the probability of an element in  $Q$  as the number of random projections in which the point lies inside a projected Convex Hull, proposed in Section III-B2.

#### C. Evaluation measurements

The evaluation measurements are overlapping and accuracy recognition (in percentage). For the accuracy analysis, we consider that a gesture is correctly detected if overlapping in the gesture sub-sequence is greater than 60% (the standard overlapping value [1]). The overlapping measure is defined by  $\frac{g \cap p}{g \cup p}$ , where  $g$  is the ground truth and  $p$  the prediction. The cost threshold for all methods was obtained by means of a stratified five-fold cross-validation. In addition, we apply the Friedman and Nemenyi tests [6] in order to look for statistical significance among the obtained performances.

Furthermore, to allow a deeper analysis of the proposed methodologies and their clinical impact, in our evaluations we use a 'Don't care' value which provides a more flexible interpretation of the results. Consider the ground truth of a certain gesture category in a video sequence as a binary vector, which activates when a sample of such category is observed in the sequence. Then, the 'Don't care' value is defined as the number of bits (frames) which are ignored at the limits of each one of the ground truth instances. Thus, by using this approach we can compensate the pessimistic overlap metric in situations when the detection is shifted some frames. An example of this situation is shown in Figure 8.

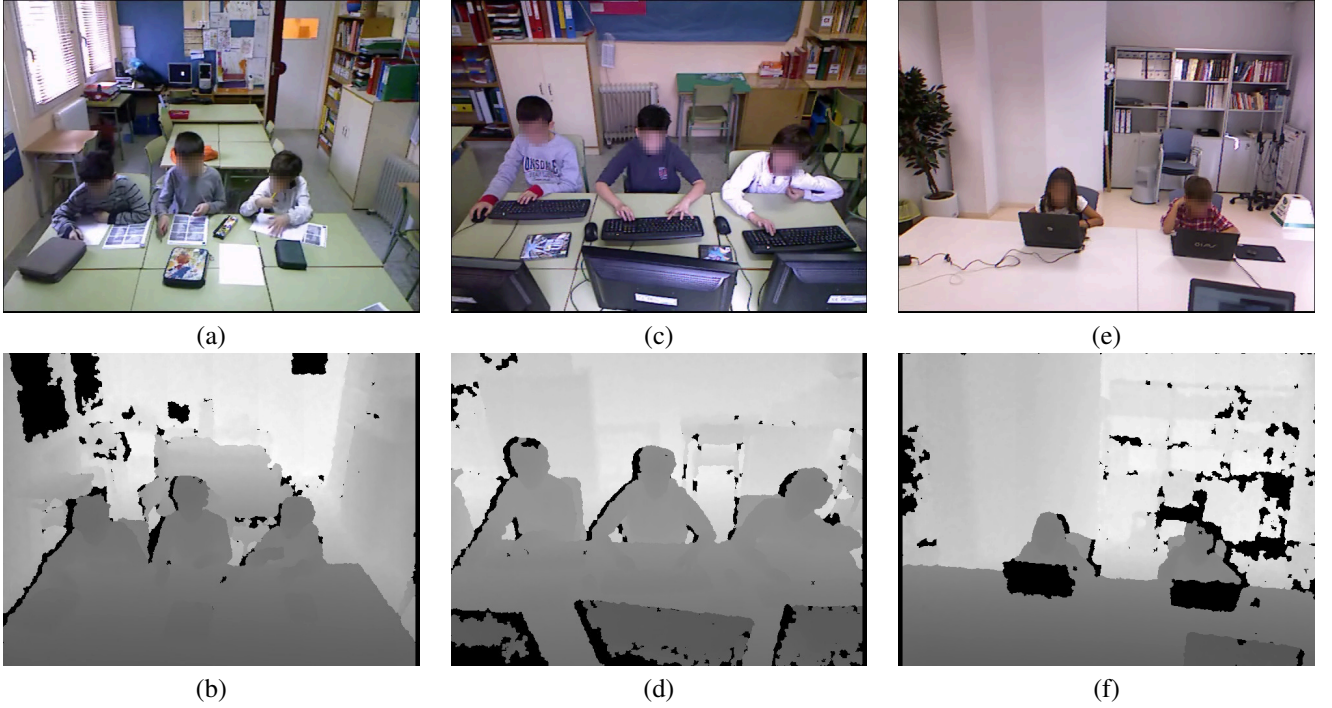


Fig. 7. (a) RGB image of the subjects diagnoses with ADHD performing mathematical tasks. (b) Depth information of ADHD subjects performing mathematical exercises. (c) RGB frame of ADHD subject in the recreational task context. (d) Depth information of ADHD subjects. (e) RGB image of the control group. (f) Depth image of the control group.

