Adaptive Real-Time Emotion Recognition from Body Movements

WEIYI WANG and VALENTIN ENESCU, AVSP-ETRO, Vrije Universiteit Brussel (VUB)
HICHEM SAHLI, AVSP-ETRO, Vrije Universiteit Brussel (VUB) and Interuniversity Microelectronics Center (IMEC)

We propose a real-time system that continuously recognizes emotions from body movements. The combined low-level 3D postural features and high-level kinematic and geometrical features are fed to a Random Forests classifier through summarization (statistical values) or aggregation (bag of features). In order to improve the generalization capability and the robustness of the system, a novel semisupervised adaptive algorithm is built on top of the conventional Random Forests classifier. The MoCap UCLIC affective gesture database (labeled with four emotions) was used to train the Random Forests classifier, which led to an overall recognition rate of 78% using a 10-fold cross-validation. Subsequently, the trained classifier was used in a stream-based semisupervised Adaptive Random Forests method for continuous unlabeled Kinect data classification. The very low update cost of our adaptive classifier makes it highly suitable for data stream applications. Tests performed on the publicly available emotion datasets (body gestures and facial expressions) indicate that our new classifier outperforms existing algorithms for data streams in terms of accuracy and computational costs.

CCS Concepts: • **Computing methodologies** → *Activity recognition and understanding*;

Additional Key Words and Phrases: Emotion recognition, random forests, semisupervised learning, online learning, real-time system

ACM Reference Format:
Weiyi Wang, Valentin Enescu, and Hichem Sahli. 2015. Adaptive real-time emotion recognition from body movements. ACM Trans. Interact. Intell. Syst. 5, 4, Article 18 (December 2015), 21 pages.
DOI: http://dx.doi.org/10.1145/2738221

1. INTRODUCTION

Human affective display is the external manifestation of internal emotional states. It represents a powerful nonverbal communication cue, which is conveyed by the continuous interplay of multimodal information such as facial expressions, vocal and physiological signals, and body gestures.

In past decades, a great effort had been made to automatically recognize emotions and affective states via different modalities or their combinations. According to de Gelder [2009], 95% of the research was conducted with face stimuli, the majority of the remaining 5% with audio, followed by bodily expressions. Despite the fact that bodily information was relatively neglected in the affective computing field, increasingly more studies had proved that body postures and gestures are important sources

The research work reported in this paper was supported by the CSC-VUB scholarship (grant number 2011688012), the VUB Interdisciplinary project "EmoApp" (grant number IRP5), and the EU FP7 project "ALIZ-E" (grant number 248116).
Authors' addresses: W. Wang, V. Enescu, and H. Sahli, Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussels, Belgium; emails: {wwang, venescu, sahli.hichem}@etro.vub.ac.be.

DOI: http://dx.doi.org/10.1145/2738221

for conveying emotionally relevant information, especially when other channels are inconsistent or unavailable to the observers [Bianchi-Berthouze et al. 2008]. It has been also shown that emotions are able to be well perceived, through bodily expressions alone (e.g., Atkinson et al. [2004] and Alaerts et al. [2011]). Furthermore, empirical studies indicated that specific patterns do exist when certain emotions are expressed via the body (e.g., Wallbott [1998] and Dael et al. [2012b]).

Based on these findings, in this work, we mainly concentrate on real-time emotion classification (of emotion categories or quantized affective dimensions) from bodily expressions with continuous input,[1] and extend our previous work [Wang et al. 2013]. In the proposed system, both statistical and bag of features, representing the dynamics of body movements, are calculated from the combined 3D low-level postural features and high-level (kinematic and geometrical) features. The UCLIC affective gesture database [Kleinsmith et al. 2006] (recorded by MoCap devices) has been employed to train a Random Forests classifier, then both sequence-based (one single emotional gesture sequence) and continuous input tests were conducted to validate the recognition capabilities of the proposed approach. The obtained recognition rate, around 78%, outperformed the best result of the apex posture-based recognition on the UCLIC database.

In order to improve the generalization capability of the prediction system, we further propose a novel semisupervised adaptive classification algorithm based on the Random Forests classifier, that could, in an unsupervised manner, update the system to adapt to the incoming data during the prediction. The model, tunable by only one parameter, is robust against outliers, and is capable of preserving the strength of the labeled training data. Experimental results on several public available emotion datasets (body gestures and facial expressions), as well as Kinect skeleton data streams of bodily expressions performed by subjects that were unseen in the training data, confirms the effectiveness of the proposed algorithm in terms of better recognition performance.

The outcome of this work could be used in real life applications, for instance child-robot interaction, where the companion humanoid is capable of interpreting the emotional status of the child and adapts its behaviors accordingly. Thus, a more natural interaction could be achieved. Another potential application is in digital entertaining. As video games, especially the recent developed body-controlled interactive games, have become a widespread form of entertainment, there is a need for the games to be sensitive to the player's subjective experiences, especially the emotional aspect [Orero et al. 2010].

In human-computer interaction, the ability for systems to understand users' nonverbal behaviors and to respond to them with an appropriate feedback is an important requirement for generating an affective interaction. Such systems are able to create a bidirectional communication with users. By analyzing their nonverbal behaviors, they are able to infer their emotional states and use this information to plan an affective (empathic) response or interaction. The recent introduction of affordable color and depth (RGB-D) sensors (e.g., Microsoft Kinect) allowing real-time human gesture recognition has drawn much attention for applications such as (i) child-robot interaction, where the robot is capable of interpreting the emotional status of the child and adapts its behaviors accordingly [Belpaeme et al. 2012], (ii) body-controlled interactive games, which are sensitive to the player subjective experiences, especially the emotional aspect [Savva et al. 2012], and (iii) measuring the aesthetic experience of dance performances. Indeed, dance choreographers have known for a long time that body posture can signal an affect-related meaning [Camurri et al. 2003].

[1]The term "continuous" here specifically means the system is acquiring and processing data streams, without segmentation, for detecting the emotions as indicated in Gunes and Schuller [2012].

The remainder of this article is organized as follows: Section 2 introduces the state-of-the-art of emotion recognition with body stimuli. Section 3 focuses on the extraction of relevant features and presents the proposed semisupervised adaptive classifier. Experimental results and their analysis are given in Section 4. Finally, in Section 5, we conclude our article and discuss several open questions and perspectives.

2. STATE-OF-THE-ART

In the following, we summarize the studies on emotion recognition from bodily gestures that are related to our work, that is, categorical emotion recognition with continuous input of body gestures. We highlight the aspects of body representations, feature design, prediction models, and publicly available databases. The interested reader is referred to Kleinsmith and Bianchi-Berthouze [2013] and Karg et al. [2013] for comprehensive surveys of recent research on emotion recognition from the bodily expressions.

In the literature, the used approaches for body movements and postures modeling, and hence bodily expressions recognition, are motivated by the way the human body is acquired and represented. Generally speaking, there are two approaches to capture body postures and movements: cameras (2D or RGB-D) or marker-based motion capture (MoCap) devices. Camera captured recordings have the advantages of easy setup, noninvasive, and lower cost. Provided the recorded videos of the body, computer vision techniques are used to track either specific parts of the body [Castellano et al. 2007c; Baltrusaitis et al. 2011], or the frontal/lateral silhouette of the body [Camurri et al. 2003; Sanghvi et al. 2011]. The main drawbacks of the vision-based approaches are the low tracking stability (highly dependent on the recording environment, clothes worn by the participants, and lighting), and the loss of information (only limited parts of the body are tracked). On the other hand, MoCap devices provide precise skeletal representations of the body in 3D (either positions or Euler angles), which is supposed to carry more information. Kleinsmith et al. [2005] and Savva and Bianchi-Berthouze [2012] captured the full body postures and motions in enacting and gaming scenarios, respectively. De Silva et al. [2006] used eight markers to capture the upper-body motions. Metallinou et al. [2011] tracked interactions instead of individual behaviors in a dyadic setting. Nevertheless, the requirement to wear a marker suit may hinder the naturalness of the expressions. Recently, the advent of increasingly mature RGB-D sensors (e.g., Microsoft Kinect sensor) opened new perspectives for online bodily expression analysis.

