CHAPTER

17 Multimodal Affect Recognition for Naturalistic Human-Computer and Human-Robot Interactions

Ginevra Castellano, Hatice Gunes, Christopher Peters, *and* Björn Schuller

Abstract

This chapter provides a synthesis of research on multimodal affect recognition and discusses methodological considerations and challenges arising from the design of a multimodal affect recognition system for naturalistic human-computer and human-robot interactions. Identified challenges include the collection and annotation of spontaneous affective expressions, the choice of appropriate methods for feature representation and selection in a multimodal context, and the need for context sensitivity and for classification schemes that take into account the dynamic nature of affect and the relationship between different modalities. Finally, two examples of multimodal affect recognition systems used in (soft) real-time naturalistic human-computer and human-robot interaction frameworks are presented.

Key Words: multimodal affect recognition, feature representation and selection, context sensitivity, human-computer interaction, human-robot interaction

Introduction

Recent work in human-computer interaction (HCI) and human-robot interaction (HRI) has shown that embodied agents and robots are increasingly being studied as partners that collaborate and do things with people (Breazeal, 2009; Schroeder et al., 2012). For example, the use of embodied agents and robots is being investigated in many HCI and HRI applications, such as providing assistance for the elderly at home, serving as tutors for children by enriching their learning experiences, and acting as therapeutic tools or as game buddies for entertainment purposes.

These applications require embodied agents and robots to be endowed with social skills. Social perception abilities include affect sensitivity—that is, the ability to recognise people's affective expressions and states, understand their social signals—and account for the context in which the interaction takes place (Castellano et al., 2010a). Affect sensitive embodied agents and robots are more likely to be able to engage with human users over extended periods of time as compared with their nonaffective counterparts (Bickmore & Picard, 2005).

Research on automatic affect recognition has contributed several studies on the design of systems capable of perceiving multimodal social, cognitive, and affective cues (e.g., facial expressions, eye gaze, body movement, physiological data, etc.) and using them to infer a user's affective and cognitive state (Calvo & D'Mello, 2010; Zeng et al., 2009).

Recently there has been a shift toward real-world HCI and HRI, which has led to the emergence of new trends in multimodal affect recognition. These include, among others, an increased focus on the automatic recognition of spontaneous and nonprototypical affective states, the development of techniques for continuous affect prediction—which allows for the dynamics of affective states to be taken into consideration, and the design of context-sensitive affect recognition systems.

Compared with systems based on a single modality, multimodal affect recognition has the potential

to achieve increased recognition performances. It is still an open question, however, why improvements of multimodal affect recognition systems over their unimodal counterparts are still relatively modest, especially when natural data is used, as shown by D'Mello and Kory (2012) in a recent meta-analysis of thirty affect recognition studies.

This chapter provides an overview of the state of the art on multimodal affect recognition, the challenges that underlie the design of a multimodal affect recognition system, and two examples of successful integration of multimodal affect recognition systems for real-time HCI and HRI.

Multimodal Affect Recognition: State of the Art and Challenges

While affect recognition systems based on one modality have been extensively investigated, studies taking into account the multimodal nature of affective states have been gaining ground only recently (Zeng et al., 2009). D'Mello and Kory (2012) showed that multimodal affect recognition systems are consistently better than their unimodal counterparts, supporting the general tendency in the literature to move toward systems that combine multiple modalities for the purpose of predicting a user's affect. Nevertheless, evidence shows that improvements over unimodal systems are still modest, suggesting the need for classifiers and fusion methods that better capture the relationships between different modalities, and for affective corpora that contain adequate samples of synchronized expressions (D'Mello & Kory, 2012).

The latest shift toward real-world HCI and HRI is driving research on multimodal affect recognition in new directions. This, in turn, has brought new challenges that need to be considered as opportunities that open up new avenues for research.

For example, of late there has been an increased interest in the automatic recognition of spontaneous, nonprototypical affective states (Castellano et al., 2012; Kleinsmith et al., 2011; Lucey et al., 2011), rather than of prototypical or basic emotions (such as anger, disgust, fear, happiness, sadness, and surprise) and a shift toward dimensional affect recognition (Gunes & Schuller, 2013; Gunes et al., 2011), which is based on a description of human affect in terms of a set of dimensions, such as valence (i.e., positive or negative affect) and arousal (i.e., affect characterized by low or high activation).

Another aspect that is receiving a lot of attention is the development of novel methods for multimodal fusion (Metallinou et al., 2013; Nicolaou et al., 2012), which should take into consideration the underlying relationships and correlation between feature sets in different modalities and affect dimensions, how different affective expressions influence each other, and how much information each of them provides about the expressed affect.

An emerging trend that addresses the need for the dynamics of affective states to be accounted for is continuous affect prediction, based on a user's input that is continuously available and analyzed over time, which aims to produce continuous values for the target affect or affect dimensions (Meng & Bianchi-Berthouze, 2011; Nicolaou et al., 2012). This is especially important for the integration of affect recognition systems in real-time HCI and HRI frameworks.

