Survey on Emotional Body Gesture Recognition

Fatemeh Noroozi, Dorota Kamińska, Ciprian Adrian Corneanu, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari,

Abstract—Automatic emotion recognition has become a trending research topic in the past decade. While works based on facial expressions or speech abound, recognising affect from body gestures remains a less explored topic. We present a new comprehensive survey hoping to boost research in the field. We first introduce emotional body gestures as a component of what is commonly known as "body language" and comment general aspects as gender differences and culture dependence. We then define a complete framework for automatic emotional body gesture recognition. We introduce person detection and comment static and dynamic body pose estimation methods both in RGB and 3D. We then comment the recent literature related to representation learning and emotion recognition from images of emotionally expressive gestures. We also discuss multi-modal approaches that combine speech or face with body gestures for improved emotion recognition. While pre-processing methodologies (e.g. human detection and pose estimation) are nowadays mature technologies fully developed for robust large scale analysis, we show that for emotion recognition the quantity of labelled data is scarce, there is no agreement on clearly defined output spaces and the representations are shallow and largely based on naive geometrical representations.

Index Terms—emotional body language, emotional body gesture, emotion recognition, body pose estimation, affective computing

1 Introduction

During conversations people are constantly changing nonverbal clues, communicated through body movement and facial expressions. The difference between words people pronounce and our understanding of their content comes from nonverbal communication also commonly called body language [1]. Some examples of body gestures and postures, key components of body language are shown in Fig. 1.

Although it is a significant aspect of human social psychology, the first modern studies concerning body language have become popular in the 1960s [2]. Probably the most important work published before the 20th century was *The Expression of the Emotions in Man and Animals* by Charles Darwin [3]. This work is the foundation of modern body language research. Many of Darwin's observations were confirmed by subsequent studies. Darwin observed that people all over the world use facial expressions in a fairly similar manner. Following this observation, Paul Ekman researched patterns of facial behavior among different cultures of the world. In 1978, Ekman and Friesen developed the Facial Action Coding System (FACS) to model human facial expressions [4]. In an updated form, this descriptive anatomical model is still being used in emotion expressions recognition.

Figure 1. Body language includes different types of nonverbal indicators such as facial expressions, body posture, gestures and eye movements. These are important markers of the emotional and cognitive inner state of a person. In this work, we review the literature on automatic recognition of body expressions of emotion, a subset of body language that focuses on gestures and posture of the human body. The images have been taken from [5].

An interesting study of the usage of body language for emotion recognition was conducted by Ray Birdwhistell who found that the final message of an utterance is affected only 35% by the actual words and 65% by nonverbal signals [6]. In the same work, analysis of thousands of negotiations recordings revealed that the body language decides the outcome of those negotiations in 60% - 80% of cases. During a phone negotiation, stronger arguments win, however during a personal meeting, decisions are made on

- *F. Noroozi was with Institute of Technology, University of Tartu, Estonia. E-mail: fatemeh.noroozi@ut.ee*
- *C. A. Corneanu and S. Escalera are with the Universitat de Barcelona and Computer Vision Center, Barcelona, Spain. E-mail: cipriancorneanu@gmail.com, sergio@maia.ub.es*
- *D. Kamińska and T. Sapiński are with Institute of Mechatronics and Information Systems, Lodz University of Technology, Lodz, Poland. E-mail: dorota.kaminska@p.lodz.pl, sapinski.tomasz@gmail.com*
- *G. Anbarjafari are with the iCV Research Group, Institute of Technology, University of Tartu, Tartu, Estonia. G. Anbarjafari is also with Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep, Turkey. E-mail: shb@ut.ee*

Manuscript received January 19, 2018; revised Xxxx XX, 201X.

the basis of what we see rather than what we hear [2]. At the present, most researchers agree that words serve primarily to convey information and the body movements to form relationships and sometimes even to substitute the verbal communication (e.g. lethal look).

Gestures are one of the most important forms of nonverbal communication. They include movements of hands, head and other parts of the body that allow individuals to communicate a variety of feelings, thoughts and emotions. Most of the basic gestures are the same all over the world: when we are happy we smile when we are upset we frown [7], [8], [9].

According to [2], gestures can be of the following types:

- *Intrinsic.* Nodding as a sign of affirmation or consent is probably innate - even people who are born blind use it;
- *Extrinsic*. Turning to the sides as a sign of refusal is a gesture we learn during early childhood. For example, babies turn their heads when they have had enough milk from their mother's breasts, or older children when they refuse a spoon with food during feeding;
- *A result of natural selection*. For example, the expansion of nostrils to oxygenate the body when preparing for battle or escape.

The ability to recognize attitude and thoughts from one's behavior was the original system of communication before speech. Understanding of emotional state enhances interaction. Although computers are now a part of human life, the relation between human and machine is not natural. Knowledge of the emotional state of the user would allow the machine to better adapt and generally improve cooperation.

While emotions can be expressed in different ways, automatic recognition has mainly focused on facial expressions and speech [10]. Considerably less works were done on body gestures and posture. With recent developments of motion capture technologies and reliability, the literature about automatic recognition of expressive movements grew significantly.

Despite the increasing interest in this topic, we are aware of just a few relevant survey papers. For example, Kleinsmith et al. [11] reviewed the literature on affective body expression perception and recognition with an emphasis on inter-individual differences, impact of culture and multimodal recognition. In another paper, Kara et al. [12] introduced categorization of movement into four types: communicative, functional, artistic, and abstract and discussed the literature associated with these types of movements.

In this work, we cover all the recent advancements in automatic emotion recognition from body gestures. The reader interested in emotion recognition from facial expressions or speech is encouraged to consult dedicated surveys [1], [13], [14]. We refer to these only marginally and only as complements to emotional body gestures. In Sec. 2 we briefly introduce key aspects of affect expression through body language in general and we discuss in-depth cultural and gender dependency. Then, we define a standard pipeline for automatically recognizing body gestures of emotion in Sec. 4 and we discuss technical aspects of each component. Furthermore, in Sec. 5 we provide a comprehensive review of publicly available databases for training automatic recognition systems. We conclude in Sec. 6 with discussions and potential future lines of research.

2 Expressing emotion through body language

According to [15], [16] body language includes different kinds of nonverbal indicators such as facial expressions, body posture, gestures, eye movement, touch and the use of personal space. The inner state of a person is expressed through elements such as iris extension, gaze direction, position of hands and legs, the style of sitting, walking, standing or lying, body posture, and movement. Some examples are presented in Fig. 1.

After the face, hands are probably the richest source of body language information [17], [18]. For example, based on the position of hands one is able to determine whether a person is honest (one will turn the hands inside towards the interlocutor) or insincere (hiding hands behind the back). Exercising open-handed gestures during conversation can give the impression of a more reliable person. It is a trick often used in debates and political discussions and people using open-handed gestures are perceived positively [2]. For example, a study of two groups of speakers in public, showed that the ones using open gestures were perceived positively by 85% of their recipients, whereas those who had their palms facing downwards were evaluated as positive only by 52% of the receivers [2].

Head positioning also reveals a lot of information about emotional state. Research [7] indicates that people are prone to talk more if the listener encourages them by nodding. The pace of the nodding can signal patience or lack of it. In neutral position the head remains still in front of the interlocutor. If the chin is lifted it may mean that the person is displaying superiority or even arrogance. Exposing the neck might be a signal of submission. In [3] Darwin noted that like animals, people tilt their heads when they are interested in something. That is why women perform this gesture when they are interested in men. An additional display of submission results in greater interest from the opposite sex, e.g. a lowered chin signals a negative or aggressive attitude.

Torso is probably the least analysed part of the body [11] [12]. However, its angle with the body is an indicative attitude. For example placing the torso frontally to the interlocutor can be considered as a display of aggression. By turning it at a slight angle one may be considered self-confident and devoid of aggression. Leaning forward, especially when combined with nodding and smiling, is the most distinct way to show curiosity [7].

The above considerations indicate that in order to correctly interpret body language as indicators of emotional state, various parts of body must be considered at the same time. According to [19], body language recognition systems may benefit from a variety of psychological behavioural protocols. An example of general movements protocol for six basic emotions is presented in Table 1.

Table 1
The general movement protocols for the six basic emotions [20], [21], [22].

Emotion	Associated body language
Fear	Noticeably high heart beat-rate (visible on the neck). Legs and arms crossing and moving. Muscle tension: Hands or arms clenched, elbows dragged inward, bouncy movements, legs wrapped around objects. Breath held. Conservative body posture. Hyper-arousal body language.
Anger	Body spread. Hands on hips or waist. Closed hands or clenched fists. Palm-down posture. Lift the right or left hand up. Finger point with right or left hand. Finger or hand shaky. Arms crossing.
Sadness	Body dropped. Shrunk body. Bowed shoulders. Body shifted. Trunk leaning forward. The face covered with two hands. Self-touch (disbelief), body parts covered or arms around the body or shoulders. Body extended and hands over the head. Hands kept lower than their normal positions, hands closed or moving slowly. Two hands touching the head and moving slowly. One hand touching the neck. Hands closed together. Head bent.
Surprise	Abrupt backward movement. One hand or both of them moving toward the head. Moving one hand up. Both of the hands touching the head. One of the hands or both touching the face or mouth. Both of the hands over the head. One hand touching the face. Self-touch or both of the hands covering the cheeks or mouth. Head shaking. Body shift or backing.
Happiness	Arms open. Arms move. Legs open. Legs parallel. Legs may be stretched apart. Feet pointing something or someone of interest. Looking around. Eye contact relaxed and lengthened.
Disgust	Backing. Hands covering the neck. One hand on the mouth. One hand up. Hands close to the body. Body shifted. Orientation changed or moving to a side. Hands covering the head.

2.1 Culture differences

It has been reported that gestures are strongly culture-dependent [23], [24]. However, due to exposure to mass-media, there is a tendency of globalization of some gestures especially in younger generations [7]. This is despite the fact that the same postures might have been used for expressing significantly different feelings by their previous generations. Consequently, over time, some body postures might change in meaning, or even disappear. For instance, the thumb-up symbol might have different meanings in different cultures. In Europe it stands for number "1" in Japan for "5", while in Australia and Greece, using it may be considered insulting. However, nowadays, it is widely used as a sign of agreement, consent or interest [25].

Facial expressions of emotion are similar across many cultures [26]. This might hold in the case of postures as well. In [27], the effect of culture and media on emotional expressions was studied. One of the conclusions was that an American and a Japanese infant present closely similar emotional expressions. Most of the studies reported on this topic in the literature inferred that intrinsic body language, gestures and postures are visibly similar throughout the world. However, a decisive conclusion still requires more in-depth exploration, which is challenging due to the variety of topics that need to be studied on numerous cultures and countries. Therefore, the researchers investigating this issue prefer to concentrate on a certain activity, and study it on various cultures, which may lead to a more understandable distinction. For example, in many cultures, holding hands resembles mutual respect, but in some others touching one another in exchanging greetings might not be considered usual [25].

2.2 Gender differences

There are some fundamental differences in the way women and man communicate through body language [28] (see Fig. 2 for some trivial examples). This may be caused by influence of culture (tasks and expectations that face both sexes), body composition, makeup and worn type of clothes.

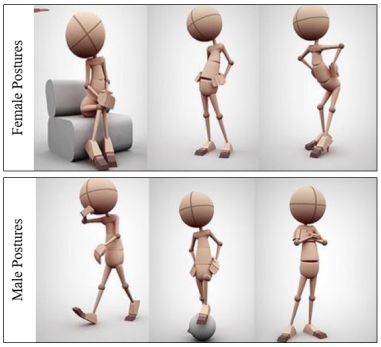

Figure 2. There are fundamental differences in the way men and women communicate through body language. In certain situations, one can easily discriminate the gender when only the body pose is shown. Illustration from [28].