With respect to designing feature vectors to describe bodily expression, several descriptors have been proposed. To analyze both categorical emotions and affective dimensions of bodliy expressions, Bianchi-Berthouze and Kleinsmith [2003] and Kleinsmith et al. [2006] used 24 low-level static features (mainly for upper body) to model acted affective postures. A similar feature set was used in a body-involved gaming scenario [Kleinsmith et al. 2011]. A set of motion features (the motion of hands, shoulders, and the head) have been also used to improve the emotion classification rate from 65% to 79% [Kleinsmith 2005]. Metallinou et al. [2012] further introduced relative features to represent the affective aspects in dyadic interactions between two participants. Camurri et al. [2004] developed an open platform "EyesWeb," which is capable of analyzing and processing the expressive gestures, providing either blob tracking features or some high-level descriptors, such as Quantity of Movement (QoM) and Contraction Index (CI) of the body. The EyesWeb platform has been employed by many studies for emotion prediction [Castellano et al. 2007c; Camurri et al. 2003]. Glowinski et al. [2011] extended the EyesWeb library by tracking the trajectories of the head and hands in 3D, from which high-level features were extracted. Body information was incorporated in Baltrusaitis et al. [2011] by calculating the angle of the shoulders. Also, body lean

angle, slouch factor of the back, contraction index, and quantity of motion were extracted in Sanghvi et al. [2011]. In order to take into account the dynamics, sequence-based statistics, such as *min*, *max*, *slope*, *peak duration*, etc., were also used [Glowinski et al. 2011; Castellano et al. 2007a; Griffin et al. 2013], along with spatiotemporal features [Shan and Gong 2007]. It is worth noting that most works in emotion classification are based on manually selected postural frames or segmented sequences (one sequence per expression), while Savva and Bianchi-berthouze [2012] used a fixed-size sliding window to carry out dynamic feature calculation, along with a majority vote strategy for emotion prediction.

For automatic emotion classification, traditional classifiers such as SVM [Shan and Gong 2007], Bayesian Network [Castellano et al. 2007b], and HMM [Bernhardt 2010] have been employed. To endow the system with the capability to incrementally learn over time without giving explicit labels, Kleinsmith [2005] proposed an incremental CALM (Categorizing and Learning Module) network. The model updates itself during testing with the ranking of possible answers with an associated probability. This learning model improved the recognition rate of the four postural affective expressions of the UCLIC database from 71% to 79%.

With respect to publicly available data, there exist several corpus that consider bodily expressions as one of the modalities for affective analysis. For example, the GEMEP (Geneva Multimodal Emotion Portrayals) corpus [Bänziger et al. 2012] provides audio-video recordings of 10 actors portraying 18 affective states. Three digital cameras were used to capture the face, the frontal upper body, and the lateral body, respectively. The FABO corpora recorded 23 participants posing nine different emotions [Gunes and Piccardi 2006]. Both facial expressions and upper-body gestures (in a sitting position) were captured by color cameras. The SEMAINE database [Mckeown et al. 2012] employed Sensitive Artificial Listener (SAL) to record audiovisual streams of 150 participants, in conversational interactions. Five color and gray scale cameras have been used to record frontal, as well as profile views. Five affective dimensions have been annotated in the SEMAINE database. The HUMAINE database [Douglas-cowie et al. 2007] is actually a collection of naturalistic and induced emotional data from different sources, such as TV chat shows, TV interviews, gestural games, etc. Selected recordings were annotated with global descriptors and dimensions as well as face/gesture descriptors. The UCLIC database [Kleinsmith et al. 2006] is originally an acted corpus featuring 13 participants, from different cultural regions, portraying four emotions (i.e., anger, fear, happiness, and sadness) with their own body languages without any constraint. In total, 183 sequences were recorded using a MoCap device with 32 markers providing 3D skeletons. The UCLIC affective gesture database was further extended to include nonacted emotional postures, by manually selecting frames from the MoCap sequences recorded when participants were playing body-controlling video games [Kleinsmith et al. 2011]. More recently, the MPI body expressions database recorded "close-to-natural" body expressions under the context of monologic narrations, using MoCap devices [Volkova et al. 2014]. Both the intended emotions (reported by the actors) and the crowd-vote emotion categories were provided in the database.

In this work, the acted MoCap UCLIC affective gesture database [Kleinsmith et al. 2006] is used to train a random forest classifier. The trained classifier is then used in a stream-based semisupervised adaptive Random Forests classifier for continuous unlabeled Kinect data classification. In our system, both statistical and bag of features, representing the dynamics of body movements, are calculated from 3D low-level postural features and high-level (kinematic and geometrical) features. The very low update cost of our adaptive classifier makes it highly suitable for real-time emotion recognition from body movements.

Table I. The Definition of the 28 Postural Features (Adapted from Kleinsmith and Bianchi-Berthouze [2007]; Added Features are Marked in Bold)

ID	Meaning	ID	Meaning
1	Euclidean Distance of Two Feet	**2**	**Euclidean Distance of Two Hands**
3	**Euclidean Distance of Two Elbows**	**4**	**Euclidean Distance of Left Hand and Head**
5	**Euclidean Distance of Right Hand and Head**	**6**	**Angle of Left Elbow**
7	**Angle of Right Elbow**	8	Orientation of Shoulders on X-Y Plane
9	Orientation of shoulders on X-Z Plane	10	Orientation of Feet on X-Y Plane
11	Right Hand - Right Shoulder in Z	12	Left Hand - Left Shoulder in Z
13	Right Hand - Right Shoulder in Y	14	Left Hand - Left Shoulder in Y
15	Right Hand - Left Shoulder in X	16	Left Hand - Right Shoulder in X
17	Right Hand - Right Elbow in X	18	Left Hand - Left Elbow in X
19	Right Elbow - Left Shoulder in X	20	Left Elbow - Right Shoulder in X
21	Right Hand - Right Elbow in Z	22	Left Hand - Left Elbow in Z
23	Right Hand - Right Elbow in Y	24	Left Hand - Left Elbow in Y
25	Right Elbow - Right Shoulder in Y	26	Left Elbow - Left Shoulder in Y
27	Right Elbow - Right Shoulder in Z	28	Left Elbow - Left Shoulder in Z

3. METHODOLOGY

3.1. Feature Extraction

Wallbott [1998] analyzed the relationship between 14 portrayed emotions and bodily expressions. He concluded that the upper-body activities, especially the behaviors of the hands/arms, and the movement quality (e.g., energy of movement, spatial extension, etc.) are important to discriminate emotions. Dael et al. [2012b] proposed a Body Action and Posture Coding System (BAP) that was developed upon the GEMEP corpus (portrayed by professional actors). 141 detailed bodily cues, including postures, actions, and other behavioral categories were defined. Statistical analysis was conducted on a subset of the BAP to discriminate 12 emotions [Dael et al. 2012a]. The results indicated that several patterns of body movement systematically occur in portrayals of specific emotions. The contribution of the previously mentioned works is twofold. First, the empirical results of Wallbott [1998] proved that bodily information (both postures and movements) has the capability to differentiate emotions. Moreover, the BAP descriptors could serve as a guide to design features to perform emotion recognition. Inspired by these works, we extract both low-level postural features and high-level kinematic and geometrical features. In contrast to the approaches proposed in Bianchi-Berthouze et al. [2008] and Kleinsmith and Bianchi-Berthouze [2007], we assume that expressive postures are evolving both spatially and temporally rather than being static. This had been investigated and supported by previous works indicating that bodily expressions could also be segmented to *onset*, *apex*, *offset* [Gunes and Piccardi 2009].

3.1.1. Postural Features. In order to represent the postural patterns of the body, we calculate the spatial distances among hands, elbows, and shoulders in each of the three dimensions, as well as the Euclidean angles of the two elbows. Moreover, we calculate the distance between feet, the orientation of feet, and the orientation of the shoulders. All these lead to 28 postural features[2] calculated on a per-frame basis as detailed in Table I.

3.1.2. High-Level Features. These features are designed to represent the high-level characteristics of the bodily expressions, such as *movement activity and power*, *body spatial*

[2]Note that, for all the feature used in this article, the skeleton's right shoulder points to the X_+ axis, the face points to the Y_+ axis, and the head points to the Z_+ axis, respectively.