Finally, some studies have started to address the issue of context-sensitive affect recognition, which takes into account the context in which the interaction takes place (e.g., task, user preferences, presence of other people, behavior of the interactant, etc.) in order to improve affect recognition performances (Castellano et al., 2012; Kapoor & Picard, 2005; Martinez & Yannakakis, 2011).

The following sections discuss in detail how the current challenges in affect recognition research are being addressed in the literature, with a specific focus on multimodal affect recognition systems.

Data Collection

BEYOND PROTOTYPICAL EMOTIONS AND ACTED AFFECTIVE EXPRESSIONS

Research on multimodal affect recognition is moving from the lab to the real world; hence the need for corpora and databases that contain spontaneous and subtle, rather than acted, prototypical and exaggerated affective expressions. While examples of naturalistic databases are gradually increasing in the literature (e.g., Lucey et al., 2011; McKeown et al., 2012, 2013), currently most affect recognition systems have been trained on databases of acted affective expressions. These often reflect stereotypes and exaggerated expressions, and they are often decontextualized. Moreover, most of the available databases contain expressions of prototypical emotions that seldom represent affective states emerging in HCI and HRI applications. This has been the case, so far, for most unimodal and multimodal affect recognition systems.

Real-world HCI and HRI require affect recognition systems trained with databases containing contextual descriptions synchronized with other modalities (Castellano et al., 2010b), a research

direction that is still underexplored. Moreover, most of the available acted databases contain expressions recorded in contexts that are not specific to a particular application. While the availability of affect databases that can be used for training affect recognition systems applicable to several interaction scenarios is a pressing need, real-world HCI and HRI scenarios require contextualized affective expressions (i.e., expressions that emerge in the same scenario of the final application) for system training and validation.

ANNOTATION

The training and testing of affect recognition systems requires ground truth data, which are usually obtained via observational assessment. While ideally one would ask the participants of an HCI or HRI experiment to rate the affective states they experienced, this is seldom a viable solution. First of all, ratings are usually collected at the end of the experiment via questionnaires, but this does not allow for affective states emerging at specific instants of the interaction to be captured. On the other hand, continuous self-annotation of affect during the experiment is not practical. Another option is to ask participants to watch videos of their experiment and label the affective states they feel they experienced; however, this may be problematic—for example, when children are involved. Alternatively, affect annotation can be performed with the help of external coders. These are usually assigned presegmented videos and asked to label each one (Castellano et al., 2010). Another approach is continuous annotation of affect dimensions from videos using tools such as Feeltrace (McKeown et al., 2012).

Affect annotation still presents open issues—for example, the difficulty of achieving good intercoder agreement and the time-consuming nature of the annotation process, which often requires enormous efforts in recruiting and training coders. Another open question concerns whether affective stimuli should be labeled by simultaneously taking into account all the modalities available to the coder rather than considering the single modalities separately. The first approach has the advantage of providing the coder with an overall perspective of the emergence of an affective state, including the context in which the affective cues are displayed.

Feature Extraction

FEATURE REPRESENTATION—FRAME- VERSUS WINDOW-BASED

Different modalities tend to operate on different time scales. In addition, the feature sampling frequency can be variable: For example, for video processing (e.g., facial expression analysis or body gestures by motion capture, depth cameras or similar), often a constant frame rate, such as 25 frames per second (fps), is chosen as a basis to calculate features such as tracked facial points, global motions, local (Gabor) binary patterns, or transformed image information, etc., for the analysis of shape- or appearance-based characteristics. This is often similar for physiological feature information. In acoustic speech analysis, suprasegmental features are calculated per word, turn, or similar entity with differing length over frame-level features typically extracted at around 100 fps. This usually includes prosodic (intonation, intensity, duration, etc.), cepstral, spectral (Mel frequency cepstral coefficients, formant information, etc.), and voice quality description (harmonicity, perturbations, etc.). Frame-level features are often referred to as *low-level descriptors* (LLDs), and the suprasegmental features are *functionals*—that is, the time series of unknown length of frame-level LLD features is projected onto a single scalar value per LLD. Such functionals comprise extremes, means, higher moments, peaks, percentiles, regression coefficients, segments, or spectral and temporal characteristics; one can also apply these in a hierarchical manner such as the extremes of means and vice versa. For linguistic feature information, it seems obvious that the sampling interval cannot be fixed, as it depends on the speech rate, and one usually has to wait until the end of linguistic entities such as words. Based on individual entities or sequences of these, one can apply knowledge from resources such as affective word lists, execute deeper linguistic analyses, or extract functional-type feature information such as bags of words, etc.

