Adaptive Body Gesture Representation for Automatic Emotion Recognition

STEFANO PIANA, ALESSANDRA STAGLIANÒ, FRANCESCA ODONE,
and ANTONIO CAMURRI, DIBRIS—Università degli Studi di Genova

We present a computational model and a system for the automated recognition of emotions starting from full-body movement. Three-dimensional motion data of full-body movements are obtained either from professional optical motion-capture systems (Qualisys) or from low-cost RGB-D sensors (Kinect and Kinect2). A number of features are then automatically extracted at different levels, from kinematics of a single joint to more global expressive features inspired by psychology and humanistic theories (e.g., contraction index, fluidity, and impulsiveness). An abstraction layer based on dictionary learning further processes these movement features to increase the model generality and to deal with intraclass variability, noise, and incomplete information characterizing emotion expression in human movement. The resulting feature vector is the input for a classifier performing real-time automatic emotion recognition based on linear support vector machines. The recognition performance of the proposed model is presented and discussed, including the tradeoff between precision of the tracking measures (we compare the Kinect RGB-D sensor and the Qualisys motion-capture system) versus dimension of the training dataset. The resulting model and system have been successfully applied in the development of serious games for helping autistic children learn to recognize and express emotions by means of their full-body movement.

Categories and Subject Descriptors: I.2.6 [**Learning**]: Knowledge Acquisition

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Body, automatic emotion recognition, motion capture, dictionary learning

ACM Reference Format:
Stefano Piana, Alessandra Staglianò, Francesca Odone, and Antonio Camurri. 2016. Adaptive body gesture representation for automatic emotion recognition. ACM Trans. Interact. Intell. Syst. 6, 1, Article 6 (March 2016), 31 pages.
DOI: http://dx.doi.org/10.1145/2818740

The automated recognition techniques described in this article are included in the Expressive Gesture Analysis library of the EyesWeb software platform (http://www.infomus.org/eyesweb_eng.php).
Recent developments and improvements on the analysis techniques of expressive movement features have been partially supported by the European Community, Horizon 2020 Programme under Grant No. 645553 (DANCE, http://dance.dibris.unige.it/).
The reviewing of this article was managed by associate editor Joseph LaViola.
The research leading to these results has received funding from the European Community, Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion, http://www.asc-inclusion.eu/).
Recent developments and improvements on the analysis techniques of expressive movement features have been partially supported by the European Community, Horizon 2020 Programme under grant agreement No. 645553 (DANCE, http://dance.dibris.unige.it/).
Authors' addresses: S. Piana and A. Camurri, DIBRIS, Univerità degli studi di Genova, Viale Causa, 13 - 16145 Genova - Italy; emails: stefano.piana@dist.unige.it, antonio.camurri@unige.it; A. Staglianò and F. Odone, DIBRIS, Univerità degli studi di Genova, Via Dodecaneso, 35 - 16146 Genova - Italy; emails: {alessandra.stagliano, francesca.odone}@unige.it.

DOI: http://dx.doi.org/10.1145/2818740

1. INTRODUCTION

Theories of emotion are generally based on facial and voice expressions. The role of full-body movement in conveying emotion and expressive gesture has been neglected and unexplored until recent years, with few exceptions in psychology (e.g., Wallbott [1998]) and in computer systems for automated analysis of expressive and affective content from full-body movement (e.g., Camurri [1995] and Camurri et al. [2003]). Recent works explain advantages of the expression and recognition of affective content by full-body gesture: For example, from a distance, bodily expressive cues are easier to perceive than subtle changes in the face [de Gelder 2009; de Gelder et al. 2010]. A recent growing amount of research in affective neuroscience demonstrates the importance of the full body in the expression and even nonconscious recognition of emotions (e.g., de Gelder and Hadjikhani [2006]). This is also evident from the emergence of multimedia technology (e.g., novel sensors) exploiting full-body movement in a growing number of applications.

Two recent surveys [Kleinsmith and Bianchi-Berthouze 2013; Karg et al. 2013] provide a detailed state of the art on the automatic recognition of affective body movements. The increasing amount of research in this field is also due to the emergence of low-cost, effective sensor technologies for full-body real-time movement analysis. One the one hand, motion-capture systems technology is improving fast and at decreasing costs. On the other hand, recent low-cost, nonintrusive game technologies (Nintendo Wii, Microsoft Kinect, and other RGB-D sensors) and sensors on mobiles (e.g., smartphones and related wearables) are enabling a growing number of successful applications involving the physical full-body participation of individuals as well as social groups. Thus, novel market niches and application areas exploiting the possibility of real-time, full-body expressive and affective analysis are emerging.

The computational framework, the experimental work, and the system that we propose in this article are the result of our research carried out within the FP7-ICT ASC-INCLUSION[1] European Project. ASC-INCLUSION aims at developing information and communications technology (ICT) solutions to assist children affected by Autism Spectrum Conditions (ASC). It focuses in particular on the development of serious games to support children to understand and express emotions. Through automatic real-time emotion recognition, our serious games aim at leading autistic children to better understand and express emotions. ASC-INCLUSION complete emotion recognition framework integrates facial expressions [Golan et al. 2006], voice [Golan et al. 2010], and full-body movements and gestures [Piana et al. 2014b]. The ultimate goal of ASC-INCLUSION is to integrate these three different modalities and help autistic children to improve their abilities in understanding and expressing emotions as well as their social integration.

In this article, we focus on a computational framework, proposing a system for the real-time automatic emotion recognition from the analysis of body movements and on serious games based on the recognition system. Such system, based on the EyesWeb XMI[2] [Camurri et al. 2007] software platform, has been adopted to develop two serious games for autistic children that will be briefly described in this article. These serious games are currently under intensive use with different institutions on autism (with high-functioning autistic children): A detailed report on the usage with this population of users will be available in an article in preparation [Piana et al. 2015].

Our research on expressive and affective gesture analysis is based on experimental psychology (e.g., de Meijer [1989], Wallbott [1998], and Boone and Cunningham

[1] http://www.asc-inclusion.eu/.

[2] http://www.infomus.org/eyesweb.

[1998]) and on humanistics, including Laban's Effort Theory [Laban and Lawrence 1947; Laban 1963]). We start from the results obtained with psychologist partners in the ASC-INCLUSION project on the individuation of relevant movement features explaining affective nonverbal behaviour. This study was based on a set of sessions where professional actors expressed different emotions under different circumstances. The acting performances have been recorded (RGB-D data) in a dataset of full-body affective behaviour [O'Reilly et al. 2014]. From the analysis of these benchmark data, we identified and developed algorithms to extract low-level and midlevel (expressive) features to capture expressive qualities of body movement. An early version of our automatic emotion recognition was presented in Piana et al. [2014b]; here we extend it to include a wider set of features and a robust and adaptive data representation, able to cope with the large intraclass variability of the considered problem. The adaptive abstraction layer is based on dictionary learning to accommodate for intraclass variability, noise, and incomplete information. The obtained representation of portions of the three-dimensional (3D) data stream is used as an input to a multiclass classifier capable to recognizing different emotions.

Besides the recordings that took place during ASC-INCLUSION [O'Reilly et al. 2014], we further acquired freely expressed full-body movements obtained by optical motion-capture systems [Qualisys 2013] and by low-cost RGB-D sensors [Kinect 2013]. These recordings appeared to be more appropriate to capture the natural intraclass variability of each considered emotion. The reason for the two types of sensors was to assess the potential of different sensing technologies in the specific application domain. The emotion datasets we built and the emotion recognition system we developed start from a given number of assumptions concerning the number and characteristics of emotions, studied by the teams of psychologist experts on autism and partners in the ASC-INCLUSION European Project. These characteristics are the main requirements for the definition of the affective dataset developed by the psychologist partners [Golan et al. 2010; O'Reilly et al. 2014].

The article is organized as follows: Section 2 gives a short review of related literature; Section 3 describes the computational framework, introducing the emotion-related dimensions, the data representation, and the classification procedure; Section 4 shows the experimental setup and the validation of the computational framework; Section 5 describes the resulting software system and briefly sketches the two serious game applications currently adopted by several autism centres. Section 7 concludes the article.

2. RELATED WORK

Over the years, research in automatic emotion recognition from implicit cues [Cowie et al. 2001] mainly focused on facial expression or voice analysis, in accordance with Ekman [1965], who pointed out how people focus more on facial expression than body gesture when they try to understand other people's emotions. However, as already pointed out in the previous section, research in experimental psychology suggests that body language constitutes a significant source of affective information [Argyle 2013; Ekman and Friesen 1974]. For example, Bull [1987] found that interest/boredom and agreement/disagreement can be associated with different body postures/movements.

Studies on the recognition of human movement from point-light display carried out by Johansson (1973) inspired further work on emotion recognition by Pollick and colleagues [Ma et al. 2006; Hill and Pollick 2000; Dittrich et al. 1996] and Atkinson et al. [2004]. These studies confirm our adoption of 3D joint information as the basic physical signals information in our approach. Pollick et al. [2001] found that, given point-light arm movements, human observers could distinguish basic emotions with an accuracy significantly above the chance level. Coulson [2004] highlighted the role of static body postures in the recognition task where artificially generated emotional-related postures where shown to people. Techniques for automated emotion recognition from full body movement were proposed by Camurri et al. [2003].

Research in experimental psychology demonstrated how some qualities of movement are related to specific emotions: for example, fear is characterized by contraction of the body, as an attempt to be self-protective and as small as possible, surprise involves turning and moving towards the object capturing our attention, joy may be characterized by openness and upward acceleration of the forearms [Boone and Cunningham 1998]. The body turning away is typical of fear and sadness; the body turning towards is typical of happiness, anger, surprise; we tend to open our arms when we are happy, angry or surprised; we can either move fast (fear, happiness, anger, surprise) or slow (sadness). de Meijer [1989] present a detailed study of how the body movements are related to the emotions. He observes the following dimensions and qualities: trunk movement: stretching–bowing; arm movement: opening–closing; vertical direction: upward–downward; sagittal direction: forward–backward; force: strong–light; velocity: fast–slow; directness: direct–indirect. de Meijer notices how various combinations of those dimensions and qualities can be found in different emotions. For instance, a joyful feeling could be characterized by a Strong force, a fast Velocity, and a direct Trajectory, but it could have a Light force as well or be an indirect (or flexible, nondirect) movement. Works inspired by these studies that involve the analysis of the movement of the entire body show that movement-related dimensions can be used in the recognition of distinct emotions [de Meijer 1989; Wallbott 1998; Camurri et al. 2003; Glowinski et al. 2011]. Kapoor et al. [2007] demonstrated a correlation between body posture and frustration in a computer-based tutoring environment. In a different study, Kapur et al. [2005] showed how four basic emotions could be automatically distinguished from simple statistical measures of motion dynamics. Balomenos et al. [2005] combined facial expressions and hand gestures for the recognition of six prototypical emotions. Recent detailed account of state of the art on bodily expressed affective content is reported in the surveys by Kleinsmith and Bianchi-Berthouze [2013] and by Karg et al. [2013].

3. THE COMPUTATIONAL FRAMEWORK

In this section, we describe our automatic emotion recognition method. Figure 1 shows our proposal of computational framework. Input data include 3D coordinates (from motion-capture systems) and 2D shape-related images. In this work, we mainly focus on 3D motion data. Given 3D data of the user, we can build a skeleton representation of her body. Each specific body joint is tracked over time, resulting in a trajectory that describes how the user moves the corresponding part of the body. Figure 2 shows the labels associated with the selected body joints. A number of different features is then extracted from the available data: Section 3.1 presents a detailed description of the extracted quantities.

As described in Section 3.2, basing on the extracted features, we segment portions of temporal streams obtaining *gesture primitives*. Then for each primitive we compute a histogram-based representation of each movement feature. A further refinement of the representation can be applied, as described in Section 3.3: An adaptive representation can be learned over the input data. This further step is optional but, as shown in Section 4, it leads to better performances as it improves the encoding of the large intraclass variability typical of the problem considered.

Finally, the representation (either histogram or dictionary based) is classified in one of N possible emotions with an N-class linear classification architecture (Section 3.4).