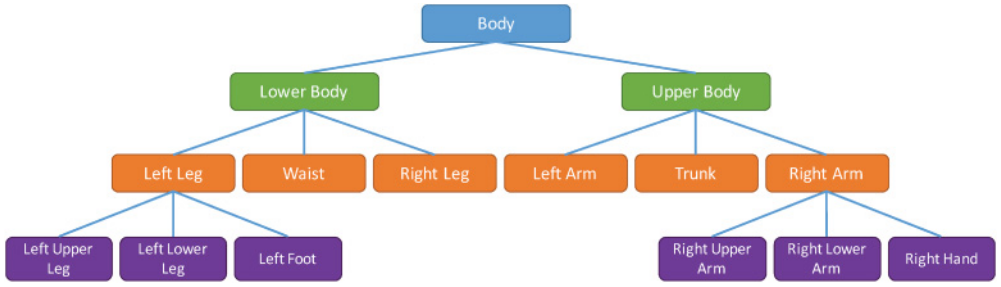

Fig. 1. Hierarchical representation of human body.

extension, and *body bending*. Some of these features correspond to the *movement quality* of Wallbott [1998].

—*Body Movement Activity and Power:* In the literature, the body movement power was often modeled as *QoM*, *Moving Speed*, and *Moving Acceleration*. For instance, Sanghvi et al. [2011] computed the difference between adjacent image frames as the QoM, while Glowinski et al. [2011] considered only the translation of two hands and the head. In this work, we consider the body movements in a more systematic and ergonomic way, taking into account the different contributions of the body parts. Specifically, human motions could be thought of as being composed of different physical segments and each segment can move independently and exhibit an independent degree of activity [Aggarwal and Cai 1997]. As illustrated in Figure 1, these body segments have a hierarchical structure, which allows estimating the body movement activity and power composed of three parameters: *Force*, *Kinetic Energy*, and *Momentum*, calculated hierarchically, from the bottom to the top of the body structure (refer to Kahol et al. [2003] for the details):

$$segment\_Force = segment\_Mass \times segment\_Acceleration, \tag{1}$$

$$segment\_KineticEnergy = 0.5 \times segment\_Mass \times segment\_Velocity^2, \tag{2}$$

$$segment\_Momentum = segment\_Mass \times segment\_Velocity, \tag{3}$$

where *segment_Mass* are estimated according to ergonomic definitions, which relate the body weight to the masses of different body segments (refer to Kroemer et al. [1994] for details). Subsequently, the *Force*, *Kinetic Energy*, and *Momentum* of the full body could be extracted at the top level.

—*Body Spatial Expansion:* As indicated in Wallbott [1998], to extract such features, we first estimate the body's bounding box ($[X, Y, Z].max$, $[X, Y, Z].min$), then three spatial extent indexes are calculated:

$$SE_{XY} = \frac{X.max - X.min}{Y.max - Y.min}, \tag{4}$$

$$SE_{YZ} = \frac{Y.max - Y.min}{Z.max - Z.min}, \tag{5}$$

$$SE_{XZ} = \frac{X.max - X.min}{Z.max - Z.min}. \tag{6}$$

—*Symmetry:* The upper body asymmetry has been considered as an indicator of relaxing attitude [Mehrabian 2007]. Moreover, it can also help in differentiating the single/double arm actions. It also indicates whether the arms are moving reversely [Dael et al. 2012a]. Thus, in this work spatial symmetric indexes are considered for

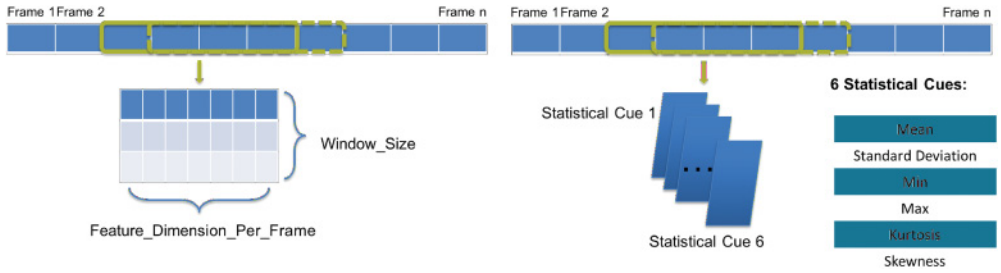

Fig. 2. *Bag of features* (left) and *statistical cues* (right).

the two hands. Moreover, to ensure body-position independency, a relative coordinate system is used:

$$Symmetry_X = \frac{X.LeftHand - X.BodyCenter}{X.BodyCenter - X.RightHand}, \tag{7}$$

$$Symmetry_Y = Y.LeftHand - Y.RightHand, \tag{8}$$

$$Symmetry_Z = Z.LeftHand - Z.RightHand \tag{9}$$

in which *BodyCenter* is defined as the central point between two hips.

—*Body Bending:* This feature is designed to capture forward body bending or backward movements (mostly happening in fear and disgust emotions).

$$Bending = Y.Head - Y.BodyCenter. \tag{10}$$

Finally, 10 high-level kinematic and geometrical features are extracted, and together with the 28 postural features, they form the frame-based 38-dimensional feature vector. Note that, as most of the previously defined features are body size dependent, we normalize them to the body size (distance of the head and feet on the z axis, in neutral poses).

3.1.3. Temporal Dynamics. Apart from the movement activity and power, using the aforementioned features of Section 3.1.2, we also model the temporal dynamics of bodily movements and propose the following features to capture the temporal characteristics of bodily expressions:

—*Bag of Features:* The features extracted at each frame of the skeletal sequence are assembled into bags, that is, short frame sequences. Such temporal feature patches capture not only the current states of each feature, but also how they fluctuate within a specific time window. Note that a similar "*10-frame window*" was defined in Savva et al. [2012], while the window was mainly used to calculate the dynamic features such as velocity, acceleration, frequency, etc.

—*Statistical Cues:* To capture the gist of the body motion, Griffin et al. [2013] used the *min*, *max*, and *range* of several postural features within the whole sequence of each expression, which required manual segmentation. Instead of that, we calculate, within a fixed-size time sliding window, six statistical cues (*mean*, *standarddeviation*, *min*, *max*, *kurtosis*, and *skewness*; see Figure 2, right), for each of the 38 features, where segmentation is not necessary.

In our current implementation, the window size is set to 30 frames; this allows covering the apex phase of the bodily expressions within the UCLIC database. Moreover, this value provides a good trade-off between the computational efficiency of the system,

and the amount of dynamic information needed to capture different bodily expressions. It has to be noted that, as we use a sliding window principle, this approach would allow capturing the temporal phases of the bodily expressions (i.e., onset, apex, offset).

3.2. Adaptive Recognition

When designing an online system for emotion recognition from body movements to be used in real-world applications, one should consider the following requirements:

—Real-time recognition with continuous high-dimensional input data where temporal segmentation is not desired.
—Adaptive recognition to unseen data. This would allow dealing with differences between persons when expressing emotions.
—Handling feature redundancy. Indeed the previously proposed features cover most of the possible bodily behaviors observed in affective expressions; however, as we do not apply any feature reduction or selection approach, the features could be redundant.
—Minimum parameters tuning.

Bearing in mind the preceding requirements, we chose the Random Forests classifier [Breiman 2001] to perform emotion recognition from body movements. According to the comprehensive experiments conducted in Caruana et al. [2008], Random Forests achieves the best recognition performance in most cases, especially when dealing with high-dimensional problems. The Random Forests classifier has several characteristics, such as (i) low computational cost, (ii) inherent support for multiclass problems, (iii) capability to deal with feature redundancy, and (iv) avoid overfitting problems, which render it suitable to our recognition problem. Additionally, Random Forests classifiers are easy to tune. The most important parameter is the size of the forest, which could be set as a large value, provided affordable computational cost. Furthermore, from the implementation point of view, as each single feature is stored, with a unique ID, in the model to determine the decision of each splitting node, temporal structure (i.e., the structure of the feature vector) will not be destroyed, which is not the case in other recognition models, such as Support Vector Machines and Gaussian Processes [Soh 2012].

In the following, we first give a brief introduction to the conventional *Random Forests* classifier as well as its *online* version, then we present the proposed *adaptive learning* algorithm that endows our system with the capability to adapt and update itself during the online recognition.

