14

BODILY EXPRESSION FOR AUTOMATIC AFFECT RECOGNITION

HATICE GUNES

School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

CAIFENG SHAN

Philips Research Eindhoven, High Tech Campus, Eindhoven, The Netherlands

SHIZHI CHEN AND YINGLI TIAN

Department of Electrical Engineering, The City College of New York, NY, USA

This chapter focuses on the why, what, and how of bodily expression analysis for automatic affect recognition. It first asks the question of 'why bodily expression?' and attempts to find answers by reviewing the latest bodily expression perception literature. The chapter then turns its attention to the question of 'what are the bodily expressions recognized automatically?' by providing an overview of the automatic bodily expression recognition literature. The chapter then provides representative answers to how bodily expression analysis can aid affect recognition by describing three case studies: (1) data acquisition and annotation of the first publicly available database of affective face-and-body displays (i.e., the FABO database); (2) a representative approach for affective state recognition from face-and-body display by detecting the space-time interest points in video and using Canonical Correlation Analysis (CCA) for fusion, and (3) a representative approach for explicit detection of the temporal phases (segments) of affective states (start/end of the expression and its subdivision into phases such as neutral, onset, apex, and offset) from bodily expressions. The chapter concludes by summarizing the main challenges faced and discussing how we can advance the state of the art in the field.

Emotion Recognition: A Pattern Analysis Approach, First Edition. Edited by Amit Konar and Aruna Chakraborty.
 Published 2015 by John Wiley & Sons, Inc.

14.1 INTRODUCTION

Humans interact with others and their surrounding environment using their visual, auditory, and tangible sensing. The visual modality is the major input/output channel utilized for next generation human–computer interaction (HCI). Within the visual modality, the body has recently started gaining a particular interest due to the fact that in daily life body movements and gestures are an indispensable means for interaction. Not many of us realize the myriad ways and the extent to which we use our hands in everyday life: when we think, talk, and work. The gaming and entertainment industry is the major driving force behind putting the human body in the core of technology design by creating controller-free human–technology interaction experiences. Consequently, technology today has started to rely on the human body as direct input by reacting to and interacting with its movement [1, 2]. One example of this is the Kinect project [2] that enables users to control and interact with a video game console (the Xbox 360 [3]) through a natural user interface using gestures and spoken commands instead of a game controller.

Bodily cues (postures and gestures) have also started attracting the interest of researchers as a means to communicate emotions and affective states. Psychologists have long explored mechanisms with which humans recognize others' affective states from various cues and modalities, such as voice, face, and body gestures. This exploration has led to identifying the important role played by the modalities' dynamics in the recognition process. Supported by the human physiology, the temporal evolution of a modality appears to be well approximated by a sequence of temporal segments called onset, apex, and offset. Stemming from these findings, computer scientists, over the past 20 years, have proposed various methodologies to automate the affect recognition process. We note, however, two main limitations to date. The first is that much of the past research has focused on affect recognition from voice and face, largely neglecting the affective body display and bodily expressions. Although a fundamental study by Ambady and Rosenthal suggested that the most significant channels for judging behavioral cues of humans appear to be the visual channels of facial expressions and body gestures, affect recognition via body movements and gestures has only recently started attracting the attention of computer science and HCI communities. The second limitation is that automatic affect analyzers have not paid sufficient attention to the dynamics of the (facial and bodily) expressions: the automatic determination of the temporal segments and their role in affect recognition are yet to be adequately explored.

To address these issues, this chapter focuses on the why, what, and how of automatic bodily expression analysis. It first asks the question of "why bodily expression?" and attempts to find answers by reviewing the latest bodily expression perception literature. The chapter then turns its attention to the question of "what are the bodily expressions recognized automatically?" by providing an overview of the automatic bodily expression recognition literature and summarizing the main challenges faced in the field. The chapter then provides representative answers to how bodily expression analysis can aid affect recognition by describing three case studies: (1) data acquisition and annotation of the first publicly available database of affective face-and-body

displays (i.e., the FABO database); (2) a representative approach for affective state recognition from face-and-body display by detecting the space-time interest points in video and using Canonical Correlation Analysis (CCA) for fusion, and (3) a representative approach for explicit detection of the temporal phases (segments) of affective states (start/end of the expression and its subdivision into phases such as neutral, onset, apex, and offset) from bodily expressions.

Due to its popularity and extensive exploration, emotion communication through facial expressions will not be covered in this chapter. The interested readers are referred to References 4–11.

14.2 BACKGROUND AND RELATED WORK

Emotion communication through bodily expressions has been a neglected area for much of the emotion research history [12, 13]. This is illustrated by the fact that 95% of the literature on human emotions has been dedicated to using face stimuli, majority of the the remaining 5% on audio-based research, and the remaining small number on whole-body expressions [12]. This is indeed puzzling given the fact that early research on emotion by Darwin [14] and James [15] has paid a considerable attention to emotion-specific body movements and postural configurations. De Gelder argues that the reason why whole-body expressions have been neglected in emotion research is mainly due to the empirical results dating from the first generation of investigations of whole-body stimuli [12]. There are potentially other reasons as to why the body may seem a less reliable source of affective information (i.e., the face bias), its cultural and ideological reasons and heritage, which have been discussed in detail in Reference 12.

Overall, the body and hand gestures are much more varied than facial changes. There is an unlimited vocabulary of body postures and gestures with combinations of movements of various body parts (with multiple degrees of freedom) [13, 16, 17]. Therefore, using bodily expression for emotion communication and perception has a number of advantages:

- Bodily expression provides a means for recognition of affect from a distance. When we are unable to tell the emotional state from the face, we can still clearly read the action from the sight of the body [12]. This has direct implications for designing affective interfaces that will work in realistic settings (e.g., affective tutoring systems, humanoid robotics, affective games).
- Some of the basic mental states are most clearly expressed by the face while others are least ambiguous when expressed by the whole body (e.g., anger and fear) [12]. Perception of facial expression is heavily influenced by bodily expression as in most situations people do not bother to censor their body movements and therefore, the body is at times referred to as the *leaky* source [18]. Consequently, bodily expression, when used as an additional channel for affect communication, can provide a means to resolve ambiguity for affect detection and recognition.

Due to such advantages, automatic recognition of bodily expressions has increasingly started to attract the attention and the interest of the affective computing researchers. In this section, we will first review existing methods that achieve affect recognition and/or temporal segmentation from body display. Second, we will summarize existing systems that combine bodily expression with other cues or modalities in order to achieve multicue and multimodal affect recognition.

14.2.1 Body as an Autonomous Channel for Affect Perception and Analysis

Human recognition of emotions from body movements and postures is still an unresolved area of research in psychology and non-verbal communication. There are numerous works suggesting various opinions in this area. Ekman and Friesen have touched upon the possibility that some bodily (and facial) cues might be able to communicate both the quantity and quality aspects of emotional experience [19]. This leads to two major perspectives regarding the emotion perception and recognition from bodily posture and movement. The first perspective claims that there are body movements and postures that mostly contribute to the understanding of the activity (and intensity) level of the underlying emotions. For instance, Wallbot provided associations between body movements and the arousal dimension of emotion. More specifically, lateralized hand/arm movements, arms stretched out to the front, and opening and closing of the hands were observed during active emotions, such as hot anger, cold anger, and interest [20]. This can somewhat be seen as contributing toward the dimensional approach to emotion perception and recognition from bodily cues. The second perspective considers bodily cues (movements and postures) to be an independent channel of expression able to convey discrete emotions. An example is De Meijer's work that illustrated that observers are able to recognize emotions from body movements alone [21].

In general, recognition of affect from bodily expressions is mainly based on categorical representation of affect. The categories happy, sad, and angry appear to be more distinctive in motion than categories such as pride and disgust. Darwin suggested that in anger, for instance, among other behaviors, the whole body trembles, the head is erect, the chest is well expanded, feet are firmly on the ground, elbows are squared [14, 20]. Wallbot also analyzed emotional displays by actors and concluded that discrete emotional states can be recognized from body movements and postures. For instance, hot anger was encoded by shoulders moving upwards, arms stretched frontally, or lateralized, the execution of various hand movements, as well as high movement activity, dynamism, and expansiveness. Analysis of the arm movements (drinking and knocking) shows that, discrete affective states are aligned with the arousal–pleasure space [22]; and arousal was found to be highly correlated with velocity, acceleration, and jerk of the movement.

To date, the bodily cues that have been more extensively considered for affect recognition are (static) postural configurations of head, arms, and legs [16, 23], dynamic hand/arm movements [20], head movements (e.g., position and rotation) [24], and head gestures (e.g., head nods and shakes) [25, 26].

14.2.1.1 Body Posture Coulson [16] presented experiments on attributing six universal emotions (anger, disgust, fear, happiness, sadness, and surprise) to static body postures using computer-generated mannequin figures. His experimental results suggested that recognition from body posture is comparable to recognition from voice, and some postures are recognized as well as facial expressions.

When it comes to automatic analysis of affective body postures the main emphasis has been on using the tactile modality (for gross bodily expression analysis) via body-pressure-based affect measurement (e.g., [27]) and on using motion capture technology (e.g., [23]). Mota and Picard [27] studied affective postures in an e-learning scenario, where the posture information was collected through a sensor chair. Kleinsmith *et al.* [23] focused on the dimensional representation of emotions and on acquiring and analyzing affective posture data using motion capture technology [23]. They examined the role of affective dimensions in static postures for automatic recognition and showed that it is possible to automatically recognize the affect dimensions of arousal, valence, potency, and avoidance with acceptable recognition rates (i.e., error rates lower than 21%).

14.2.1.2 Body Movement Compared to the facial expression literature, attempts for recognizing affective body movements are few and efforts are mostly on the analysis of posed bodily expression data. Burgoon *et al.* discussed the issue of emotion recognition from bodily cues and provided useful references in Reference 28. They claimed that affective states are conveyed by a set of cues and focus on the identification of affective states such as positivity, anger, and tension in videos from body and kinesics cues. Meservy *et al.* [29] focused on extracting body cues for detecting truthful (innocent) and deceptive (guilty) behavior in the context of national security. They achieved a recognition accuracy of 71% for the two-class problem (i.e., guilty/innocent). Bernhardt and Robinson analyzed non-stylized body motions (e.g., walking, running) for affect recognition [30] using kinematic features (e.g., velocity, acceleration, and jerk measured for each joint) and reported that the affective states angry and sad are more recognizable than neutral or happy.