In this work, we consider in particular six different emotions, widely accepted and recognized as “basic” (from the body gestures point of view) and cross-cultural [Cornelius 1996, 2000]: happiness, sadness, anger, fear, disgust, and surprise. This choice is also confirmed by the needs emerged in the ASC-INCLUSION EU project

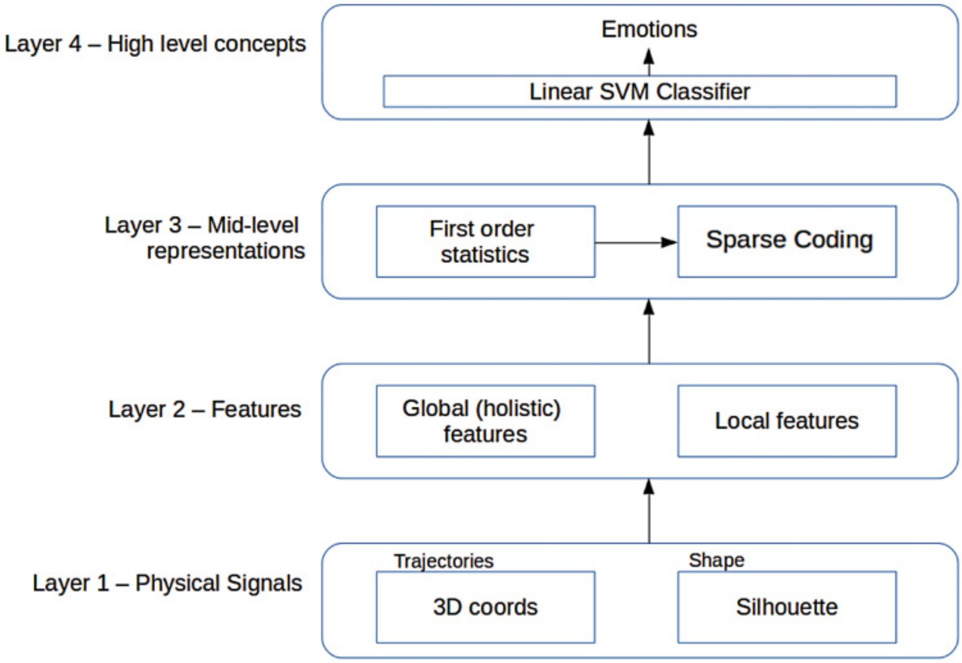

Fig. 1. The proposed Conceptual Framework, derived from Camurri et al. [2005]: 2D and 3D input data are processed to extract the relevant features (Layer 2), and then a statistical representation of such features is built (Layer 3); a further sparse coding enhances the expressive power of the representations. A feature vector is obtained by concatenating all the histograms or sparse codes. A linear SVM classifier (Layer 4) is adopted for the final emotion recognition.

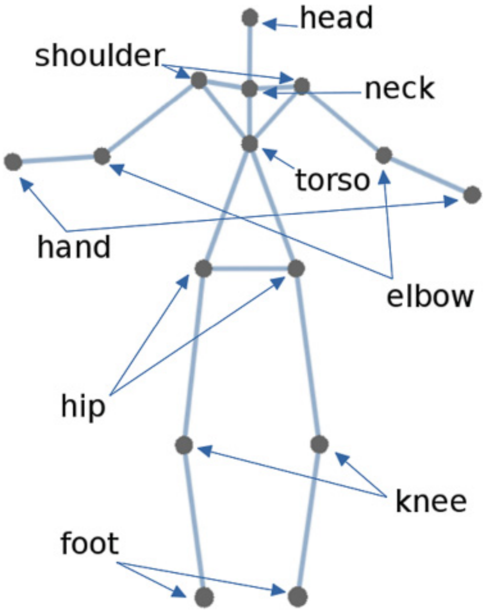

Fig. 2. 3D skeleton and associated labels.

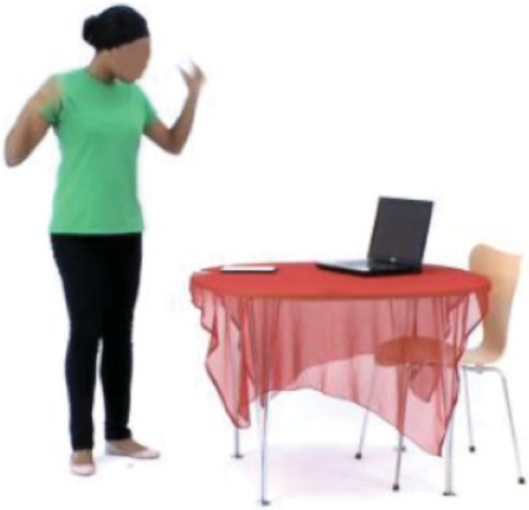

Fig. 3. Actress expressing anger with body gestures during one of the ASC-INCLUSION recording sessions.

where we aim at building a system that can teach autistic children to recognize emotions by using unique body movement. Experts on autism selected the list of emotions to be used in this task [O'Reilly et al. 2014].

3.1. Movement Features for Emotion Recognition

It is well known how the body is able to convey and communicate emotions and, in general, implicit information. It is important to understand which movement features are the most significant in discriminating the different emotions. The first part of this study has focused on this topic: identifying a set of relevant low- and midlevel movement related features to be implemented in our pipeline.

This body of work constituted one of the core achievements of the studies conducted in ASC-INCLUSION by clinical partners on the problem of teaching autistic children to recognize and express emotions: The fulfilment of such a task required audiovisual material to demonstrate these emotional displays. Furthermore, to demonstrate how these emotions might occur within a social context, a number of social scenarios portraying each emotion was needed. For this purpose, 20 actors were filmed in September 2012 [O'Reilly et al. 2014]. Short clips of actors expressing emotions were recorded. Each actor was provided with a storyboard. The storyboard just set a time span and the intensity of the emotion to be displayed and helped ensure consistency across actors. Three modalities (facial expressions, voice tones, and body gestures) were recorded and validated. The recordings were made with professional cameras, microphones, and a Kinect sensor. Figure 3 shows a frame of an example clip in which an actress was expressing anger with body gestures. The recorded clips were validated through surveys submitted to typically developing adults (a detailed description of the validation process is given in Section 4). A total of 82 body gesture stimuli passed the validation process.

The validated body gesture clips were used to verify literature and previous studies and to identify a set of relevant movement characteristic that help in the cognitive process of discriminating emotions. The result of such studies are emotion-related dimensions that help in discriminating between different expressed feelings. We analysed these dimensions and built a computational model explaining the corresponding

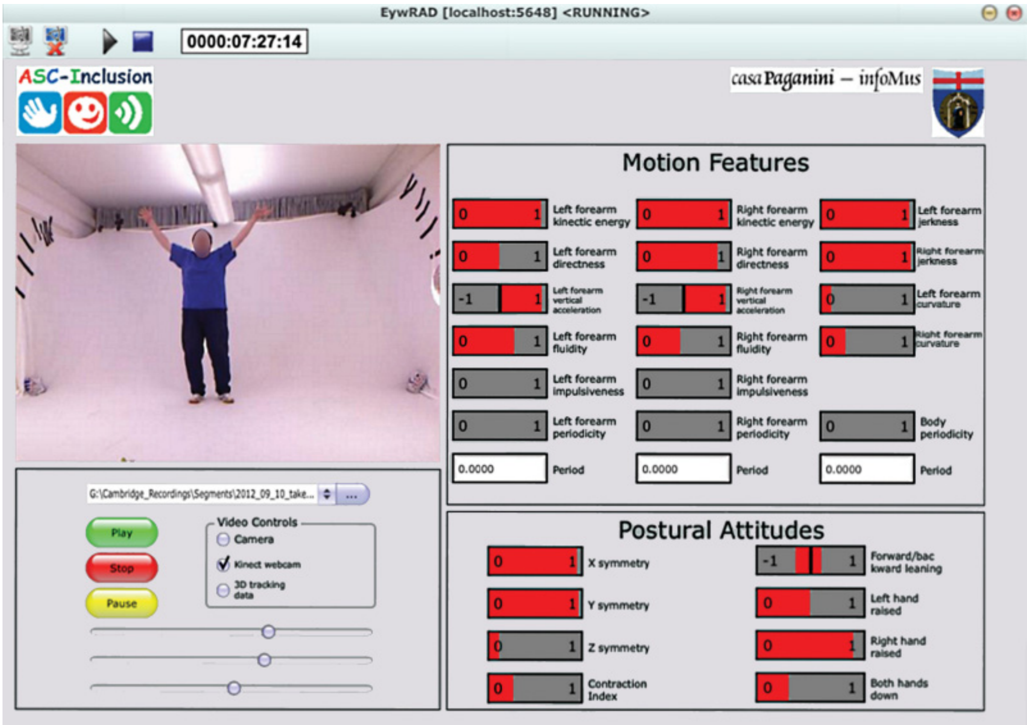

Fig. 4. EyesWeb Gesture Analyser, used to evaluate the nonverbal full-body affective dataset recorded at Cambridge University [O'Reilly et al. 2014]. The user can see the video running (on the left) and see the evolution of the motion and postural features (on the right).

movement qualities. Figure 4 shows the software system we developed in ASC-INCLUSION to evaluate different features on different types of emotions.

As a result of this analysis, we identified a set of descriptive features, also inspired by Laban's Effort Theory. Further, based on previous literature [Glowinski et al. 2011; Wallbott 1998] we mainly focus on the upper body (head, shoulders, elbows, hands, and torso joints).

A summary of the extracted features, the joints on which they are computed, and the emotional dimension to which they relate is given in Table I. A detailed description of the selected features is given the next subsections.

3.1.1. Holistic Features. We define *Holistic Features* as those quantities that are computed over a group of joints considered as a single entity or the whole body.

Kinetic Energy. The Kinetic Energy is the overall energy spent by the user during the movement, estimated as the total amount of displacement in all of the tracked joints. Given the three-dimensional position of the i-th joint at time frame f let $v_i(f) = \sqrt{\dot{x}_i^2(f) + \dot{y}_i^2(f) + \dot{z}_i^2(f)}$ denote the speed of the joint at time frame f. We then define $KE(f)$, the Kinetic Energy index, as an approximation of the body kinematic energy, the weighted sum of the kinetic energy of each joint:

$$KE(f) = \frac{1}{2}\sum_{i=1}^{n} m_i v_i^2(f). \tag{1}$$

Table I. Movement and Expressive Features

Feature	Type	Relative joint(s)	Related dimension(s)
Kinetic energy	Holistic, Kinematic	All	Force, Velocity
Contraction index (static)	Holistic, Postural	All	Trunk & Arm Movement
Contraction index (dynamic)	Holistic, Postural	All	Trunk & Arm Movement, Force
3D symmetry	Holistic, postural	All	Trunk & Arm Movement
Triangle symmetries	Holistic, Postural	(head/shoulders), (head/hands)	Trunk Movement Arm Movement
Leaning position	Holistic, Postural Local	Trunk's joints head	Trunk Movement
Leaning velocity	Holistic, Postural Local	Trunk's joints Head	Trunk Movement
Periodicity	Holistic, Postural Local	All Shoulders, Hands	Trunk Movement Arms Movement
Direction	Holistic, Kinematic Local, Kinematic	All Hands, Elbows	Trunk & Arm Movement, Vertical & Sagittal Dir. Arm Movement, Vertical & Sagittal Dir.
Velocity	Local, Kinematic	Hands, Shoulders, Elbows	Arm Movement, Force, Velocity
Acceleration	Local, Kinematic	Hands, Shoulders, Elbows	Arm Movement, Force, Velocity
Jerk	Local, Kinematic	Hands, Shoulders, Elbows	Arm Movement, Force, Velocity
Fluidity	Local, Kinematic	Hands, Shoulders	Arm Movement, Force, Velocity
Impulsiveness	Local, Kinematic	Hands	Arm Movement, Force
Curvature	Local, Trajectory	Hands, Elbows	Arm Movement, Directness
Smoothness	Local, Trajectory	Hands	Arm Movement, Directness
Distance from body	Local, Postural	Hands, Elbows	Arm Movement
Distance from head	Local, Postural	Hands	Arm Movement

The energy contribution for each joint is weighted according to its mass (m_i) estimated using anthropometric tables (see McConville et al. [1980]). This quantity is important for differentiating emotions, as stated by Camurri et al. [2003], that showed how movement activity is a relevant feature in recognizing emotion from full-body movement. The highest values of energy are related to anger, joy, and terror, and the lowest values correspond to sadness and boredom.

Direction. The direction of the joints movement can be helpful in discriminating emotions: For example, joyful movements tend to have an upward acceleration of forearms [Boone and Cunningham 1998]. Here, upwards is relative to self, the reference in this case is centred in the trunk of the person, and the upward direction is the (vertical) direction identified by the barycentre of the body and the barycentre of the head (we assume that the users mostly assume natural postures and postural attitudes). We extract the overall movement direction of the tracked joints, which we consider as an approximation of the global movement direction of the body. This feature is also considered locally, since we also extract the principal direction of the hands, shoulders, and elbows individually.

Postural Attitudes. The way a person occupies the surrounding space with her body (see Laban's *personal space* [Laban 1963; Laban and Lawrence 1947]). For instance, a scared subject will tend to assume a self-defensive posture (i.e., keeping hands near the trunk or the head), whereas an angry person will tend to perform wide movements to look aggressive. We measure these behaviours through the *Contraction Index*: Given the set of 3D joints of a person at frame f, we compute the contraction index as the ratio between the minimum parallelepiped surrounding the torso of the person BV_t and the minimum parallelepiped surrounding the whole set of joints BV_w as defined in

Equation (2):

$$CI(f) = \frac{BV_t(f)}{BV_w(f)}. \tag{2}$$

The Contraction Index is a quantity that ranges from 0 to 1 where values near 0 mean very spread postures, while values close to 1 represent very contracted postures (i.e., crouching). Looking at the variation of the CI we can compute the Dynamic Contraction Index as the

$$DCI(f) = \dot{C}I(f). \tag{3}$$

Concerning posture, we consider relevant qualities the leaning velocity and position of the trunk. They can be measured starting from the degree of inclination, forward or backward, of the trunk joints (neck and torso). Let B_t be a simplified representation of the trunk portion of the spinal cord the trunk leaning L_t at frame f can be computed as the angle between the vertical axis Y and B_t:

$$L_t(f) = \angle Y B_t(f), \tag{4}$$

Similarly to the dynamic contraction index, we can compute the variation of the leaning as:

$$DL_t(f) = \dot{L}_t(f). \tag{5}$$

The neck leaning and dynamic leaning are computed with respect to the person's trunk as follows:

$$L_n(f) = \angle B_t B_n(f), \tag{6}$$

$$DL_n(f) = D\dot{L}_n(f), \tag{7}$$

where B_n represents the cervical portion of the spinal cord.