3.2.1. Random Forests. A Random Forests is basically an ensemble of decision trees [Breiman 2001], within which each tree is grown and evaluated independently from the others. During the training, each tree is fed with a bootstrapped subset of the training set by sampling with replacement. When growing a specific tree, each node runs a series of random tests with a randomly sampled subset of the original features. The best scored test (scores are often computed as the *Information Gain* or the *Gini Index*) and its splitting threshold are stored in the node. A variant of Random Forests, the Extremely Randomized Forests [Geurts et al. 2006], implements a random selection of the splitting threshold. Owing to the random characteristics, although each tree is a weak predictor, the whole forest could provide powerful capacities.

3.2.2. Online Random Forests. To cope with sequential data, Saffari et al. [2009] proposed an online version of the Random Forests. The approach allows solving the two main problems related to the use of the Random Forests in an online fashion:

—How to perform the bagging with the incoming samples?
—How to grow the trees on-the-fly?

The first question was solved by modeling a *Poisson* distribution on each incoming sample. Specifically, each tree is updated p times, where p is generated by $Poisson(\lambda)$. λ could be simply set to 1. As for the second question, the binary splitting on a node is modified as follows: if enough samples, determined by a parameter β indicating the minimum required number of samples to split a node, already arrived in the note, a splitting into a left child node and a right child node occurs according to the best scored test and its threshold. The same strategy is followed by all terminal nodes (leaf nodes). It has to be noted that building and updating a tree starts from a root node with a set of randomly selected samples. Experiments on several public datasets reported that the online classification error would converge to the one of its batch counterpart [Saffari et al. 2009].

3.2.3. Adaptive Random Forests. Dealing with emotion recognition from bodily expressions, where emotional displays and expressive gestures vary between persons, even for the same person, the performance of supervised learning algorithms is highly dependent on the size and goodness of the labeled training data. Often only limited datasets are available (e.g., the UCLIC database contains only 183 sequences), which might lead to a weak descriptive power of the trained model, hence low recognition performance. To overcome this problem, in this work we propose a *semisupervised Adaptive Random Forests*, based on the Online Random Forests classifier of Saffari et al. [2009].

ALGORITHM 1: Adaptive Random Forests

Input: A Random Forest, $\mathcal{F}$, trained in a batch mode, with the labeled training data
Input: The unlabeled sample, $\vec{x}$
Input: The proportion of trees that could be updated, α
Input: The confidence threshold to update the forest, β
Input: Randomly generated index sequence, I, the length of I equals to the number of trees; the number of elements in I with the value *TRUE* is determined by α
Output: Updated forest, $\mathcal{F}'$
Output: Predicted label for the sample, L

```
prediction = F'(x).evaluate.label;
confidence = F'(x).evaluate.confidence;
if confidence > β then
    for each tree f_n in F' do
        if I(n) == TRUE then
            f_n.update(x, prediction);
        end
    end
end
```

The pseudoalgorithm of the proposed Adaptive Random Forests is described in Algorithm 1. We first pretrain a Random Forests, $\mathcal{F}$, in a batch fashion using a labeled dataset. $\mathcal{F}$ is then used in an online fashion; for an incoming unlabeled sample its prediction results (label) and confidence are output. If the confidence exceeds the threshold β, we consider this sample and the corresponding prediction reliable and further feed them to update the forest in an online mode. In order to preserve the strength of the training data and keep the system healthy during update, we introduce an extra parameter, α, which controls the number of trees that could be updated by the unlabeled samples. Updatable trees are selected randomly in the forest and are fixed during the training.

Drifting is always the primary issue for most semisupervised learning algorithms [Tang et al. 2007]. If the model keeps receiving updating samples with wrong labels, it

Table II. Recognition Results (Weighted Classification Accuracy) on the AVEC-2011 Database

Recognition Rate (%)	Activity	Expectancy	Power	Valence	Average
Baseline	60.2	58.3	56.0	63.6	59.5
Meng and Bianchi-Berthouze [2014]	67.2	54.4	53.7	65.0	60.1
Random Forests	62.4	60.6	54.8	67.6	61.8
Adaptive RF	**67.8**	**61.5**	**57.7**	**68.4**	**63.9**

may gradually lose its prediction capability. To overcome this problem, in the proposed algorithm, we make use of two parameters, α and β. As indicated previously, parameter α controls the percentage of the trees that could be updated during recognition. $\alpha = 0$ stands for no update, and $\alpha = 1$ for a complete update. β, on the other hand, guarantees that only "inlier" data can feed the model.

The parameters α and β are set empirically, according to the properties of the specific problem in hand, reliability of the training data, and the predicting power of the pretrained model. The setting of α is relatively straightforward. If the system could highly rely on the training data and lower adaptability is required, we set α to a small value, and vice versa. In our experiments we simply set $\alpha = 0.5$, indicating that half of the trees are updated. For setting the value of β, the user can observe the predicted results and the corresponding confidence during the training phase, then fix the value of β to balance between the quality of the update and the adaptability of the system.

4. EXPERIMENTS

4.1. Adaptive Random Forests Evaluation on AVEC 2011 Dataset

Before evaluating the proposed methodologies on emotion recognition from body movements, in this section, we assess the effectiveness of the semisupervised Adaptive Random Forests proposed in Section 3.2.3. The assessment was conducted on the AVEC (Audio/Visual Emotion Challenge) 2011 dataset [Valstar et al. 2011], considering the fact that it provides ready-to-use features and labels, as well as baseline recognition rates. The AVEC-2011 database is a subset of the SEMAINE corpus [Mckeown et al. 2012]; it is composed of upper-body video sequences of dyadic interactions between human subjects and virtual agents. All frames are annotated with four emotion dimensions, that is, *Activity*, *Expectancy*, *Power*, and *Valence*, in continuous time and continuous values. These values were then dichotomized as 0 and 1 based on the mean of each dimension. This leads to a binary classification problem for each emotion dimension. The dataset is divided into training, development, and test subsets. In our experiments, we use the training set to train the Adaptive Random Forests, and the development set for the recognition performance analysis. Considering the large amount of data (> 1.3 million frames), in our experiments only $1,000$ frames from the training partition and $1{,}000$ frames from the development partition, along with their Local Binary Patterns (LBP) features, are used (refer to Valstar et al. [2011] for the details on sampling and feature extraction). Table II summarizes the recognition results (i.e., weighted classification accuracy) of the proposed Adaptive Random Forests, compared to the baseline Support Vector Machine of AVEC-2011, the hierarchical HMM-based approach [Meng and Bianchi-Berthouze 2014], as well as the conventional Random Forests classifier.

As can be seen, the conventional Random Forests classifier outperformed the baseline except on the *Power* dimension. The proposed Adaptive Random Forests further enhances the recognition rate, and reached the highest performance on all of the four emotion dimensions. For these experiments we used a forest of 100 trees, and the parameters were set to $\alpha = 0.75$ and $\beta = 0.5$. Since the main objective of this experiment is to evaluate the effectiveness of the proposed adaptive learning method in emotion prediction domain, rather than to improve the prediction accuracy by all means, we used

Fig. 3. An example of fear emotion at four time points from the UCLIC database.

exactly the same precalculated features, data partition, and frame sampling scheme as provided in the dataset [Valstar et al. 2011]. Although the first prize of the AVEC-2011 video subchallenge achieved the classification accuracy of 66.9% in average [Ramirez et al. 2011], considering that a completely different feature set was used in this work, the results are not directly comparable.

4.2. Emotion Prediction from Bodily Expressions

In this work, the UCLIC database [Kleinsmith et al. 2006] is used to valuate the features and the learning model. Although affective postures were the main research target of the UCLIC database, the acquired continuous frames, which contain onset, apex, and offset phases of each expression, make it also valuable for temporal analysis. In the UCLIC database, two sets of labels are provided: (a) observers' labels (observers with different cultural backgrounds were asked to assign emotion labels according to their own perception, provided one apex skeletal frame of each sequence); (b) actors' portrayals (the emotions that the actors were asked to act). In our experiments, the actual portrayals are used as our ground truth. Figure 3 depicts several key frames of a sample sequence of the "fear" emotion.