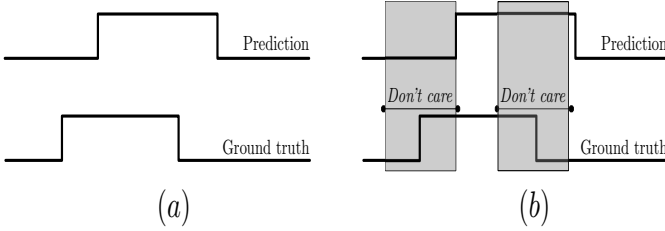


Fig. 8. (a) Example of overlapping between a prediction and the ground truth. (b) Example where the *Don't Care* value is used to soften the overlap metric.

### D. Experimental Results

Figure 9 shows the overlapping and accuracy percentages of each one of the compared methods and for each one of the defined behavioural patterns.

In order to present a more reduced and understandable version of the results, we selected specific 'Don't care' values and performed an analysis on those cases. In Tables III and IV we show the overlapping and accuracy values per behavioural pattern and method for certain 'Don't Care' values.

Finally, Table V shows the mean rank per each methodology and the final mean rank.

Once all the rankings are computed, in order to reject the null hypothesis that the measured performance ranks differ from the mean performance rank, and that the performance ranks are affected by randomness in the results, we use the Friedman test. Thus, with  $h = 4$  methods to compare and  $U = 4$  behavioural patterns  $\times 4$  *Don't care* values (1,50,100,150)  $\times 2$  metrics (overlapping and accuracy) = 32, the Friedman

TABLE III  
PERFORMANCE OF THE COMPARED METHODOLOGIES IN TERMS OF OVERLAPPING.

<i>Head Turn</i>	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.1012	0.0581	<b>0.1015</b>	0.0942
DC 50	<b>0.2314</b>	0.1352	0.1998	0.1924
DC 100	<b>0.2960</b>	0.1753	0.2673	0.2582
DC 150	<b>0.3257</b>	0.2179	0.3096	0.2954
<i>Torso in Table</i>	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.0979	0.1737	0.0966	<b>0.2521</b>
DC 50	0.1412	0.2165	0.1415	<b>0.2901</b>
DC 100	0.1675	0.2402	0.1895	<b>0.3134</b>
DC 150	0.1964	0.2628	0.2293	<b>0.3364</b>
<i>Classmate Inv.</i>	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.2830	<b>0.3610</b>	0.2796	0.3198
DC 50	0.3308	<b>0.4164</b>	0.3266	0.3573
DC 100	0.3666	<b>0.4603</b>	0.3649	0.3893
DC 150	0.4019	<b>0.4975</b>	0.4005	0.4174
<i>Movement</i>	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.1028	0.0789	0.1682	<b>0.2521</b>
DC 50	0.2683	0.2121	0.3718	<b>0.3945</b>
DC 100	0.3826	0.2981	0.4429	<b>0.4651</b>
DC 150	0.4551	0.3672	0.5044	<b>0.5215</b>

statistic value is computed as follows, where  $V$  is the mean rank:

$$X_F^2 = \frac{12U}{h(h+1)} \left[ \sum_j V_j^2 - \frac{h(h+1)^2}{4} \right]. \quad (16)$$

In our case, with  $h = 4$  DTW methods to compare,  $X_F^2 = 14.8875$ . Since this value is undesirable conservative, Iman and Davenport proposed a corrected statistic:

$$F_F = \frac{(U-1)X_F^2}{U(h-1) - X_F^2}. \quad (17)$$



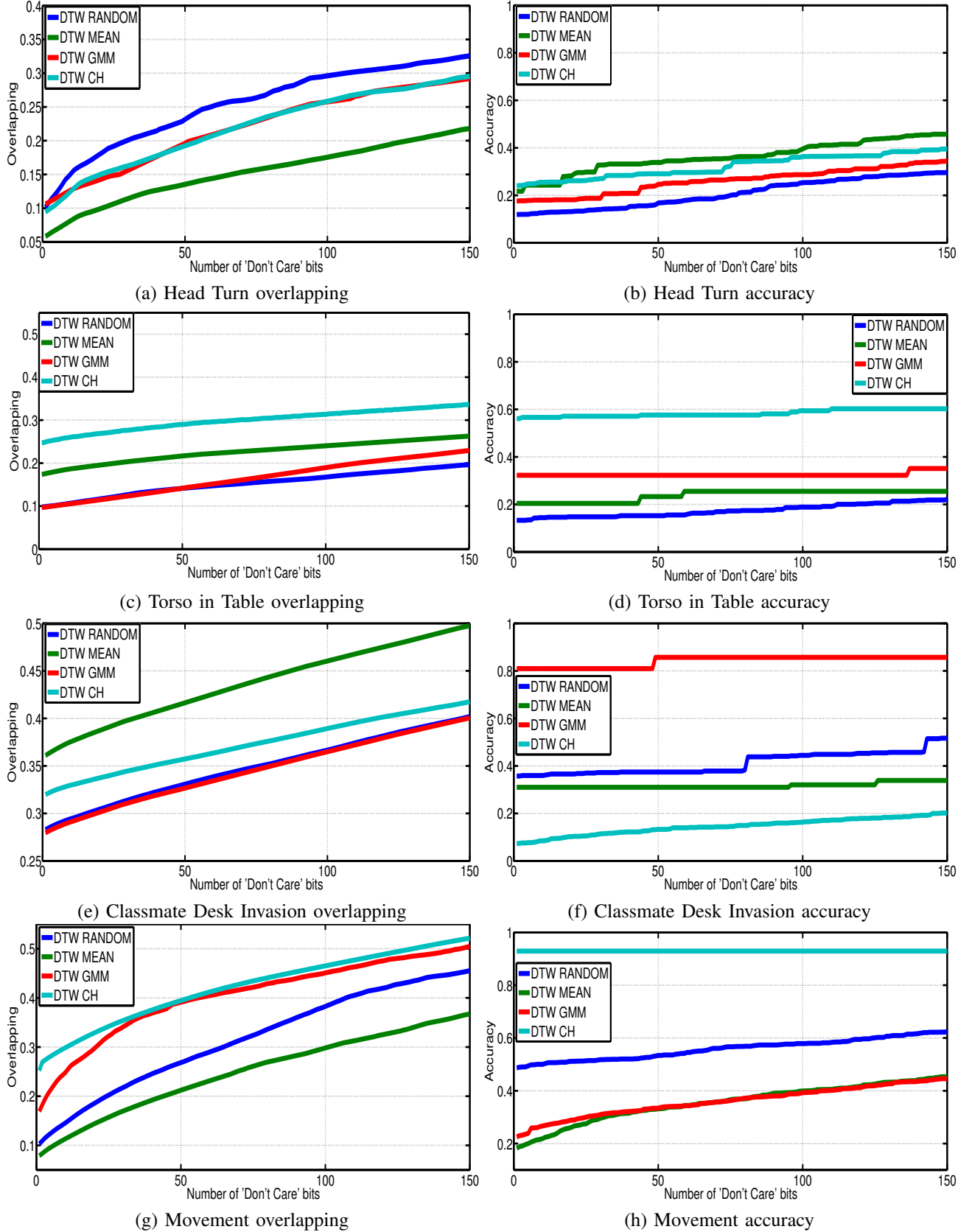


Fig. 9. (a) Overlapping metric per method and number of 'Don't Care' bits for the *Head Turn* behavioural pattern. (b) Accuracy value for each one of the compared methods and number of 'Don't Care' bits for the *Head Turn* pattern. (c) Overlapping metric for each method and number of 'Don't Care' bits for the *Torso in Table* pattern. (d) Accuracy metric per each method and number of 'Don't Care' bits for the *Torso in Table* behavioural pattern. (e) Overlapping metric and number of 'Don't Care' bits for the *Classmate Desk Invasion* behavioural pattern. (f) Accuracy value per each compared method and number of 'Don't Care' bits for the *Classmate Desk Invasion* behavioural pattern. (g) Overlapping metric per method and number of 'Don't Care' bits for the *Movement* pattern. (h) Accuracy value and number of 'Don't Care' bits for the *Movement* behavioural pattern.