Women wearing skirts often sit with crossed legs or ankles. But it is not the sole reason of this gesture applying almost exclusively to women. As a result of body composition, most men are not able to sit that way, thus this position became a symbol of femininity. Another good example is the pose with wide-spread legs (a wide standing pose), which is mainly attributed to men. Generally men use this gesture unconsciously, it demonstrates their courage and domination. A similar position has been also observed among monkeys [2].

Women show emotions and their feelings more willingly than men [29]. For example women tend to display overt signs of sadness while men tend to withdraw such expressions [30]. On the other hand, man are more likely to display power and dominance [29]. It has been shown that while being angry with equal intensity and frequency, men tend to manifest indicators of anger more evidently than women [30]. However, nowadays these general tendencies start to faint and are considered gender stereotypes [31].

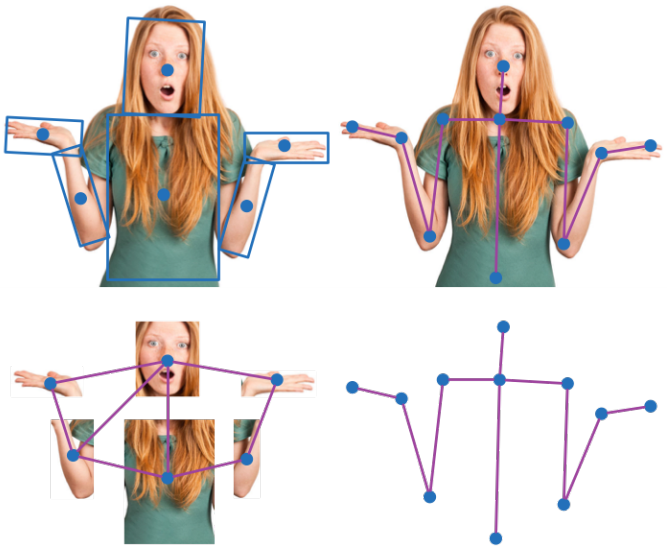

Figure 3. The two most common ways of modelling the human body in automatic body gesture recognition are either as an ensemble of body parts or as a kinematic model. An ensemble of body parts (left) groups different parts of the body which are independently detected. Soft restrictions can be imposed to refine these detections. A kinematic model (right) is a collections of interconnected joints with predefined degrees of freedom similar to the human skeleton.

3 Models of the Human Body and Emotion

Before discussing the main steps for automatically recognizing emotion from body gestures, (details in Sec. 4), we will first introduce the modelling of the input and output of such systems. The input will be an abstraction of the human body (and its dynamics) that we would like to map through machine learning methods to a certain predefined abstraction of emotion. Deciding the appropriate way of modelling the human body and emotion is an essential design decision. We begin by discussing abstractions of the human body (Sec. 3.1) and then main models of emotion used in affective computing (Sec. 3.2).

3.1 Models of the human body

Human body has evolved such that it can perform complex actions, which require coordination of various organs. Therefore, many everyday actions present unique spatio-temporal movement structures [32]. In addition, some pairs of body actions, e.g. walking and drinking, may be performed at the same time and their expression might not be additive [33]. The two main lines for abstracting the human body have been following either a constrained composition of human body parts or a kinematic logic based on the skeletal structure of the human body (see Fig. 3).

Part Based Models. In a part based approach the human body is represented as flexible configuration of body parts. Body parts can be detected independently (face, hands, torso) and priors can be imposed using domain knowledge of the human body structure to refine such detection. Some examples of ensemble of parts models of the human body are pictorial structures and grammar models. Pictorial structures are generative 2D assemblies of parts, where each part is detected with its specific detector. Pictorial structures

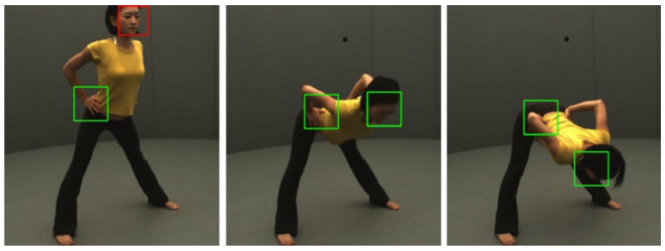

Figure 4. Example of body pose estimation and tracking using ensemble of parts namely, head and hands [35].

are a general framework for object detection widely used for people detection and human pose estimation [34]. An example of body pose estimation using pictorial structures is shown in Fig. 4.

Grammar models provide a flexible framework for detecting objects, which was also applied for human detection in [36]. Compositional rules are used to represent objects as a combination of other objects. In this way, human body could be represented as a composition of trunk, limbs and face; as well composed by eyes, nose and mouth.

Kinematic Models. Another way of modelling the human body is by defining a collection of interconnected joints also known as kinematic chain models. This is usually a simplification of the human skeleton and its mechanics. A common mathematical representation of such models is through a cyclical tree graphs which also present the advantage of being computationally convenient. Contrary to part based approach [34], nodes of structure trees represent joints, each one parameterized with its degrees of freedom. Kinematic models can be planar, in which case they are a projection in the image plane or depth information can be considered as well. Richer, more realistic variants can be defined for example as a collection of connected cylinders or spheroids or 3D meshes [37], [38]. Some examples of body pose detection using kinematic models and deep learning methods can be seen in Fig. 5.

3.2 Models of emotion

The best way of modelling affect has been subject of debate for a long time and many perspectives upon the topic were proposed. The most influential models (and in general most relevant for affective computing applications) can be classified in three main categories: categorical, dimensional and componential [45] (see Fig. 6 for examples of each category).

Categorical models. Classifying emotions into a set of distinct classes that can be recognized and described easily in daily language has been common since at least the time of Darwin. More recently, influenced by the research of Paul Ekman [46], [49] a dominant view upon affect is based on the underlying assumption that humans universally express and recognize a set of discrete primary emotions which include happiness, sadness, fear, anger, disgust, and surprise. Mainly because of its simplicity and its universality claim, the universal primary emotions hypothesis has been used intensively in affective computing research.

Dimensional models. Another popular approach is to model emotions along a set of latent dimensions [47], [50], [51]. These dimensions include valence (how pleasant or

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874986, IEEE Transactions on Affective Computing

(a) (b) (c) (d) (e) (f) (g)

Figure 5. Examples of single-person pose estimation (a) [39], (b) [40], (c) [41], (d) [42], multi-person pose estimation (e) [43], multi-person pose estimation and tracking (f) [44] and 3D shape reconstruction [38]

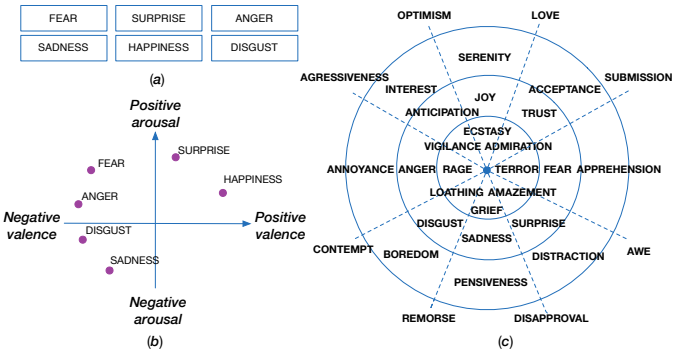

Figure 6. There are three main categories of emotion models in affective computing. Categorical (a) (here a universal set of emotions as define by Ekman [46]), (b) dimensional (Russell's model depicted) [47], and componential models (c) (Plutchik's model shown) [48].

unpleasant a feeling is) arousal[1] (how likely is the person to take action under the emotional state) and control (the sense of control over the emotion). Due to their continuous nature, such models can theoretically describe more complex and subtle emotions. Unfortunately, the richness of the space is more difficult to use for automatic recognition systems because it can be challenging to link such described emotion to a body expression of affect. This is why, many automatic systems based on dimensional representation of emotion simplify the problem by dividing the space in a limited set of categories like positive vs negative or quadrants of the 2D space [52].

Componential models. Somehow in-between categorical and dimensional models in terms of descriptive capacity, componential models of affect arrange emotions in a hierarchical fashion where each superior layer contains more complex emotions which can be composed of emotions of previous layers. One of the best known examples of componential models was proposed by Plutchik [48]. According to his theory, more complex emotions are pairs of more basic emotions called dyads and the more complex an emotion is, the lower the probability to occur. While this model is rarely used in affective computing compound emotions have gained relative visibility [53]. In a similar fashion, the compound emotion model is breaking from the former restrictive framework of categorical universal emotions by proposing to combine basic component categories to construct new emotions like Happily Surprised and Angrily Surprised. A richer emotion and emotional display can be defined in this way. In general, these models provide an effective compromise between ease of interpretation and expressive capacity which could be useful in building discriminative computational models of affective display.

1. or activation

4 Body Gesture Based Emotion Recognition

In this section, we present the main components of what we call an Emotional Body Gesture Recognition (EBGR) system. For a detailed depiction see Fig. 7. An important preparation step, which influences all the subsequent design decisions for such an automatic pipeline is the determination of the appropriate modelling of input (human body) and targets (emotion). Depending on the type of the model that has been chosen, either a publicly accessible database can be utilized, or a new one needs to be created. Similarly, other elements of the system need to be selected and configured such that they are compatible with each other, and overall, provide an efficient performance. Regardless of the foregoing differences between various types of EBGR systems, the common first step is to detect the body as a whole, i.e. to subtract the background from every frame which represents a human presenting a gesture. We will briefly discuss the main literature for human detection in Sec. 4.1. The second step is detection and tracking of the human pose in order to reduce irrelevant variation of data caused by posture (we dedicate Sec. 4.2 to this). The final part of the pipeline, which we discuss in Sec. 4.3, consists in building an appropriate representation of the data and applying a learning technique (usually classification or regression) to map this representation to the targets. We conclude this section with a presentation of the most important applications of automatic recognition of emotion using body gesture.

4.1 Human Detection

Human detection in images usually consists in determining rectangular bounding boxes that enclose humans. It can be a challenging task because of the non-rigid nature of the human body, pose and clothing, which result in high variation of appearance. In uncontrolled environments changes of illumination and occlusions add to the complexity of the problem.