In the following, we first evaluate the usefulness of the proposed body movement features of Section 3.1, and we further evaluate the recognition performances of the proposed Adaptive Random Forests in several settings.

4.2.1. Feature Evaluation

—*Postural Features vs. High-Level Features*. The proposed feature set is composed of two categories as described in Section 3.1, namely, low-level postural features and high-level features. In order to compare and analyze the contribution of the two subsets, we conducted a 10-fold cross validation on the 183 single expressive gesture sequences of the UCLIC database. A conventional Random Forests classifier with 100 trees is used in these experiments. As baseline features, the 24 postural features as defined in Kleinsmith and Bianchi-Berthouze [2007] were extracted from the apex frame (representing the most visually informative frame of the expressive gesture, provided by the UCLIC database) of each sequence. Using these features, the Random Forests classifier achieved an overall recognition rate of 75.41% (138 correctly classified out of 183), considered hereafter as baseline classifier.

The recognition performance is presented in Table III. When using only postural features, we could reach a higher classification accuracy as compared to the case where the high-level features are used. As can be seen, the high-level features are not sufficient by themselves to differentiate expressive gestures. This is also supported by a χ^2 ranking (performed using the WEKA toolbox [Hall et al. 2009]) and variable importance ranking (given by the Random Forests classifier), as reported in Table IV. As one can see, the high-level features are relatively ranked lower. This could also

Table III. Recognition Accuracy when Using Postural Features and High-Level Features Separately. The Last Row Gives the Baseline Recognition Using the Features of Kleinsmith and Bianchi-Berthouze [2007] on the Apex Frames Only

	Postural Features				High-Level Features			
	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad
Fear	28	10	8	3	21	15	8	5
Happy	6	36	4	1	11	32	1	3
Angry	5	3	28	5	4	14	16	7
Sad	1	5	6	34	4	3	3	36
Rate	68.85%				57.38%			
Baseline	75.41%							

be partly explained by the *acted nature* of the UCLIC database. It has to be noted that in these experiments the classification is made using the full sequence.

—*Bag of Features* vs. *Statistical Cues*. We then further analyzed the recognition rate when using all frame-based features (i.e., postural and high-level features) combined with the statistical cues or bag of features. Table V summarizes the classification results.

One can see that, by introducing the statistical cues or bag of features, the recognition performance is improved to a certain extent. Note that only the apex frames (the most expressive frame in each expression) were used in the baseline result. As manual selection of the apex frame is not required in our approach, our method can be used on continuous unsegmented data streams. Also, comparing Table III and Table V, one can notice that by introducing the temporal information a significant improvement was obtained to distinguish *fear* and *happy* emotions.

Another conclusion from this experiment is that the statistical cues and bag of features reach similar recognition rates. This is because the statistical cues could be considered as abstract descriptions of the bag of features, that is, they convey similar information. Finally, although the feature set dimensionality would largely increase (six times for the statistical cues and 30 times for the bag of features), the *Random Forests* classifier still delivers results in real time due to its efficient tree-searching nature. This is a great merit of our approaches when they are used in real-life applications.

—*Postural Features vs. High-Level Features with Temporal Dynamics*. Knowing the introduced temporal information has enhanced the distinguishing performance of different emotions; it is also interesting to know how the low-level postural and high-level features would gain a better descriptive power by incorporating temporal dynamics, respectively. Thus, we further compared the recognition accuracy similar to the first experiment in Section 4.2.2, while six statistical cues were added to each one of the frame-based features. The results are given in Table VI.

By comparing the results from Table III and Table VI, we can obviously conclude that both low-level postural and high-level feature set could benefit from the extra temporal dynamics, among which, the postural features achieved more improvement than its counterpart. A possible reason could be that the body movement activity and power features have already contained temporal information intrinsically.

4.2.2. Emotion Recognition

—*Online Recognition*. In these experiments, the Online Random Forests classifier of Saffari et al. [2009] (Section 3.2.2), trained on the UCLIC database, is used for continuous Kinect data classification. Although there are differences between the definition of joints in the Kinect SDK and the ones defined in the UCLIC database (due to the physical constraints of the markers placement), during our experiments

Table IV. χ^2 Rankings and Variable Importance Rankings (Given by the Random Forests Classifier) of the Full Frame-Based Feature Set

χ^2 Rank	RF Rank	Feature
1	10	Euclidean Distance of Two Feet
2	1	Left Hand - Left Shoulder in Z
3	7	Orientation of Feet on X-Y Plane
4	4	Right Hand - Right Shoulder in Z
5	3	Left Hand - Left Elbow in Z
6	6	Right Hand - Right Elbow in Z
7	8	Euclidean Distance of Left Hand and Head
8	9	Angle of Left Elbow
9	12	Angle of Right Elbow
10	15	Right Elbow - Right Shoulder in Z
11	2	Head Leaning
12	13	Euclidean Distance of Right Hand and Head
13	21	Left Elbow - Left Shoulder in Z
14	25	Momentum of Body
15	30	Left Hand - Right Shoulder in X
16	5	Orientation of Shoulders on X-Y Plane
17	14	Left Elbow - Left Shoulder in Y
18	11	Spatial Index on Y-Z Plane
19	38	Euclidean Distance of Two Hands
20	18	Left Hand - Left Elbow in X
21	32	Right Hand - Right Elbow in X
22	29	Energy of Body
23	17	Symmetry Index on Y-Z Plane
24	27	Right Hand - Right Elbow in Y
25	24	Left Hand - Left Elbow in Y
26	26	Spatial Index on X-Y Plane
27	37	Right Hand - Left Shoulder in X
28	23	Symmetry Index on X-Y Plane
29	36	Euclidean Distance of Two Elbows
30	22	Symmetry Index on X-Z Plane
31	16	Right Elbow - Right Shoulder in Y
32	35	Force of Body
33	19	Orientation of Shoulders on X-Z Plane
34	33	Spatial Index on X-Z Plane
35	34	Right Hand - Right Shoulder in Y
36	28	Left Elbow - Right Shoulder in X
37	20	Left Hand - Left Shoulder in Y
38	31	Right Elbow - Left Shoulder in X

we observed that it is not necessary to match them perfectly since the system has a certain tolerance.

For the evaluation, we asked six participants to randomly portray, as naturally as possible, the four emotions contained in the UCLIC database (see Figure 4 for an example of the interface). To illustrate the expected bodily expressions, we presented to the participants a randomly selected posture image from the UCLIC database. To avoid nervousness and "overacted" artifact, the participants rehearsed as many times as they needed before the real experiments. Since the UCLIC database used the "T-pose" as the neutral posture, from which each expression started, we also asked the participants to perform the "T-pose" in the beginning and the end of each portrayal.

Table V. Recognition Accuracy Combining the Frame-Based Features and Statistical Cues or Bag of Features. The Last Row Gives the Baseline Recognition Rate Using the Features of Kleinsmith and Bianchi-Berthouze [2007] on Apex Frames Only

	Frame-Based Features				+Statistical Cues				+Bag of Features			
	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad
Fear	31	11	4	3	34	5	5	5	35	7	3	4
Happy	5	36	4	2	4	38	2	3	4	39	3	1
Angry	4	5	29	3	4	3	30	4	3	4	30	4
Sad	2	3	4	37	3	2	3	38	2	2	3	39
Rate	72.68%				76.50%				**78.14%**			
Baseline	75.41%											

Table VI. Recognition Accuracy when Using Postural Features and High-Level Features Separately, Both Enriched with Temporal Statistical Cues

	Postural Features + Statistical Cues				High-Level Features + Statistical Cues			
	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad
Fear	31	7	8	3	24	12	9	4
Happy	6	38	2	1	13	32	1	2
Angry	4	4	28	5	4	15	15	7
Sad	2	4	4	36	3	1	4	38
Rate	72.68%				59.56%			

Fig. 4. Illustration of the data acquisition using the Kinect sensor and two different bodily expressions (*angry* on the left, *fear* on the right).