Symmetry Related Features. Lateral symmetry has long been studied in facial expressions, resulting in valuable insights about a general hemisphere dominance in the control of emotional expression. An established example is the expressive advantage of the left hemiface that has been demonstrated with chimeric face stimuli, static pictures of emotional expressions with one side of the face replaced by the mirror image of the other. A study by Roether et al. on human gait demonstrated pronounced lateral asymmetries also in human emotional full-body movement [Roether et al. 2008]. We extract a symmetry estimation of the whole set of joints with respect to the sagittal, tranverse, and coronal body planes; it is measured evaluating the limbs' spatial symmetry with respect to the body on the tree planes. Each partial symmetry (SI_s, SI_t, SI_c) is computed from the position of the centre of mass and the left and right joints (e.g., hands shoulders, foots, knees) as described below:

$$SI_s = \frac{(x_{Li} - x_B) - (x_{Ri} - x_B)}{|x_{Li} - x_B| - |x_{Ri} - x_B|} \qquad i = 0, 1, \ldots, n, \tag{8}$$

$$SI_t = \frac{(y_{Li} - y_B) - (y_{Ri} - y_B)}{|y_{Li} - y_B| - |y_{Ri} - y_B|} \qquad i = 0, 1, \ldots, n, \tag{9}$$

$$SI_c = \frac{(z_{Li} - z_B) - (z_{Ri} - z_B)}{|z_{Li} - z_B| - |z_{Ri} - z_B|} \qquad i = 0, 1, \ldots, n, \tag{10}$$

where x_B, y_B, and z_B are the coordinates of the centre of mass; x_{Li}, y_{Li}, and z_{Li} are the coordinates of a left joint (e.g., left hand, left shoulder, left foot, etc.); and x_{Ri}, y_{Ri}, and z_{Ri} are the coordinates of the corresponding right joint (e.g., right hand, right shoulder, right foot, etc.). The three partial indexes are then combined into a normalized index that expresses the overall estimated symmetry.

We also look at symmetry indexes obtained considering the distances between the head and the shoulders and then the head and the hands as follows: Given three joints J_c, J_l, and J_r, where J_r is a joint of the right part of the body (i.e., the right hand), J_l is the corresponding left joint, and J_c is a central joint (i.e., head, neck, torso), we can compute the triangle $J_cJ_lJ_r$ (with the central joint J_c considered as its vertex) and compute a symmetry index S_t at frame f from $TS_1(f)$ and $TS_2(f)$ as follows:

$$TS_1(f) = |\overline{J_cJ_l}(f) + \overline{J_cJ_r}(f)| \times (1 - |\overline{J_cJ_l}(f) - \overline{J_cJ_r}(f)|), \tag{11}$$

where $\overline{J_cJ_l}$ and $\overline{J_cJ_l}$ represent the length of the segments connecting the left and the right joint to the central one:

$$TS_2(f) = |\overline{J'_cJ_l}(f) + \overline{J'_cJ_r}(f)| \times (1 - |\overline{J'_cJ_l}(f) - \overline{J'_cJ_r}(f)|). \tag{12}$$

$TS_2(f)$ is the component of the symmetry projected on the opposite side of the triangle where $\overline{J'_cJ_l}$ and $\overline{J'_cJ_l}$ represent the length of the segments connecting the left and the right joint to the projection of the central one on the opposite side of the $J_cJ_lJ_r$ triangle using the body planes as projection directions. The triangular symmetry $S_t(f)$ will be calculated as:

$$S_t(f) = \frac{TS_1(f) + TS_2(f)}{\overline{J_cJ_lJ_r}(f)}, \tag{13}$$

where $\overline{J_cJ_lJ_r}$ represents perimeter of the triangle $J_cJ_lJ_r$.

Periodicity. Periodicity is a concept related to how often a certain movement is repeated in a sequence. Our estimation of the periodicity of a movement is based on the *Periodicity Transform* [Sethares and Staley 1999]. We evaluate the periodicity of the centre of mass of the body with respect to the sagittal, coronal, and transverse planes of the whole body.

The Periodicity Transform looks for the best periodic characterization of the sequence x of length N. The underlying technique is to project x onto some periodic subspace P_p. This periodicity is then removed from x leaving the residual r stripped of its p-periodicities. A sequence of real numbers is called p-periodic if there is an integer p with $x(k+p) = x$ for all k integers. In our case the sequence x contains the coordinates of the centre of mass of the body.

3.1.2. Local Features. Local features can be computed over a single joint varying over time. The same local feature can be extracted from different joints, as summarized in Table I.

Speed Related Features. We extract the Velocity of a joint and its derivatives (Acceleration and Jerk) in different cases, because the expressiveness of a movement is influenced by the speed and its variation. Low velocity is generally related to sadness, while high velocity is related to happiness, anger, and fear. In particular, we extract such information from head, hands, shoulders, and elbows.

Trajectory Related Features. Each joint describes a trajectory in space over time, and the qualities of such trajectories are important to discriminate among emotions. For instance, smooth trajectories are related to a calm state of mind, while the fluidity is low if we are afraid or angry.

The curvature $ki(f)$ of the trajectory of a joint i is a quantity related to the trajectory's smoothness, and the curvature is computed at time frame f as follows:

$$k_i(f) = \frac{\dot{r}_i(f) \times \ddot{r}_i(f)}{\dot{r}_i^3(f)}, \tag{14}$$

where $\dot{r}_i(f)$ is the velocity of the trajectory of the i-th point and $\ddot{r}_i(f)$ is its acceleration.

We define our algorithm for computing smoothness by taking inspiration from Todorov and Jordan [1998], that is, we compute correlation between trajectory curvature and velocity. We consider the Pearson correlation coefficient for two variables: $log(k_i)$ and $log(v_i)$:

$$\rho_i(k_i, v_i) = \frac{\sigma_{log(k_i),log(v_i)}}{\sigma_{log(k_i)}\sigma_{log(v_i)}}. \tag{15}$$

k_i and v_i are computed over a "short" time window, in such a way that covariance $\sigma_{log(k_i),log(v_i)}$ can be approximated with 1, as the k_i and v_i variate (or not) approximately at the same time:

$$\rho_i'(k_i, v_i) = \frac{1}{\sigma_{log(k_i)}\sigma_{log(v_i)}}, \tag{16}$$

where $\rho_i'(k_i, v_i)$ represents the Smoothness Index SmI_i for joint i.

Theories on human motion and trajectory planning [Hoff 1992; Hogan 1984; Flash and Hogan 1985] showed that trajectories of human limbs can be modelled by the *minimum jerk law*, and, inspired by these theories, we define the fluidity index FI_i of the trajectory of joint i as:

$$FI_i = \frac{1}{\int (j_i + 1)dt}, \tag{17}$$

where j_i represents the *jerk* of joint i

We extract the smoothness from the trajectories of hands, curvature from hands and elbows, and fluidity from hand, elbows, and shoulders.

Postural Attitudes. Considering the distance between the different parts of the body as local postural-related features, we compute the following quantities: the distance between hands, hands-head distance, hands-torso distance, and elbows-torso distance. These indexes are easily computed starting from the streams of 3D points. We also take as local features the leaning velocity and position of the head joint, which as in the case of global leaning, are computed starting from the degree of inclination (forward or backward) of the head.

Impulsiveness. In human motion analysis, impulsiveness can be defined as a temporal perturbation of a regime motion (Heiser et al. [2004]). Impulsiveness refers to the physical concept of impulse as a variation of the momentum. From psychological studies [Evenden 1999; Nagoshi et al. 2006] an impulsive gesture lacks premeditation, that is, it is performed without a significant preparation phase. We extract the impulsiveness of hands as follows:

ALGORITHM 1: Impulsiveness

```
let $\delta_t = 0.45$ sec;
let energyThreshold = 0.02;
**if** $KE >$ energyThreshold **then**
    **if** $0 \le dt \le \delta_t$ **then**
        evaluate energy peaks and movement direction;
        **if** the energy peak is solitary in the given direction **then**
            $II = \frac{\bar{KE}}{dt}$
        **end if**
    **end if**
**end if**
```

The algorithm is derived from Mazzarino and Mancini [2009], where a gesture is considered an impulse if it is characterized by a short duration and high magnitude.

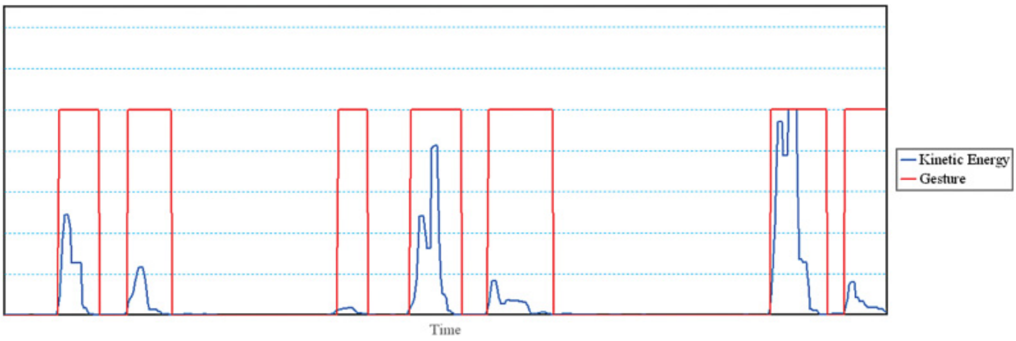

Fig. 5. Kinetic energy-based segmentation of gestures primitives.

In addition to the observations made by Mancini and Mazzarino, we included the condition that, to be recognized as impulsive, a gesture has to be performed without preparation, that is, without movements of antagonist muscles in the opposite direction of the intended movement. This includes sudden change of the movement direction and intensity. We consider dt as the gesture duration, KE is the kinetic energy, and $\bar{KE}$ is the mean of the kinetic energy computed over the gesture duration. The values of the proposed thresholds have been empirically estimated through perception tests on videos portraying people who performed highly and lowly impulsive gestures.

Periodicity. As for the holistic periodicity feature, the *local periodicity* is related to the repetition of a given movement. The local periodicity is extracted from the hands joints with respect to the three body planes, while we extract only the component related to the transverse plane from the shoulders.

3.2. Data Representation

We now discuss the operations we made on the extracted features in order to have a meaningful representation of each gesture. Raw data are composed by relatively long sequences including repetitions of a certain emotion.

The body joints are tracked along the entire sequence and the resulting trajectories are segmented into a set of single gestures (or primitives, clips, or segments) by applying a threshold over some features. In our experiments we observed that we obtain a good segmentation by thresholding the kinetic energy. Figure 5 shows the kinetic energy of a sequence of seven different gesture primitives. Basing the segmentation on low values of the kinetic energy it is possible to separate each single gesture correctly.

We now have N primitive gestures $\mathbf{x}_i \in \mathbb{R}^{f \times T_i}$, for $i = 1, \ldots, N$, where f is the number of features (in particular, $f = 86$) and T_i is the temporal length of the i-th gesture. In general, such time series will have variable lengths, depending on the different duration, speed, and complexity of the gesture. At the same time, from the machine learning stand point, we usually assume that points are feature vectors living in a fixed d-dimensional space. This is an issue common to problems that deal with temporal data and many solutions have been proposed. It is common practice to map initial representations on an intermediate (fixed-length) representation [Noceti et al. 2011; Noceti and Odone 2012]. There are different ways of doing so, for example, sub-sampling variable length series with a fixed number of samples (but in this case we may miss important information), clipping a time series into a subset of fixed-length subsequences (and in this case the size of the subsequences will be a crucial parameter), and zero-padding of the shorter sequences (in this case there is the possibility to alter the information contained in the sequence, and the maximum length of the sequences would be crucial again). Alternatively, one may resort to alternative descriptions, such as first- or second-order statistics. In this work, after a preliminary experimental analysis, we rely on first-order statistics and compute histograms of each time series. Once an appropriate quantization is chosen, each temporal sequence $\mathbf{x}_i$ will be represented by a histogram of a fixed length b. In this way we lose temporal information in favour of a compact, easy-to-compute, fixed-length representation.

Figure 6 confirms that the histogram-based representation still carries important information characterizing emotions: The figure reports cumulative histograms of the Contraction Index of 100 gesture primitives for the six basic emotions. The histograms show different shapes for different emotions, also suggesting that the Contraction Index is significant to discriminate emotions.