Castellano *et al.* [31] presented an approach for the recognition of acted emotional states based on the analysis of body movement and gesture expressivity. They used the non-propositional movement qualities (e.g. amplitude, speed, and fluidity of movement) to infer emotions (anger 90%, joy 44%, pleasure 62%, sadness 48%). A similar technique was used to extract expressive descriptors of movement (e.g., quantity of motion of the body and velocity of the head movements) in a music performance and to study the dynamic variations of gestures used by a pianist [32]. They found that the timing of expressive motion cues (i.e., the attack and release of the temporal profile of the velocity of the head and the quantity of motion of the upper body) is important in explaining emotional expression in piano performances. Reference 33 presents a framework for analysis of affective behavior starting with a reduced amount of visual information related to human upper-body movements. The work uses the EyesWeb Library (and its extensions) for extracting a number of expressive gesture features (e.g., smoothness of gesture, gesture duration) by tracking

of trajectories of head and hands (from a frontal and a lateral view), and the GEMEP corpus (120 posed upper body gestures for 12 emotion classes from 10 subjects) for validation. The authors conclude that for distinguishing bodily expression of different emotions dynamic features related to movement quality (e.g., smoothness of gesture, duration of gesture) are more important than categorical features related to the specific type of gesture.

A number of researchers have also investigated how to map various visual signals onto emotion dimensions. Cowie *et al.* [25] investigated the emotional and communicative significance of head nods and shakes in terms of Arousal and Valence dimensions, together with dimensional representation of *solidarity*, *antagonism*, and *agreement*. Their findings suggest that both head nods and shakes clearly carry information about arousal. However, their significance for evaluating the valence dimensions is less clear (affected by access to words) [25]. In particular, the contribution of the head nods for valence evaluation appears to be more complicated than head shakes (e.g., "I understand what you say, and I care about it, but I don't like it").

14.2.1.3 Gait Gait, in the context of perception and recognition, refers to a person's individual walking style. Therefore, gait is a source of dynamic information by definition. Emotion perception and recognition from gait patterns is also a relatively new area of research [34,35]. Janssen *et al.* [34] focused on emotion recognition from human gait by means of kinetic and kinematic data using artificial neural nets. They conducted two experiments: (1) identifying participants' emotional states (normal, happy, sad, angry) from gait patterns and (2) analyzing effects on gait patterns of listening to different types of music (excitatory, calming, no music) while walking. Their results showed that subject-independent emotion recognition from gait patterns is indeed possible (up to 100% accuracy). Karg *et al.* [35] focused on using both discrete affective states and affective dimensions for emotion modeling from motion capture data. Person-dependent recognition of motion capture data reached 95% accuracy based on the observation of a single stride. This work showed that gait is a useful cue for the recognition of arousal and dominance dimensions.

14.2.1.4 Temporal Dynamics An expression is a dynamic event, which evolves from neutral, onset, apex to offset [36], a structure usually referred to as *temporal dynamics* or *temporal phases*. Evolution of such a temporal event is illustrated, for a typical facial expression, in Figure 14.1. The neutral phase is a plateau where there are no signs of muscular activation and the face is relaxed. The *onset* of the action/movement is when the muscular contraction begins and increases in intensity and the appearance of the face changes. The *apex* is a plateau usually where the intensity reaches a stable level and there are no more changes in facial appearance. The *offset* is the relaxation of the muscular action. A natural facial movement evolves over time in the following order: neutral(N) $\longrightarrow$ onset(On) $\longrightarrow$ apex(A) $\longrightarrow$ offset(Of) $\longrightarrow$ neutral(N). Other combinations such as multiple-apex facial actions are also possible.

Similarly, the temporal structure of a body gesture consists of (up to) five phases: preparation $\longrightarrow$ (pre-stroke) hold $\longrightarrow$ stroke $\longrightarrow$ (post-stroke) hold $\longrightarrow$ retraction. The *preparation* moves to the stroke's starting position and the *stroke* is the most

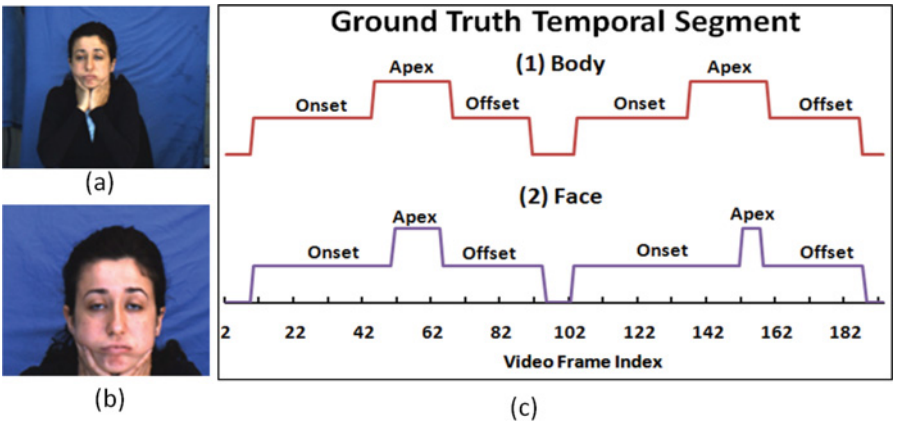


FIGURE 14.1 (a) Sample image of boredom expression to extract body gesture feature (body camera), (b) sample image of boredom expression to extract facial feature (face camera), and (c) the corresponding temporal segments from body gesture and facial features respectively. Taken with permission from the FABO database.

energetic part of the gesture. *Holds* are optional still phases which can occur before and/or after the stroke. The *retraction* returns to a *rest* pose (e.g., arms hanging down, resting in lap, or arms folded). Some gestures (e.g., finger tapping) have multiple strokes that include small beat-like movements that follow the first stroke, but seem to belong to the same gesture [37].

Studies demonstrate that the temporal dynamics play an important role for interpreting emotional displays [38, 39]. It is believed that information about the time course of a facial action may have psychological meaning relevant to the intensity, genuineness, and other aspects of the expresser's state. Among the four temporal phases of neutral, onset, apex, and offset, features during the apex phase result in maximum discriminative power for expression recognition. Gunes and Piccardi showed that, during automatic affect recognition from facial/bodily gestures, decoupling temporal dynamics from spatial extent significantly reduces the dimensionality of the problem compared to dealing with them simultaneously and improves affect recognition accuracy [40]. Thus, successful temporal segmentation can not only help to analyze the dynamics of an (facial/bodily) expression, but also improve the performance of expression recognition. However, in spite of their usefulness, the complex spatial properties and dynamics of face and body gestures also pose a great challenge to affect recognition. Therefore, interest in the temporal dynamics of affective behavior is recent (e.g., [11, 40–42]). The work of Reference 41 temporally segmented facial action units (AUs) using geometric features of 15 facial key points from profile face images. In Reference 37, a method for the detection of the temporal phases in natural gesture was presented. For body movement, a finite-state machine (FSM) was used to spot multiphase gestures against a rest state. In order to detect the gesture phases, candidate rest states were obtained and evaluated. Three variables were used to model the states: distance from rest image, motion magnitude, and duration. Other

approaches have exploited dynamics of the gestures without attempting to recognize their temporal phases or segments explicitly (e.g., [31, 43] and [29]).

14.2.2 Body as an Additional Channel for Affect Perception and Analysis

Ambady and Rosenthal reported that human judgment of behaviors based jointly on face and body proved 35% more accurate than those based on the face alone [44]. The face and the body, as part of an integrated whole, both contribute in conveying the emotional state of the individual. A single body gesture can be ambiguous. For instance, the examples shown in the second and fourth rows in Figure 14.2 have similar bodily gestures, but the affective states they express are quite different, as shown by the corresponding facial expressions. In light of such findings, instead of looking at the body as an independent and autonomous channel of emotional expression, researchers have increasingly focused on the relationship between bodily postures and movement with other expressive channels such as voice and face (e.g., [40, 45, 46]).

FIGURE 14.2 Example images from the FABO database recorded by the face (top) and body (bottom) cameras separately. Representative images of non-basic facial expressions (a1–h1) and their corresponding body gestures (a2–h2): (a) neutral, (b) negative surprise, (c) positive surprise, (d) boredom, (e) uncertainty, (f) anxiety, and (g) puzzlement. Taken with permission from the FABO database.

It is important to state that automatic affect recognition does not aim to replace one expression channel (e.g., the facial expressions) as input by another expression channel (e.g., bodily expressions). Instead, the aim is to explore various communicative channels more deeply and more fully in order to obtain a thorough understanding of cross-modal interaction and correlations pertaining to human affective display. An example is the work of Van den Stock *et al.* investigating the influence of whole-body expressions of emotions on the recognition of facial and vocal expressions of emotion [47]. They found that recognition of facial expression was strongly influenced by the bodily expression. This effect was a function of the ambiguity of the facial expression. Overall, during multisensory perception, judgments for one modality seem to be influenced by a second modality, even when the latter modality can provide no information about the judged property itself or increase ambiguity (i.e., cross-modal integration) [48, 49]. Meeren *et.al.* [50] showed that the recognition of facial expressions is strongly influenced by the concurrently presented emotional body language and that the affective information from the face and the body start to interact rapidly, and the integration is a mandatory automatic process occurring early in the processing stream. Therefore, fusing facial expression and body gesture in video sequences provides a potential way to accomplish improved affect analysis.

When it comes to using the body as an additional channel for automatic analysis, the idea of combining face and body expressions for affect recognition is relatively new. Balomenos *et al.* [51] combined facial expressions and hand gestures for the recognition of six prototypical emotions. They fused the results from the two subsystems at a decision level using pre-defined weights. An 85% accuracy was achieved for emotion recognition from facial features alone. An overall recognition rate of 94.3% was achieved for emotion recognition from hand gestures. Karpouzis *et al.* [52] fused data from facial, bodily, and vocal cues using a recurrent network to detect emotions. They used data from four subjects and reported the following recognition accuracies for a 4-class problem: 67% (visual), 73% (prosody), 82% (all modalities combined). The fusion was performed on a frame basis, meaning that the visual data values were repeated for every frame of the tune. Neither work has focused on explicit modeling and detection of the (facial/bodily) expression temporal segments. Castellano *et al.* considered the possibility of detecting eight emotions (some basic emotions plus irritation, despair, etc.) by monitoring facial features, speech contours, and gestures [45]. Their findings suggest that incorporating multiple cues and modalities helps with improving the affect recognition accuracy, and the best channel for affect recognition appears to be the gesture channel followed by the audio channel.

Hartmann *et al.* [53] defined a set of expressivity parameters for the generation of expressive gesturing for virtual agents. The studies conducted on perception of expressivity showed that only a subset of parameters and a subset of expressions were recognized well by users. Therefore, further research is needed for the refinement of the proposed parameters (e.g., the interdependence of the expressivity parameters). Valstar *et al.* [54] investigated separating posed from genuine smiles in video sequences using facial, head, and shoulder movement cues, and the temporal correlation between these cues. Their results seem to indicate that using video data from

face, head, and shoulders increases the accuracy, and the head is the most reliable source, followed closely by the face. Nicolaou *et al.* capitalize on the fact that the arousal and valence dimensions are correlated, and present an approach that fuses spontaneous facial expression, shoulder gesture, and audio cues for dimensional and continuous prediction of emotions in valence-arousal space [55]. They propose an output-associative fusion framework that incorporates correlations between emotion dimensions. Their findings suggest that incorporating correlations between affect dimensions provides greater accuracy for continuous affect prediction. Audio cues appear to be better for predicting arousal, and visual cues (facial expressions and shoulder movements) appear to perform better for predicting valence.