In total, 75 expressive segments were collected and used for recognition. The upper plot of Figure 5 depicts the recognition results for a long sequence containing a random repetition of the four emotion segments. The recognition is made in real time giving the feedback of *neutral state* or the recognized emotions. In the UCLIC database, the neutral gesture was defined to be a *T-Pose*, which could be simply detected by measuring both the body movement energy and body spatial extension. More sophisticated approaches could be introduced, but it is out of the scope of this article. A simple window-based smoothing filter was applied to remove sharp errors. As can be seen from the left part of Table VII, for the 75 expressive segments, 54 were correctly recognized, giving a recognition rate of 72%. It has to be noted that, even if the subjects did not appear in the training data, the expressive gestures could still be well classified.

—*Adaptive Real-Time Emotion Prediction*. We finally applied the proposed Adaptive Random Forests on the continuous unlabeled Kinect data. The middle plot of Figure 5 illustrates the obtained recognition results. One can notice, that the *angry*

Table VII. Confusion Matrix of the Prediction Using the Kinect Skeleton Streams, with the Model Trained Using the UCLIC Database

	Online Random Forests				Adaptive Random Forests			
	Fear	Happy	Angry	Sad	Fear	Happy	Angry	Sad
Fear	12	4	1	1	13	3	1	1
Happy	2	12	3	1	2	12	3	1
Angry	1	4	14	0	0	2	17	0
Sad	0	1	3	16	0	1	3	16
Rate	72.00%				**77.33%**			

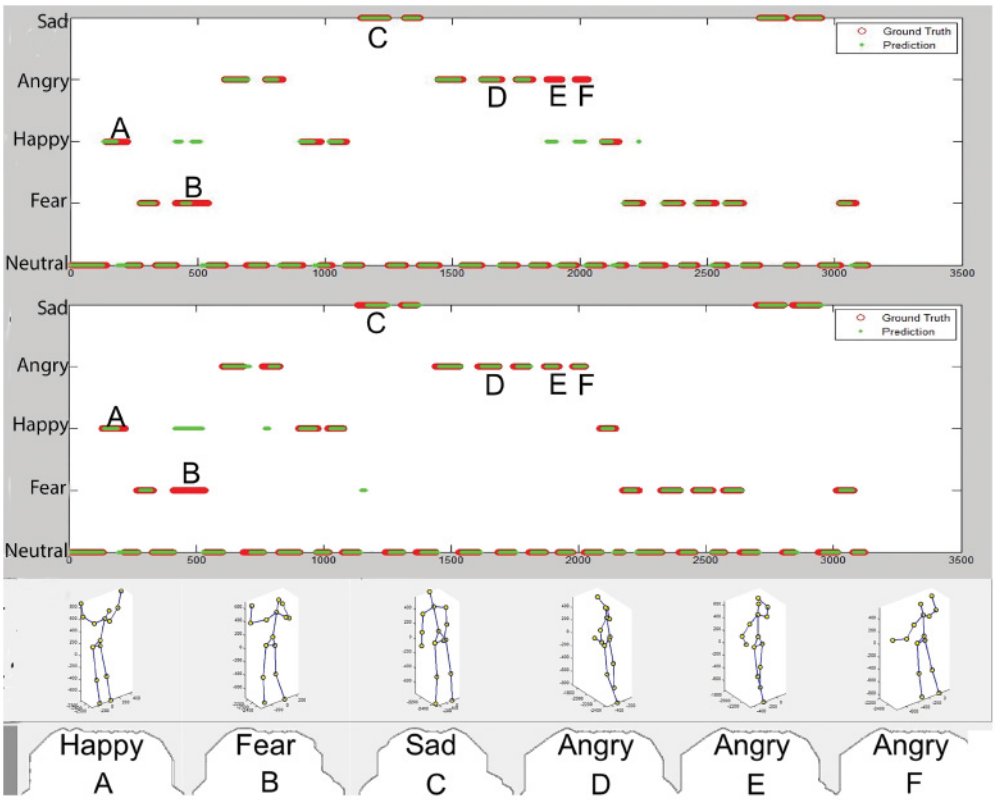

Fig. 5. Continuous recognition result of a sequence recorded by the Kinect sensor. Red bars are labels and green bars are the predicted results. Overlaps represent correctly recognized frames. The upper plot illustrates the results using the Online Random Forests classifier; the middle plot illustrates the result using the proposed Adaptive Random Forests classifier. The lower plot depicts six skeletal views (A to F) illustrating particular bodily expressions. Note that the "Neutral" category represents a "T-pose," as in the UCLIC database.

expressions, *E* and *F*, are correctly recognized by the adaptive algorithm, which have been misclassified (*happy*) by the online random forests. From the right part of Table VII the overall recognition accuracy is also improved.

5. DISCUSSION

There are three important criteria to assess the automatic emotion analyzers: *correctness* (to match the labels and annotations), *robustness* (low sensitivity to different

subjects and the changing conditions), and *efficiency* (fast processing and responses) [Gunes and Schuller 2012]. To meet these three criteria, in this article, we presented a complete framework that aims at taking 3D bodily motion stream as input, and instantly giving the predicted emotions. By combining the low-level postural features and high-level kinematic and geometrical features, together with the temporal cues, we promoted the descriptive power for emotion recognition from bodily expressions. The novel Adaptive Random Forests algorithm was introduced to update the model with the new data, which could provide stable and reliable prediction without tedious model tuning and optimization. Eventually we built an adaptive real-time continuous emotion recognition system. It takes Microsoft Kinect stream as the input signal, and gives the predicted emotion labels on the screen, in real time. Owing to the computational efficiency, the system could be practically applied to real-life interactions, such as child-robot interaction or digital entertaining.

In the following, we will discuss several interrelated aspects regarding emotion classification from bodily expressions with time-continuous input and output, as in most real-life scenarios:

—2D or 3D? Reviewing previous works, we have found that the studies focusing on the body modality often opted for various MoCap systems or depth sensors to capture the bodily expressions (e.g., Savva et al. [2012], Kapur et al. [2005], and Müller et al. [2015]), since the high precision and stability of the body tracking is usually desired. On the other hand, the mixed-modality works tended to analyze the bodily behaviors using multiple video cameras, with different zooms and setting angles (e.g., Glowinski et al. [2008], Gunes and Piccardi [2005], and Castellano et al. [2007b]), where the bodily signal was often a complementary channel, hence the loss of body information could be compensated by the rich information provided by other modalities. Although the extracted body information is largely influenced by the computer vision algorithms, recording environment, clothing, body parts/objects occlusion, etc., the video-based 2D representations are still preferable in those cases mainly due to the data acquisition convenience. In recent years, the advent of increasingly mature RGB-Depth sensors (e.g., the new generation Microsoft Kinect sensor,[3] Leap Motion sensor[4]) opened new perspectives to capture body movements. Thanks to the improved depth channel quality, as well as the reduced size, we could foresee that those sensors would be better integrated into real-life entertaining devices or companion humanoids, to provide richer and stable body behavioral information compared to the conventional 2D cameras, while eliminating the uncomfortableness introduced by MoCap systems.

—Feature Extraction: As already comprehensively studied in the literature, both the low-level body form features and high-level descriptors are capable of distinguishing emotions, under various circumstances. Furthermore, psychological experiments (e.g., Wallbott [1998] and de Meijer [1989]) and the body action coding systems (e.g., Dael et al. [2012b] and Gunes and Piccardi [2005]) indicated that the low-level and high-level body descriptors could both contribute to differentiate emotions. Hence, in our platform, we propose to combine the two sets of features in emotion classification tasks. A significant higher classification rate was achieved in our experiment, as expected.

As indicated in the study of Atkinson et al. [2007], the static form and the dynamic motion information complement each other for the emotion recognition tasks. The necessity of incorporating temporal information was increasingly raised in recent

[3]https://www.microsoft.com/en-us/kinectforwindows.
[4]https://www.leapmotion.com.

works. The most straightforward approach is to compute some statistical measures of each feature over the period of each expression. This has been used in the representative works of Castellano et al. [2007c] and Glowinski et al. [2008], where a set of statistical cues were used to model the temporal transitions of the original features, after segmentation of the input sequence. Such segmentation is not suitable for real-time applications, where continuous input-output mapping is desired. Therefore, in our work, we propose to use overlapped sliding windows to capture the temporal dynamics, either by concatenating the values in recent history, to form the temporal patches, or to compute the statistical measures over the windowed period. Both methods further enhanced the recognition rate, while the output continuity was kept. The window-based methods also serve as a low-pass filter, hence producing a smooth continuous prediction, as demonstrated in Figure 5.