TABLE IV  
PERFORMANCE OF THE COMPARED METHODOLOGIES BASED ON THE ACCURACY METRIC.

Head Turn	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.1198	0.2174	0.1755	<b>0.2399</b>
DC 50	0.168	<b>0.3386</b>	0.2469	0.2909
DC 100	0.2534	<b>0.3955</b>	0.2871	0.3638
DC 150	0.2963	<b>0.4573</b>	0.3448	0.3951
Torso in Table	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.1331	0.2043	0.3227	<b>0.5614</b>
DC 50	0.1525	0.2328	0.3227	<b>0.5760</b>
DC 100	0.1885	0.2551	0.3227	<b>0.5941</b>
DC 150	0.2215	0.2551	0.3512	<b>0.6024</b>
Classmate Inv.	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.3575	0.3100	<b>0.8096</b>	0.0729
DC 50	0.3743	0.3100	<b>0.8572</b>	0.1330
DC 100	0.4445	0.3204	<b>0.8572</b>	0.1636
DC 150	0.5168	0.3389	<b>0.8572</b>	0.2015
Movement	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	0.1827	0.2254	0.4870	<b>0.9291</b>
DC 50	0.3321	0.3340	0.5339	<b>0.9291</b>
DC 100	0.3992	0.3933	0.5789	<b>0.9291</b>
DC 150	0.4536	0.4458	0.6238	<b>0.9291</b>

TABLE V  
MEAN RANKS FOR EACH METHOD AND CERTAIN 'DON'T CARE VALUES'.

Mean rank	DTW Random	DTW Mean	DTW GMM	DTW CH
DC 1	3.1250	2.7500	2.3750	<b>1.7500</b>
DC 50	3.1250	2.625	2.3750	<b>1.8750</b>
DC 100	3.0000	2.7500	2.3750	<b>1.8750</b>
DC 150	3.0000	2.7500	2.3750	<b>1.8750</b>
Overall mean	3.0625	2.7187	2.3750	<b>1.8437</b>

Applying this correction we obtain  $F_F = 5.68$ . With four methods and 32 experiments,  $F_F$  is distributed according to the  $F$  distribution with 3 and 91 degrees of freedom. The critical value of  $F(3, 93)$  for 0.05 is 0.12. As the value of  $F_F$  is higher than 0.12 we can reject the null hypothesis.

Furthermore, we perform a Nemenyi test in order to check if any of these methods can be singled out [6], the Nemenyi statistic is obtained as follows:

$$CD = q_\alpha \sqrt{\frac{h(h+1)}{6U}}. \quad (18)$$

In our case, for  $k = 4$  DTW methods to compare and  $N = 32$  experiments the critical value for a 95% of confidence is  $CD_{0.95} = 2.569 \cdot \sqrt{\frac{20}{192}} = 0.8291$ . As a result non of the standard DTW methods intersect with our proposal of DTW GMM or DTW CH which is the best in mean ranking. This results are highly desirable since they supports the fact that the *proposed methodologies obtain a statistically significant improvement* in performance when compared to standard DTW approaches. For completion, we also compute the  $CD_{0.90}$  and  $CD_{0.75}$ ; results are shown in Figure 10.

These results support the fact that our proposal DTW APE is statistically better than the standard DTW approaches, obtaining very satisfying results while keeping similar computational complexity. In addition, though our contribution can be applied to any general purpose gesture recognition system, from a clinical point of view, the presented analyses were reported as relevant by physicians involved in the project and specialists on ADHD from hospitals in the area of Catalonia.



Fig. 10. Mean rank and confidence interval per method.

## V. CONCLUSIONS AND FUTURE WORK

In this paper we presented an extension of the DTW algorithm in order to handle the intra-class variability of a gesture class. This variability was encoded using one-class classifiers, such as, GMMs and APEs. In order to be able to embed these classifiers in the DTW context, the association cost was redefined to take into account the properties of such classifiers. We applied this extension in a real world problem, detecting ADHD behavioural patterns to support clinicians in diagnose purposes. In our experiments, on a novel multi-modal ADHD dataset, the proposed methodology obtained statistically significant improvements with respect to DTW techniques while obtaining relevant classification rates from a clinical point of view.

The results of this study motivate the use of the proposed techniques with a much broader set of ADHD behavioural patterns in order to provide additional information to the clinician. Moreover, the presented methodology represents a significant contribution for general purpose Human Behaviour Analysis systems.