A number of systems use the tactile modality for gross bodily expression analysis via body-pressure-based affect measurement (measuring participants' back and seat pressure) [56, 57]. Kapoor and Picard focused on the problem of detecting the affective states of high interest, low interest, and refreshing in a child who is solving a puzzle [57]. They combined sensory information from the face video, the posture sensor (a chair sensor) and the game being played in a probabilistic framework. The classification results obtained by Gaussian Processes for individual modalities showed that affective states are best classified by the posture channel (82%), followed by the features from the upper face (67%), the game (57%), and the lower face (53%). Fusion significantly outperformed classification using the individual modalities and resulted in 87% accuracy. D'Mello and Graesser [56] considered a combination of facial features, gross body language, and conversational cues for detecting some of the learning-centered affective states. Classification results supported a *channel*∗*judgment*-type interaction, where the face was the most diagnostic channel for spontaneous affect judgments (i.e., at any time in the tutorial session), while conversational cues were superior for fixed judgments (i.e., every 20 seconds in the session). The analyzers also indicated that the accuracy of the multichannel model (face, dialog, and posture) was statistically higher than the best single-channel model for the fixed but not spontaneous affect expressions. However, multichannel models reduced the discrepancy (i.e., variance in the precision of the different emotions) of the discriminant models for both judgment types. The results also indicated that the combination of channels yielded enhanced effects for some states but not for others.

14.2.3 Bodily Expression Data and Annotation

Communication of emotions by body gestures is still an unresolved area in psychology. Therefore, the number of databases and corpus that contain expressive bodily gestures and are publicly available for research purposes is scarce, and there exists no annotation scheme commonly used by all researchers in the field.

Data. To the best of our knowledge there exist three publicly available databases that contain expressive bodily postures or gestures. *The UCLIC Database of Affective Postures and Body Movements* [58] contains acted emotion data (angry, fearful, happy, and sad) collected using a VICON motion capture system, and non-acted affective states (frustration, concentration, triumphant, and defeated) in a computer game

setting collected using a Gypsy5 (Animazoo UK Ltd.) motion capture system. *The GEMEP Corpus* (The Geneva Multimodal Emotion Portrayals Corpus) [59] contains 120 posed face and upper-body gestures (head and hand gestures), for 12 emotion classes (pride, joy, amusement, interest, pleasure, relief, hot anger, panic fear, despair, irritation, anxiety, and sadness) from 10 subjects recorded by multiple cameras (e.g., frontal and lateral view). The Bimodal Face and Body Gesture Database (the FABO database) comprises of face-and-body expressions [60] and will be reviewed in detail in the next sections.

Annotation. Unlike the facial actions, there is not one common annotation scheme that can be adopted by all the research groups [13] to describe and annotate the body AUs that carry expressive information. Therefore, it is even harder to create a common benchmark database for affective gesture recognition. The most common annotation has been command-purpose annotation, for instance calling the gesture as rotate or click gesture. Another type of annotation is based on the gesture phase, for example, start of gesture stroke-peak of gesture stroke-end of gesture stroke. Rudolf Laban was a pioneer in attempting to analyze and record body movement by developing a systematic annotation scheme called Labanotation [61]. Traditionally Labanotation has been used mostly in dance choreography, physical therapy, and drama for exploring natural and choreographed body movement. Despite the aforementioned effort of Laban in analyzing and annotating body movement, a more detailed annotation scheme, similar to that of Facial Action Coding Scheme (FACS) is needed. A gesture annotation scheme, possibly named as Body Action Unit Coding System (BACS), should include information and description as follows: body part (e.g., left hand), direction (e.g., up/down), speed (e.g., fast/slow), shape (clenching fists), space (flexible/direct), weight (light/strong), time (sustained/quick), and flow (fluent/controlled) as defined by Laban and Ullman [61]. Additionally, temporal segments (neutral-start of gesture stroke-peak of gesture stroke-end of gesture stroke-neutral) of the gestures should be included as part of the annotation scheme. Overall, the most time-costly aspect of current gesture manual annotation is to obtain the onset-apex-offset time markers. This information is crucial for coordinating facial/body activity with simultaneous changes in physiology or speech [62].

14.3 CREATING A DATABASE OF FACIAL AND BODILY EXPRESSIONS: THE FABO DATABASE

The Bimodal Face and Body Gesture Database (the FABO database, henceforth) was created with the aim of using body as an additional channel, together with face, for affect analysis and recognition. The goal was to study how affect can be expressed, and consequently analyzed, when using both the facial and the bodily expression channels simultaneously. Details on the recordings and data annotation are described in the following sections.

Recordings. We recorded the video sequences simultaneously using two fixed cameras with a simple setup and uniform background. One camera was placed to specifically capture the face alone and the second camera was placed in order to

capture face-and-body movement from the waist above. Prior to recordings subjects were instructed to take a neutral position, facing the camera and looking straight to it with hands visible and placed on the table. The subjects were asked to perform face and body gestures simultaneously by looking at the facial camera constantly. The recordings were obtained by using a *scenario approach* that was also used in previous emotion research [63]. In this approach, subjects are provided with situation vignettes or short scenarios describing an emotion-eliciting situation. They are instructed to imagine these situations and act out as if they were in such a situation. In our case the subjects were asked what they would do when "it was just announced that they won the biggest prize in lottery" or "the lecture is the most boring one and they can't listen to it anymore," etc. More specifically, although the FABO database was created in laboratory settings, the subjects were not instructed on emotion/case basis as to how to move their facial features and how to exactly display the specific facial expression. In some cases the subjects came up with a variety of combinations of face and body gestures. As a result of the feedback and suggestions obtained from the subjects, the number and combination of face and body gestures performed by each subject varies. A comprehensive list is provided in Table 14.1. The FABO database contains around 1900 gesture sequences from 23 subjects in age from 18 to 50 years. Figure 14.2 shows example images of non-basic facial expressions and their corresponding body gestures for neutral, negative surprise, positive surprise, boredom, uncertainty, anxiety, and puzzlement. Further details on the FABO database recordings can be found in Reference 60.

Annotation. We obtained the annotations for face and body videos separately, by asking human observers to view and label the videos. The purpose of this annotation was to obtain independent interpretations of the displayed face and body expressions and evaluate the performance (i.e., how well the subjects were displaying the affect they intended to communicate using their face and bodily gesture) by a number of human observers from different ethnic and cultural background. To this aim, we developed a survey for face and body videos separately, using the labeling schemes for affective content (e.g., happiness) and signs (e.g., how contracted the body is) by asking six independent human observers. We used two main labeling schemes in line with the psychological literature on descriptors of emotion: (a) verbal categorical labeling (perceptually determined, i.e., happiness) in accordance with Ekman's theory of emotion universality [64] and (b) broad dimensional labeling: arousal/activation (arousal–sleep/activated–deactivated) in accordance with Russell's theory of arousal and valence [65]. The participants were first shown the whole set of facial videos and only after finishing with the face they were shown the corresponding body videos. For each video they were asked to choose one label only, from the list provided: sadness, puzzlement/thinking, uncertainty/"I don't know," boredom, neutral surprise, positive surprise, negative surprise, anxiety, anger, disgust, fear, and happiness. For the temporal segment annotation, one human coder repeatedly viewed each face and body sequence, in slowed and stopped motion, to determine when (in which frame) the neutral–onset–apex–offset–neutral phases start and end [66]. Further details on the FABO data annotation can be found in Reference 49.

TABLE 14.1 List of the affective face and upper-body gestures performed for the recordings of FABO database

Expression	Face gesture	Body gesture
Neutral	Lips closed, eyes open, muscles relaxed	Hands on the table, relaxed
Uncertainty and puzzlement	Lip suck, lid droop, eyes closed, eyes turn right/left/up/down	Head tilt left/right/up/down, shoulder shrug, palms up, palms up + shoulder shrug, right/left hand scratching the head/hair, right/left hand touching the right/left ear, right/left hand touching the nose, right/left hand touching the chin, right/left hand touching the neck, right/left hand touching the forehead, both hands touching the forehead, right/left hand below the chin, elbow on the table, two hands behind the head
Anger	Brows lowered and drawn together; lines appear between brows; lower lid tense/may be raised; upper lid tense/may be lowered due to brows' action; lips are pressed together with corners straight or down or open; nostrils may be dilated	Open/expanded body; hands on hips/waist; closed hands/clenched fists; palm-down gesture; lift the right/left hand up; finger point with right/left hand; shake the finger/hand; crossing the arms
Surprise	Brows raised; skin below brow stretched not wrinkled;horizontal wrinkles across forehead; eyelids opened; jaw drops open or stretching of the mouth	Right/left hand moving toward the head; both hands moving toward the head; moving the right/left hand up; two hands touching the head; two hands touching the face/mouth; both hands over the head; right/left hand touching the face/mouth; self-touch/two hands covering the cheeks; self-touch/two hands covering the mouth; head shake; body shift/backing
Fear	Brows raised and drawn together; forehead wrinkles drawn to the center; upper eyelid is raised and lower eyelid is drawn up; mouth is open; lips are slightly tense or stretched and drawn back	Body contracted; closed body/closed hands/clenched fist; body contracted; arms around the body; self-touch (disbelief)/covering the body parts/arms around the body/shoulders; body shift-backing; hand covering the head; body shift-backing; hand covering the neck; body shift-backing; hands covering the face; both hands over the head; self-touch (disbelief) covering the face with hands

(*continued*)

TABLE 14.1 (*Continued*)

Expression	Face gesture	Body gesture
Anxiety	Lip suck; lip bite; lid droop; eyes closed; eyes turn right/left/up/down	Hands pressed together in a moving sequence; tapping the tips of the fingers on the table; biting the nails; head tilt left/right/up/down
Happiness	Corners of lips are drawn back and up; mouth may or may not be parted with teeth exposed or not; cheeks are raised; lower eyelid shows wrinkles below it; and may be raised but not tense; wrinkles around the outer corners of the eyes	Body extended; hands clapping; arms lifted up or away from the body with hands made into fists
Disgust	Upper lip is raised; lower lip is raised and pushed up to upper lip or it is lowered; nose is wrinkled; cheeks are raised; brows are lowered	Hands close to the body; body shift-backing; orientation changed/moving to the right or left; backing; hands covering the head; backing; hands covering the neck; backing; right/left hand on the mouth; backing; move right/left hand up
Bored	Lid droop, eyes closed, lip suck, eyes turn right/left/up/down	Body shift; change orientation; move to the right/left; hands behind the head; body shifted; hands below the chin, elbow on the table
Sadness	Inner corners of eyebrows are drawn up; upper lid inner corner is raised; corners of the lips are drawn downwards	Contracted/closed body; dropped shoulders; bowed head; body shift-forward leaning trunk; covering the face with two hands; self-touch (disbelief)/covering the body parts/arms around the body/shoulders; body extended+hands over the head; hands kept lower than their normal position, hands closed, slow motion; two hands touching the head move slowly; one hand touching the neck, hands together closed, head to the right, slow motion.