A human detection pipeline follows the general pipeline of object detection problems and consists of extracting potential candidate regions, representing those regions, classifying the regions as human or non-human, and merging

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874986, IEEE Transactions on Affective Computing

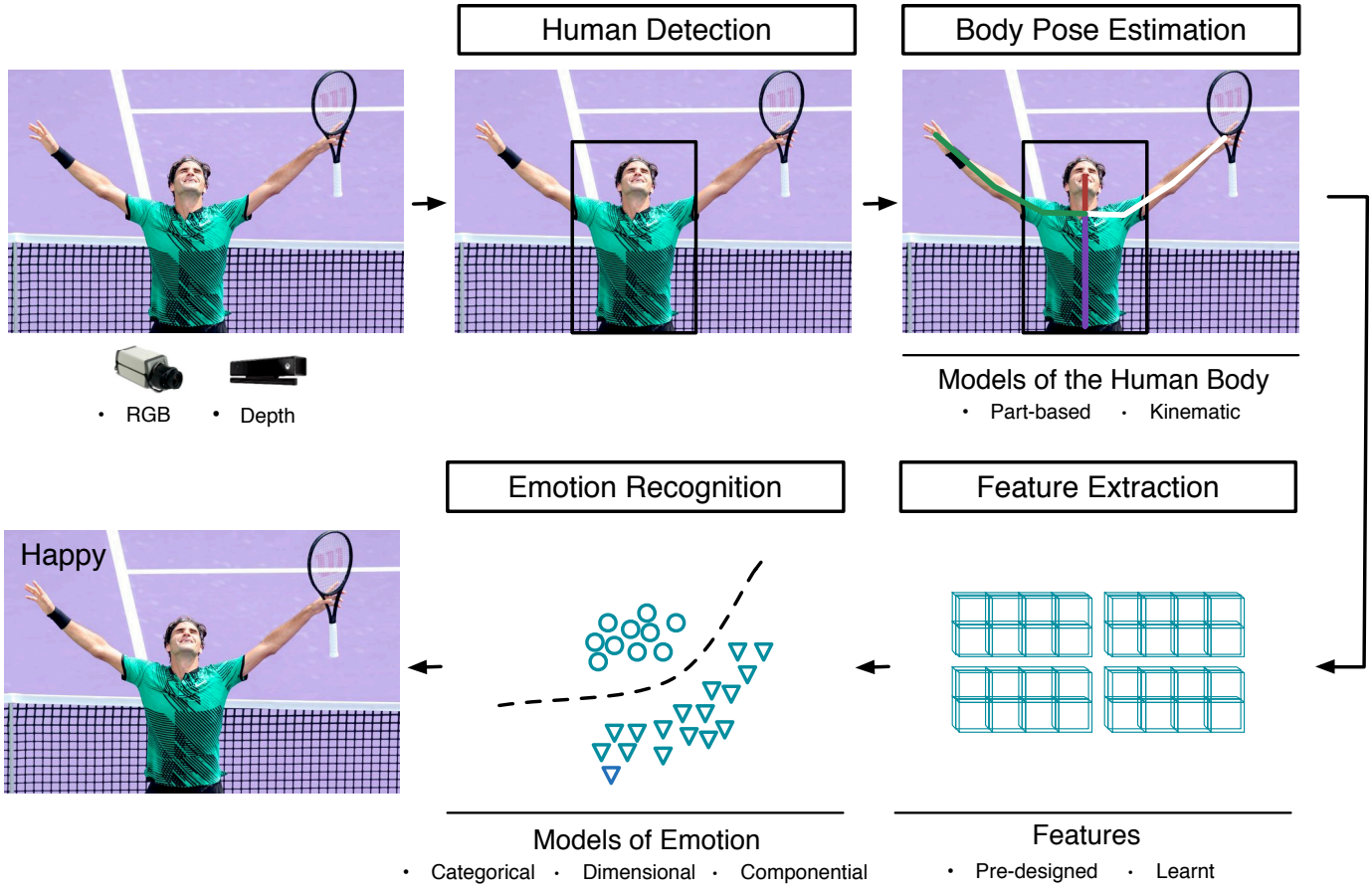

Figure 7. General overview of an Emotional Body Gesture Recognition System. After detecting persons in the input for background extraction, a common step is to estimate the body pose. This is done either by detecting and tracking different parts of the body (hands, head, torso, etc.) or by mapping a kinematic model (a skeleton) to the image. Based on the extracted model of the human body, a relevant representation is extracted or learned in order to map the input to a predefined emotion model using automatic pattern recognition methods.

positives into final decisions [54]. If depth information is available, it can be used to limit the search space and considerably simplify the background substraction problem [54]. Modern techniques might not exactly follow this modularization, either by jointly learning representation and classification or by directly proposing detection regions from input.

One of the first relevant methods for human detection was propose by Viola and Jones [55]. Following a method previously applied to face detection, it employs a cascade structure for efficient detection, and utilizing AdaBoost for automatic feature selection [55].

An important advancement in performance came with the adoption of gradient-based features for describing shape. Dalal and Triggs, popularized the so called histogram of oriented gradient (HOG) features for object detection by showing substantial gains over intensity based features [56]. Since their introduction, the number of variants of HOG features has proliferated greatly with nearly all modern detectors utilizing them in some form [57].

Earlier works on human detection assumed no prior knowledge over the structure of the human body. Arguably one of the most important contributions in this direction was the Deformable Part Models (DPM) [58]. A DPM is a set of parts and connections between the parts which relate to a geometry prior. In the initial proposal by Felzenswalb et al. a discriminative part based approach models unknown part positions as latent variables in a support vector machine (SVM) framework. Local appearance is easier to model than global appearance and training data can be shared across deformations. Some authors argued that there is still no clear evidence for the necessity of components and parts, beyond the case of occlusion handling [59].

Lately, a spectacular rise in performance in many pattern recognition problems was brought by training deep neural networks (DNN) with massive amounts of data. While providing good performance, earlier DNN models tended to be slow, especially when used as sliding-window classifiers. A considerable amount of work focused on accelerating and improving the potential region proposal process. One way to alleviate this problem was to use several networks in a cascaded fashion. For example, a smaller, almost shallow network can be trained to greatly reduce the initially large number of candidate regions produced by the sliding window. Then in a second step, only high confidence regions were passed through a deep network obtaining in this way a trade-off between speed and accuracy [60]. The idea of cascading any kind of features of different complexity, including deeply learned features was addressed by seeking an algorithm for optimal cascade learning under a crite-

rion that penalizes both detection errors and complexity. This made possible to define quantities such as complexity margins and complexity losses, and account for these in the learning process. This algorithm was shown to select inexpensive features in the early cascade stages, pushing the more expensive ones to the later stages [61]. Currently, the usage of regional based CNNs has become the standard in human detection. One of the most successful techniques originated with the Fast R-CNN [62] which proposed to use a single state training, a multi-tasking loss (combining classification and region location) and a shared feature space. Additionally, Faster-RCNN [63] and its adaptation to human detection [64] introduced a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. While initially, these methods would achieve 5fps using a VGG-16, current versions can achieve real-time human detection with high performance using very deep architectures like ResNet [65].

For a comprehensive survey of the human detection literature, the interested reader is referred to [54] and [57], [59].

4.2 Body Pose Estimation

Due to the high dimensions of the search space and the large number of degrees of freedom, as well as variations of cluttered background, body parameters and illumination, human pose estimation is a challenging task [66], [67]. It also demands avoiding body part penetration and impossible positions.

It can be performed using either model fitting or learning. Model-based methods fit an expected model to the captured data, as an inverse kinematic problem [68], [69]. In this context, the parameters can be estimated based on the tracked feature points, using gradient space similarity matching and the maximum likelihood resulted from the Markov Chain Monte Carlo approach [70]. However, model-based methods are not robust against local extrema, and require initialisation [66].

Performing pose estimation using learning is computationally expensive, because of the high dimensions of the data, and requires a large database of labelled skeletal data. However, recent advances in computational power and availability of data made training large capacity models feasible such that starting with the DeepPose [71], the field experienced a considerable change from using traditional approaches to deep neural networks. Currently, state-of-the-art in human pose estimation from single monocular images [39], [40], [41], [42], [72], [73], [74], [75] and from videos [76], or multi-person pose estimation [77], [78], [79] and 3D pose estimation [80] are all based on deep learning methods.

One of the main directions of research in recent years in deep learnt pose estimation models focused on how to efficiently combine local with global information for both precise joint localisation and overall skeletal consistency. One way was to combine deep neural networks with graphical models for learning spatial relationship between joints [39]. Due to better appearance consistency, parts of the body like the head and the shoulders are detected with increased accuracy than other parts of the body. Learning local spatial context can be used to sequentially improve these predictions by leveraging the fact that parts occur in consistent geometric configurations [42]. Similarly, Bulat et al. proposed a cascaded network to explicitly infer part relationships to improve inter-joint consistency. It contains a part detection network and deep regression network responsible for regressing the location of all parts and impose consistency. It is trained via confidence map regression [75]. Motivated by the same idea, several methods propose to capture information at multiple scales. A particularly successful proposal is the so-called Hourglass Network that uses skip-connections to promote multi-scale feature learning [40]. Several of these modules are stacked, feeding the output of one as input into the next. This provides a mechanism for repeated bottom-up, top-down inference. In this way, initial estimates and features can be reevaluated and improved across the whole image [40].

Based on the observation that pooling mechanisms in convolutional networks reduce localisation accuracy, an alternative dropout implementation was introduced with improved performance [72]. Spatial precision localization is recovered by using a fine-to-coarse architecture. Instead of predicting the outputs in one go, Carreira et al. propose a self-correcting model that focuses on predicting what is wrong with the current estimate in order to progressively improve an initial solution [41].

All these approaches assume that only one person is present in the image and they cannot handle realistic cases where several people appear in the scene, and interact with each other. Due to this, in the past years, body pose estimation has shifted towards multi-person pose estimation [74], [78], [79]. Multi-person pose estimation introduces significantly more challenges, since the number of persons in an image is not known a priori. Moreover, it is natural that persons occlude each other during interactions, and may also become partially truncated to various degrees.

Multi-person body pose estimation methods mainly fall into two categories. Top-down approaches first detect persons and then their pose while bottom-up approaches detect body parts and then subsequently associate these parts to human instances.

In recent literature bottom-up methods [74], [79] are dominant. A representative example of bottom-up approaches is the method proposed by Insafutdinov et al. that propose to jointly solve the subset partitioning and labelling problem. This can be challenging as people can be partially visible, significant overlap of bounding box regions of people is frequent, and there exists an apriori unknown number of persons in an image. Robust multi-person pose estimation is achieved by first proposing initial part detections and pairwise terms between all parts. These detections are then jointly clustered for determining the belonging to persons. A slightly different approach combines bottom-up and top-down inference [79]. It joins a variation of the unary joint detector architecture with a part affinity field regression to enforce inter-joint consistency [79]. A greedy algorithm is employed to generate person instance proposals in a bottom-up fashion. Their best results are obtained in an additional top-down refinement process in which they run a standard single-person pose estimator on the person instance box proposals generated by the bottom-up stage.

Human actions can differ greatly in their dynamics. According to [81], they can be either periodic (e.g. running, walking or waving) or nonperiodic (e.g. bending), and either stationary (e.g. sitting) or nonstationary/transitional (e.g. skipping as a horizontal motion, and jumping or getting up as vertical motions). Despite significant progress of single frame based multi-person pose estimation, the problem of articulated multi-person body joint tracking in monocular video remains largely unaddressed. In general, pose estimation in videos mainly aim to improve pose estimation by utilizing temporal smoothing constraints but they are not directly applicable to videos with multiple potentially occluding persons.

One recent example of multi-person pose estimation and tracking is Arttrack [44]. They achieve state-of-the-art results by sparsifying the body-part relationship graph and by using a feed-forward convolutional architecture that is able to detect and associate body joints of the same person even in clutter. This model is used to generate proposals for body joint locations and formulate articulated tracking as spatio-temporal grouping of such proposals. This allows to jointly solve the association problem for all people in the scene by propagating evidence from strong detections through time and enforcing constraints that each proposal can be assigned to one person only. Another recent example is PoseTrack [82] that follows the same idea of representing body joint detections in a video using a spatio-temporal graph. By solving an integer linear program to partition the graph into sub-graphs that correspond to plausible body pose trajectories for each person jointly state-of-the-art multi-person pose estimation and tracking can be achieved.

Instead of inferring the projected pose in the image plane, some methods infer an additional depth dimension [83], [84], [85], [86]. Also, another interesting recent development is 3D dense pose estimation [87] and related topics like 3D shape reconstruction [80], 3D shape correspondence [38] and 3D shape completion [37]. For state-of-the-art benchmarking in all these related topics, of particular interest is a recent competition and associated database [88].