The overall first-order statistics is described by the concatenation of all the f histograms in a feature vector $\mathbf{h} \in \mathbb{R}^{fb}$.

3.3. Learning Adaptive Data Representations to Describe Emotions

In this section, we propose a novel way of learning adaptive data representations for emotion recognition, starting from the basic representations provided above. This processing layer is motivated by the observation that automatic emotion recognition is a difficult problem, mainly because the same emotion can be expressed in very different ways (large intra class variability), because of the amount of noise and redundant information within the available data, and because of the limited amount of data available. Starting from these observation we consider an alternative data representation to the one presented in Section 3.2. This new *adaptive* representation, based on dictionary learning, allows us to couple prior information available from psychological studies with the intrinsic information derived by data.

The idea behind sparse and adaptive dictionary learning is to represent a certain family of signals, in our case histograms of emotion related features, as linear combination of a few elements selected from a dictionary of basic signals, called *atoms*. Both the atoms and the coefficients of the linear combinations are learned from the input data.

3.3.1. Sparse Data Representations. Sparse Dictionary Learning (DL) aims at building an adaptive data representation by decomposing each datum into a linear combination of *a few* components selected from a dictionary of basic elements or *atoms*. The adaptiveness of the dictionary is achieved through the process of learning the atoms directly from the input data instead of using a fixed dictionary derived analytically such as Wavelets.

From the computational standpoint, Dictionary Learning methods attempt to solve an optimization problem whose objective function is based on the reconstruction error obtained by reconstructing the training data as a linear combination of the unknown atoms. Sparsity is induced by adding a prior on the decomposition coefficients. In general, most of the DL methods follow the schema of minimizing the cost function in two steps, the first fixing the dictionary and finding the coefficients and the second using these coefficients to find the optimal dictionary. Different formulations can be found, for instance, in the work of Chen and Maggioni [2010] and Allard et al. [2012].

We now describe how we employ dictionary learning to our purposes. For a more general account on the dictionary learning theory, we refer the interested reader to Rubinstein et al. [2010].

Given n signals, $\{h_1, \ldots, h_n\} \in \mathbb{R}^d$, we aim at finding a dictionary of K atoms $\{d_1, \ldots, d_K\} \in \mathbb{R}^d$ such that h_i^m, the m-th component of the vector h_i, is given approximately by

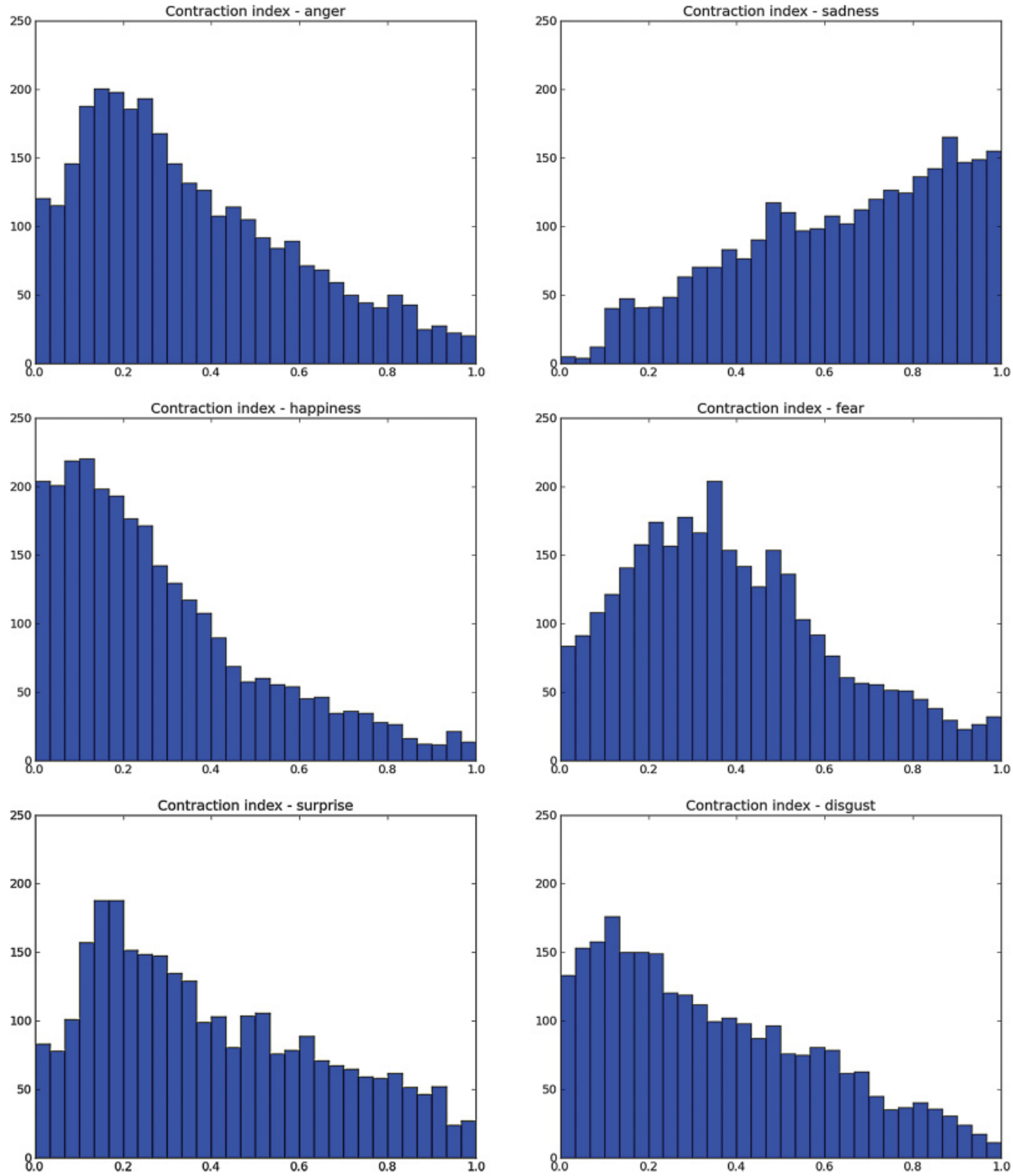

Fig. 6. Thirty bins of cumulative histograms of the Contraction Index produced from the dataset recordings of adults. The histogram peaks show the most recurrent values assumed by the CI for each emotion: CI tends to assume low values (body posture expanded) in anger (a), happiness (b), surprise (c), and disgust (f), midvalues in fear (e), and high values (body posture contracted) in sadness (d).

$$h_i^m \approx \sum_{j=1}^{K} u_i^j d_j^m \tag{18}$$

with only a few nonnull u_j^i for each h_i and where the superscripts run through rows and subscripts through columns. Thus, $u_i = (u_i^1, \ldots, u_i^K)^\top$ is the coefficient vector for the histogram h_i corresponding to the i-th row of the $K \times n$ *code* matrix U. In a more convenient matrix notation, Equation (18) can thus be rewritten as $H \approx DU$ with H denoting the $d \times n$ *data* matrix and D the $d \times K$ *dictionary* matrix.

The idea is to learn both D and U from the available examples by solving the optimization problem

$$\min_{D,U} \|H - DU\|_F^2 + \tau\|U\|_1 \tag{19}$$

for $\tau > 0$ and subject to the normalization constraints $\|d_j\|_2 \leq 1$ with $j = 1, \ldots, K$ for each atom d_j. In Equation (19), $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius and the ℓ_1-norm, respectively.

This problem has been proposed in Lee et al. [2006], and in the literature it is referred to as ℓ_1 dictionary learning. The first term in the summation is a data-fidelity term that ensures a small average reconstruction error. The second term is a penalty term inducing sparsity on the coefficients. The parameter τ determines the tradeoff between the two terms, while the number of atoms K is fixed *a priori*.

The joint minimization in Equation (19) is nonconvex and nondifferentiable, but both the subproblems obtained by keeping U or D fixed and minimizing with respect to D or U are convex. Several block coordinate descent schemes have thus been proposed; see Lee et al. [2006] and Mairal et al. [2010], for example. These schemes proceed by alternate minimization according to the following steps:

(i) initialize D and U;
(ii) solve problem in Equation (19) for fixed D;
(iii) solve problem in Equation (19) for fixed U;
(iv) go to step (ii) until convergence.

In this work, we use PADDLE,[3] an optimization algorithm proposed in Basso et al. [2011] and based on first-order proximal methods, for both steps (ii) and (iii). The overall scheme stops either upon reaching the maximum number of iterations or if the functional value changes by less than a fixed tolerance.

The dictionary matrix D is initialized by randomly picking K columns from the data matrix H in the first loop of alternate optimization between D and U.

Starting from the histograms, we apply a dictionary learning step to smooth the representations and capture the most representative patterns.

In this article, we consider both the classical ℓ_1 dictionary learning and its variant based on estimating the encoding matrix $\mathbf{C}$. In this case, the minimization problem to be solved becomes

$$\min_{D,U} \|H - DU\|_F^2 + \tau\|U\|_1 + \eta\|U - CH\|_F^2. \tag{20}$$

Thus, apart from $\mathbf{D}$ and $\mathbf{U}$, we will compute also a matrix $\mathbf{C} \in \mathbb{R}^{K\times d}$ such that $\mathbf{U} \approx \mathbf{CX}$. The encoding matrix $\mathbf{C}$ can be seen as a bank of filters that, when applied to the input data $\mathbf{H}$, projects it in the space of the coefficients. Thus, once a dictionary is learned, instead of minimizing the functional with respect to $\mathbf{u}$ each time we get a new datum, it will be sufficient to apply the matrix $\mathbf{C}$ to $\mathbf{h}$. This approach may be very beneficial from the computational point of view, concerning the decomposition of the data with respect to a fixed dictionary, learned previously, and is recommended for the real-time performances required by the interactive games.

[3] The open-source Python implementation is freely available at http://slipguru.disi.unige.it/research/PADDLE, Retrieved October 12, 2011.

3.3.2. Different Choices for the Input Data. So far we have considered a generic choice of the input data $h_i \in \mathbb{R}^d$, and now we discuss various possibilities. We start off from the data representation summarized in Section 3.2 and derive an adaptive representation with the goal of capturing the data nuances. Since the initial representation is rather heterogeneous, we consider different possible sets of features, summarized as follows:

- a dictionary learned over the data matrix composed by the whole vectors $\mathbf{h}$ obtained concatenating the histograms of all the features together (ALL)
- grouping together the features related to a certain part of the body: head, torso, shoulders, hands, elbows, all (G1), and learning a dictionary per group
- grouping together features related to a certain quality of the movement: velocity, position, qualities, corners, symmetries (G2), and learning a dictionary per group
- processing all the features independently and learning one different dictionary for each of them (1DPF)

The final representation step is to build the complete feature vector concatenating each subvector obtained representing suitably each feature or group of features. Let $\mathbf{v}_{i_j}$ be the representation of the j-th feature of the i-th gesture. The gesture $\mathbf{x}_i$ will be represented by the feature vector $\mathbf{v}_i = [\mathbf{v}_{i_1}|\mathbf{v}_{i_2}|\ldots|\mathbf{v}_{i_f}]$.

Notice how, thanks to the sparsity constraint, this final representation will include only a subset of atoms to approximate accurately each input datum. In any case, these atoms are not directly related to the initial features, as they derive from the intermediate histogram based representation. For this reason it is not possible to associate them any clear semantic meaning, such as an understanding of what are the most important features for the problem.

3.4. Classification

In this phase, we assume we have at disposal a training set of input and output elements (example primitives and the associated labels) from which we learn a relationship which will hopefully generalize on yet to be seen data. We first represent our data following Section 3.2 or Section 3.3, obtaining a training set of $(\mathbf{v}_i, y_i)$, where y_i encodes one of M possible classes (emotions).

Each basic emotion represent one of six classes, and we want to classify correctly the gestures. Assuming to have a labelled training set, we can train a multiclass classifier, building a model that is sample based. We need this model to have a generalization of the properties of each single class, so, basing its answers on this model, our framework will be able to classify new unlabelled data, performing the automatic emotion recognition.

We base our classification step on a one-versus-one Support Vector Machine (SVM) [Vapnik 1998], followed by an Error Correcting Output Coding (ECOC) [Klautau et al. 2002], that is, building a matrix that models the correlation between the different classes. ECOC allows us to attenuate the effect of having a severely nonseparable problem and improves the overall classification results from 3 to 5%.

4. METHOD ASSESSMENT

After a review of the available (full-body, 3D, nonverbal) emotion datasets [Kleinsmith and Bianchi-Berthouze 2013; Karg et al. 2013], given our main focus on autism rehabilitation and related psychological requirements, we decided to build a new corpus of recordings.