14.4 AUTOMATIC RECOGNITION OF AFFECT FROM BODILY EXPRESSIONS

14.4.1 Body as an Autonomous Channel for Affect Analysis

In this section, we first investigate affective body gesture analysis in video sequences by approaching the body as an autonomous channel. To this aim, we exploit

spatial–temporal features [67], which makes few assumptions about the observed data, such as background, occlusion, and appearance.

14.4.1.1 Spatial–Temporal Features In recent years, spatial–temporal features have been used for event detection and behavior recognition in videos. We extract spatial–temporal features by detecting space-time interest points [67]. We calculate the response function by application of separable linear filters. Assuming a stationary camera or a process that can account for camera motion, the response function has the form

$$R = (I * g * h_{\text{ev}})^2 + (I * g * h_{\text{od}})^2 \tag{14.1}$$

where $I(x, y, t)$ denotes images in the video, $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions (x, y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{\text{ev}}(t; \tau, \omega) = -\cos(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2}$ and $h_{\text{od}}(t; \tau, \omega) = -\sin(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2}$. In all cases we use $\omega = 4/\tau$ [67]. The two parameters σ and τ correspond roughly to the spatial and temporal scales of the detector. Each interest point is extracted as a local maxima of the response function. As pointed out in Reference 67, any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response, while region undergoing pure translational motion, or areas without spatially distinguishing features, will not induce a strong response.

At each detected interest point, a cuboid is extracted which contains the spatio-temporally windowed pixel values. See Figure 14.3 for examples of cuboids extracted. The side length of cuboids is set as approximately six times the scales along each dimension, so containing most of the volume of data that contribute to the response function at each interest point. After extracting the cuboids, the original video is discarded, which is represented as a collection of the cuboids. To compare two cuboids, different descriptors for cuboids have been evaluated in Reference 67, including normalized pixel values, brightness gradient and windowed optical flow, followed by a conversion into a vector by flattening, global histogramming, and local histogramming. As suggested, we adopt the flattened brightness gradient as the cuboid descriptor. To reduce the dimensionality, the descriptor is projected to a lower dimensional PCA space [67]. By clustering a large number of cuboids extracted from the training data using the K-Means algorithm, we derive a library of cuboid prototypes. So each cuboid is assigned a type by mapping it to the closest prototype vector. Following Reference 67, we use the histogram of the cuboid types to describe the video.

14.4.1.2 Classifier We adopt the support vector machine (SVM) classifier to recognize affective body gestures. SVM is an optimal discriminant method based on the Bayesian learning theory. For the cases where it is difficult to estimate the density model in high-dimensional space, the discriminant approach is preferable to the generative approach. SVM performs an implicit mapping of data into a higher dimensional feature space and then finds a linear separating hyperplane with the maximal margin to separate data in this higher dimensional space. SVM allows domain-specific

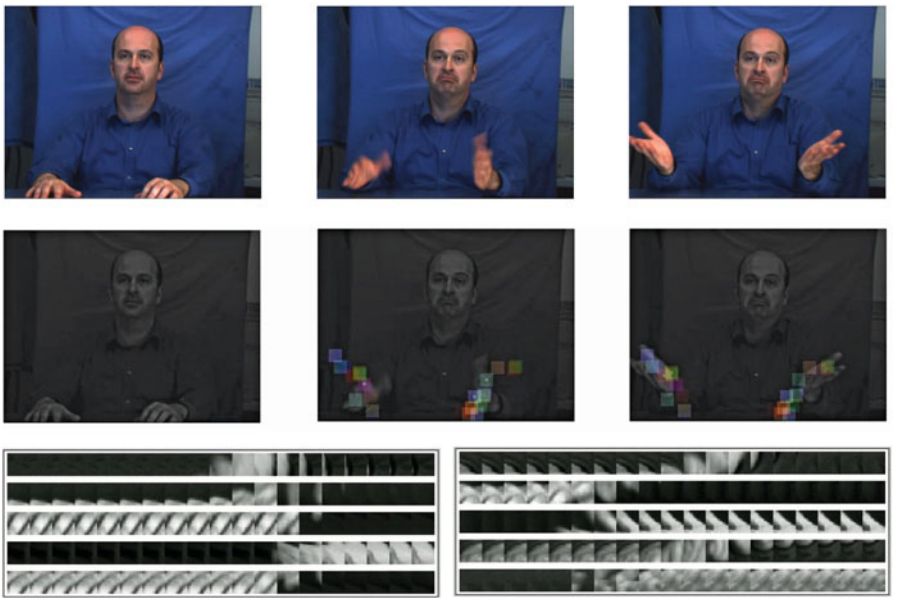


FIGURE 14.3 (Best viewed in color) Examples of spatial–temporal features extracted from videos. The first row is the original input video. Taken with permission from the FABO database. The second row visualizes the cuboids extracted, where each cuboid is labeled with a different color; the third row shows some cuboids, which are flattened with respect to time. Color version of the figure is available in the internet edition.

selection of the kernel function, and the most commonly used kernel functions are the linear, polynomial, and radial basis function (RBF) kernels.

14.4.2 Body as an Additional Channel for Affect Analysis

In this section, we investigate how body contributes to the affect analysis when used as an additional channel. For combining the facial and bodily cues, we exploit CCA, a powerful statistical tool that is well suited for relating two sets of signals, to fuse facial expression and body gesture at the feature level. CCA derives a semantic "affect" space, in which the face and body features are compatible and can be effectively fused.

We propose to fuse the cues from the two channels in a joint feature space, rather than at the decision level. The main difficulties for the feature-level fusion are the features from different modalities may be incompatible, and the relationship between different feature spaces is unknown. Here we fuse face and body cues at the feature level using CCA. Our motivation is that, as face and body cues are two sets of measurements for affective states, conceptually the two modalities are correlated, and their relationship can be established using CCA.

14.4.2.1 Canonical Correlation Analysis CCA [68] is a statistical technique developed for measuring linear relationships between two multidimensional

variables. It finds pairs of base vectors (i.e., canonical factors) for two variables such that the correlations between the projections of the variables onto these canonical factors are mutually maximized.

Given two zero-mean random variables $\mathbf{x} \in R^m$ and $\mathbf{y} \in R^n$, CCA finds pairs of directions $\mathbf{w}_x$ and $\mathbf{w}_y$ that maximize the correlation between the projections $x = \mathbf{w}_x^T\mathbf{x}$ and $y = \mathbf{w}_y^T\mathbf{y}$. The projections x and y are called *canonical variates*. More formally, CCA maximizes the function

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^T\mathbf{x}\mathbf{y}^T\mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^T\mathbf{x}\mathbf{x}^T\mathbf{w}_x]E[\mathbf{w}_y^T\mathbf{y}\mathbf{y}^T\mathbf{w}_y]}} = \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}} \quad (14.2)$$

where $\mathbf{C}_{xx} \in R^{m\times m}$ and $\mathbf{C}_{yy} \in R^{n\times n}$ are the *within-set covariance matrices* of $\mathbf{x}$ and $\mathbf{y}$, respectively, while $\mathbf{C}_{xy} \in R^{m\times n}$ denotes their *between-sets covariance matrix*. A number of at most $k = \min(m, n)$ canonical factor pairs $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ can be obtained by successively solving $\arg\max_{\mathbf{w}_x^i, \mathbf{w}_y^i}\{\rho\}$ subject to $\rho(\mathbf{w}_x^j, \mathbf{w}_x^i) = \rho(\mathbf{w}_y^j, \mathbf{w}_y^i) = 0$ for $j = 1, \ldots, i-1$, that is, the next pair of $\langle \mathbf{w}_x, \mathbf{w}_y \rangle$ are orthogonal to the previous ones.

The maximization problem can be solved by setting the derivatives of Equation. 14.2, with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, equal to zero, resulting in the eigenvalue equations as

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x = \rho^2\mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y = \rho^2\mathbf{w}_y \end{cases} \quad (14.3)$$

Matrix inversions need to be performed in Equation 14.3, leading to numerical instability if $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are rank deficient. Alternatively, $\mathbf{w}_x$ and $\mathbf{w}_y$ can be obtained by computing principal angles, as CCA is the statistical interpretation of principal angles between two linear subspaces.

14.4.2.2 Feature Fusion of Facial and Bodily Expression Cues Given $B = \{\mathbf{x}|\mathbf{x} \in R^m\}$ and $F = \{\mathbf{y}|\mathbf{y} \in R^n\}$, where $\mathbf{x}$ and $\mathbf{y}$ are the feature vectors extracted from bodies and faces, respectively, we apply CCA to establish the relationship between $\mathbf{x}$ and $\mathbf{y}$. Suppose $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ are the canonical factors pairs obtained, we can use d $(1 \le d \le k)$ factor pairs to represent the correlation information. With $\mathbf{W}_x = [\mathbf{w}_x^1, \ldots, \mathbf{w}_x^d]$ and $\mathbf{W}_y = [\mathbf{w}_y^1, \ldots, \mathbf{w}_y^d]$, we project the original feature vectors as $\mathbf{x}' = \mathbf{W}_x^T\mathbf{x} = [x_1, \ldots, x_d]^T$ and $\mathbf{y}' = \mathbf{W}_y^T\mathbf{y} = [y_1, \ldots, y_d]^T$ in the lower dimensional correlation space, where x_i and y_i are uncorrelated with the previous pairs x_j and $y_j, j = 1, \ldots, i-1$. We then combine the projected feature vectors $\mathbf{x}'$ and $\mathbf{y}'$ to form the new feature vector as

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x^T\mathbf{x} \\ \mathbf{W}_y^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x & 0 \\ 0 & \mathbf{W}_y \end{pmatrix}^T \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \quad (14.4)$$

This fused feature vector effectively represents the multimodal information in a joint feature space for affect analysis.