4.3 Feature Extraction and Emotion Recognition

The final stage of an EBGR process is building a relevant representation and using it to learn a mapping to the corresponding targets. Depending on the nature of the input, the representation can be static, dynamic or both. Also representation can be geometrical or could include appearance information and can focus on different parts of the body. Moreover, the mapping will then need to be taken into account in order to decide on the most probable class for a given input sample, i.e. to recognize it, which can be performed by using various classification methods. The foregoing topics will be discussed in what follows.

4.3.1 Feature Extraction

Gunes et al. [89], [90] detected face and the hands based on the skin color information, and the hand displacement to neutral position was calculated according to the motion of the centroid coordinates. They used the information from the upper body. For example, in a neutral gesture, there is no movement, but in a happy or sad gesture, the body gets extended, and the hands go up, and get closer to the head than normal. More clearly, they defined motion protocols in order to distinguish between the emotions. In the first frame, the body was supposedly in its neutral state, i.e. the hands were held in front of the torso. In the subsequent frames, the in-line rotations of the face and the hands were analyzed. The actions (body modeling) were first coded by two experts. The first and the last frames from each body gesture, which stand for neutral and peak emotional states, respectively, were utilized for training and testing.

Vu et al. [91] considered eight body action units, which represent the movements of the hands, head, legs and waistline. Kipp et al. [92] provided an investigation of a possible correlation between emotions and gestures. The analysis was performed on static frames extracted from videos representing certain emotional states, as well as emotion dimensions of pleasure, arousal and dominance. The hand shape, palm orientation and motion direction were calculated for all the frames as the features. The magnitudes and directions of the correlations between the expected occurrences and the actual ones were evaluated by finding the correspondences between the dimension pairs, and calculating the resulting deviations.

Glowinski et al. [93] focused on the hands and head as the active elements of an emotional gesture. The features were extracted based on the attack and release parts of the motion cue, which refer to the slope of the line that connects the first value to the first relative extremum and the slope of the line that connects the last value to the last relative extremum, respectively. They also extracted the number of local maxima of the motion cue and the ratio between the maximum and the duration of the largest peak, which were used to estimate the overall impulsiveness of the movement.

Kessous et al. [94] extracted the features from the body and hands. Based on silhouette and hands blobs, they extracted the quantity of motion, silhouette motion images (SMIs) and the contraction index (CI). Velocity, acceleration and fluidity of the handâĂŹs barycenter were also computed. Glowinski et al. [95] successfully extended their work using the same database as in [93], where the 3D position, velocity, acceleration and jerk were extracted from every joint of the skeletal structure of the arm. Kipp and Martin [92] used a dimensional method to represent an affect emotional gesture along a number of continuous axes. Three independent bipolar dimensions namely, pleasure, arousal and dominance, were considered in order to define the affective states. The locations of 151 emotional terms were obtained.

In [96], dynamic features were extracted in order to obtain a description of the submotion characteristics, including initial, final and main motion peaks. It was suggested that the timing of the motions greatly represents the properties of emotional expressions. According to [32], these features can be handled based on the concept of motion primitives, i.e. dynamic features can be represented by a number of subactions.

Hirota et al. [97] used the information about the hands, where dynamic time warping (DTW) was utilized to match the time series. Altun et al. [98] considered force sensing resistor (FSR) and accelerometer signals for affect recognition. Lim et al. [99] captured 3D points corresponding to 20 joints

at 30 frames per second (fps), where in the recognition stage, 100 previous frames were analyzed in case of every frame.

Saha et al. [100] created skeleton models representing the 3D coordinates of 20 upper body joints, i.e. 11 joints corresponding to the hands, head, shoulders and spine were considered in order to calculate nine features based on the distances, accelerations and angles between them. The distance between the hands and spine, the maximum acceleration of the hands and elbows and the angle between the head, shoulder center and spine were considered as features, making use of static and dynamic information simultaneously.

Camurri et al. [101] utilized five motion cues, namely, QoM, CI, velocity, acceleration and fluidity. Piana et al. [102] proposed 2D and 3D features for dictionary learning. The 3D data were obtained by tracking the subjects, and the 2D data from the segmentation of the images. The spacial data included 3D CI, QoM, motion history gradient (MHG) and barycentric motion index (BMI). Patwardhan et al. [103] utilized 3D static and dynamic geometrical features (skeletal) from the face and upper-body. Castellano et al. [104] considered the velocity and acceleration of the trajectory followed by the hand's barycenter, which was extended in [105], adopting multiple modalities (face, body gesture, speech), and in [93], considering 3D features instead. Vu and et al. [91] used the AMSS [106] in order to find the similarity between the gesture templates and the input samples.

Unfortunately more complex representations are very scarce in emotional body gesture recognition. Chen et al. [107] used HOG on the motion history image (MHI) for finding the direction and speed, and Image-HOG features from bag of words (BOW) to compute appearance features. Another example is the usage of a multichannel CNN for learning a deep representation from the upper part of the body [108]. Finally, Botzheim et al. [109] used spiking neural networks for temporal coding. A pulse-coded neural network approximated the dynamics with the ignition phenomenon of a neuron and the propagation mechanism of the pulse between neurons.

4.3.2 Gesture Based Emotion Recognition

Gunes and Piccardi [90] used feature vector containing displacement measures between two frames: a frame with a neutral expression and an expressive one. The features were calculated from the upper-body to classify gestures into 6 emotional categories (anger-disgust, anger-fear, anger-happiness, fear-sadness-surprise, uncertainty-fear-surprise and uncertainty-surprise). A set of standard classifiers was trained and Bayesian Net provided the best classification results. Similar approach is presented in [104].

Castellano et al. compared different methods such as 1-nearest-neighbor with dynamic time warping (DTW-1NN), J48 decision tree and the Hidden Naive Bayes (HNB) to classify dynamic representations of body gestures into 4 emotional categories. Best results were obtained using DTW-1NN.

Modelling body parts independently for action analysis is a common approach in many studies [32]. For example, in [110], the authors independently modeled actions of arms, head and torso. Contrarily, a structural body model was proposed in [111]. The authors defined a tree-based description of the body, where each activity corresponds to a node representing the part engaged in performing it.

Kleinsmith et al. [112] proposed an automatic recognition model that recognize four affective states such as concentrating, defeated, frustrated and triumphant, based on non-acted body postures of Nintendo gamers. Corpora labelling was obtained using online posture evaluation survey. They outside observersâĂŹ judgments were based on computer avatar stimuli to create a non-gender and non-culturally specific faceless humanoid. The most frequent affective label assigned by observers was defined as a ground truth. Each silhouette was described by low-level posture configuration features and multilayer perceptron was used as the classifier of this system.

The same type of database was investigated by Savva et al. in [113]. They extracted time-related features such as body rotations, angular velocity, frequency and acceleration, body directions, and amount of movements from Nintendo gamer's whole body. For such selected features, recurrent neural network (RNN) was matched as a clarifier.

Griffin et al. [114] investigated laughter-related body movements and their significance in human communication. They analysed different laughter states such as hilarious, social, awkward, fake, and non-laughter and their impact on body movement. The analysis showed significant differences in torso and limb movements between laughter and and non-laughter expression. Social and hilarious laughter can be distinguished from each other by the amount of: spine bending, shoulder rotation and hand movements. To distinguish all above mentioned laughter types, they used hand movements (distance of hands from hip and head), shoulder movements and spine and neck bending. They compared results obtained for several types of classifiers, the most effective one was non-parametric model RF. In their research, Venture et al. [115] focused on emotional gaits performed by 4 professional actors, recorded by Vicon motion capture system. The actors were asked to walk while expressing one of the following emotions: neutral attitude, joy, anger, sadness and fear. Basing on geometric models the inverse kinematics was computed in order to obtain 34 degrees of freedom (DOF) model. They confirmed that 12 DOF is enough to recognize emotions in gait: motion of the lower torso and variation of trunk and head inclination are the most important features.

Saha et al. [100] investigated gestures reflecting five basic human emotional states from skeletal geometrical features. They compared binary decision tree, ensemble tree, k-nearest neighbour and SVM, obtaining best results using ensemble tree.

In more recent study, Samadani et al. [116] focused on a stochastic model of the affective movement dynamics using HMMs. The output of HMMs were used to derive a Fisher score movement representation and next used to optimize affective movement recognition using SVM. Moreover, to obtain a minimal discriminative representation of the movements, the authors used supervised PCA, which are based on Hilbert-Schmidt independence criterion in the Fisher score space. The effectiveness of proposed method was validated using two different datasets: a full-body and handâĂŞarm movements only corpora.

Glowinski et al. [95] expanded the previous work analyzing

meaningful groups of emotions related to four quadrants of valence/arousal space, and described them using trajectories of head and hands as well as frontal and lateral view of the body. A compact representation was grouped into clusters of four classes namely high-positive (amusement, pride), high-negative (hot-anger, fear, despair), low-negative (pleasure, relief, interest) and low-negative (cold anger, anxiety, sadness).

Fourati and Pelachaud [117] proposed a deeper analysis of emotional movements expression using a wide range of features from the whole body. They described movement on anatomical, directional and posture/movement level. Random forest approach was used to classify 8 emotional states (joy, anger, panic fear, anxiety, sadness, shame, pride and neutral) expressed by actors in various daily actions such as walking, sitting or knocking.

Senecal et al. [118] proposed a system for continuous emotional behavior recognition during theater performance. They represented the whole human body using Laban Movement Analysis (LMA) [119] mapped onto Russell Circumplex Model (RCM) [120]. LMA is an efficient method for interpreting, describing, visualizing and notating human movement. The efficiency of LMA as a body motion descriptor, studied and improved for over a decade, was proven by a variety of case studies. The proposed system, based on neural network, is able to project the emotion transition of the actorâĂŹs body as a trajectory on the RCM diagram during theater performances with sufficient accuracy.

Kaza et al. [121] used a set of kinematic and geometrical features extracted from joint-oriented skeleton tracking during gameplay scenarios. The efficiency of chosen features was evaluated using deep learning algorithms such as multilayer perceptron,Restricted Boltzmann Machines (RBMs) and proposed stacked RBM, compared with classic classifiers such as NaÃŕve Bayes, Linear MultiClass SVM, Non-Linear SVM. Stacked RMB outperformed all other classification methods.

Li et al. [122] proposed a method for emotion recognition from human gait using Microsoft Kinect cameras. They extracted features from 3-dimensional coordinates of 14 main body joints using Fourier transformation and PCA. Naive Bayes, Random Forests, SVM and Sequential Minimal Optimization (SMO) were used as classifiers to recognizing anger and happiness from neutral state. Piana et al. [123] extracted features corresponding to kinematics of single joints and psychological theories such as impulsiveness and contraction index from the whole human body. They used real-time SVM as a classifier to create an interactive game for autistic children. Currently, the game is validated among autistic children to analyse if it may serve as an effective tool for learning both emotions recognition and emotions expression.

In their system, Arunnehru et al. [124] used orientation, elongation, solidity, rectangularity of shape, distance and speed as motion-based features. SVM, NaÃŕve Bayes, and dynamic time wrapping (DTW) were used as classifiers. The best results were obtained using DTW with the average recognition rate of 93.39%. This high performance may be caused by recognizing only three emotional states: happy, angry and fearful.

Kosti et al. in [125] presented a method for emotion recognition based on images containing people in context in non-controlled environments. They trained a two low-rank filter CNN that jointly analysed the person and the whole scene to recognize emotional state. The analysed images depicted people annotated with 26 emotional categories as well as the continuous dimensions valence, arousal, and dominance. By this research they emphasized the importance of considering the context for recognizing peopleâĂŹs emotions in images.