4.1. The Data

The experimental analysis we provide in this article is based on a dataset of people (nonactors) freely expressing emotions with their body: Starting from the protocols of the recordings collected with professional actors [O'Reilly et al. 2014], we recorded

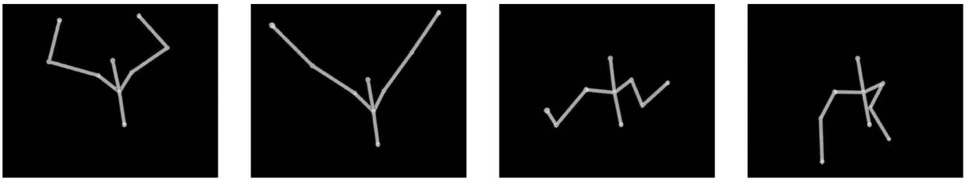

Fig. 7. An example of the stimuli used for data validation with humans: Four frames of a sequence of a happy person.

an in-house dataset of expressive gestures with two different systems: the Qualisys motion-capture system [Qualisys 2013] and a Microsoft Kinect [Kinect 2013]. In both cases, we recorded sequences of 3D coordinates (or 3D skeletons), corresponding to the same body joints.

Qualisys is a sophisticated motion-capture system, composed of many nine in our setup) infrared cameras that capture the light reflected by markers. The data it provides are precise and reliable (its spatial measurement error is under 1mm), its frame rate is 100fps, but it also has many drawbacks: It is very expensive and it requires a complex calibration procedure and a large space (and, for these reasons, it has limited portability). Finally, it requires the user to wear markers, making it inappropriate for users with Autism Spectrum conditions. Instead, Microsoft Kinect, as a commercial product, is very easy to acquire and to install, and it has very few requirements on the acquisition environment, but provides noisier and lower-resolution data, since its frame rate is only 30fps and its spatial measurement error at the optimal distance from the sensor (2m) is in the order of 1cm.

The two acquisition systems differ considerably and are indeed meant for distinct purposes: Qualisys has been used at an early stage of development, as a proof of concept of the proposed methodology, and Kinect has been adopted later to test the applicability of the method in real-world applications and serious games for autistic children.

The two datasets include 12 participants, 4 females and 8 males, aged between 24 and 60, expressing the six basic emotions with their body: To help them in the task they were presented with a short story related to the target emotion. Each participant repeated the expression of the same emotion for a number of times ranging from 3 to 7. We obtained a total of about 100 videos with the Kinect and 70 videos with Qualisys, which have been segmented according to the Kinetic Energy values to separate segments of expressive gestures. Each segment/datum has been manually associated with a label, obtained by the actor stating what type of emotion he or she was expressing. Overall, we obtain two datasets of 310 segments for six emotions with Qyalisys and 579 segments for six emotions with Kinect.

4.2. Human Data Validation

To evaluate the degree of difficulty of the task, the obtained segments have been validated by humans (60 people who had not participated in the data acquisition process) through an on-line survey.

Each participant was shown 10 video segments displaying the temporal evolution of 3D skeletons: For each video, the participant was asked to guess which emotion was being expressed among the six alternatives or to choose “I don’t know.” A gesture segment was associated to a specific emotion label and considered valid if at least 60% of the answers from all participants associated with such segment were coherent. A simplified version of the survey was submitted in a subsequent moment to the same participants: In this case, the emotion set was reduced to four emotions: Happiness, Sadness, Anger, and Fear. Figure 7 shows an example of the input stimulus, where the limited amount of information conveyed by the data is clear. Indeed, the goal

Table II. Human Validation Results

	Six emotions	Four emotions
Happiness	81.3%	87.5%
Fear	48.5%	81.2%
Disgust	37.2%	—
Sadness	86.7%	94:9%
Anger	73.9%	82.0%
Surprise	35.2%	—
Average	61.9%	85.2%

First column: emotion label. Second column: Correct recognition rate with six emotions. Third column: Correct recognition rate with four emotions.

Table III. Confusion Matrix of the Human Validation Results for the Four-Classes Problem

	Unknown	Happiness	Fear	Sadness	Anger
Happiness	4.9%	**87.5%**	3.5%	2.9%	1.1%
Fear	4.2%	5.1%	**81.2%**	8.6%	0.9%
Sadness	2.1%	0.7%	0.9%	**94.9%**	1.4%
Anger	5.3%	7.2%	3.3%	2.1%	**82.0%**

Table IV. Confusion Matrix of the Human Validation Results for the Six Classes Problem

	Unknown	Happiness	Fear	Disgust	Sadness	Anger	Surprise
Happiness	5.7%	**81.3%**	2.4%	1.9%	0.9%	4.8%	2.9%
Fear	6.3%	7.2%	**48.5%**	13.6%	0.9%	8.1%	15.4%
Disgust	4.8%	5.1%	10.9%	**37.2%**	15.8%	13.7%	12.4%
Sadness	3.8%	0.9%	1.6%	2.8%	**86.7%**	1.5%	2.7%
Anger	6.6%	8.2%	3.3%	4.9%	2.3%	**73.9%**	0.8%
Surprise	9.5%	7.9%	8.3%	11.0%	10.2%	17.7%	**35.2%**

of this experiment was to understand to what extent a human observer is able to understand emotions simply by looking at the body movements. The sole 3D skeleton is a guarantee that the user is not exploiting other information, such as insights from facial expressions or the context.

The results of the human validation, with respect to the data ground truth, are reported in Table II and clearly testify to the difficulty of the recognition problem we are addressing. The human validation showed that three of the six basic emotions (happiness, sadness, and anger) are clearly recognizable from body movements, while the other three (surprise, disgust, and fear), as shown in Table IV, are easily confused with one another. For this reason, a subproblem of four classes (happiness, sadness, anger, and fear) has also been taken into account: Table III shows that in the simplified problem the ability of people in recognizing the emotions greatly increases, especially for the fear emotion that was often confused with disgust or surprise in the six-emotions problem.

4.3. Automatic Emotion Recognition Results

We now evaluate the appropriateness of our framework (discussed in Section 3) with and without the adaptive encoding proposed in Section 3.3.

Our analysis is carried out considering an experimental protocol which exploits the available dataset in two different ways:

—**Random Split (RS)**: The dataset is split randomly in a training and a test set of equal size. To provide statistical significance to the obtained results, the procedure has been repeated 50 times; averages over all the repetitions are reported.

Table V. Comparison of the Early Results Obtained with Data Captured with Qualisys and Kinect, Represented with the Whole Vector **h**

Qualisys			Kinect		
	Four classes	Six classes		Four classes	Six classes
RS	79.2 ± 0.226	62.3 ± 0.253	RS	82.0 ± 0.135	68.5 ± 0.185
LOSO	74.6 ± 0.270	54.2 ± 0.291	LOSO	78.7 ± 0.225	61.6 ± 0.231

—**Leave One Subject Out (LOSO)**: In this case the classifiers are trained and tuned over $N-1$ (11 in our case) subjects and tested over the subject left out. The procedure is repeated leaving all subjects out and then averaging the N outcomes.

The two procedures have a different meaning: RS provides us an insight into how the system performances trend as we increase the number of subjects in the training set, and LOSO allows us to discuss the system ability in generalizing with respect to new subjects.

The Impact of a Specific Acquisition Device. We first compare how a specific acquisition device may influence the obtained results. To this purpose we perform two different experiments, one on the Qualisys and one on the Kinect. Qualisys provides very precise 3D skeletons, with the tradeoff of a rather slow and cumbersome acquisition procedure. Conversely, Kinect provides noisier results, more easily affected by environmental conditions, but the acquisition process is simple and straightforward.

Table V shows the preliminary results obtained using **h** as feature vectors on Qualisys and Kinect data, both on a six-emotions and four-emotions problem (in the latter case we restrict ourselves to a dataset of 213 segments for Qualisys and 398 for Kinect). First, we notice a consistent significant improvement if we discard the more ambiguous emotions. The four-emotion recognition problem reports higher performances than its six-emotions counterpart, with an increase of more 10%. This confirms the intuition we obtained on the human data validation: Some emotions are more clearly expressed by body movements than others.

Also we notice that, despite the initial data quality, Kinect data produce consistently better results than Qualisys data. This allows us to draw two important conclusions: The amount of noise of input data does not affect the system performances, while the size of the training set does. Indeed, we argue the reason for the lower performances of Qualisys data is the fact the dataset was smaller. This speaks in favour of adopting a simpler acquisition procedure, as the Kinect one, to obtain a larger set of data should they be needed.

Comparing the results, we get better results on the Kinect data, despite they fact that they are less reliable, because of the dimension of the dataset. A small dataset is not enough to capture the high variability of the data typical of the problem we are addressing. The results we show in the remainder of the section are obtained on the bigger Kinect dataset.

Parameters Tuning for Adaptive Data Representations. Before we start comparing different representations, let us briefly discuss parameter tuning. In the case of classical sparse coding via Equation (19), the parameter to choose is τ, which is related to the amount of sparsity of the desired encoding. The parameter choice is not particularly crucial for the obtained reconstruction, as it affects the obtained results only very mildly. This observation is confirmed by the results reported in Table VI (four-emotions recognition, RS protocol): We notice a very limited variability of the results as τ changes in its range $[0, 1]$. In the case of sparse coding with encoding matrix estimation (Equation (20)), we also need to tune parameter η. This parameter regulates

Table VI. The Effect of Different Parameters Choices on the Dictionary-Based Representations (Four Emotions Recognition, RS Protocol)

τ	$\eta = 0$	$\eta = 0.7$
0.1	77.4 ± 0.141	83.6 ± 0.103
0.3	76.6 ± 0.134	82.3 ± 0.097
0.5	77.4 ± 0.137	82.3 ± 0.102
0.7	77.4 ± 0.138	81.9 ± 0.112
0.9	78.6 ± 0.134	82.0 ± 0.107

how much the approximation $\mathbf{Ch}$ affects the functional, producing a smoothing of the representations.

Table VI also compares the accuracy obtained by using the coding vector $\mathbf{u}$ as a representation, obtained by minimizing Equation (20) with $\eta = 0$ (thus without considering the $\mathbf{C}$ matrix in the functional) and with $\eta = 0.7$, with τ ranging between 0.1 and 0.9.

It is clear that the classification results are positively influenced by the inclusion of the term depending on $\mathbf{C}$. We experimentally observe how the choice of η (provided it differs from zero) does not influence significantly the obtained performances. From now on we consider $\eta = 0.7$ as a best-performing value over an appropriately chosen validation set.

We conclude this section by observing that a further element to consider is the size of the dictionary K. Different K may lead to different results but, more generally, affect the computational performances of the method (smaller representations means fewer computations). Thus, small K will be favoured. Experiments in the reminder of the section will report results for different small values of K.

Comparing Different Representations. We now compare the performances obtained by changing the data representation. We consider in particular the global histogram $\mathbf{h}$, the coding vector $\mathbf{u}$ obtained by the minimization procedure (19), or the coding obtained by computing directly $\mathbf{Ch}$ (minimizing Equation (20)). Figure 8 considers the RS protocol, $\eta = 0.7$, τ span between 0.1 and 0.9, and $K = \{20, 30, 40\}$. We notice an overall improvement of the classification accuracy with adaptive representations. The improvement is significant and consistent on the four-emotions recognition problem, in some cases leading to a performance boost of about 3%. In the six-emotions recognition problem (which is a harder classification problem) the effect of the parameter choice is more meaningful; in any case, for an appropriate choice of parameter τ, we obtain a significant improvement in the results, regardless of K. As for $\mathbf{u}$s and $\mathbf{Ch}$s, in general, $\mathbf{u}$ appears to be more appropriate, with a small performance gain (but a high computational cost loss, since the former vector is computed minimizing an iterative problem, while to compute the latter only a matrix-vector multiplication is needed).

Similarly, Figure 9 shows a comparison between the classification results obtained following the LOSO protocol. In this case, the superiority of adaptive representations is more consistent. In gereral, $\mathbf{u}$s look more stable with respect to the different parameters. Tables VII and VIII summarize the results obtained by the best combination of parameters for the RS and LOSO protocols, respectively.

Here again, Tables VII and VIII present the best classification percentage obtained using the different data representations. In this case, adaptive representations appear to be consistently superior.