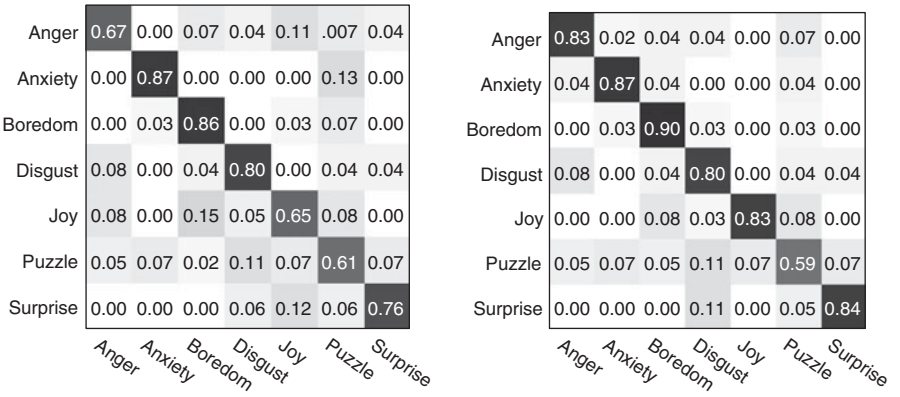


FIGURE 14.4 Confusion matrices for affect recognition from bodily gestures (*left*) and facial expressions (*left*).

14.4.2.3 Experiments and Results In our experiments we used the FABO database [60]. We selected 262 videos of seven emotions (Anger, Anxiety, Boredom, Disgust, Joy, Puzzle, and Surprise) from 23 subjects. To evaluate the algorithms' generalization ability, we adopted a fivefold cross-validation test scheme in all recognition experiments. That is, we divided the data set randomly into five groups with roughly equal number of videos and then used the data from four groups for training, and the left group for testing; the process was repeated five times for each group in turn to be tested. We report the average recognition rates here. In all experiments, we set the soft margin C value of SVMs to infinity so that no training error was allowed. Meanwhile, each training and testing vector was scaled to be between –1 and 1. In our experiments, the RBF kernel always provided the best performance, so we report the performance of the RBF kernel. With regard to the hyper-parameter selection of RBF kernels, as suggested in Reference 69, we carried out grid-search on the kernel parameters in the fivefold cross-validation. The parameter setting producing the best cross-validation accuracy was picked. We used the SVM implementation in the publicly available machine learning library SPIDER[1] in our experiments. To see how the body contributes to the affect analysis when used as an additional channel, we extracted the spatial–temporal features from the face video and the body video and then fused the cues from the two channels at the feature level using CCA.

We first report the classification performance (the confusion matrix) based on bodily cues only in Figure 14.4 (left). The average recognition rate of the SVM classifier using the bodily cues is 72.6%. When we look at the affect recognition using the facial cues only, the recognition rate obtained is 79.2%. Looking at the confusion matrix shown in Figure 14.4, we observe that the emotion classification based on facial expressions is better than that of bodily gesture. This is possibly because there are much variation in affective body gestures.

We then fused facial expression and body gesture at the feature level using CCA. Different numbers of CCA factor pairs can be used to project the original face and

[1] http://kyb.tuebingen.mpg.de/bs/people/spider/index.html

TABLE 14.2 Experimental results of affect recognition by fusing body and face cues

Feature fusion	CCA	Direct	PCA	PCA + LDA
Recognition rate	88.5%	81.9%	82.3%	87.8%

body feature vectors to a lower dimensional CCA feature space, and the recognition performance varies with the dimensionality of the projected CCA features. We report the best result obtained here. We compared the CCA feature fusion with another three feature fusion methods: (1) Direct feature fusion, that is, concatenating the original body and face features to derive a single feature vector. (2) PCA feature fusion: the original body and face features are first projected to the PCA space respectively, and then the PCA features are concatenated to form the single feature vector. In our experiments, all principle components were kept. (3) PCA+LDA feature fusion: for each modality, the derived PCA features are further projected to the discriminant LDA space; the LDA features are then combined to derive the single feature vector. We report the experimental results of different feature fusion schemes in Table 14.2. The confusion matrices of the CCA feature fusion and the direct feature fusion are shown in Figure 14.5. We can see that the presented CCA feature fusion provides best recognition performance. This is because CCA captures the relationship between the feature sets in different modalities, and the fused CCA features effectively represent information from each modality.

14.5 AUTOMATIC RECOGNITION OF BODILY EXPRESSION TEMPORAL DYNAMICS

Works focusing on the detection of the expression temporal segments modeled temporal dynamics of facial or bodily expressions by extracting and tracking geometric or appearance features from a set of fixed interest points (e.g., [40, 70]). However,

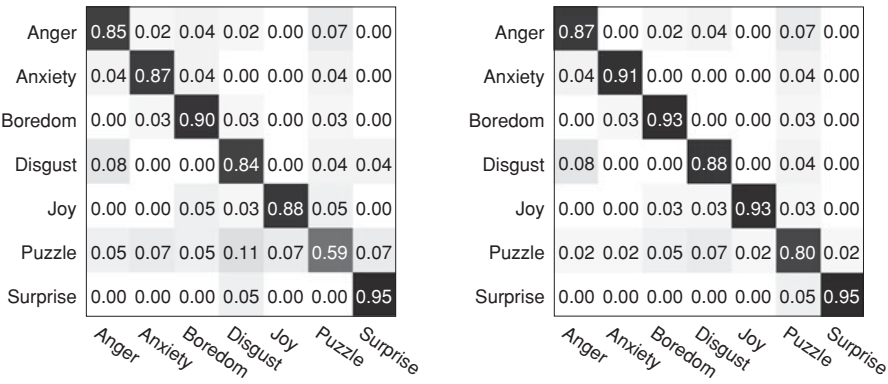


FIGURE 14.5 Confusion matrices of affect recognition by fusing facial expression and body gesture. (*Left*) Direct feature fusion; (*right*) CCA feature fusion. Taken with permission from the FABO database.

such approaches have two limitations. First, the selection of the fixed interest points requires human expertise and mostly needs human intervention. Second, tracking is usually sensitive to occlusions and illumination variations (e.g., the facial point tracking will fail when the hands touch the face). Inaccuracy in tracking will significantly degrade the temporal segmentation performance. To mitigate the aforementioned issues, we propose two types of novel and efficient features in this section, that is, motion area and neutral divergence, to simultaneously segment and recognize temporal phases of an (facial/bodily) expression. The motion area feature is calculated by simple motion history image (MHI) [71, 72], which does not rely on any facial points tracking or body tracking, and the neutral divergence feature is based on the differences between the current frame and the neutral frame.

14.5.1 Feature Extraction

The motion area and the neutral divergence features are extracted from both facial and body gesture information without any motion tracking, so the approach avoids losing informative apex frames due to the unsynchronized face and body gesture temporal phases. Furthermore, both features are efficient to compute.

14.5.1.1 Motion Area We extract the motion area based on the MHI, which is a compact representation of a sequence of motion movement in a video [71, 72]. Pixel intensity of MHI is a function of the motion history at that location, where brighter values correspond to more recent motions. The intensity at pixel (x, y) decays gradually until a specified motion history duration t and the MHI image can be constructed using the equation

$$\begin{aligned}\mathrm{MHI}_\tau(x,y,t) = {} & D(x,y,t) * \tau + [1 - D(x,y,t)] * U[\mathrm{MHI}_\tau(x,y,t-1) - 1] \\ & * [\mathrm{MHI}_\tau(x,y,t-1) - 1])\end{aligned} \tag{14.5}$$

where $U[x]$ is a unit step function and t represents the current video frame index. $D(x, y, t)$ is a binary image of pixel intensity difference between the current frame and the previous frame. $D(x, y, t) = 1$ if the intensity difference is greater than a threshold, otherwise, $D(x, y, t) = 0$. τ is the maximum motion duration. In our system, we set *threshold* $= 25$ and $\tau = 10$. Figure 14.6b shows the generated MHI of a *surprise* expression. The motion area of each video frame is the total number of the motion pixels in the corresponding MHI image. The motion pixels are defined as the pixels with non-zero intensity in the MHI image. The calculation of the motion area $\mathrm{MA}_\tau(t)$ can be described by the following equation:

$$\mathrm{MA}_\tau(t) = \sum_{x=1}^{W} \sum_{y=1}^{H} U[\mathrm{MHI}_\tau(x,y,t) - e]) \tag{14.6}$$

where $0 < e < 1$, $U[x]$ is a unit step function, and W and H are the width and the height of the MHI image.

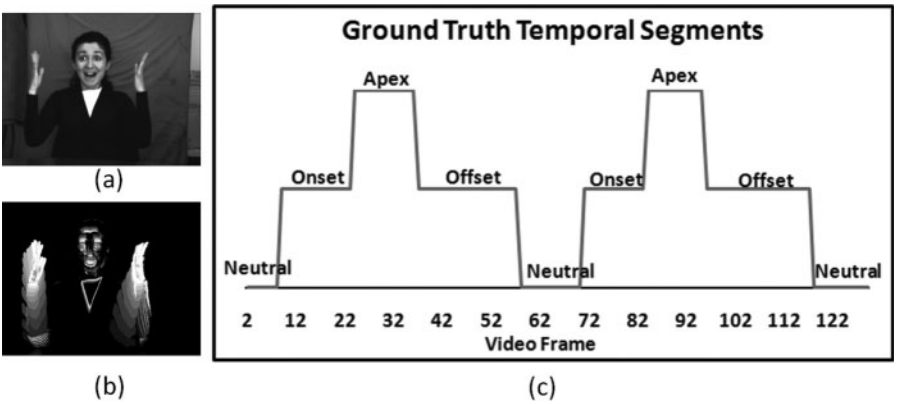


FIGURE 14.6 (a) A *surprise* expression, (b) MHI of the *surprise* expression shown in (a), and (c) ground truth temporal segments of the expression. Part (a) is taken with permission from the FABO database.