4.3.3 Gestures in Multimodal Emotion Recognition

Although body gestures are important part of human communication, often they are a supplement of other reflexive behavior forms such as facial expression, speech, or context. Studies in applied psychology showed that human recognition of facial expressions is influenced by body expression and context [126]. Expanding the focus to several expression forms can facilitate research on emotion recognition as well as human-machine interaction.

Works combining facial and body display for emotion recognition are rather scarce. Historical works usually focused on simple fusion techniques of predefined body and face representations [127], [128]. More recently, Psaltis et al. [129] introduced a multi-modal late fusion structure that could be used for stacked generalization on noisy databases. Basic emotions like surprise, happiness, anger, sadness and fear are classified from facial and body gestures representations with better recognition performance than each of the mono-modal correspondents.

Uncovering the interrelation between speech and body gestures for emotion recognition has also been of interest [91], [130]. For example, using prosody and audio spectral features for modeling the interaction dynamics of speech with three types of body representations: head motion, lower and upper body motions has been proposed [130]. In [91] on the other hand, the authors presented a bi-modal approach (gestures and speech) for recognition of four emotional states: happiness, sadness, disappointment, and neutral. Gestures recognition module fused video and 3D acceleration sensors. The outputs from speech and gestures based recognition were fused by using weight criterion and best probability and majority vote. Fifty Japanese words (or phrases) and 8 types of gestures recorded from five participants were used to validate the system. Performance of the classifier indicated better results for bi-modal than each of the uni-modal recognition system.

Tri-modal approaches combining face, body and speech for emotion recognition also exist [94] [131]. In [94], a database consisting of audio-video recordings of people interacting with an agent in a specific scenario was proposed. Ten people of different gender, using several different native languages including French, German, Greek and Italian pronounced a sentence in 8 different emotional states. Facial expression, gesture and acoustic features were used with an automatic system based on a Bayesian classifier. Results obtained from each modality were compared with the fusion of all modalities. Combining features into multi-modal sets resulted in increases by more than 10% when compared to the most successful uni-modal system. Furthermore, the best results were obtained when merging gesture and speech.

Another interesting category of works use gestures only implicitly, usually for complementing facial information with more context. This is the case in group level emotion recognition [132] where a limited set of affect labels (positive, negative, neutral) are classified from images of groups of people where body information is visible. Another example is apparent personality recognition of people talking to a camera. The upper part of the body is visible and gesture information is implicitly represented as part of the scene [133]. Nevertheless as far as we are aware in neither of the two problems, gestures haven't been used explicitly. This could be an interesting problem for future research.

4.4 Applications

Applications of emotional body gesture recognition are mainly of three types [134], [135], [136]. The first consist of systems that detect the emotions of the users. The second includes actual or virtual animated conversational agents, such as robots and avatars. They are expected to act similarly to humans when they are supposed to have a certain feeling. The third includes systems that really feel the emotions. For example, these systems have applications in video telephony [137], video conferencing and stress-monitoring tool, violence detection [138], [139], [140], video surveillance [141], and animation or synthesis of life-like agents [139] and automatic psychological research tools [141]. All the three types have been extensively discussed in the literature. However, this paper concentrates on affect detection only.

Automatic multi-modal emotion recognition systems can utilize sources of information that are based on face, voice and body gesture, at the same time. Thus they can constitute an important element of perceptual user interfaces, which may be utilized in order to improve the ease of use of online shops. They can also have applications in pervasive perceptual man-machine interfaces, which are used in intelligent affective machines and computers that understand and react to human emotions [142]. If the system is capable of combining the emotional and social aspects of the situations for making a decision based on the available cues, it can be a useful assistant for humans [143].

5 Data

We further present main public databases of gesture based expressions of affect useful for training EGBR systems. We discuss RGB, Depth and bi-modal of RGB + Depth databases in Sec. 5.1, 5.2 and 5.3, respectively. The reader is referred to Table 2 for an overview of the main characteristics of the databases and to Fig. 8 for a selection of database samples.

5.1 RGB

One of the first body language databases with affect annotations was made publicly available by Gunes and Piccardi [90]. The database contains 206 samples with six basic emotions, as well as four more states, namely, neutral, anxiety, boredom and uncertainty. 156 samples were used for training, and 50 samples for the test.

Castellano et al. [104], [105] collected affective body language data consisting of 240 gestures [150]. Their database is a part of the HUMAINE database [146]. There were six male and four female participants, i.e. 10 in total. They acted eight emotions, i.e. anger, despair, interest, pleasure, sadness, irritation, joy and pride, equally distributed in the valence arousal space. However, they focused on four emotions, i.e. anger, joy, pleasure and sadness. A camera filmed the full body of the subjects from the front view at a rate of 25 fps. In order to accelerate the silhouette extraction, they used a uniform dark background.

The Geneva Multi-modal Emotion Portrayals (GEMEP) database [93] contains more than 7000 audio-video portrayals of emotional expressions. It includes 18 emotions portrayed by 10 actors. 150 portrayals were systematically chosen based on ratings by experts and non-experts, which resulted in the best recognition of the emotional intentions. Their analysis was on the basis of 40 portrayals selected from the mentioned set. They were chosen such that they represent four emotions, namely, anger, joy, relief and sadness. Each of these emotions is from one quadrant of the two main affective dimensions, i.e. arousal and valence.

The Theater corpus was introduced by Kipp and Martin [92], based on two movie versions of the play *Death of a Salesman*, namely, DS-1 and DS-2.

Vu et al. [91] considered eight types of gestures that are present in the home party scenario of the mascot robot system. The database involved five participants, i.e. four males and one female. Their ages ranged from 22 to 30 years. They were from three different nationalities: Japanese, Chinese, and Vietnamese.

A subset of the LIRIS-ACCEDE video database [148] was created in [147], which contains upper bodies of 64 subjects, including 32 males and 32 females, with six basic emotions. Their ages were between 18 and 35 years.

5.2 Depth

The GEMEP-FERA database, which was introduced by Baltrušaitis et al. [144], is a subset of the GEMEP corpus. The training database was created by 10 actors. In the test database, six actors participated. From these actors, three were common with the training database, but the other three were new. The database consists of short videos of the upper body of the actors. The average length of the videos is 2.67 seconds. The videos do not start with a neutral state.

The database created by Saha et al. [100] involved 10 subjects. The age of the subjects ranged from 20 to 30. The subjects were stimulated by five different emotions, namely, anger, fear, happiness, sadness, and relaxation. These emotions caused the subjects to take different gestures accordingly. Each subject was filmed at a frame rate of 30 fps, for 60 seconds. Next, the Cartesian coordinates of the body joints were processed.

In [103], six basic emotions, namely, anger, surprise, disgust, sad, happy and fear, were acted by 15 subjects. The subjects were between 25 to 45 years old. Five subjects were female, and the rest were male. In addition, five subjects were Americans, and the rest were Asians. The lighting conditions were controlled, and the poses of the bodies of the subjects were completely frontal. The subjects' distances from the camera were from to 1.5 to 4 meters.

The UCFKinect [151] was collected using Kinect and skeleton estimation from [152]. 16 subjects, including 13

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874986, IEEE Transactions on Affective Computing

(a) (b) (c) (d) (e) (f)

Figure 8. Selected samples from databases containing gesture based expressions of affect: (a) FABO [21], (b) GEMEP-FERA [93], [144], [145], (c) Theater [92], (d) HUMAINE [94], [104], [105], [146], (e) LIRIS-ACCEDE [147], [148], (f) MSR-Action 3D [149].

males and three females, participated in the recordings. All the subjects were between 20 and 35 years old. Each of them performed 16 actions such as balance, punch or run and repeated it five times. In total, 1280 actions were recorded. The 3D coordinates of 15 joints were calculated for each frame. The data on the background and the clothes were not included in the calculations, and only the data on the skeleton was extracted.

The MSR Action 3D consists of twenty actions, namely, high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw.

5.3 Bi-modal: RGB + Depth

The database created by Psaltis et al. [129] contains facial expressions that frequently appear in games. The subjects acted five basic emotions, namely, anger, fear, happiness, sadness and surprise. They considered a neutral emotion class for labeling the samples that do not present any motion, and cannot be classified under any of the basic emotions. 15 subjects participated to create 450 videos. Each video starts with an almost neutral state, and evolves toward a peak emotional state. The labels were assigned based on the emotion that the subject was asked to perform, not on the actually performed movements. The whole duration of every video is 3 seconds. Before acting each emotion, a video presenting the required emotion was shown to the subjects, and then the subjects performed the movements with their own styles five times. The database is divided into three parts. One part only contains facial expressions, another part only contains body gestures, and the last part contains both face and body data. They used a dense-ASM tracking algorithm for tracking the features and extracting the AUs. In order to evaluate the performance of the proposed method, they applied it to the FERA database as well.

The emoFBVP database [153] includes multi-modal recordings of actors, i.e. face, body gesture, voice and physiological signals. Audiovisual information of three different expressions, i.e. intensities, of 23 emotions are included, as well as tracking of facial features and skeletal tracking. 10 professional actors participated in acquiring the data. Each recording was repeated six times. Three recordings were in a standing position, and the others in a seated position. To date, this database offers the most diverse range of emotional body gestures in the literature, but right now it is not available. Finally, the specifications of all available databases are summarized in Table 2. The list of emotions that have been considered in each of the databases is provided in Table 3. A sample image from each of the databases can be seen in Fig. 8.

6 Discussion

In this section we discuss the different aspects of automatic emotional body gesture recognition presented in this work. We start with the collections of data currently available for the community. Then, the discussion mainly focuses on representation building and emotion recognition from gestures. This includes the categories of mostly used features, taking advantage of complementarity by using multi-modal approaches and most common pattern recognition methods and target spaces.

6.1 Data

Majority of freely accessible data sets contain acted expressions. This type of material is usually composed of high quality recordings, with clear undistorted emotion expression. The easiness of acquiring such recordings opens a possibility of obtaining several samples from a single person. The conventional approach to collect acted body language databases is to let actors present a scene portraying particular emotional states. Professionals are able to immerse in an emotion they perform, which may be difficult for ordinary people. This kind of samples are free from uncontrollable influences and usually they do not require additional evaluation and labeling processes. However, some researchers emphasize that these types of recordings may lead to creating a set of many redundant samples, as there is a high dependency on actors skills and his or her ability to act out the same emotional state differently. Another argument against such recordings states that they do not reflect real world conditions. Moreover, acted emotions usually comprise of basic emotions only, whereas in real life emotions are often weak, blurred, occur as combinations, mixtures, or compounds of primary emotions. Wherefore current trends indicate that spontaneous emotions are preferable for research. There is another method to record emotional body movements in natural situations. One can use movies, TV programs such as talk shows, reality shows or live coverage. This type of material might not always be of satisfactory quality (background noise, artifacts, overlapping, etc.) and may obscure the exact nature of recorded emotions. Moreover, collections of spontaneous samples must be evaluated by human decision

Table 2
Main characteristics of a selected list of publicly available databases for recognizing gesture based expression of affect. C: controlled (in-the-lab), U: uncontrolled (in-the-wild)

Reference	Name	Device	Body parts	Modality	Context	#Emotions	#Gestures	#Subjects	#Females	#Males	#Sequences	#Samples	FR[2] (fps)	Background	AVL[3] (s)
Gunes et al., 2006 [21]	FABO	Digital camera	Face and body	Visual	C	10	NA	23	12	11	23	206	15	Uniform blue	∼3600
Glowinski et al., 2008 [93]	GEMEP	Digital camera	Face and body	Audiovisual	C	18	NA	10	5	5	1260	>7000	25	Uniform dark	NA
Castellano et al., 2007 [104]	HUMAINE	Camera	Face and body	Audiovisual	C	8	8	10	4	6	240	240	25	Uniform dark	NA
Gavrilescu, 2015 [147]	LIRIS-ACCEDE	Camera	Face and upper body	Visual	C	6	6	64	32	32	NA	NA	NA	Nonuniform	60
Kipp et al., 2009 [92]	THEATER	Camera (movie clips)	Body	Audiovisual	U	8	NA	NA	NA	NA	NA	258	NA	Nonuniform	NA
Fourati et al., 2014 [154]	EMILYA	Motion capture and 4 cameras	Body	Visual	C	8	7	11	6	5	23	7084	NA	NA	5.5 Âś 3

Table 3
Labels included in a selected list of the databases. F = FABO [21], G = GEMEP [93], T = T heater [92], H = HUMAINE [104], LA = LIRIS-ACCEDE [147], GF = GEMEP-FERA [144].