Generalization Performances on Different Groups of Features. In this section, we analyse the appropriateness of keeping all features together before learning an overall

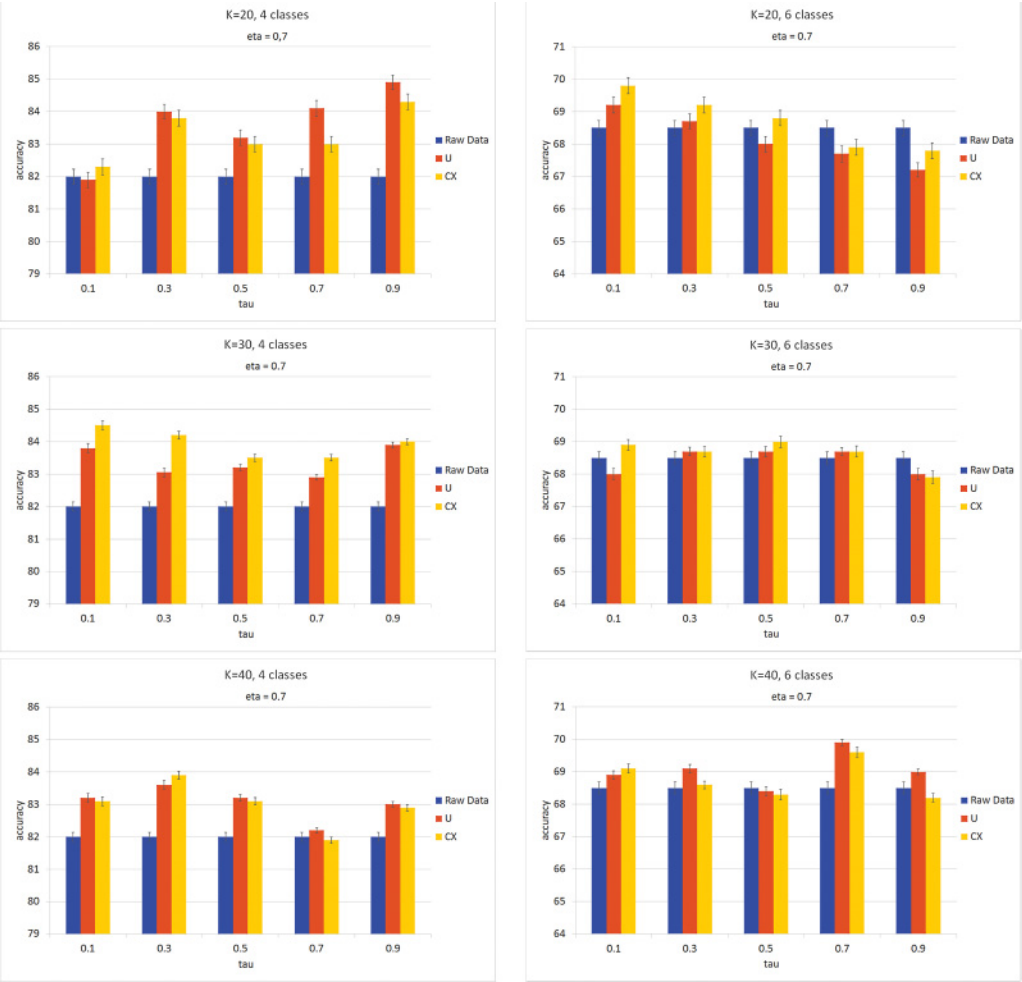

Fig. 8. RS: Comparison of the classification results obtained with histograms (blue bars), **U**s (red bars), and **Ch**s (yellow bars) for the four-classes (left) and six-classes (right) problems.

dictionary for all features or grouping features according to some principle, obtaining different dictionaries and different partial representations (see Section 3.3).

The experiments are carried out on a new test set of 66 segments, extracted automatically from data collected with the help of four new participants, two males and two females. Therefore, we are also evaluating the generalization abilities of our method both on new users and new acquisition settings (this new acquisition campaign has been carried out about one year after the first one. Thus, we fix the parameters as the best-performing ones in the previous section.

Table X shows classification results obtained on a six-classes problem that considers the six emotions. In the case of G1, we obtain six dictionaries, and thus we concatenate six different coding vectors, and in the case of G2 five dictionaries. Adaptive representations applied to this problem show comparable results to the raw representation, and the best performances, in this case, are reached using GRP2 representation where single features are grouped based on their quality.

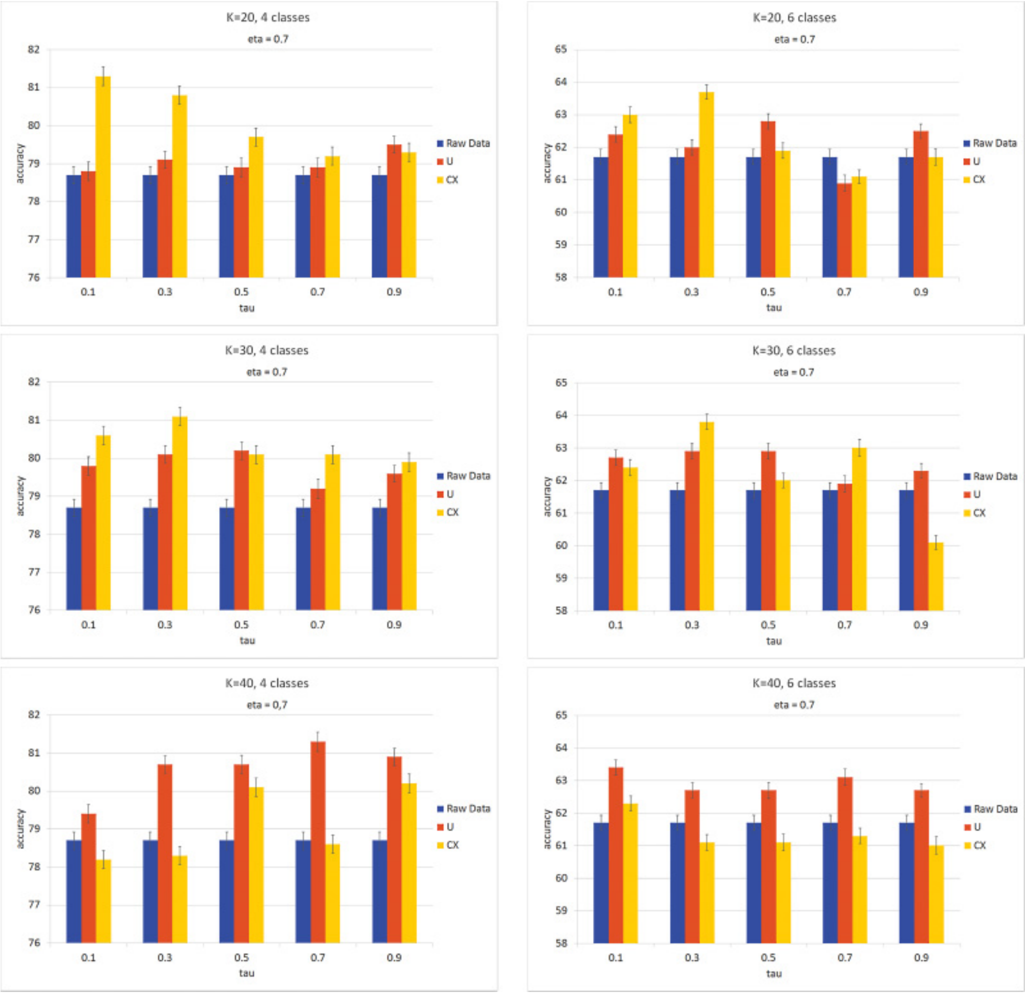

Fig. 9. LOSO: Comparison of the classification results obtained with histograms (blue bars), **U**s (red bars), and **Ch**s (yellow bars) for the four-classes (left) and six-classes (right) problems.

Table VII. RS: Summary of the Best Classification Rates

	Four classes	Six classes
Raw Data	82.0 ± 0.135	68.5 ± 0.185
U	84.9 ± 0.084	69.9 ± 0.090
Ch	84.5 ± 0.140	69.8 ± 0.191

Table VIII. LOSO: Summary of the Best Classification Rates

	Four classes	Six classes
Raw data	78.7 ± 0.225	61.6 ± 0.231
U	81.3 ± 0.255	63.4 ± 0.230
Ch	81.3 ± 0.247	63.7 ± 0.247

Table IX. 4 Emotions Problem. Generalization Performances (as a Reference Raw Histograms Provide 79.5% ± 0.147 Recognition Rate)

	1DPF	ALL	G1	G2
U	85.6 ± 0.137	79.5 ± 0.123	76.6 ± 0.132	80.2 ± 0.106
CX	85.2 ± 0.129	80.3 ± 0.131	79.7 ± 0.125	78.2 ± 0.119

Table X. 6 Emotions Problem. Generalization Performances (as a Reference Raw Histograms Provide 69.3% ± 0.233 Recognition Rate)

	1DPF	ALL	G1	G2
U	67.9 ± 0.226	64.6 ± 0.219	70.6 ± 0.231	70.8 ± 0.228
CX	68.8 ± 0.217	63.7 ± 0.213	62.2 ± 0.221	66.3 ± 0.219

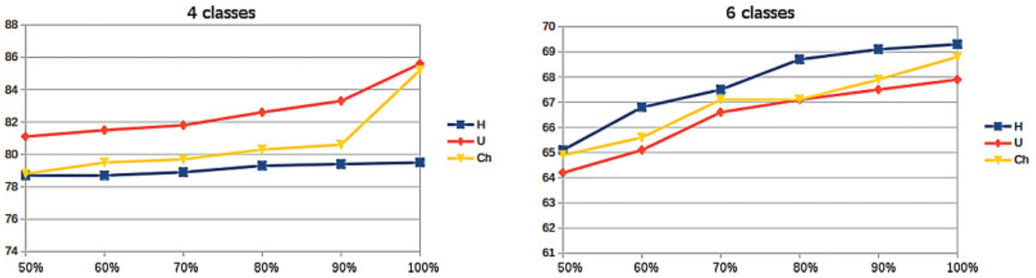

Fig. 10. Classification results as a function of varying the training set size. Experiments performed on different data representations for the four-classes problem (left) and the six-classes problem (right).

Results obtained on the four-emotions problem are shown in Table IX. In this case, the adaptive representations have generally better performances with a peak of 85.6% accuracy of the 1DPF.

The Importance of the Training Set Size. To stress the importance of the amount of data in this problem, we performed another set of classification experiments, fixing the test set and varying the training set size. Figure 10 shows that as the training set grows we obtain better classification performances. In the case of the four-classes problem, which is simpler, the available dataset provides stable performances with the baseline histogram-based representation. Instead, the two adaptive representations produce higher performances for larger training sets. Instead, for the more difficult sic-classes problem, we observe an increase in the performances for all representations, with a steeper trend for adaptive ones (U and Ch). This suggest that even the full set of available data (100%) is too limited for the problem and a larger dataset would allow for higher performances. Based on this experimental evidence, we recently started a new acquisition campaign, involving a group of children. We are currently carrying out an evaluation and validation of the system with autistic children, which we will report in future works.

5. DESIGNING A SYSTEM FOR EMOTION RECOGNITION

Our system is inspired to the multilayered conceptual framework for nonverbal multimodal interaction developed by Camurri et al. [2003, 2004] and shown in Figure 1. Figure 11 shows how the system is designed; it addresses (i) the synchronisation of different input streams received from sensors, (ii) the analysis of the relevant features, (iii) the generation of a significant representation of the extracted information, and, finally, (iv) the analysis of such data to infer the user's emotional state. It is based on the

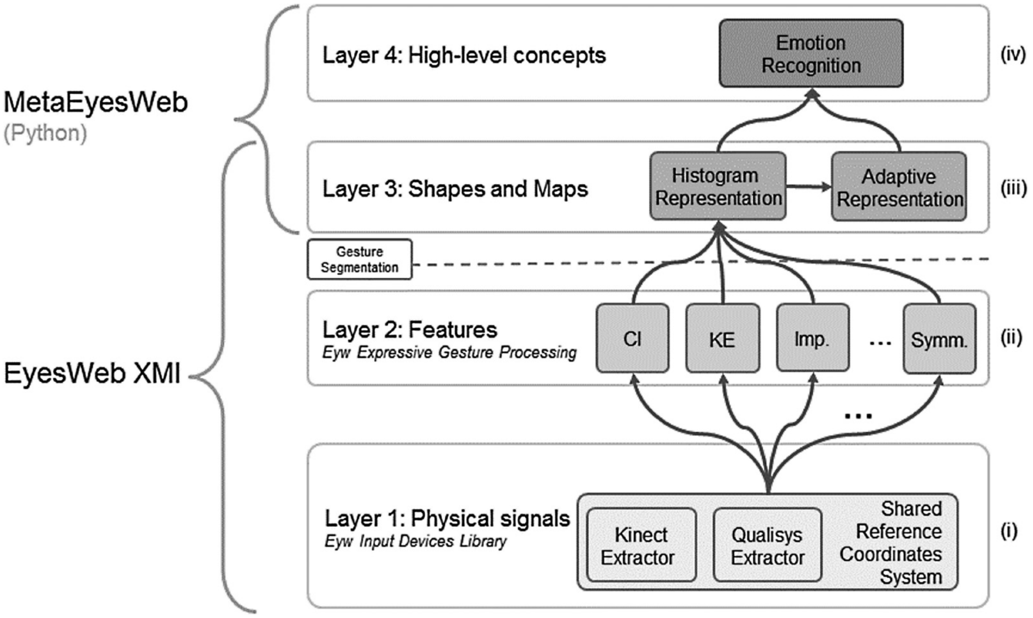

Fig. 11. The components of the system implementing the framework for emotion recognition of Figure 1.

EyesWeb XMI software platform[4] [Camurri et al. 2007]. EyesWeb XMI is a modular distributed development environment that supports the design and implementation of interactive systems and applications. The EyesWeb Environment can be directly interfaced with various input devices such as webcams, microphones, and Motion Capture systems. We used the *input Devices Library* (shown in the first layer from the bottom of Figure 11), to capture and perform early processing on the input streams (e.g., images, motion-capture data) coming from either a Kinect sensor or the Qualisys motion-capture system.