Figure 14.7 (left) illustrates how the (normalized) motion area of the *surprise* expression (shown in Figure 14.6) is obtained. The expression starts from the neutral (frames $0-10$, hands on desk) followed by the onset (frames $11-24$, hands move up), the apex, the offset and back to the neutral. As shown in Figure 14.7 (left), the motion area $\mathrm{MA}_{\tau}(t)$ is almost 0 at neutral phase and increases and finally reaches the peak at frame 15. As the expression approaches its *apex*, the motion begins to slow down, which causes $\mathrm{MA}_{\tau}(t)$ to decrease between frame 15 to frame 24. The *apex* occurs between frames 25 and 34 in Figure 14.7 (left). During the *apex* phase, the expression reaches its maximum spatial extent and lasts for some time. Hence, there is relatively small (or no motion) during that phase. During the *offset* phase,

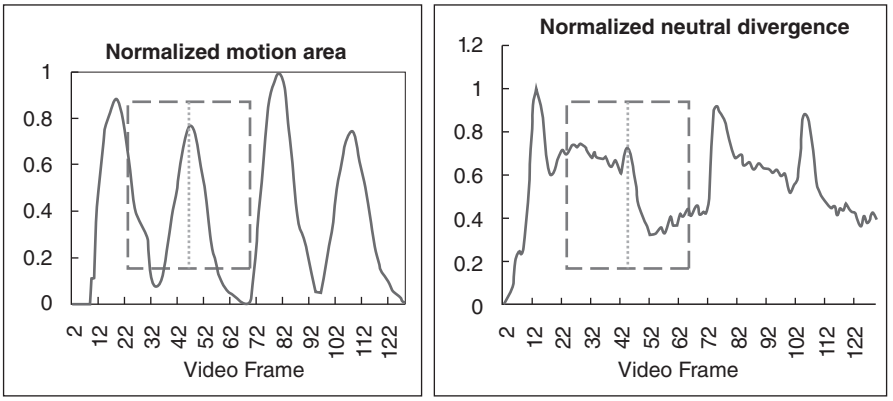


FIGURE 14.7 The motion area feature representation of the current frame is a vector of normalized motion area (*left*), and the neutral divergence feature representation of the current frame is a vector of normalized neutral divergence (*right*).

both the facial expression and the body gesture are moving from the *apex* phase back to the *neutral* phase. This is illustrated in Figure 14.7 (left) between frames 35 and 54. Finally, the expression enters its *neutral* phase between frames 55 and 70 (with very small motion area). The motion area is further normalized to the range of [0, 1] with maximum motion area corresponding to 1. The normalization is done in order to handle variation due to different expressions or subjects.

14.5.1.2 Neutral Divergence The neutral divergence feature measures the degree of difference between the current frame and the neutral frame of an expressive display. Since all videos in the FABO database [60] start from a neutral position, the current frame's neutral divergence ND(t) is calculated by summing up the absolute intensity difference between the current frame image $I(x, y, d, t)$ and the neutral frame image $I(x, y, d, t0)$ over three color channels, as shown in Equation 14.7:

$$\mathrm{ND}(t) = \sum_{d=1}^{3} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathrm{abs}[I(x, y, d, t) - I(x, y, d, t0)] \tag{14.7}$$

where d is the number of color channels of the frame.

Figure 14.7 (right) plots the (normalized) neutral divergences of the *surprise* expression shown in Figure 14.6. Similar to the motion area normalization, the neutral divergence is also normalized to the range of [0, 1]. The neutral divergence is 0 at the *neutral* phase. During the *onset* phase, the neutral divergence increases (as can be observed in Figure 14.7 (right)). During the *apex* phase, the neutral divergence remains relatively stable (with a large neutral divergence value) as there is little movement in the facial expression or the body gesture. However, the *apex* phase is quite different from the *neutral* phase. The neutral divergence decreases at the *offset* phase. When the expression enters its *neutral* phase again, between frames 55 and 70, as shown in Figure 14.7 (right), the neutral divergence does not go back to 0 as would be expected. This indicates that the facial and bodily parts do not return back to their exact starting position. Nevertheless, the difference between the final *neutral* phase and the *apex* phase using the neutral divergence feature is still recognizable (see Figure 14.7 (right)).

14.5.2 Feature Representation and Combination

14.5.2.1 Feature Representation The normalized motion area and the neutral divergence are extracted for every frame in an expression video. To recognize the expression phases of the current frame, we employ a fixed-size temporal window with the center located at the current frame as shown in Figure 14.7a (left). The normalized motion area of every frame within the temporal window is extracted (forming a vector in chronological order). Similar to the motion area feature, as shown in Figure 14.7b (right), the normalized neutral divergence of every frame within the temporal window is also extracted (forming a vector in chronological order). In our experiments, we

set the temporal window size to 31, for both the motion area features and the neutral divergence features.

14.5.2.2 Feature Combination The motion area and the neutral divergence features provide complementary information regarding temporal dynamics of an expression. The motion area is able to separate the onset/offset from the apex/neutral phases, since the onset/offset generates large movements. However, the motion area can neither distinguish the *apex* phase from the *neutral* nor the *onset* phase from the *offset*. Nevertheless, the *apex* phase has large intensity deviation from the initial neutral frame. Therefore, the neutral divergence is able to separate the *neutral* phase from the *apex* phase. During the *onset* phase, the neutral divergence is increasing (the opposite occurs during the *offset* phase). Consequently, the neutral divergence is able to separate the *onset* phase from the *offset* phase as well. The combination of both features is obtained by simply concatenating the motion area feature vector with the neutral divergence feature vector.

14.5.2.3 Classifier We employ SVM with an RBF kernel as our multiclass classifier [73]. SVM is used to find a set of hyper-planes which separate each pair of classes with a maximum margin. The temporal segmentation of an expression phase can be considered as a multiclass classification problem. In other words each frame is classified into neutral, onset, apex, and offset temporal phases.

14.5.3 Experiments

14.5.3.1 Experimental Setup We conducted experiments using the FABO database [60]. We chose 288 videos where the ground truth expressions from both the face camera and the body camera were identical. We used 10 expression categories, including both basic expressions (disgust, fear, happiness, surprise, sadness, and anger) and non-basic expressions (anxiety, boredom, puzzlement, and uncertainty). For each video, there are two to four complete expression cycles. Videos of each expression category are randomly separated into three subsets. Then two of these subsets are chosen for training and the remaining subset is kept for testing. Due to the random separation process, the subjects may overlap between the training and the testing sets.

14.5.3.2 Experimental Results We first perform a threefold cross-validation by combining the motion area and the neutral divergence features. Two subsets are used for training, and the remaining subset is used for testing. The procedure is repeated three times, each of the three subsets being used as the testing data exactly once. The accuracy is calculated by averaging the true positive rate of each class (i.e., the neutral, the onset, the apex, and the offset). The average accuracy obtained by the threefold cross validation is 83.1%.

Figure 14.8 shows the temporal segmentation results of the *surprise* expression video shown in Figure 14.6. The ground truth temporal phase of each frame in the expression video is indicated by the solid line, while the corresponding predicted

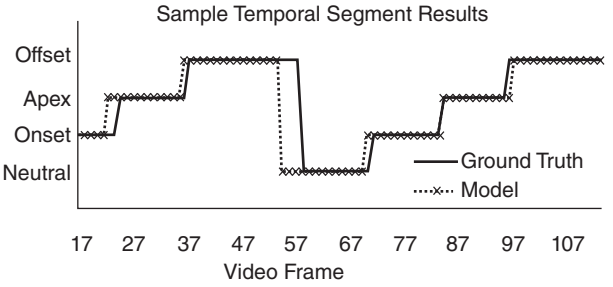


FIGURE 14.8 Temporal segmentation results corresponding to the *surprise* expression video shown in Figure 14.6.

temporal phase is plotted using the dash line with *x*. The predicted temporal segmentation of the expression video matches the ground truth temporal phase quite well (except at the phase transition frames). For example, frames 22 and 23 are predicted as the apex phase while the ground truth indicates that they are the onset frames right before the apex.

Table 14.3 shows the confusion matrix resulting from the temporal phase segmentation. Each row is the ground truth temporal segment while the columns are the classified temporal segments. Based on the confusion matrix, both the *onset* and the *offset* phase appear to be confused mostly with the *apex* phase. The *apex* phase is temporally adjacent to both the *onset* phase and the *offset* phase. This is mainly due to the fact that the temporal boundary between these phases is not straightforward. However, as shown in the last column of Table 14.3, the overall performance of each temporal phase is fairly stable.

We also conducted an experiment in order to evaluate the effectiveness of the combined feature set (combining the motion area and the neutral divergence). This experiment uses the first subset of expression videos as the testing data and the other two subsets as the training data. Using the motion area alone, the temporal phase detection rate is 68.5%. The neutral divergence feature alone achieves 74.1% detection rate. By combining both the motion area and the neutral divergence, the expression phase segmentation performance has boosted up to 82%. In order to understand why the combined feature set significantly improves the performance, we compare the confusion matrices obtained from the motion area (alone) and the combined feature set. Table 14.4 (top) reports the confusion matrix using the motion area feature alone.

TABLE 14.3 Summary of the threefold cross validation results

True/model	Neutral	Onset	Apex	Offset	Accuracy (%)
Neutral	2631	121	208	161	84.3
Onset	113	2253	324	31	82.8
Apex	187	282	4365	251	85.8
Offset	171	70	227	2539	84.4

TABLE 14.4 Confusion matrices using motion area feature alone (top), and using the combined feature set (bottom)

True/model	Neutral	Onset	Apex	Offset
Neutral	1739	75	757	125
Onset	73	1745	288	429
Apex	691	222	3079	225
Offset	102	472	278	1792
Neutral	2213	107	226	150
Onset	106	2037	246	146
Apex	261	227	3553	176
Offset	150	104	245	2145

Rows indicate the ground truth temporal phases while columns indicate the recognized temporal phases.

From the matrix, we can see that the apex frames are mostly confused with the neutral frames. As an example, there are 757 neutral frames misclassified as the *apex* phase, while there are 691 *apex* frames misclassified as the *neutral* phase. Similarly, the *onset* phase is mostly confused with the *offset* phase. Therefore, we conclude that the motion area can neither distinguish the *apex* phase from the *neutral*, nor the *onset* phase from the *offset*.

As can be observed in Table 14.4 (bottom), combining the motion area and the neutral divergence features reduces the confusion between the *neutral* phase and the *apex* phase, significantly. For instance, there are only 226 neutral frames misclassified as apex, and 261 apex frames misclassified as neutral. Similarly, the confusion between the *onset* phase and the *offset* phase is also reduced. These comparisons confirm the effectiveness of combining both the motion area and the neutral divergence features on the temporal segmentation. The neutral divergence and the motion area provide complementary information for identifying the temporal dynamics of an expression.

14.6 DISCUSSION AND OUTLOOK

Human affect analysis based on bodily expressions is still in its infancy. Therefore, for the interested reader we would like to provide a number of pointers for future research as follows.

Representation-related issues. According to research in psychology, three major approaches to affect modeling can be distinguished [74]: *categorical*, *dimensional*, and *appraisal-based* approach. The categorical approach claims that there exist a small number of emotions that are basic, hard wired in our brain and recognized universally (e.g., [7]). This theory has been the most commonly adopted approach in research on automatic measurement of human affect from bodily expressions. However, a number of researchers claim that a small number of discrete classes may not reflect the complexity of the affective state conveyed [65]. They advocate the use of *dimensional description* of human affect, where affective states are not independent from one another; rather, they are related to one another in a systematic manner (e.g.,

[65, 74–76]). The most widely used dimensional model is a circular configuration called *Circumplex of Affect* introduced by Russell [65]. This model is based on the hypothesis that each basic emotion represents a bipolar entity being a part of the same emotional continuum. The proposed polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). Another well-accepted and commonly used dimensional description is the 3D emotional space of pleasure–displeasure, arousal–nonarousal and dominance–submissiveness [75], at times referred to as the *PAD emotion space*. Scherer and colleagues introduced another set of psychological models, referred to as *componential models* of emotion, which are based on the appraisal theory [74, 76, 77]. In the appraisal-based approach, emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world (relevant concerns/needs) [74, 76–78]. Although pioneering efforts have been introduced by Scherer and colleagues (e.g., [79]), how to use the appraisal-based approach for automatic measurement of affect is an open research question as this approach requires complex, multicomponential, and sophisticated measurements of change. Overall, despite the existence of such diverse affect models, there is still not an agreement between researchers on which model should be used for which affect measurement task, and for each modality or cue.