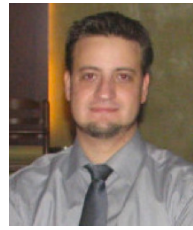
## ACKNOWLEDGEMENTS

This work is partly supported by projects IMSERSO-Ministerio de Sanidad 2011 Ref. MEDIMINDER and RECER-CAIXA 2011 Ref. REMEDI, SUR-DEC of the Generalitat de Catalunya and FSE, and TIN2013-43478-P. The work of Antonio is supported by an FPU fellowship from the Ministerio de Educacion of Spain.

## REFERENCES

- [1] A comparison of affine region detectors. *International Journal of Computer Vision*, 65:43–72, 2005.
- [2] R. Bakeman and V. Quera. *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press, 2011.
- [3] M. Bautista, A. Hernández-Vela, V. Ponce, X. Perez-Sala, X. Baró, O. Pujol, C. Angulo, and S. Escalera. Probability-based dynamic time warping for gesture recognition on rgb-d data. In *International Conference on Pattern Recognition Workshops, WDIA*. Springer, 2012.
- [4] J. Carletta. Squibs and discussions assessing agreement on classification tasks: The kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.
- [5] P. Casale, O. Pujol, and P. Radeva. Approximate convex hulls family for one-class classification. *Multiple Classifier Systems*, pages 106–115, 2011.
- [6] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, 2006.
- [7] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *ICCV*, pages 415–422, nov. 2011.
- [8] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. *SPM, IEEE*, 22(2):38–51, 2005.

- [9] J. Hashemi, T. V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro. Computer vision tools for the non-invasive assessment of autism-related behavioral markers. *arXiv preprint arXiv:1210.7014*, 2012.
- [10] A. Hernández-Vela, M. Reyes, V. Ponce, and S. Escalera. Grabcut-based human segmentation in video sequences. *Sensors*, 12(11):15376–15393, 2012.
- [11] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.
- [12] J. López-Ibor. Cie-10: Trastornos mentales y del comportamiento. Madrid: Meditor, 1992.
- [13] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.
- [14] J. Nair, U. Ehimare, B. Beitman, S. Nair, A. Lavin, et al. Clinical review: evidence-based diagnosis and treatment of adhd in children. *Missouri medicine*, 103(6):617, 2006.
- [15] A. P. A. T. F. on DSM-IV. *DSM-IV draft criteria*. Amer Psychiatric Pub Inc, 1993.
- [16] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
- [17] A. Pentland. Socially aware computation and communication. *Computer*, 38:33–40, 2005.
- [18] B. R., M. Vanrell, and R. Baldrich. A data set for fuzzy colour naming. *Color Research and Application*, 31(1):48–56, 2006.
- [19] M. Reyes, G. Dominguez, and S. Escalera. Feature weighting in dynamic time warping for gesture recognition in depth data. *ICCV*, 2011.
- [20] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, pages 309–314, 2004.
- [21] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.
- [22] M. Svensn and C. M. Bishop. Robust bayesian mixture modelling. *ESANN*, 64:235–252, 2005.
- [23] B. Yang, J. Cui, H. Zha, and H. Aghajan. Visual context based infant activity analysis. In *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*, pages 1–6. IEEE, 2012.
- [24] H.-D. Yang, S. Sclaroff, and S.-W. Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE TPAMI*, 31(7):1264–1277, 2009.
- [25] F. Zhou, F. D. la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE TPAMI*, 2010.



**Sergio Escalera** Sergio Escalera received the B.S. and M.S. degrees from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2003 and 2005, respectively. He obtained the Ph.D. degree on Multi-class visual categorization systems at Computer Vision Center, UAB. He obtained the 2008 best Thesis award on Computer Science at Universitat Autònoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at University of Barcelona. His research interests include, between others, machine learning, statistical pattern recognition, visual object recognition, and human computer interaction systems, with special interest in human pose recovery and behavior analysis.



**Laura Igual** Laura Igual received the degree in Mathematics from Universitat de Valencia in 2000. She developed her Ph.D Thesis in the program of Computer Science and Digital Communication at the Department of Technology of the Universitat Pompeu Fabra. She obtained her Ph.D Thesis in January 2006 and since then she is a research member at the Computer Vision Center (CVC) of Barcelona. Since 2009, she is a lecturer at the Department of Applied Mathematics and Analysis of the Universitat de Barcelona. She is a member of the Perceptual Computing Lab and a consolidated research group of Catalonia. Her research interests include medical imaging, with focus on neuroimaging, computer vision, machine learning, and mathematical models and variational methods for image processing.