To compute features on the received streams, EyesWeb XMI provides a vast library of modules for expressive gesture analysis and machine learning [Camurri et al. 2004], which has been updated and extended to extract the movement *Features* described in detail in Section 3.1; some of these feature are listed in layer 2 (ii) of Figure 11.

Physical Signals (i) and Features (ii) are used to compute *Shapes and Maps* (iii): In our system, on this layer, Histograms and adaptive descriptors (Sections 3.2, 3.3) are computed from the available features. Histogram representations are calculated by EyesWeb's machine learning library. To learn the adaptive representations, we use EyesWeb and *PADDLE* [Basso et al. 2011], which is developed in PYTHON, and to do, so we rely on *MetaEyesweb*, an API that allows communication, data transfer, and synchronization between the EyesWeb environment and PYTHON applications.

The *High-level concepts* layer (iv), in our application, has the task of understanding user's emotional states based on the previously described models. This task is performed by a *MetaEyesweb* application that gets the representation of an input sample from the EyesWeb environment and gives an evaluation of the expressed emotion as output.

In our system, each specific task is carried out by an Eyesweb XMI Application (*Patch*). A *Patch* is a graphs of linked software modules (blocks), and each *Block*

[4] http://www.infomus.org/eyesweb.

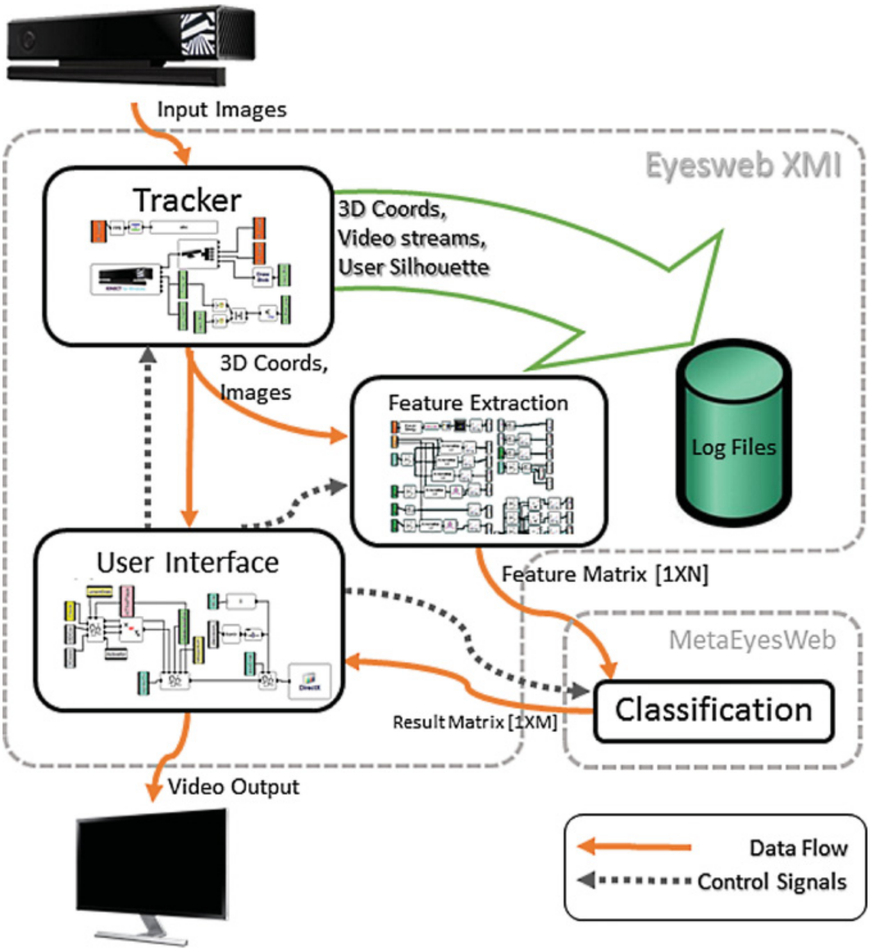

Fig. 12. The diagram shows the data flow between the system's patches: The tracker gets image streams from an input device and computes motion-captured data and user shapes, logs the captured data to a file, and sendd them to the other components. The Feature extractor, derived from the module shown in Figure 4, computes expressive features from the data, builds the feature vector, and sends it to the classifier component. The classifier, given the feature vector, produces the classification results and sends them to the user interface; the user interface controls and monitors the interaction with the user, synchronizes all the other components, receives 3D coordinates and images from the tracker and inferred emotions from the classifier, and produces outputs (i.e., audiovisual output) for the user.

performs a single operation (i.e., capture images from sensors); linking different blocks together allow us to easily and quickly build systems that can solve various and complex problems. Figure 12 depicts a detailed view of the designed system. A group of *Patches* is shown, and each single Patch has a specific purpose: capturing data from the sensors and tracking the user, performing feature extraction and computing the data representation, logging sensor's data and usage statistics to file, generating the user interface and controlling the evolution of the serious games, and communicating with *MeatEyesWeb* applications. The boxes of Figure 12 offer a detailed view of the designed patches: The one on the top shows the blocks that interface with the Kinect sensor and extract motion-capture data, and the right one shows a portion of the feature extraction patch where the overall *kinetic energy* and kinematic features of the hands are computed. Finally, the bottom one shows a portion of the patch that generates the user interface. The links that connect the various blocks represent the exchanged data and signals that control the evolution of the system.

6. SYSTEM APPLICATION: SERIOUS GAMES FOR AUTISTIC CHILDREN

The system architecture described above has been successfully applied in the design of two serious games to help autistic children learn to recognize and to express emotions through full-body movement within the framework of the ASC-Inclusion FP7 ICT Project[5] [Schuller et al. 2013].

The goal of the project is to develop and evaluate an online expression analyser, showing children with Autism Spectrum Condition how they can improve their context-dependent body emotional expressiveness. Based on the movements of the child acquired as 3D streams by RGB-D sensors (as the Microsoft Kinect), the set of emotionally relevant features is extracted and analysed to infer the emotional state of the child in real time.

Finally, we provide an appropriate feedback to the user by means of a graphical user interface.

We designed two game prototypes, a one-player and a two-player game. Both the games perform real-time automatic emotion recognition described in the previous sections and interact with the user by asking to guess an emotion and to express it with his or her body. The user interacts with the graphical user interface through body gesture and, depending on the turn, has to guess an emotion or perform it with the body.

6.1. Guess the Emotion

Guess the Emotion (Figure 13) is a simple single-player game that is composed of two phases. At first, the system shows the player a short video of a person that expresses an emotion. To help to focus just on the information carried by the body movements and discard other stimuli (context, facial expression, voice), a black-and-white silhouette or three-dimensional skeleton representation of a person is shown. The player is then asked to pick from a list the emotion that in her opinion the person in the video was feeling (before answering, the player may also ask the system to see again the video, in case of doubt). If the player's answer is correct, then she gains one point. Then the game asks to the player to express the same emotion of the video with her own body. While the player tries to express the emotion, the recognizing system will try to understand which emotion is being expressed by the player. If the recognized emotion matches the required one the player gains a point, if it differs from the correct one the system will ask to the player if she wants to try again. She will have up to three attempts: If the system answers correctly, or the user spends all three attempts, then the game ends and the user is given a final score that includes the points acquired during the two phases of the game.

6.2. Emotional Charades

In *Emotional Charades* [Piana et al. 2014a] (Figure 14), two players are involved but they do not see each other. When the game starts, a role will be assigned to each player: One player will be the *Actor* and the other the *Observer*. The *Actor* decides which emotion to express, and then she will move her body in front of the sensor to do it. The *Observer* will see the moving silhouette of the *Actor*, and, just observing the body movements, will try to guess the correct emotion. The computer will try to do the same. The guesses of both the computer and the *Observer* will be shown to the *Actor,* and she will say whether the *Observer* or the computer gave the correct answer.

[5] http://asc-inclusion.eu/.

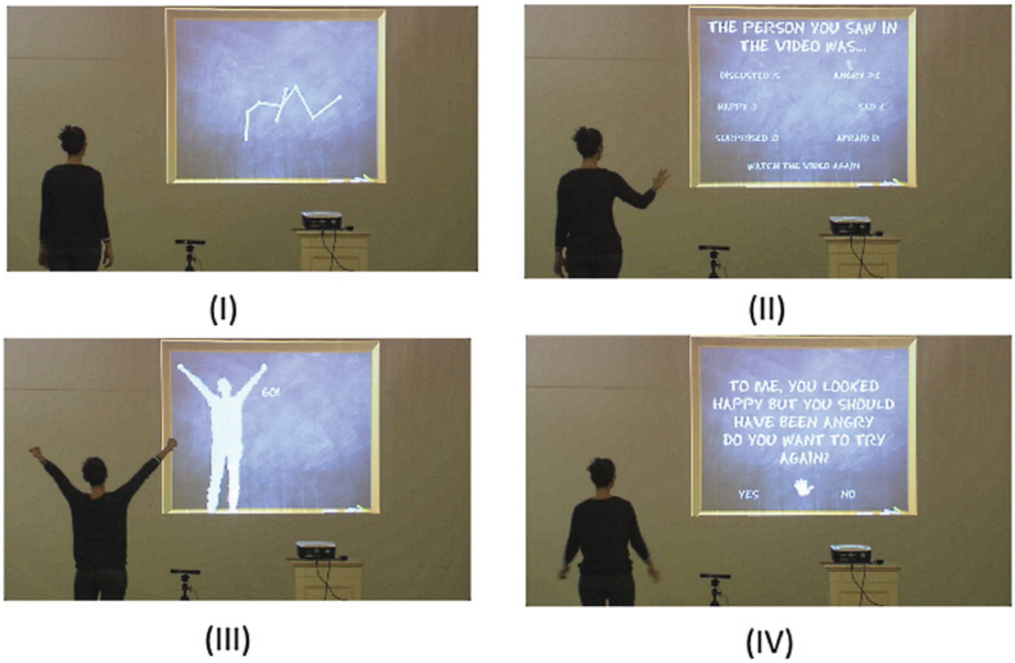
THE PERSON YOU SAW IN THE VIDEO WAS...
DISGUSTED :S
ANGRY >:(
HAPPY :)
SAD :(
SURPRISED :O
AFRAID D:
WATCH THE VIDEO AGAIN
GO!
TO ME, YOU LOOKED HAPPY BUT YOU SHOULD HAVE BEEN ANGRY DO YOU WANT TO TRY AGAIN?
YES
NO
(I)
(II)
(III)
(IV)

Fig. 13. A description of the different stages of Guess the Emotion: (I) The user sees a short video of a person expressing an emotion and then (II) chooses an emotion she recognised and (III) tries to express the same emotion with her body. Finally (IV) the game gives feedback on the user's emotion expression.

This is a collaborative game. The *Actor* will gain more points if both the computer and the *Observer* give the correct answer. In particular, she will gain 2 points if both the *Observer* and the computer guessed right, 1 point if only one of them is right, and no points if they are both wrong. The *Observer* will gain 2 points if her answer is the only one correct, 1 point if the computer gives the same answer as her (whether it is correct or wrong), and no points if only the computer gives the correct answer. The game goes on with inverted roles.

7. CONCLUSIONS

In this article, we presented a complete framework to monitor and process body movements to automatically recognize emotions. We started from 3D motion data of full-body movements, acquired by a motion-capture systems or low-cost RGB-D sensors. We identified and developed algorithms to extract a set of emotion-related features to describe the gestures belonging to different emotions, and we proposed gesture representations based on such features and summarized by means of first-order statistics. To improve the generalization performances of our representations and cope with the large intraclass variability, we also considered a sparse coding layer. Finally, emotion classification was performed by means of a linear SVMs architecture where the information on the relationships between the different classes is captured by a ECOC matrix.

We assessed the performance of our system over different input data, problems, and representations. The system's performance in terms of recognition accuracy was comparable to the ones reached by humans provided with point-light display stimuli.

As a by-product of the research conducted, we obtained further evidence of the appropriateness of sparse adaptive representations for large dimensional problems.