Context. Context usually refers to the knowledge of who the subject is, where she is, what her current task is, and when the observed behavior has been shown. Majority of the works on automated affect analysis from bodily expressions focused on context-free, acted, and emotional expressions (e.g., [40, 45, 80]). More recently, a number of works started exploring automatic analysis of bodily postures in an application-dependent and context-specific manner in non-acted scenarios. Examples include recognizing affect when the user is playing a body-movement-based video game [81] and detecting the level of engagement when the user is interacting with a game companion [82]. Defining and setting up a specific context enables designing automatic systems that are realistic and are sensitive to a specific target user group and target application. Defining a context potentially simplifies the problem of automatic analysis and recognition as the setup chosen may encourage the user to be in a controlled position (e.g., sitting in front of a monitor or standing in a predefined area), wearing specific clothes (e.g., wearing bright-colored t-shirts [82] or a motion capture suit [81]), etc. Overall, however, how to best incorporate and model context for affect recognition from bodily expressions needs to be explored further.

Data acquisition protocol. Defining protocols on how to acquire benchmark data for affective bodily posture and gesture analysis is an ongoing research topic. Currently it is difficult to state whether it is sufficient (or better) to have body-gesture-only databases (e.g., The UCLIC Affective Posture database) or whether it is better to record multiple cues and modalities simultaneously (e.g., recording face and upper body as was done for the FABO database and the GEMEP Corpus). Overall, data acquisition protocols and choices should be contextualized (i.e., by taking into account the application, the user, the task, etc.).

Modeling expression variation. Emotional interpretation of human body motion is based on understanding the action performed. This does not cause major issues when classifying stereotypical bodily expressions (e.g., clenched fists) in terms of

emotional content (e.g., anger, sadness). However, when it comes to analyzing natural bodily expressions, the same emotional content (category) may be expressed with similar bodily movements but with some variations or with very different bodily movements. This presents major challenges to the machine learning techniques trained to detect and recognize the movement patterns specific to each emotion category. This, in turn, will hinder the discovery of underlying patterns due to emotional changes. To mitigate this problem, recent works have focused on using explicit models of action patterns to aid emotion classification (e.g., [83]).

Multiple cues/modalities and their dynamics. Although body has been investigated as an additional channel for affect analysis and recognition, it is still not clear what role it should play when combining multiple cues and modalities: Should it be given higher or lower weight? Can it be the primary (or only) cue/modality? In which context? When can gait be used as an additional modality for affect recognition? How does it relate to, or differ from other bodily expression recognition? These questions are likely to stir further investigations. Additionally, when dealing with multiple cues, it is highly likely that the temporal segments of various cues may not be aligned (synchronized) as illustrated in Figure 14.1c where the apex frames for the bodily expression constitute the onset segment for the facial expression. One noteworthy study that investigated fully the automatic coding of human behavior dynamics with respect to both the temporal segments (onset, apex, offset, and neutral) of various visual cues and the temporal correlation between different visual cues (facial, head, and shoulder movements) is that of Valstar *et al.* [54], who investigated separating posed from genuine smiles in video sequences. However, in practice, it is difficult to obtain accurate detection of the facial/bodily key points and track them robustly for temporal segment detection, due to illumination variations and occlusions (see examples in Figure 14.2). Overall, integration, temporal structures, and temporal correlations between different visual cues are virtually unexplored areas of research, ripe for further investigation.

14.7 CONCLUSIONS

This chapter focused on a relatively understudied problem: bodily expression for automatic affect recognition. The chapter explored how bodily expression analysis can aid affect recognition by describing three case studies: (1) data acquisition and annotation of the first publicly available database of affective face-and-body displays (i.e., the FABO database); (2) a representative approach for affective state recognition from face-and-body display by detecting the space-time interest points in video and using CCA for fusion, and (3) a representative approach for explicit detection of the temporal phases (segments) of affective states (start/end of the expression and its subdivision into phases such as neutral, onset, apex, and offset) from bodily expressions. The chapter concluded by summarizing the main challenges faced and discussing how we can advance the state of the art in the field.

Overall, human affect analysis based on bodily expressions is still in its infancy. However, there is a growing research interest driven by various advances and demands

(e.g., real-time representation and analysis of naturalistic body motion for affect-sensitive games, interaction with humanoid robots). The current automatic measurement technology has already started moving its focus toward naturalistic settings and less-controlled environments, using various sensing devices, and exploring bodily expression either as an autonomous channel or as an additional channel for affect analysis. The bodily cues (postures and gestures) are much more varied than face gestures. There is an unlimited vocabulary of bodily postures and gestures with combinations of movements of various body parts. Despite the effort of Laban in analyzing and annotating body movement [61], unlike the facial expressions, communication of emotions by bodily movement and expressions is still a relatively unexplored and unresolved area in psychology, and further research is needed in order to obtain a better insight on how they contribute to the perception and recognition of the various affective states. This understanding is expected to pave the way for using the bodily expression to its full potential.

ACKNOWLEDGMENTS

The work of Shizhi Chen and YingLi Tian was partially developed under an appointment to the DHS Summer Research Team Program for Minority Serving Institutions, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and U.S. Department of Homeland Security (DHS). ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. It has not been formally reviewed by DHS. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DHS, DOE, or ORAU/ORISE. DHS, DOE, and ORAU/ORISE do not endorse any products or commercial services mentioned in this article.

REFERENCES

[1] EyeToy, http://en.wikipedia.org/wiki/eyetoy (last accessed May 8, 2011).

[2] Kinect, http://en.wikipedia.org/wiki/kinect (last accessed May 8, 2011).

[3] Xbox 360, http://en.wikipedia.org/wiki/xbox_360 (last accessed May 8, 2011).

[4] R. A. Calvo and S. D'Mello, "Affect detection: an interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, 1(1): 18–37, 2010.

[5] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman, "Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification," in Proceedings of the IEEE International Conference on Multimodal Interfaces, 2002, pp. 491–496.

[6] P. Ekman and W. V. Friesen, *The Facial Action Coding System: A Technique for Measurement of Facial Movement*, Consulting Psychologists Press, San Francisco, CA, 1978.

[7] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*, Prentice Hall, New Jersey, 1975.

[8] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, New York, 1997.

[9] M. Pantic and M. S. Bartlett, "Machine analysis of facial expressions," in *Face Recognition* (eds K. Delac and M. Grgic), I-Tech Education and Publishing, Vienna, Austria, 2007, pp. 377–416.

[10] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12): 1424–1445, 2000.

[11] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31: 39–58, 2009.

[12] B. de Gelder, "Why bodies? Twelve reasons for including bodily expressions in affective neuroscience," *Philos. Trans. R. Soc. B: Biol. Sci.*, 364: 3475–3484, 2009.

[13] M. Mortillaro and K. R. Scherer, "Bodily expression of emotion," in *The Oxford Companion to Emotion and the Affective Sciences*, Oxford University Press, 2009, pp. 78–79.

[14] C. Darwin, *The Expression of the Emotions in Man and Animals*, John Murray, London, 1872.

[15] W. James, *Principles of Psychology*, H. Holt & Co., 1890.

[16] M. Coulson, "Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence," *Nonverbal Behav.*, 28(2): 117–139, 2004.

[17] D. B. Givens, *The Nonverbal Dictionary of Gestures, Signs and Body Language Cues*, Center for Nonverbal Studies Press, Washington, 2010.

[18] P. Ekman, "Darwin, deception, and facial expression," *Ann. N. Y. Acad. Sci.*, 2003.

[19] P. Ekman and W. V. Friesen, "Origin, usage and coding: the basis for five categories of nonverbal behavior," in the Symposium on Communication Theory and Linguistic Models in the Social Sciences, Buenos Aires, Argentina, 1967.

[20] H. G. Wallbott, "Bodily expression of emotion," *Eur. J. Soc. Psychol.*, 28: 879–896, 1998.

[21] M. DeMeijer, "The contribution of general features of body movement to the attribution of emotions," *J. Nonverbal Behav.*, 13(4): 247–268, 1989.

[22] F. E. Pollick, H. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, 82: 51–61, 2001.

[23] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in Proceedings of the ACII, 2007, pp. 48–58.

[24] J. F. Cohn, L. I. Reed, T. Moriyama, X. Jing, K. Schmidt, and Z. Ambadar, "Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles," in Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 129–135.

[25] R. Cowie, H. Gunes, G. McKeown, L. Vaclau-Schneider, J. Armstrong, and E. Douglas-Cowie, "The emotional and communicative significance of head nods and shakes in a naturalistic database," in Proceedings of the LREC International Workshop on Emotion, 2010, pp. 42–46.

[26] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in Proceedings of the International Conference on Intelligent Virtual Agents, 2010, pp. 371–377.

[27] S. Mota and R. W. Picard, "Automated posture analysis for detecting learner's interest level," in Proceedings of the IEEE CVPR Workshops, 2003.

[28] J. K. Burgoon, M. L. Jensen, T. O. Meservy, J. Kruse, and J. F. Nunamaker, "Augmenting human identification of emotional states in video," in Proceedings of the International Conference on Intelligent Data Analysis, 2005.

[29] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon Jr., J. F. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intell. Syst.*, 20(5): 36–43, 2005.

[30] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in Proceedings of the ACII, 2007, pp. 59–70.

[31] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in Proceedings of the ACII, 2007, pp. 71–82.

[32] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. R. Scherer, "Automated analysis of body movement in emotionally expressive piano performances," *Music Percept.*, 26: 103–119, 2008.

[33] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Towards a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, 2(2): 106–118, 2011.

[34] D. Janssen, W. I. Schöllhorn, J. Lubienetzki, K. Fölling, H. Kokenge, and K. Davids, "Recognition of emotions in gait patterns by means of artificial neural nets," *J. Nonverbal Behav.*, 32(2): 79–92, 2008.

[35] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Transactions on Systems, Man and Cybernetics-Part B*, 40: 1050–1061, 2010.

[36] P. Ekman, "About brows: emotional and conversational signals," in *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium* (eds M. V. Cranach, K. Foppa, W. Lepenies, and D. Ploog), Cambridge University Press, New York, 1979, pp. 169–248.

[37] A. D. Wilson, A. F. Bobick, and J. Cassell, "Temporal classification of natural gesture and application to video coding," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 948–954.