**Oriol Pujol** Oriol Pujol Vila obtained the degree in Telecommunications Engineering in 1998 from the Universitat Politècnica de Catalunya (UPC). The same year, he joined the Computer Vision Center and the Computer Science Department at Universitat Autònoma de Barcelona (UAB). In 2004 he received the Ph.D. in Computer Science at the UAB on work in deformable models, fusion of supervised and unsupervised learning and intravascular ultrasound image analysis. In 2005 he joined the Dept. of Matemàtica Aplicada i Anàlisi at Universitat de Barcelona where he became associate professor. He is member of the BCN Perceptual Computing Lab. He has been since 2004 an active member in the organization of several activities related to image analysis, computer vision, machine learning and artificial intelligence



**Miguel Ángel Bautista** received his B. Sc. and M. Sc. degrees in Computer Science and Artificial Intelligence from Universitat de Barcelona and Universitat Politècnica de Catalunya respectively in 2010. He is a research member at Computer Vision Center at Universitat Autònoma de Barcelona, Applied Math and Analysis Dept. at Universitat de Barcelona and BCN Perceptual Computing Lab and Human Pose Recovery and Behavior Analysis Group at University of Barcelona. In 2010 Miguel Angel received the first prize from the Catalan Association of Artificial Intelligence Thesis Awards. Currently Miguel Angel is pursuing a Ph. D in Error Correcting Output Codes as a theoretical framework to treat multi-class and multi-label problems. His interests are, between others, Machine Learning, Computer Vision, Convex Optimization and its applications into Human Gesture analysis.



**Antonio Hernández-Vela** received his Bachelor degree in Computer Science and M.S. degree in Computer Vision and Artificial Intelligence at Universitat Autònoma de Barcelona (UAB) in 2009 and 2010, respectively. He is currently a research member at the Computer Vision Center (UAB) and PhD student at University of Barcelona. He is also member of the BCN Perceptual Computing Lab research group and Human Pose Recovery and Behavior Analysis Group. He is mainly interested in the application of Computer Vision and Artificial Intelligence techniques to projects that can help impaired people to improve their life quality, especially in the area of human pose recovery and behaviour analysis.



**Josep Moya** Doctor in Medicine, Psychiatry and Psychoanalyst. He is with the Mental Health Department at Parc Taulí (Barcelona), he also is the leader of the Observatory of Communitarian Mental Health of Catalonia. He is a teacher in the Department of Social Wellness and Family of the Generalitat de Catalunya, and he also is teaching in the Center for Legal Studies and Specialized Training at the Department of Justice of the Generalitat de Catalunya. He is the president of CRAPPSI (Private Catalan Foundation for Research and Evaluation of Psychoanalytic Practice). Currently he leads a research project on the Impact of the Economic Crisis on the Mental Health of the Population. He has published several articles on ADHD.



**Verónica Violant** obtained her Ph. D in Psychology from the Ramon Llull University. She is a tenured professor at University of Barcelona, currently at the Didactic and Educational Organization Department. She leads the graduate course on Pedagogics, Childhood and Disease at University of Barcelona. She is a member of the research group for Socio-educational Interventions in Childhood and Youth. Her research interests are hospital pedagogics. Concretely, in paediatrics and neonatology. She is author of various publications on the attentiveness on diseases in childhood and youth. In 2012 she was awarded with the Diamond Prize of research of the International Awards of the Eureka Sciences.



**Maria Teresa Anguera** obtained her Ph.D in Philosophy and Humanities (Psychological section) at University of Barcelona. Maria Teresa Anguera also holds a Degree in Law from University of Barcelona. Maria Teresa is a distinguished professor at the Department of Behavioral Science Methodologies at University of Barcelona since 1986. Maria Teresa has a long teaching trajectory at University of Barcelona together with several research participations at foreign universities. Maria Teresa has advised several Ph.D dissertations and has published more than 100 journal papers on psychology. She is an academic at the Spanish Royal Academy of Medicine. She has also been vice-rector of Scientific Politics at University of Barcelona. Since 2011 she is a member of the Steering Doctorate Committee.