Database	F	G	T	H	LA	GF	Frequency
Sadness	●	●	●	●	●	●	6
Anger	●		●	●	●	●	5
Anxiety	●	●	●				3
Disgust	●	●			●		3
Fear	●				●	●	3
Surprise	●	●					3
Boredom	●		●				2
Happiness	●					●	2
Interest		●		●			2
Contempt		●					2
Despair		●		●			2
Irritation		●		●			2
Joy				●		●	2
Pleasure		●		●			2
Relief		●				●	2
Admiration		●	●				1
Neutral		●					1
Pride		●		●			1
Shame		●					1
Aghastness		●					1
Amazement			●				1
Amusement		●					1
Boldness				●			1
Comfort				●			1
Dependency			●				1
Disdain			●				1
Distress			●				1
Docility			●				1
Elation		●					1
Excitement			●				1
Exuberance			●				1
Fatigue	●						1
Gratefulness			●				1
Hostility			●				1
Indifference			●				1
Insecurity			●				1
Nastiness			●				1
Panic Fear		●					1
Rage		●					1
Relaxation			●				1
Respectfulness			●				1
Satisfaction			●				1
Tenderness		●					1
Uncertainty	●						1
Unconcern			●				1

makers or professional behaviorists to determine the gathered emotional states. Nonetheless it does not guarantee objective, genuinely independent assessments. Additionally, copyright reasons might make it difficult to use or disclose movies or TV recordings. An accurate solution for sample acquisition may be provoking an emotional reaction using staged situations, which has been already used in emotion recognition from speech or mimics. Appropriate states may be induced using imaging methods (videos, images), stories or computer games. This type of recordings are preferred by psychologists, although the method can not provide desirable effects as reaction to the same stimuli may differ. Similarly to spontaneous speech recordings, triggered emotional samples should be subjected to a process of labeling. Ethical or legal reasons often prohibit to use or make them publicly available. Taking into account above mentioned issues, real-life emotion databases are rarely available to the public, and a good way of creating and labelling such samples is still open to question.

The process of choosing appropriate representation of emotional states is intricate. It is still debatable how detailed and which states should be covered. Analyzing Table 3 one can observe how broad affective spectrum has been used in various types of research. Most authors focus on sets containing six basic emotions (according to EkmanâĂŹs model). Sadness and anger occur in majority of databases. Fear, surprise and disgust are also commonly used. However, there are quite a lot of affective states that are not consistently represented in the available databases. Some examples (see Tab. 3) are uncertainty, unconcern, aghastenss, shame, tenderness, etc.

There is a lack of consistency in the taxonomy used for naming the affective states. For example both joy and happiness, are used interchangeably depending on the database. It is difficult to evaluate whether these are the same or different states. Joy is more beneficial, as it is less transitory than happiness and is not tied to external circumstances. Therefore, it is possible that there is a misunderstanding in naming: while happiness may be caused by down to earth experiences, material objects, joy needs rather spiritual experiences, gratitude, and thankfulness, thus may be difficult to evoke and act. These misunderstandings may be also a result of translations. Such issues will reoccur until a consistent taxonomy of emotions will be presented, so far there is no agreement among experts even on the very definition of primary states. Moreover, due to the heterogeneity of described databases, comparison of their quality is problematic. With just several public accessible emotional databases and with the addition of the above described issues, comparison of detection algorithms becomes a challenging task. There is clearly space and necessity of creation of more unified emotional state databases.

6.2 Feature Extraction and Emotion Recognition

Feature Extraction. The large majority of the methods developed to recognize emotion from body gestures use geometrical representations. A great part of these methods build

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874986, IEEE Transactions on Affective Computing

Table 4
Summary of a few multi-modal emotion recognition methods.
S=Speech, F=Face, H=Hands, B=Body.

Reference	Modalities	#samples	#emotions	Representation
Gunes Piccardi [90]	F + B	206	6	Motion protocols
Castellano's et al. [104]	B	240	4	Multi cue
Castellano et al. [105]	B	240	4	Multi cues
Glowinski et al. [93]	B	40	4	Multi cues
Kipp Martin [92]	B	119	6	PAD
Kessous et al. [94]	S + F + B	NA	8	Multi cues
Vu et al. [91]	S + B	5	4	Motion protocols
Gavrilescu [147]	B + H	384	6	Motion protocols

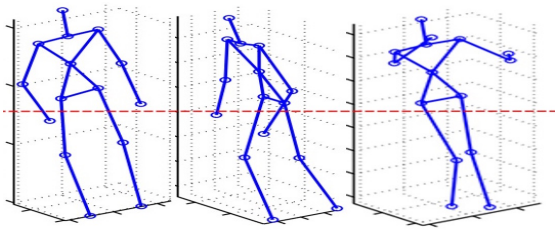

Figure 9. Sample upper and lower body postures [155].

simple static or dynamic features related to the coordinates of either joints of kinematic models or of parts of the body like head, hands or torso. Some of the most used features are displacements [90], orientation of hands [92] motion cues like velocity and acceleration [93], [95], [96], [98], [101], [104], shape information and silhoutte [94], smoothness and fluidity, periodicity, spatial extent and kinetic energy, among others. While most descriptors are very simple there are also examples of slightly more advanced descriptors like Quantity of Motion (QoM measures of the amount of motion in a sequence), Silhoutte Motion Images (SMI contains information about the changes of the shape and position of the silhouette), Contraction Index (CI measures the level of contraction or expansion of the body), and Angular Metrics for Shape Similarity (AMSS) [91], [102].

Considering dynamic features such as acceleration, movement gain and velocity, or at least combining them with static features, usually leads to higher recognition rates than relying solely on the latter, since they result in a richer representation and since some emotional traits are expressed mostly in the dynamics of the human body.

Most of the methods proposed focus on the upper body (head, neck, shoulders, arms, hands and hand fingers) [92], [97], [100], hands [93], arm [95], body and hands [94], full body [91], [102]. Upper body and lower body parts are represented in Fig. 9.

Among all different parts of the body, in the context of body gesture recognition, numerous studies have focused on hand gestures, which requires hand segmentation and tracking. The features that can be extracted from the hands include palm orientation, hand shape, elbow, wrist, palms and shoulder joints, hand shape and motion direction, which are analyzed independently from the body, in order to calculate the motion of the hand and the individual fingers along each of the axes. The motions of the hand are measured with the body as the reference. The motions of the body itself are found in terms of the changes of the pose of the upper body, i.e. its inclinations to the left, right, forward or backward.

Complex learnt representations for recognizing emotion from body gestures are very scarce, mostly because there is a lack of big volumes of labelled data (see Tab. 2) for learning such representations in a supervised way. Two of the very few works that uses deep learning representations for body emotion recognition are multichannel CNN from upper body [108] and spiking neural networks for temporal coding [109]. As previously discussed in Sec. 6.1 there is a lack of consistent taxonomy for the output spaces in the various databases published to date. This results in considerable fragmentation of the data and makes transfer learning techniques difficult. Even though not explored yet in the literature, unsupervised learning might be interesting for pre-training general representations of the moving human body, before tuning to more specific emotion oriented models.

Emotion Recognition. There is a tendency in the literature to reduce the output spaces for simplifying the recognition problem. This has been done either by grouping emotions into quadrants of a dimensional emotion space [95] or by grouping emotions based on similarity of their apperance [90]. In general, most methods have focused on recognizing basic emotions like Anger, Joy, Pleasure, Sadness and Fear [100], [104]. Though not dominant, methods that target richer output spaces also exist [125].

Another popular approach is to show extensive comparison between sets of standard classifiers like decision trees, k-NNs and SVMs. The results of using numerous classifiers and different numbers of emotion classes based on two different databases are summarized in Table 5 and 6, respectively.

According to Table 5, J48 [156] has the best performance between three used classifiers tested on HUMAINE database. This work used just four labels of the mentioned database. According to Table 6 the best performance on a different database, with 10 subjects and 5 labels, is achieved by ensemble tree classification methods. Moreover, the different methods with their performances are represented as a chart for emotional gesture recognition in Fig. 10.

More complete representations of the body can also be used in a more meaningful way. Particularly interesting are structural models where different parts of the body are indepentedly representated and contribute to a final decision over the emotion which takes into account predefined priors [111]. Going even further, additional information from the context, like the background could be used as well to refine final decision [125].

Quiet recent investigated works in this study which are around two year ego, show the whole body has been used to propose the emotion recognition systems [118], [121], [123], [124], [125] They focused on two component of these system, features and classification methods. The features such as mapped LMA onto RMC [118], 3D geometrical and kinematics of tracked joints [121], 3D dimensional coordinates by applying Fourier transformation and PCA [122], kinematics of single joints [123] and dynamics features such as the velocity, orientation, elongation, stability, and rectangularity rectangularity measures [124] have been extracted which can be mentioned as more reliable feature since using of these features have been preferred

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874986, IEEE Transactions on Affective Computing

Table 5
Comparison of the effect of using various classification methods and different numbers of emotion classes based on HUMAINE database in the framework of the EU-IST Project [104].

Classifier	Performance (%)	#classes
1NN-DTW	53.70	4
J48 or Quinlan's C4.5	56.48	4
Hidden Naive Bayes	51.85	4

Table 6
Comparison of the effect of using various classification methods and different numbers of emotion classes based on the recorded samples by Kinect. The database included by 10 subjects in the age group of 25±5 years [100].

Classifier	Performance (%)	#classes
Ensemble tree	90.83	5
Binary decision tree,	76.63	5
K-NN	86.77	5
SVM	87.74	5
Neural network	89.26	5

by new researches. Basically new works are willing to find the ways to improve the previous proposed systems. And also neural network [118], Deep learning [121], RF [122] and SVM, NaÃŕve Bayes, and DTW [123], [124] are the most recent classification methods which have been used. According to [122] RF have better performance rather than naive Bayes, LibSVM and SMO based on same propose method and dataset and according to [124] DTW method, works better tham SVM with polynomial kernel and naive-Bayes classifiers. Different body language components (gestures, faces) together with speech carry affective information and complementary processing have obvious advantages. A consistent part of the literature uses multiple representations of the body in a complementary way to recognize emotion. For example, there are works that combine body with speech [91], [130] and with face [127], [128]. Regardless of the fusion techniques used, all these methods report improvements of results backing the hypothesis that there is considerable complementarity in different modalities and its exploration is fruitfull. Also, it has already been previously commented that a more complete body representation is also helpful in this respect (for example, upper and lower body considered together). Unfortunately research in multimodal emotion recognition remains rather scarce and simplistic. The few works that exists mostly focus in simplistic fusion techniques from shallow representations of body and face or body and speech. Even though all methods report important improvements over monomodal equivalents, this potential remains largely unexplored. The reader is refered to Table 4 for a selected set of studies that have used body representations together with representations of other modalities for recognizing emotion.