[38] J. A. Russell and J. D. Fernndez Dols, *The Psychology of Facial Expression*, Cambridge University Press, Cambridge, 1997.

[39] K. L. Schmidt and J. F. Cohn, "Human facial expressions as adaptations: evolutionary questions in facial expression research," *American Journal of Physical Anthropology*, 116(S33): 3–24, 2001.

[40] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 39(1): 64–84, 2009.

[41] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man and Cybernetics-Part B*, 36(2): 433–449, 2006.

[42] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in Proceedings of the IEEE CVPR Workshops, 2006, pp. 149–154.

[43] A. Camurri, I. Lager, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *Int. J. Hum.-Comput. Stud.*, 59: 213–225, 2003.

[44] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis," *Psychol. Bull.*, 11(2): 256–274, 1992.

[45] 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.

[46] K. R. Scherer and H. Ellgring, "Multimodal expression of emotion," *Emotion*, 7: 158–171, 2007.

[47] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion*, 7(3): 487–494, 2007.

[48] J. Driver and C. Spence, "Multisensory perception: beyond modularity and convergence," *Curr. Biol.*, 10(20): 731–735, 2000.

[49] H. Gunes and M. Piccardi, "Creating and annotating affect databases from face and body display: a contemporary survey," in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2006, pp. 2426–2433.

[50] H. K. Meeren, C. C. Van Heijnsbergen, and B. De Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences of the United States of America*, 102: 16518–16523, 2005.

[51] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, and K. Karpouzis, "Emotion analysis in man–machine interaction systems," in *Lecture Notes in Computer Science*, vol. 3361, Sprinter, 2005, pp. 318–328.

[52] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal and bodily expressions recognition," in *Artificial Intelligence for Human Computing*, Springer, 2007, pp. 91–112.

[53] B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud, "Design and evaluation of expressive gesture synthesis for embodied conversational agents," in Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems, 2005.

[54] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in Proceedings of the ACM International Conference on Multimodal Interfaces, 2007, pp. 38–45.

[55] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, 2011.

[56] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Model. User-Adapt. Interact.*, 10: 147–187, 2010.

[57] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in Proceedings of the ACM International Conference on Multimedia, 2005, pp. 677–682.

[58] The UCLIC Database of Affective Postures and Body Movements, http://www.ucl.ac.uk/uclic/people/n_berthouze/research (last accessed May 8, 2011).

[59] T. Bänziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: the GEMEP Corpus," in Proceedings of the ACII, 2007, pp. 476–487.

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

[61] R. Laban and L. Ullmann, *The Mastery of Movement*, 4th revision edn, Princeton Book Company Publishers, 1988.

[62] J. Allman, J. T. Cacioppo, R. J. Davidson, P. Ekman, W. V. Friesen, C. E. Izard, and M. Phillips. NSF report—facial expression understanding, 2003, http://face–and–emotion.com/dataface/nsfrept/basic_science.html (last accessed Sept. 19, 2014).

[63] H. G. Wallbott and K. R. Scherer, "Cues and channels in emotion recognition," *J. Personal. Soc. Psychol.*, 51: 690–699, 1986.

[64] P. Ekman, *Emotions Revealed*, Weidenfeld and Nicolson, 2003.

[65] J. A. Russell, "A circumplex model of affect," *J. Personal. Soc. Psychol.*, 39: 1161–1178, 1980.

[66] H. Gunes and M. Piccardi, "Observer annotation of affective display and evaluation of expressivity: face vs. face-and-body," in Proceedings of the HCSNet Workshop on the Use of Vision in Human–Computer Interaction, 2006, pp. 35–42.

[67] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in Proceedings of the VS-PETS, 2005, pp. 65–72.

[68] H. Hotelling, "Relations between two sets of variates," *Biometrika*, 8: 321–377, 1936.

[69] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, A practical guide to support vector classification, 2003.

[70] M. Pantic and I. Patras, "Temporal modeling of facial actions from face profile image sequences," in Proceedings of the IEEE International Conference on Multimedia and Expo, vol. 1, 2004, pp. 49–52.

[71] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23: 257–267, 2001.

[72] J. Davis, "Hierarchical motion history images for recognizing human motion," in Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video, 2001, pp. 39–46.

[73] C. C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. on Intelligent System and Technology*, 2(3): 27, 2011.

[74] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Conscious. Cogn.*, 17(2): 484–495, 2008.

[75] A. Mehrabian, "Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament," *Curr. Psychol.: Dev. Learn. Personal., Soc.*, 14: 261–292, 1996.

[76] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford University Press, Oxford/New York, 2001.

[77] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth, "The world of emotion is not two-dimensional," *Psychol. Sci.*, 18: 1050–1057, 2007.

[78] N. H. Frijda, *The Emotions*, Cambridge University Press, 1986.

[79] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Netw.*, 18(4): 317–352, 2005.

[80] P. R. DeSilva, A. Kleinsmith, and N. Bianchi-Berthouze, "Towards unsupervised detection of affective body posture nuances," in Proceedings of the ACII, 2005, pp. 32–39.

[81] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Transactions on Systems, Man and Cybernetics-Part B*, 2011.

[82] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction, 2011.

[83] D. Bernhardt and P. Robinson, "Detecting emotions from connected action sequences," in Proceedings of the International Conference on Visual Informatics, 2009.

[84] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Oxford, 1988.

[85] C. Bartneck, "Integrating the OCC model of emotions in embodied characters," in Proceedings of the Workshop on Virtual Conversational Characters, 2002, pp. 39–48.

[86] H. P. Espinosa, C. A. R. Garcia, and L. V. Pineda, "Features selection for primitives estimation on emotional speech," in Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, 2010, pp. 5138–5141.

[87] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional audio-visual speech synthesis based on pad," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3): 570–582, 2011.

[88] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in Proceedings of the IEEE International Conference on Multimedia and Expo, July 2010, pp. 1079–1084.

[89] M. Pitt and N. Shephard, "Filtering via simulation: auxiliary particle filtering," *J. Am. Stat. Assoc.*, 94: 590–599, 1999.

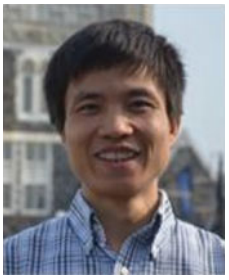
[90] P. Ravindra De Silva, M. Osano, and A. Marasinghe, "Towards recognizing emotion with affective dimensions through body gestures," in Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2006.

AUTHOR BIOGRAPHIES

Hatice Gunes is a lecturer at the School of Electronic Engineering and Computer Science, Queen Mary University of London (QMUL), UK. She received her Ph.D. in Computing Sciences from the University of Technology, Sydney (UTS), Australia. Prior to joining QMUL, she was a postdoctoral researcher at Imperial College, London, UK, working on SEMAINE, a European Union (EU-FP7) award-winning project that aimed to build a multimodal dialogue system, which can interact with humans via a virtual character and react appropriately to the user's nonverbal behavior, and MAHNOB that aimed at multimodal analysis of human naturalistic nonverbal behavior. Dr. Gunes (co-)authored over 45 technical papers in the areas of affective computing, visual and multimodal information analysis and processing, human–computer interaction, machine learning, and pattern recognition. She is a guest editor of special issues in *International Journal of Synthetic Emotions* and *Image and Vision Computing Journal*, a member of the Editorial Advisory Board for *International Journal of*

Computer Vision & Signal Processing and the *Affective Computing and Interaction Book* (IGI Global, 2011), a cochair of the EmoSPACE Workshop at IEEE FG 2011, and a reviewer for numerous journals and conferences in her areas of expertise. Dr. Gunes was a recipient of the Outstanding Paper Award at IEEE FG 2011, the Best Demo Award at IEEE ACII 2009, and the Best Student Paper Award at VisHCI 2006. She is a member of IEEE, ACM, and the HUMAINE Association.

Caifeng Shan is a senior scientist with Philips Research, Eindhoven, The Netherlands. He received the Ph.D. degree in Computer Vision from the Queen Mary University of London, UK. His research interests include computer vision, pattern recognition, image/video processing and analysis, machine learning, multimedia, and related applications. He has authored around 40 refereed scientific papers and 5 pending patent applications. He has edited two books *Video Search and Mining* (Springer, 2010) and *Multimedia Interaction and Intelligent User Interfaces: Principles, Methods and Applications* (Springer, 2010). He has been the Guest Editor of *IEEE Transactions on Multimedia* and *IEEE Transactions on Circuits and Systems for Video Technology*. He has chaired several international workshops at flagship conferences such as IEEE, ICCV, and ACM Multimedia. He has served as a program committee member and reviewer for many international conferences and journals. He is also a member of IEEE.



Shizhi Chen is a Ph.D. student in the Department of Electrical Engineering at the City College of New York. His research interests include facial expression recognition, scene understanding, machine learning, and related applications. He received the B.S. degree of Electrical Engineering from SUNY, Binghamton, in 2004, and the M.S. degree of Electrical Engineering and Computer Science from UC, Berkeley, in 2006. From 2006 to 2009, he worked as an engineer in several companies including Altera, Supertex, Inc., and the US Patent and Trademark Office. He is a member of Eta Kappa Nu (electrical engineering honor society), and a member of Tau Beta Pi (engineering honor society). He has also received numerous scholarships and fellowships, including Beat the Odds scholarship, Achievement Rewards for College Scientists (ARCS) Fellowship, and NOAA CREST Fellowship.

YingLi Tian is an associate professor in the Department of Electrical Engineering at the City College of New York since 2008. She received her Ph.D. from the Department of Electronic Engineering at the Chinese University of Hong Kong in 1996 and her B.S. and M.S. from TianJin University, China, in 1987 and 1990. She is experienced in computer vision topics ranging from object recognition, photometric modeling, and shape from shading, to human identification, 3D reconstruction, motion/video analysis,

multi-sensor fusion, and facial expression analysis. After she worked in National Laboratory of Pattern Recognition at the Chinese Academy of Sciences, Beijing, China, Dr. Tian joined the Robotics Institute in Carnegie Mellon University as a postdoctoral fellow. She focused on automatic facial expression analysis. From 2001 to 2008, Dr. Tian was a research staff member at IBM T. J. Watson Research Center, Hawthorne, New York. She focused on moving object detection, tracking, and event and activity analysis. She was one of the inventors of the IBM Smart Surveillance Solutions (SSS) and was leading the video analytics team. She received the IBM Invention Achievement Awards every year from 2002 to 2007. She also received the IBM Outstanding Innovation Achievement Award in 2007. As an adjunct professor at Columbia University, she co-taught a course on Automatic Video Surveillance (Spring 2008). Dr. Tian has published more than 90 papers in journals and conferences and has filed more than 30 patents. She is also a senior member of IEEE.