—Prediction Model: Continuous emotion prediction under the classification scheme (to distinguish discrete emotion categories or quantized affective dimensions) is a relatively less addressed question in the research community, despite the applicational demand from, for instance, human-computer interaction, human-robot interaction, serious games, e-learning, clinical assistance, etc. Gunes and Schuller [2012]. The major issue is how to produce a stable and nonjerky output along time. Conventionally, the emotion classifications were carried out upon either single frames (e.g., Kleinsmith and Bianchi-Berthouze [2007]) or presegmented expressions (e.g., Castellano et al. [2007c]). Various classifiers were used, such as SVM, MLP-NN, HMM, Logistic Regression, CRF, etc. (refer to Section 6.7 of Gunes and Schuller [2012] for a comprehensive review). However, those methods could not be directly applied to real-time applications, due to either low computational efficiency (especially on high dimensionality) or poor capability to deal with temporal dynamics. Noticeably, a recent work of Meng and Bianchi-Berthouze [2014] proposed to use a hierarchical structure with a HMM model at the top layer to take into account the temporal dependency of the input. In our work, we used a more straightforward approach by feeding a Random Forests classifier with the window-strapped feature vectors. Although the feature dimension was considerably high, the computational cost remained at a very low level without sacrificing the stability and accuracy. Since each element in the concatenated vector is treated independently during both training and prediction, the spatial-temporal structure in the window could be well conserved. Furthermore, a novel Adaptive Random Forests classifier, which updates the model itself during prediction without labels, was proposed to improve the generalization capabilities, which is a less addressed problem in emotion prediction research. Shifting phenomenon is avoided by conserving the power of the original labeled training data. The proposed model requires minimum tuning effort to reach a good performance, unlike most other sophisticated algorithms. This approach is especially useful in two cases that are common in emotion recognition applications: (1) there are not sufficient labeled training data; and (2) the prediction needs to be gradually adapted to new users.

—Affective data: High-quality body emotion database with well-designed scenario and context, natural and realistic expressions, and reliable annotation, could not only train a good emotion predictor, but also serve as a means to evaluate methodologies. Constrained by the availability of 3D databases, in our experiments, for demonstrating and evaluating the proposed methodologies, we made used of a posed categorical emotion database. A recent work of Volkova et al. [2014] has reviewed the published emotion database that focused on the bodily expressions. The paper highlighted the urgent need in the affective computing community to create a naturalistic emotion corpus that includes 3D body modality. The authors made an effort to approach this objective, by obtaining "close-to-natural" expressions under narrative scenarios,

where the participants were not aware of the end purpose of the recordings. Indeed, much more believable emotional expressions were delivered, compared to the enacted ones, owing to several experimental settings (e.g., facial expressions and speech were also recorded to prevent the participants only focusing on the body language; amateur actors were recruited rather than professionals to reduce the exaggeration). However, as the participants were encouraged to emphasize their emotional expressions, artifacts were still not avoidable. Another attempt using a MoCap system to capture the naturalistic body behaviors during body-controlled console games was made in Savva et al. [2012]. Under this specific context, the authors were not able to record long expressive sequences containing multiple emotions and various expressions. Indeed, the body motions during the considered game were action-related without much freedom, thus not useful to other applications. To answer both issues of naturalness and longer recording, our recent study proposed to record a naturalistic multimodal emotion corpus in the child-robot turn-taking gaming scenario, and introduced the EMO-CHILD database [Wang et al. 2014]. The EMO-CHILD database intends to provide spontaneous bodily information in both 3D and 2D, as well as facial expressions, for various research purposes. A dual-Kinect setup, to capture the 3D skeletal motions, along with multiview cameras, to capture facial expressions and upper body movements, have been used. This nonintrusive recording setting ensured naturalistic expression and motion tracking precision. One of our future works is to label the occurrence of specific emotion categories, meeting the demand of a spontaneous and continuous emotional body expression corpus.

ACKNOWLEDGMENTS

The reviewing of this article was managed by Albert Ali Salah, Hayley Hung, Oya Aran, Hatice Gunes, and Matthew Turk. We also would like to thank the referees for their comments and suggestions, which helped improve this paper considerably.

REFERENCES

Jake K. Aggarwal and Quin Cai. 1997. Human motion analysis: A review. In *Proceedings of Nonrigid and Articulated Motion Workshop*. 90–102.

Kaat Alaerts, Evelien Nackaerts, Pieter Meyns, Stephan P. Swinnen, and Nicole Wenderoth. 2011. Action and emotion recognition from point light displays: An investigation of gender differences. *PLoS ONE* 6, 6 (Jan. 2011), e20989. DOI: http://dx.doi.org/10.1371/journal.pone.0020989

Anthony P. Atkinson, Winand H. Dittrich, Andrew J. Gemmell, and Andrew W. Young. 2004. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception* 33, 6 (2004), 717–746.

Anthony P. Atkinson, Mary L. Tunstall, and Winand H. Dittrich. 2007. Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures. *Cognition* 104 (2007), 59–72.

Tadas Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. el Kaliouby, P. Robinson, and R. Picard. 2011. Real-time inference of mental states from facial expressions and upper body gestures. In *Proceedings of 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG'11)*. 909–914. DOI: http://dx.doi.org/10.1109/FG.2011.5771372

Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer. 2012. Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion* 12, 5 (Oct. 2012), 1161–1179. DOI: http://dx.doi.org/10.1037/a0025827

Tony Belpaeme, Paul E. Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, Rosemarijn Looije, Mark Neerincx, Yiannis Demiris, Raquel Ros-Espinoza, Aryel Beck, Lola Cañamero, Antione Hiolle, Matthew Lewis, Ilaria Baroni, Marco Nalin, Piero Cosi, Giulio Paci, Fabio Tesser, Giacomo Sommavilla, and Remi Humbert. 2012. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction* 1, 2 (2012), 33–53. DOI: http://dx.doi.org/10.5898/jhri.v1i2.62

Daniel Bernhardt. 2010. *Emotion Inference From Human Body Motion*. Technical Report 787. Computer Laboratory, University of Cambridge, Cambridge, 227 pages.

Nadia Bianchi-Berthouze, P. Cairns, and A. L. Cox. 2008. On posture as a modality for expressing and recognizing emotions. In *Joint Proceedings of the 2005, 2006, and 2007 International Workshops on Emotion in HCI*. Citeseer, 74–80.

Nadia Bianchi-Berthouze and Andrea Kleinsmith. 2003. A categorical approach to affective gesture recognition. *Connection Science* 15, 4 (Dec. 2003), 259–269. DOI:http://dx.doi.org/10.1080/09540090310001658793

Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

Antonio Camurri, Paolo Coletta, Alberto Massari, Barbara Mazzarino, Massimiliano Peri, Matteo Ricchetti, Andrea Ricci, Gualtiero Volpe, and Max Msp. 2004. Toward real-time multimodal processing: EyesWeb 4.0. In *Proceedings of 2014 Convention on Motion, Emotion and Cognition*, Vol. 1.

Antonio Camurri, Ingrid Lagerlöf, and Gualtiero Volpe. 2003. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies* 59, 1–2 (July 2003), 213–225. DOI:http://dx.doi.org/10.1016/S1071-5819(03)00050-8

Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. 2008. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning* (2008), 96–103. DOI:http://dx.doi.org/10.1145/1390156.1390169

Ginevra Castellano, Antonio Camurri, Barbara Mazzarino, and Gualtiero Volpe. 2007a. A mathematical model to analyse the dynamics of gesture expressivity. In *Proceedings of 2007 Convention on Artificial and Ambient Intelligence*. Newcastle upon Tyne, UK, 1–6.

Ginevra Castellano, Loic Kessous, and George Caridakis. 2007b. Multimodal emotion recognition from expressive faces, body gestures and speech. *Affective Computing and Intelligent Interaction* 4738 (2007), 71–82.

Ginevra Castellano, S. Villalba, and Antonio Camurri. 2007c. Recognising human emotions from body movement and gesture dynamics. In *Proceedings of 2nd International Conference on Affective Computing and Intelligent Interaction*. Springer, 71–82.