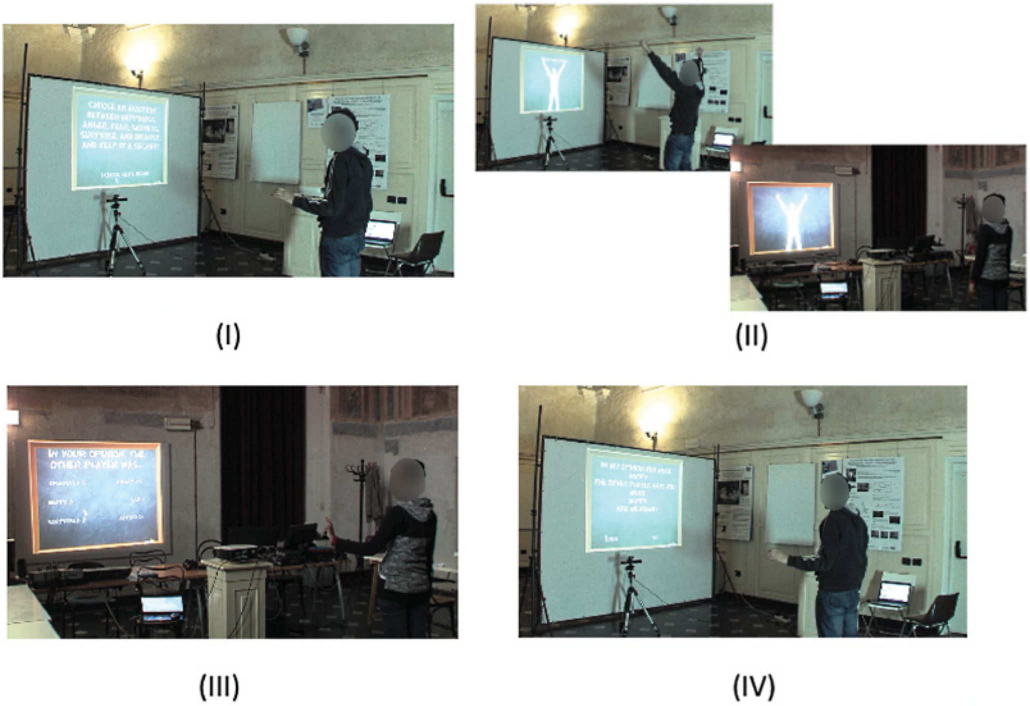

Fig. 14. A description of the different steps of Emotional Charades: (I) the *Actor* chooses an emotion to express from a list of candidates; (II) the *Actor* expresses the chosen emotion while the *Observer* sees the *Actor's* silhouette; (III) the *Observer* guesses the expressed emotion; (IV) the *Actor* chooses the winner between the *Observer* and the computer.

Moreover, an important observation was made that possibly deserves further investigation: The proposed system obtained a higher emotion recognition rate using low-quality sensors such as Kinect than with the motion-capture system. This is possibly due to the larger size of the Kinect dataset, which allowed us to better address the intrinsic noise and variabilities of the considered classes.

The system was used as a building block in the design of serious games developed within the ASC-Inclusion project. In particular, we designed two game prototypes for one or two players.

We are carrying out an extensive validation of the serious games with ASC-affected children: The models and the serious games have been updated with new recordings of typically developing children. This work will be presented in a article currently in preparation [Piana et al. 2015]

Future work will focus on further developments of the system architecture, including the extension toward a continuous dimensional representation of emotions (i.e., a projection on the Valence/Arousal plan), by extending to continuous emotion classification, for example, adopting the PAD model.

The promising results of this work brought us to develop another case study with dancers, and to apply the same computational framework to recognize different movement and dance qualities: current work includes the extension of the system in the framework of the EU Horizon 2020 ICT Project DANCE n.645553, to investigate sensory substitution: how expressive movement qualities (in individuals as well as groups) can be translated in the auditory domain by means of interactive sonification.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their precious comments and our colleagues at Casa Paganini–InfoMus research centre for discussions and concrete contributes to this research project.

REFERENCES

W. K. Allard, G. Chen, and M. Maggioni. 2012. Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Appl. Comput. Harmonic Anal.* 32, 3 (2012), 435–462.

M. Argyle. 2013. *Bodily Communication*. Routledge, London.

A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, A. W. Young, and others. 2004. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception (London)* 33 (2004), 717–746.

T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. 2005. Emotion analysis in man-machine interaction systems. In *Machine Learning for Multimodal Interaction*. Springer, Berlin, 318–328.

C. Basso, M. Santoro, A. Verri, and S. Villa. 2011. PADDLE: Proximal algorithm for dual dictionaries learning. In *Proceedings of the International Conference on Artificial Neural Networks and Machine Learning (ICANN'11)*. Springer, Berlin, 379–386.

R. T. Boone and J. G. Cunningham. 1998. Children's decoding of emotion in expressive body movement: The development of cue attunement. *Dev. Psychol.* 34 (1998), 1007–1016.

P. E. Bull. 1987. *Posture and Gesture.* Pergamon Press, London.

A. Camurri. 1995. Interactive dance/music systems. In *Proceedings of the 1995 International Computer Music Conference.* 245–225.

A. Camurri, P. Coletta, G. Varni, and S. Ghisio. 2007. Developing multimodal interactive systems with eyesweb XMI. In *Proceedings of the Conference on New Interfaces for Musical Expression (NIME), 2007*. ACM, New York, NY, 302–305.

A. Camurri, I. Lagerlöf, and G. Volpe. 2003. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *Int. J. Human-Comput. Stud.* 59, 1 (2003), 213–225.

A. Camurri, B. Mazzarino, and G. Volpe. 2004. Analysis of expressive gesture: The eyesweb expressive gesture processing library. In *Gesture-Based Communication in Human-Computer Interaction*. Springer, Berlin, 460–467.

A. Camurri, G. Volpe, G. De Poli, and M. Leman. 2005. Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia* 12, 1 (2005), 43–53.

G. Chen and M. Maggioni. 2010. Multiscale geometric wavelets for the analysis of point clouds. In *Proceedings of the 44th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 1–6.

R. R. Cornelius. 1996. *The Science of Emotion: Research and Tradition in the Psychology of Emotions.* Prentice-Hall, Piscataway, NJ.

R. R. Cornelius. 2000. Theoretical approaches to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

M. Coulson. 2004. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *J. Nonverb. Behav.* 28, 2 (2004), 117–139.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* 18, 1 (2001), 32–80.

B. de Gelder. 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philos. Trans. Roy. Soc. B: Biol. Sci.* 364, 1535 (2009), 3475–3484.

B. de Gelder and N. Hadjikhani. 2006. Non-conscious recognition of emotional body language. *Neuroreport* 17, 6 (2006), 583–586.

B. de Gelder, J. Van den Stock, H. K. M. Meeren, C. Sinke, M. E. Kret, and M. Tamietto. 2010. Standing up for the body. Recent progress in uncovering the networks involved in the perception of bodies and bodily expressions. *Neurosci. Biobehav. Rev.* 34, 4 (2010), 513–527.

M. de Meijer. 1989. The contribution of general features of body movement to the attribution of emotions. *J. Nonverb. Behav.* 13, 4 (1989), 247–268.

W. H. Dittrich, T. Troscianko, S. Lea, and D. Morgan. 1996. Perception of emotion from dynamic point-light displays represented in dance. *Perception (London)* 25, 6 (1996), 727–738.

P. Ekman. 1965. Differential communication of affect by head and body cues. *J. Person. Soc. Psychol.* 2, 5 (1965), 726.

P. L. Ekman and W. V. Friesen. 1974. Detecting deception from the body or face. *J. Person. Soc. Psychol.* 29, 3 (1974), 288.

J. L. Evenden. 1999. Varieties of impulsivity. *Psychopharmacology* 146, 4 (1999), 348–361.

T. Flash and N. Hogan. 1985. The coordination of arm movements: An experimentally confirmed mathematical model. *J. Neurosci.* 5, 7 (1985), 1688–1703.

D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. 2011. Toward a minimal representation of affective gestures. *IEEE Trans. Affect. Comput.* 2, 2 (2011), 106–118.

O. Golan, E. Ashwin, Y. Granader, S. McClintock, K. Day, V. Leggett, and S. Baron-Cohen. 2010. Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces. *J. Autism Dev. Disorders* 40, 3 (2010), 269–279.

O. Golan, S. Baron-Cohen, and J. Hill. 2006. The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *J. Autism Dev. Disorders* 36, 2 (2006), 169–183.

P. Heiser, J. Frey, J. Smidt, C. Sommerlad, P. M. Wehmeier, J. Hebebrand, and H. Remschmidt. 2004. Objective measurement of hyperactivity, impulsivity, and inattention in children with hyperkinetic disorders before and after treatment with methylphenidate. *Eur. Child Adolescent Psychiat.* 13, 2 (2004), 100–104.

H. Hill and F. E. Pollick. 2000. Exaggerating temporal differences enhances recognition of individuals from point light displays. *Psychol. Sci.* 11, 3 (2000), 223–228.

B. R. Hoff. 1992. *A Computational Description of the Organization of Human Reaching and Prehension*. Ph.D. Dissertation. University of Southern California.

N. Hogan. 1984. An organizing principle for a class of voluntary movements. *J. Neurosci.* 4, 11 (1984), 2745–2754.

G. Johansson. 1973. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics* 14, 2 (1973), 201–211.

A. Kapoor, W. Burleson, and R. W. Picard. 2007. Automatic prediction of frustration. *Int. J. Human-Comput. Stud.* 65, 8 (2007), 724–736.

A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen. 2005. Gesture-based affective computing on motion capture data. In *Affective Computing and Intelligent Interaction*. Springer, Berlin, 1–7.

M. Karg, A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic. 2013. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Trans. Affect. Comput.* 4, 4 (2013), 341–359.

A. Klautau, N. Jevtic, and A. Orlitsky. 2002. Combined binary classifiers with applications to speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Kinect. 2013. Kinect for Windows. Retrieved from http://www.microsoft.com/en-us/kinectforwindows/.

A. Kleinsmith and N. Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* 4, 1 (2013), 15–33.

R. Laban. 1963. *Modern Educational Dance*. Macdonald & Evans, London.

R. Laban and F. C. Lawrence. 1947. *Effort*. Macdonald & Evans, London.

H. Lee, A. Battle, R. Raina, and A. Ng. 2006. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*. 801–808.

Y. Ma, H. M. Paterson, and F. E. Pollick. 2006. A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Beh. Res. Methods* 38, 1 (2006), 134–141.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. 2010. Online learning for matrix factorization and sparse coding. *J. Machine Learn. Res.* 11 (2010), 19–60.

B. Mazzarino and M. Mancini. 2009. The need for impulsivity & smoothness-improving HCI by qualitatively measuring new high-level human motion features. In *Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP)*. 62–67.

J. T. McConville, C. E. Clauser, T. D. Churchill, J. Cuzzi, and I. Kaleps. 1980. *Anthropometric Relationships of Body and Body Segment Moments of Inertia*. Technical Report. DTIC Document.

C. T. Nagoshi, J. R. Wilson, and L. A. Rodriguez. 2006. Impulsivity, sensation seeking, and behavioral and emotional responses to alcohol. *Alcohol. Clin. Exp. Res.* 15, 4 (2006), 661–667.

N. Noceti and F. Odone. 2012. Learning common behaviors from large sets of unlabeled temporal series. *Image and Vision Computing* 30, 11 (2012), 875–895.

N. Noceti, M. Santoro, and F. Odone. 2011. Learning behavioral patterns of time series for video-surveillance. In *Machine Learning for Vision-Based Motion Analysis*. Springer, Berlin, 275–304.

H. O'Reilly, D. Pigat, S. Berggren, S. Fridenson, S. Tal, O. Golan, S. Bolte, S. Baron-Cohen, and D. Lundqvist. 2014. The EU-Emotion Stimulus Set: A Validation Study. Retrieved from http://www.autismresearchcentre.com/projects/Emoticons.aspx.

S. Piana, A. Staglianò, F. Odone, and A. Camurri. 2014a. Emotional charades. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI)*.

S. Piana, A. Staglianò, F. Odone, A. Verri, and A. Camurri. 2014b. Real-time automatic emotion recognition from body gestures. arXiv:1402.5047 (2014).

S. Piana, A. Staglianò, S. Semino, F. Odone, C. Usai, and A. Camurri. 2015. Evaluation of the emotional game for ASC children. (2015). In preparation (2015).

F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2 (2001), B51–B61.

Qualisys. 2013. Qualisys Motion Capture Systems. Retrieved from http://www.Qualisys.com.

C. L. Roether, L. Omlor, and M. A. Giese. 2008. Lateral asymmetry of bodily emotion expression. *Curr. Biol.* 18, 8 (2008), R329–R330.

R. Rubinstein, A. M. Bruckstein, and M. Elad. 2010. Dictionaries for sparse representation modeling. *Proc. IEEE* 98, 6 (2010), 1045–1057.

B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, P. Robinson, I. Davies, O. Golan, S. Fridenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, S. Bölte, D. Lundqvist, S. Berggren, A. Baranger, and N. Sullings. 2013. ASC-inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions. In *Proceedings of the 1st International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI'13) held in conjunction with the 8th Foundations of Digital Games 2013 (FDG)*, B. Schuller, L. Paletta, and N. Sabouret (Eds.). ACM, SASDG (Chania, Greece, May 2013).

W. A. Sethares and T. W. Staley. 1999. Periodicity transforms. *IEEE Trans. Signal Process.* 47, 11 (1999), 2953–2964.

E. Todorov and M. I. Jordan. 1998. Smoothness maximization along a predefined path accurately predicts the speed profiles of complex arm movements. *J. Neurophysiol.* 80, 2 (1998), 696–714.

V. Vapnik. 1998. *Statistical Learning Theory*. Wiley, New York, NY.

H. G. Wallbott. 1998. Bodily expression of emotion. *Eur. J. Soc. Psychol.* 28, 6 (1998), 879–896.

Received December 2014; revised September 2015; accepted November 2015