The number of emotion classes affects the performance of a given classifier as well. Usually, reducing the number of classes from a given database should increase the performance. The best recognition rate, i.e. 93%, is obtained by considering five emotion classes and using neural networks. It should be noted that low-quality samples or features

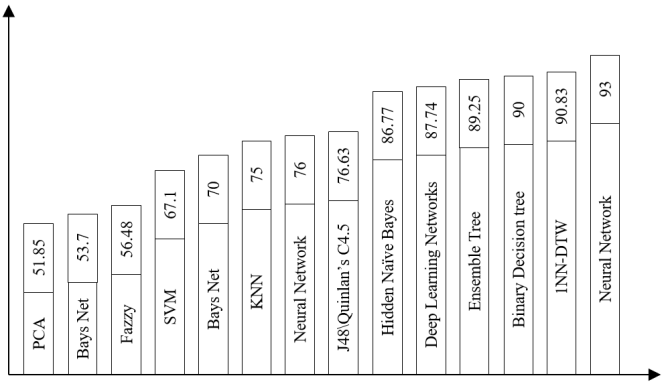

Figure 10. Performances (%) of different emotion recognition methods based on the different databases.

may degrade the performance, and cause a violation of the expected trend.

Approaches to train and test emotional gesture recognition systems are investigated based on the existing literature, where a certain portion of the database is used for training, and the rest is left for testing. Some of the proposed techniques present superior performances on specific databases, i.e. they have led to accuracy rates higher than 90%. However, in order to ensure that the system is reliable, it needs to be tested against different types of data, including various conditions of background, e.g. dark, light, uniform and nonuniform. Moreover, it is worth paying attention that different training and testing strategies may result in different performance rates.

7 Conclusion

In this paper we defined a general pipeline of Emotion Body Gesture Recognition methods and detailed its main blocks. We have briefly introduced important pre-processing concepts like person detection and body pose estimation and detailed a large variety of methods that recognize emotion from body gestures grouped along important concepts such as representations learning and emotion recognition methods. For introducing the topic and broadening its scope and implications we defined emotional body gestures as a component of body language, an essential type of human social behavior. The difficulty and challenges of detecting general patterns of affective body language are underlined. Body language varies with gender and has important cultural dependence vital issues for any researcher willing to publish data or methods in this field.

In general the representations used remain shallow. Most of them are naive geometrical representations, either skeletal or based on independently detected parts of the body. Features like motion cues, distances, orientations or shape descriptors abound. Even though recently we can see deep meaningful representations being learned for facial analysis for affect recognition a similar approach for a more broader affective expression of humans is still to be developed in the case of body analysis. For sure the scarcity of body gesture and multimedia affective data is playing a very important role, problem that recently is starting to be overcome in the case of facial analysis. An additional problem is that while

in the case of facial affective computing there has been a quite clear consensus of the output space (primitive facial expressions, facial Action Units and recently more comprehensive output spaces) in the case of general affective expressions in a broader sense such consensus does not exist. A proof in this sense is the variety of labels proposed in the multitude of publicly available data, some of them following redundant or confusing taxonomies.

In general, for comprehensive affective human analysis from body language, emotional body gesture recognition should learn from emotional facial recognition and clearly agree on sufficiently simple and well defined output spaces based on which to publish large high quality amounts of labelled and unlabelled data that could serve for learning rich deep statistical representations of the way the affective body language looks like.

ACKNOWLEDGMENTS

This work is supported Estonian Research Council Grant (PUT638), Estonian-Polish Joint Research Project, the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund, the Spanish Project TIN2016-74946-P (MINECO/FEDER, UE), CERCA Programme / Generalitat de Catalunya and the Scientific and Technological Research Council of Turkey (TÜBÄřTAK) (Proje 1001 - 116E097). This work is partially supported by ICREA under the ICREA Academia programme. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[2] A. Pease and B. Pease, *The definitive book of body language. Peace International*. Peace International, 2004.

[3] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[4] P. Ekman and F. Wallace, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologist Press, 1978.

[5] N. Smrtić, "Asertivna komunikacija i komunikacija u timu," Ph.D. dissertation, Polytechnic of Međimurje in Čakovec. Management of tourism and sport., 2015.

[6] K. Brow, *Kinesics. Encyclopedia of Language and Linguistics 2nd Edition*. Elsevier Science, 2005.

[7] B. Pease and A. Pease, *The definitive book of body language*. Bantam, 2004.

[8] D. Rosenstein and H. Oster, "Differential facial responses to four basic tastes in newborns," *Child development*, pp. 1555–1568, 1988.

[9] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[10] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[11] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *TAC*, vol. 4, no. 1, pp. 15–33, 2013.

[12] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *TAC*, vol. 4, no. 4, pp. 341–359, 2013.

[13] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[14] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[15] J. F. Iaccino, *Left brain-right brain differences: Inquiries, evidence, and new approaches*. Psychology Press, 2014.

[16] H. Ruthrof, *The body in language*. Bloomsbury Publishing, 2015.

[17] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 1–7.

[18] A. Pease and B. Pease, *The Definitive Book of Body Language: how to read othersâĂŹ attitudes by their gestures*. Hachette UK, 2016.

[19] A. W. Siegman and S. Feldstein, *Nonverbal behavior and communication*. Psychology Press, 2014.

[20] H. Gunes and M. Piccardi, "Fusing face and body gesture for machine recognition of emotions," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. IEEE, 2005, pp. 306–311.

[21] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1148–1153.

[22] H. Gunes, C. Shan, S. Chen, and Y. Tian, "Bodily expression for automatic affect recognition," *Emotion recognition: A pattern analysis approach*, pp. 343–377, 2015.

[23] D. Efron, "Gesture and environment." 1941.

[24] A. Kendon, "The study of gesture: Some remarks on its history," in *Semiotics 1981*. Springer, 1983, pp. 153–164.

[25] " Dimension of body language ," http://westsidetoastmasters.com/resources/book_of_body_language/toc.html/, [Online; accessed 19-June-2017].

[26] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[27] L. A. Camras, H. Oster, J. J. Campos, K. Miyake, and D. Bradshaw, "Japanese and american infants' responses to arm restraint." *Developmental Psychology*, vol. 28, no. 4, p. 578, 1992.

[28] https://www.udemy.com/body-language-basics-in-business-world, accessed: 2017-12-15.

[29] R. W. Simon and L. E. Nath, "Gender and emotion in the united states: Do men and women differ in self-reports of feelings and expressive behavior?" *American journal of sociology*, vol. 109, no. 5, pp. 1137–1176, 2004.

[30] S. H. Kennedy, G. Einstein, and J. Downar, "Gender/sex differences in emotions," *Medicographia*, vol. 35, no. 3, pp. 271–280, 2013.

[31] U. Hess, S. Senécal, G. Kirouac, P. Herrera, P. Philippot, and R. E. Kleck, "Emotional expressivity in men and women: Stereotypes and self-perceptions," *Cognition & Emotion*, vol. 14, no. 5, pp. 609–642, 2000.

[32] D. Bernhardt, "Emotion inference from human body motion," Ph.D. dissertation, University of Cambridge, 2010.

[33] T.-y. Wu, C.-c. Lian, and J. Y.-j. Hsu, "Joint recognition of multiple concurrent activities using factorial conditional random fields," in *Proc. 22nd Conf. on Artificial Intelligence (AAAI-2007)*, 2007.

[34] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on computers*, vol. 100, no. 1, pp. 67–92, 1973.

[35] T. Tung and T. Matsuyama, "Human motion tracking using a color-based particle filter driven by optical flow," in *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08*, 2008.

[36] P. F. Felzenszwalb and D. McAllester, "Object detection grammars." in *ICCV Workshops*, 2011, p. 691.

[37] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia, "Deformable shape completion with graph convolutional autoencoders," *arXiv preprint arXiv:1712.00268*, 2017.

[38] N. Verma, E. Boyer, and J. Verbeek, "Feastnet: Feature-steered graph convolutions for 3d shape analysis," in *CVPR 2018-IEEE Conference on Computer Vision & Pattern Recognition*, 2018.

[39] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose

estimation," in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

[40] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[41] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4733–4742.

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[43] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Advances in Neural Information Processing Systems*, 2017, pp. 2274–2284.

[44] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated multi-person tracking in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 4327. IEEE, 2017.

[45] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Modeling emotions for affect-aware applications," *Cover and title page designed by ESENCJA Sp. z oo*, p. 55, 2015.

[46] P. Ekman, "Universal and cultural differences in facial expression of emotion," *Nebr. Sym. Motiv.*, vol. 19, pp. 207–283, 1971.

[47] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Research in Personality*, vol. 11, pp. 273–294, 1977.

[48] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[49] P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique," *Psychol. Bull.*, vol. 115, no. 2, pp. 268–287, 1994.

[50] M. Greenwald, E. Cook, and P. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *J. Psychophysiology*, no. 3, pp. 51–64, 1989.

[51] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *JPSP*, vol. 54, pp. 1063–1070, 1988.

[52] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *TPAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[53] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[54] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: a survey," *Pattern Recognition*, vol. 51, pp. 148–175, 2016.

[55] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *null*. IEEE, 2003, p. 734.

[56] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*. Springer, 2006, pp. 428–441.

[57] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *TPAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[58] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*. IEEE, 2008, pp. 1–8.

[59] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *European Conference on Computer Vision*. Springer, 2014, pp. 613–627.

[60] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. S. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades." in *BMVC*, 2015, pp. 32–1.

[61] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3361–3369.

[62] R. Girshick, "Fast r-cnn," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1440–1448.

[63] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[64] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in *European Conference on Computer Vision*. Springer, 2016, pp. 443–457.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[66] A. Kar, "Skeletal tracking using microsoft kinect," *Methodology*, vol. 1, pp. 1–11, 2010.

[67] G. Anbarjafari, S. Izadpanahi, and H. Demirel, "Video resolution enhancement by using discrete and stationary wavelet transforms with illumination compensation," *Signal, Image and Video Processing*, vol. 9, no. 1, pp. 87–92, 2015.

[68] C. Barron and I. A. Kakadiaris, "Estimating anthropometry and pose from a single image," in *CVPR*, vol. 1. IEEE, 2000, pp. 669–676.

[69] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *CVPR*, vol. 1. IEEE, 2000, pp. 677–684.

[70] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," in *CVPRW*. IEEE, 2010, pp. 1–8.

[71] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[72] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[73] P. Hu and D. Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified gaussians," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5600–5609.

[74] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.

[75] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *European Conference on Computer Vision*. Springer, 2016, pp. 717–732.

[76] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3063–3072.

[77] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[78] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, vol. 3, no. 4, 2017, p. 6.

[79] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[80] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," *arXiv preprint arXiv:1803.04758*, 2018.

[81] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1646–1661, 2007.

[82] U. Iqbal, A. Milan, and J. Gall, "Posetrack: Joint multi-person pose estimation and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[83] H. Coskun, "Human pose estimation with cnns and lstms," Master's thesis, Universitat Politècnica de Catalunya, 2016.

[84] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 332–347.

[85] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *CVPR*, vol. 2, no. 5, 2017, p. 6.

[86] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," *3DV*, vol. 1, no. 2, p. 5, 2017.

[87] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[88] M. Andriluka, U. Iqbal, A. Milan, E. Insafutdinov, L. Pishchulin, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5167–5176.

[89] H. Gunes, M. Piccardi, and T. Jan, "Face and body gesture recognition for a vision-based multimodal analyzer," in *Proceedings of the Pan-Sydney area workshop on Visual information processing*. Australian Computer Society, Inc., 2004, pp. 19–28.

[90] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE international conference on systems, man and cybernetics*, vol. 4. IEEE, 2005, pp. 3437–3443.