Nele Dael, Marcello Mortillaro, and Klaus R. Scherer. 2012a. Emotion expression in body action and posture. *Emotion* 12, 5 (Oct. 2012), 1085–1101. DOI:http://dx.doi.org/10.1037/a0025737

Nele Dael, Marcello Mortillaro, and Klaus R. Scherer. 2012b. The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior* 36, 2 (Jan. 2012), 97–121. DOI:http://dx.doi.org/10.1007/s10919-012-0130-0

Beatrice de Gelder. 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Tranactions of the Royal Society B: Biological Sciences* 364 (2009), 3475–3484.

Marco de Meijer. 1989. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior* 13, 4 (1989), 247–268.

P. Ravindra De Silva, Ajith P. Madurapperuma, Ashu Marasinghe, and Minetada Osano. 2006. A multi-agent based interactive system towards child's emotion performances quantified through affective body gestures. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. 1236–1239.

Ellen Douglas-cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of Affective Computing and Intelligent Interaction*, Vol. 4738. 488–500.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63 (2006), 3–42.

Donald Glowinski, Antonio Camurri, Gualtiero Volpe, Nele Dael, and Klaus R. Scherer. 2008. Technique for automatic emotion recognition by body gesture analysis. In *Proceedings of IEEE Workshops on Computer Vision and Pattern Recognition*. 1–6. DOI:http://dx.doi.org/10.1109/CVPRW.2008.4563173

Donald Glowinski, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus R. Scherer. 2011. Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing* 2, 2 (April 2011), 106–118. DOI:http://dx.doi.org/10.1109/T-AFFC.2011.7

Harry J. Griffin, Min S. H. Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. 2013. Laughter type recognition from whole body motion. In *Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 349–355. DOI:http://dx.doi.org/10.1109/ACII.2013.64

Hatice Gunes and Massimo Piccardi. 2005. Fusing face and body gesture for machine recognition of emotions. In *Proceedings of IEEE International Workshop on Robot and Human Interactive Communication*. 306–311.

Hatice Gunes and Massimo Piccardi. 2006. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proceedings of the International Conference on Pattern Recognition*, Vol. 1. 1148–1153.

Hatice Gunes and Massimo Piccardi. 2009. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics. Part B* 39, 1 (Feb. 2009), 64–84. DOI: http://dx.doi.org/10.1109/TSMCB.2008.927269

Hatice Gunes and Björn Schuller. 2012. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* (July 2012). DOI: http://dx.doi.org/10.1016/j.imavis.2012.06.016

Ian H. Hall, Mark Frank, Eibe Holmes, Geoffrey Pfahringer, Bernhard Reutemann, and Peter Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

Kanav Kahol, Priyamvada Tripathi, and Sethuraman Panchanathan. 2003. Gesture segmentation in complex motion sequences. In *Proceedings of the International Conference on Image Processing*.

Asha Kapur, Ajay Kapur, Naznin Virji-babul, George Tzanetakis, and Peter F. Driessen. 2005. Gesture-based affective computing on motion capture data. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII'05)*. 1–7.

Michelle Karg, Ali-akbar Samadani, Rob Gorbet, K. Kolja, Jesse Hoey, and Dana Kulic. 2013. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing* 4, 4 (2013), 341–359.

Andrea Kleinsmith. 2005. An incremental and interactive affective posture recognition system. In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors*.

Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2007. Recognizing affective dimensions from body posture. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*. 45–58.

Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing* 4, 1 (Jan. 2013), 15–33. DOI: http://dx.doi.org/10.1109/T-AFFC.2012.16

Andrea Kleinsmith, Nadia Bianchi-Berthouze, and Anthony Steed. 2011. Automatic recognition of non-acted affective postures: A video game scenario. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society* 41, 4 (Jan. 2011), 1027–1038. DOI: http://dx.doi.org/10.1109/TSMCB.2010.2103557

Andrea Kleinsmith, P. R. De Silva, and Nadia Bianchi-Berthouze. 2005. Recognizing emotion from postures: Cross-cultural differences in user modeling. *Lecture Notes in Artificial Intelligence* (2005), 50–59.

Andrea Kleinsmith, P. Ravindra De Silva, and Nadia Bianchi-Berthouze. 2006. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers* 18, 6 (2006), 1371–1389.

Karl H. E. Kroemer, Henrike B. Kroemer, and Katrin E. Kroemer-Elbert. 1994. *Ergonomics: How to Design for Ease and Efficiency*. Prentice Hall.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2012), 5–17.

Albert Mehrabian. 2007. Nonverbal communication.

Hongying Meng and Nadia Bianchi-Berthouze. 2014. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Systems, Man, and Cybernetics Part B* 44, 3 (2014), 315–328.

Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. 2012. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing* (Sept. 2012). DOI: http://dx.doi.org/10.1016/j.imavis.2012.08.018

Angeliki Metallinou, Athanassios Katsamanis, Yun Wang, and Shrikanth Narayanan. 2011. Tracking changes in continuous emotion states using body language and prosodic cues. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2. 2288–2291.

Philipp M. Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *Proceedings of the 6th Affective Computing and Intelligent Interaction (ACII'15)*.

Joseph Onderi Orero, Florent Levillain, Marc Damez-Fontaine, Maria Rifqi, and Bernadette Bouchon-Meunier. 2010. Assessing gameplay emotions from physiological signals. In *Proceedings of the International Conference on Kansei Engineering and Emotion Research*. 1684–1693.

Geovany A. Ramirez, Tadas Baltrusaitis, and Louis-Philippe Morency. 2011. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, Vol. 2.

Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. 2009. On-line random forests. In *Proceedings of the IEEE 12th Internatinal Conference on Computer Vision Workshops*. 1393–1400. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5457447.

Jyotirmay Sanghvi, Ginevra Castellano, Ana Paiva, and Peter W. Mcowan. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion categories and subject descriptors. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction*.

Nikolaos Savva and Nadia Bianchi-Berthouze. 2012. Automatic recognition of affective body movement in a video game scenario. In *Proceedings of the International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*. 149–158.

Nikolaos Savva, Alfonsina Scarinzi, and Nadia Bianchi-Berthouze. 2012. Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 3 (Sept. 2012), 199–212. DOI:http://dx.doi.org/10.1109/TCIAIG.2012.2202663

Caifeng Shan and Shaogang Gong. 2007. Beyond facial expressions: Learning human emotion from body gestures. In *Proceedings of the British Machine Vision Conference*.

Harold Soh. 2012. Online spatio-temporal Gaussian process experts with application to tactile classification. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Feng Tang, Shane Brennan, Qi Zhao, and Hai Tao. 2007. Co-tracking using semi-supervised support vector machines. In *Proceedings of the IEEE 11th International Conference on Computer Vision* (2007), 1–8. DOI:http://dx.doi.org/10.1109/ICCV.2007.4408954

Michel Valstar, Florian Eyben, Gary Mckeown, Roddy Cowie, and Maja Pantic. 2011. AVEC 2011 the first international audio/visual emotion challenge. In *Proceedings of Affective Computing and Intelligent Interaction*.

Ekaterina Volkova, Stephan de la Rosa, Heinrich H. Bülthoff, and Betty Mohler. 2014. The MPI emotional body expressions database for narrative scenarios. *PLoS ONE* 9, 12 (2014), e113647. DOI:http://dx.doi.org/10.1371/journal.pone.0113647

Harald G. Wallbott. 1998. Bodily expression of emotion. *European Journal of Social Psychology* 28, 6 (1998), 879–896.

Weiyi Wang, Geogiors Athanasopoulos, Selma Yilmazyildiz, Geogiors Patsis, Valentin Enescu, Hichem Sahli, Werner Verhelst, Antoine Hiolle, Matthew Lewis, and Lola Cañamero. 2014. Natural emotion elicitation for emotion modeling in child-robot interactions. In *Proceedings of the 4th Workshop on Child-Computer Interaction (WOCCI'14)*.

Weiyi Wang, Valentin Enescu, and Hichem Sahli. 2013. Towards real-time continuous emotion recognition from body movements. In *Proceedings of Human Behavior Understanding 2013*. Lecture Notes in Computer Science, Vol. 8212. 235–245.

Received March 2014; revised September 2015; accepted September 2015