[91] H. A. Vu, Y. Yamazaki, F. Dong, and K. Hirota, "Emotion recognition based on human gesture and speech information using rt middleware," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, 2011, pp. 787–791.

[92] M. Kipp and J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?" in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–8.

[93] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *CVPRW*. IEEE, 2008, pp. 1–6.

[94] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 33–48, 2010.

[95] D. Glowinski, M. Mortillaro, K. Scherer, N. Dael, and G. V. A. Camurri, "Towards a minimal representation of affective gestures," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 498–504.

[96] G. Castellano, "Movement expressivity analysis in affective computers: from recognition to expression of emotion," *Unpublished doctoral dissertation). Department of Communication, Computer and System Sciences, University of Genoa, Italy*, 2008.

[97] K. Hirota, H. A. Vu, P. Q. Le, C. Fatichah, Z. Liu, Y. Tang, M. L. Tangel, Z. Mu, B. Sun, F. Yan *et al.*, "Multimodal gesture recognition based on choquet integral," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, 2011, pp. 772–776.

[98] K. Altun and K. E. MacLean, "Recognizing affect in human touch of a robot," *Pattern Recognition Letters*, vol. 66, pp. 31–40, 2015.

[99] A. Lim and H. G. Okuno, "The mei robot: towards using motherese to develop multimodal emotional intelligence," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 126–138, 2014.

[100] S. Saha, S. Datta, A. Konar, and R. Janarthanan, "A study on emotion recognition from body gestures using kinect sensor," in *Communications and Signal Processing (ICCSP), 2014 International Conference on*. IEEE, 2014, pp. 056–060.

[101] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, and G. Volpe, "Toward real-time multimodal processing: Eyesweb 4.0," in *Proc. Artificial Intelligence and the Simulation of Behaviour (AISB) 2004 Convention: Motion, Emotion and Cognition*. Citeseer, 2004, pp. 22–26.

[102] S. Piana, A. Stagliano, A. Camurri, and F. Odone, "A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition," in *IDGEI International Workshop*, 2013.

[103] A. Patwardhan and G. Knapp, "Augmenting supervised emotion recognition with rule-based decision model," *arXiv preprint arXiv:1607.02660*, 2016.

[104] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 71–82.

[105] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.

[106] T. Nakamura, K. Taki, H. Nomiya, and K. Uehara, "Amss: A similarity measure for time series data," *IEICE Transactions on Information and Systems*, vol. 91, pp. 2579–2588, 2008.

[107] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," *Image and Vision Computing*, vol. 31, no. 2, pp. 175–185, 2013.

[108] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks*, vol. 72, pp. 140–151, 2015.

[109] J. Botzheim, J. Woo, N. T. N. Wi, N. Kubota, and T. Yamaguchi, "Gestural and facial communication with smart phone based robot partner using emotional model," in *World Automation Congress (WAC), 2014*. IEEE, 2014, pp. 644–649.

[110] F. Lv and R. Nevatia, "Recognition and segmentation of 3-d human action using hmm and multi-class adaboost," *ECCV*, pp. 359–372, 2006.

[111] S. Vacek, S. Knoop, and R. Dillmann, "Classifying human activities in household environments," in *Workshop at the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[112] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 4, pp. 1027–1038, 2011.

[113] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience," *IEEE Transactions on Computational Intelligence and AI in games*, vol. 4, no. 3, pp. 199–212, 2012.

[114] H. J. Griffin, M. S. Aung, B. Romera-Paredes, C. McLoughlin, G. McKeown, W. Curran, and N. Bianchi-Berthouze, "Laughter type recognition from whole body motion," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 349–355.

[115] G. Venture, H. Kadone, T. Zhang, J. Grèzes, A. Berthoz, and H. Hicheur, "Recognizing emotions conveyed by human gait," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 621–632, 2014.

[116] A.-A. Samadani, R. Gorbet, and D. Kulić, "Affective movement recognition based on generative and discriminative stochastic dynamic models," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 454–467, 2014.

[117] N. Fourati and C. Pelachaud, "Multi-level classification of emotional body expression," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[118] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, "Continuous body emotion recognition system during theater performances," *Computer Animation and Virtual Worlds*, vol. 27, no. 3-4, pp. 311–320, 2016.

[119] R. Laban, "The mastery of movement. revised and enlarged by lisa ullman," *Plymouth: Northcote House*, 1988.

[120] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[121] K. Kaza, A. Psaltis, K. Stefanidis, K. C. Apostolakis, S. Thermos, K. Dimitropoulos, and P. Daras, "Body motion analysis for emotion recognition in serious games," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2016, pp. 33–42.

[122] S. Li, L. Cui, C. Zhu, B. Li, N. Zhao, and T. Zhu, "Emotion recognition using kinect motion capture data of human gaits," *PeerJ*, vol. 4, p. e2364, 2016.

[123] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 1, p. 6, 2016.

[124] J. Arunnehru and M. K. Geetha, "Automatic human emotion recognition in surveillance video," in *Intelligent Techniques in Signal Processing for Multimedia Security*. Springer, 2017, pp. 321–342.

[125] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[126] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice." *Emotion*, vol. 7, no. 3, p. 487, 2007.

[127] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[128] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2007, pp. 375–388.

[129] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal affective state recognition in serious games applications," in *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*. IEEE, 2016, pp. 435–439.

[130] Z. Yang and S. S. Narayanan, "Analysis of emotional effect on speech-body gesture interplay," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[131] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A survey of autonomous human affect detection methods for social robots engaged in natural hri," *Journal of Intelligent & Robotic Systems*, vol. 82, no. 1, pp. 101–133, 2016.

[132] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 524–528.

[133] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *European Conference on Computer Vision*. Springer, 2016, pp. 400–418.

[134] R. W. Picard and R. Picard, *Affective computing*. MIT Press, 1997, vol. 252.

[135] R. W. Picard, "Affective computing for hci." in *HCI (1)*, 1999, pp. 829–833.

[136] R. W. Picard, "Affective computing: from laughter to ieee," *TAC*, vol. 1, no. 1, pp. 11–17, 2010.

[137] J. Cassell, "A framework for gesture generation and interpretation," *Computer vision in human-machine interaction*, pp. 191–215, 1998.

[138] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *TPAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[139] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[140] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

[141] A. Pentland, "Looking at people: Sensing for ubiquitous and wearable computing," *TPAMI*, vol. 22, no. 1, pp. 107–119, 2000.

[142] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *TPAMI*, vol. 23, no. 10, pp. 1175–1191, 2001.

[143] B. Reeves and C. Nass, *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK, 1996.

[144] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, and R. Picard, "Real-time inference of mental states from facial expressions and upper body gestures," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 909–914.

[145] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 966–979, 2012.

[146] E. Douglas-Cowie, C. Cox, J.-C. Martin, L. Devillers, R. Cowie, I. Sneddon, M. McRorie, C. Pelachaud, C. Peters, O. Lowry *et al.*, "The humaine database," in *Emotion-Oriented Systems*. Springer, 2011, pp. 243–284.

[147] M. Gavrilescu, "Recognizing emotions from videos by studying facial expressions, body postures and hand gestures," in *Telecommunications Forum Telfor (TELFOR), 2015 23rd*. IEEE, 2015, pp. 720–723.

[148] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *TAC*, vol. 6, no. 1, pp. 43–55, 2015.

[149] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPRW*, 2010, pp. 9–14.

[150] I. Humaine, "Human-machine interaction network on emotion, 2004-2007," 2008.

[151] S. Z. Masood, C. Ellis, A. Nagaraja, M. F. Tappen, J. J. LaViola, and R. Sukthankar, "Measuring and reducing observational latency when recognizing actions," in *ICCVW*, 2011, pp. 422–429.

[152] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[153] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[154] N. Fourati and C. Pelachaud, "Emilya: Emotional body expression in daily actions database." in *LREC*, 2014, pp. 3486–3493.

[155] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 420–436, 2013.

[156] H. Chauhan and A. Chauhan, "Implementation of decision tree algorithm c4. 5," *International Journal of Scientific and Research Publications*, vol. 3, no. 10, 2013.

Fatemeh Noroozi received her B.Sc. in Computer Engineering, Software, from Shiraz University, Iran. Her thesis was entitled "Modeling of Virtual Organizations Integrating on Physical Core based on a Service-oriented Architecture". Afterwards, she received her M.Sc. in Mechatronics Engineering from the University of Tehran, Iran. Her thesis was entitled "Developing a Real-time Virtual Environment for Connecting to a Touching Interface in Dental Applications ". Currently, she is a PhD student at the University of Tartu, Estonia, working on "multi-modal Emotion Recognition based Human-robot Interaction Enhancement".

Dorota Kamińska graduated in Automatic Control and Robotics and completed postgraduate studies in Biomedical image processing and analysis at ÅĄodz University of Technology. She received her PhD degree from Faculty of Electrical, Electronic, Computer and Control Engineering at ÅĄodz University of Technology in 2014. The topic of her thesis was âĂIJEmotion recognition from spontaneous speechâĂİ. She gained experience during the TOP 500 Innovators programme at Haas School of Business, University of California in Berkeley. Currently she is an educator and scientist at Institute of Mechatronics and Information Systems. She is passionate about biomedical signals processing for practical appliances. As a participant of many interdisciplinary and international projects, she is constantly looking for new challenges and possibilities of self-development.

Ciprian Adrian Corneanu got his BSc in Telecommunication Engineering from Télécom SudParis in 2011 and his MSc in Computer Vision from Universitat Autónoma de Barcelona in 2015. Currently he is a Ph.D. student at the Universitat de Barcelona and a fellow of the Computer Vision Center, UAB. His main research interests include face and behavior analysis, affective computing, social signal processing and human computer interaction.

Tomasz Sapiński received his M.Sc. degree in Computer Science from Faculty of Technical Physics, Information Technology and Applied Mathematics at ÅĄodz University of Technology. Currently he is Ph.D. student at Institute of Mechatronics and Information Systems, ÅĄodz University of Technology. His main research topics are: multi-modal emotion recognition and practical applications of virtual reality.

Sergio Escalera obtained the P.h.D. degree on Multi-class visual categorization systems at Computer Vision Center, UAB. He obtained the 2008 best Thesis award on Computer Science at Universitat AutÃšnoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB, CVC, and the Barcelona Graduate School of Mathematics. He is an associate professor at the Department of Mathematics and Informatics, Universitat de Barcelona. He is an adjunct professor at Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University. He has been visiting professor at TU Delft and Aalborg Universities. He is also a member of the Computer Vision Center at UAB. He is series editor of The Springer Series on Challenges in Machine Learning. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. His research interests include, between others, statistical pattern recognition, affective computing, and human pose recovery and behavior understanding, including multimodal data analysis.

Gholamreza Anbarjafari heads the intelligent computer vision (iCV) research lab in the Institute of Technology at the University of Tartu. He is an IEEE Senior member and the Chair of the Signal Processing / Circuits and Systems / Solid-State Circuits Joint Societies Chapter of the IEEE Estonian section. He received the Estonian Research Council Grant (PUT638) and the Scientific and Technological Research Council of Turkey (116E097) in 2015 and 2017, respectively. He has been involved in many international industrial projects. He is expert in computer vision, human-robot interaction, graphical models and artificial intelligence. He is an associated editor of several journals such as SIVP, Information and JIVP and have been lead guest editor of several special issues on human behaviour analysis. He has supervised over 10 MSc students and 7 PhD students. He has published over 100 scientific works. He has been in the organizing committee and technical committee of conferences such as ICOSST, ICGIP, SIU, SampTA, FG and ICPR. He is organizing a challenge and a workshop on in FG17, CVPR17, and ICCV17.