At some point, however, some form of synchronization will be needed—either to unite the feature information or to come to a decision at a certain moment in time informed by the diverse modalities (see Stream Fusion, p. 251). One option to reach this goal is the application of functionals to the diverse LLDs from different modalities and also to the processing of video, physiological, or other multimodal information on a suprasegmental level. Decisive in this case is the unit of analysis of interest, which can be linguistic entities if linguistic analysis is involved. This allows a recognition system to directly attach affective information to these units—such as words, turns or similar—which may be well suited from an application point of view. In case of absence of speech, or a multimodal fusion without availability of such linguistically motivated information, fixed

intervals at a larger "macro" window size can be a good choice. Again, the application scenario will have an influence on the choice of the window length as a compromise between reasonably fast update, sufficient LLD feature information, and "stationary emotion" (i.e., the emotion can be assumed not to change over the unit of analysis) within this macro window. A typical value can be around 1 second, as was used in the SEMAINE project (see Multimodal Affect Detection for a Sensitive Artificial Listener, p. 252); however, more research will be needed to identify an optimal value.

FEATURE SELECTION IN A MULTIMODAL CONTEXT

Obviously one can optimize the feature space individually per information stream, such as acoustic or video features. However, the multimodal context allows for a combined feature selection, in particular in the case of a feature-level fusion (see Early Feature-Level Fusion, p. 251). This can lead to improved performance (Schuller et al., 2008). In fact, an individual optimisation per stream followed by a combined selection process can be an interesting choice: At first, the often highly correlated information should be reduced individually per modality. Then, a secondary optimisation process can lead to further improvements reducing cross-modal redundancy (Schuller et al., 2008). As unimodal data are usually available in larger amounts than multimodal data, the optimization per single modality can partly benefit from more available data. Likewise, if the number of considered modalities or feature streams is greater than two, selection of subgroupings of modalities could be considered.

Context Sensitivity

To correctly understand a phenomenon—an affective display for example—it is often necessary to move beyond the phenomenon in isolation to consider broader circumstances and aspects. These can be thought of as surrounding the phenomenon in both space (a smile interpreted in the context of the movements of the rest of the face) and time (dialogue preceding the smile), and can potentially include many factors relating to the interactants (personality, gender, culture, preferences, moods, goals), their impressions of each other and even themselves (how others perceive them, their goals, and so on), and the state of the interaction (commencing, maintained, closing). For example, ratings of the behavior of individuals may vary depending on their accompanying background (Ennis et al., 2011). Context is therefore of great importance in attempting to improve the performance and robustness of affect recognition systems, especially when other modalities are not sufficient or lead to nonmeaningful interpretations. Context is typically difficult to account for, however, as it necessarily involves the identification of those features, from among a large number of potential candidates, that are most relevant to understanding and interpreting an unfolding situation.

While some efforts have been reported in the literature, only a limited number of studies have addressed the problem of context-sensitive affect recognition. Kapoor and Picard (Kapoor & Picard, 2005), for example, proposed an approach for the recognition of interest in a learning environment by combining nonverbal cues and information about the learner's task (e.g., level of difficulty and state of the game). Peters and colleagues (Peters et al., 2010) used eye gaze and head direction to model user engagement with a virtual agent. Interpretation of the quality of user engagement with the interaction is contextualized by accounting for gaze toward relevant objects at appropriate times in the interaction. In this case, participant gaze toward an object when it has not been part of the recent discussion is deemed to signal less engagement than participant gaze toward an object that has just been described by the system. Context sensitivity is also a vital basis for determining the novelty of events and objects (Grandjean & Peters, 2011), which is fundamental to social attention, recollection, and learning capabilities in artificial social entities.

A key challenge is contextual feature representation—that is, how to model and encode relationships between different types of context and between context and other modalities. Morency et al. (2008) proposed a context-based recognition framework that integrates information from human participants engaged in a conversation to improve visual gesture recognition. They proposed the idea of an encoding dictionary, a technique for contextual feature representation that models different relationships between a contextual feature and visual gestures. Martinez and Yannakakis (2011) proposed a method for the fusion of physiological signals and game-related information for automatic affect recognition in a game scenario. Their approach uses frequent sequence mining to extract sequential features that combine events across different user input modalities. Castellano and colleagues (2012) investigated contextual feature representation in a game-based HRI scenario and explored how to encode task and game context and

their relationships in a timely manner for automatic engagement prediction. They investigated the use of overall features, which capture game and social context in an independent way at the interaction level, and turn-based features, which encode the interdependencies of game and social context at each turn of the game. They found that the integration of game- and social context–based features with features encoding their interdependencies leads to higher recognition performances.

Classification Schemes

Classification methods for affect recognition can be viewed under two schemes: static versus dynamic classification and discrete versus continuous recognition.

STATIC VERSUS DYNAMIC MODELING

Analysis of automatic human nonverbal behavior can be performed either by using the features from one frame at a time or by considering the sequential nature of the frame sequence, as in a time series. These two approaches are referred to as *static or frame-based* and *dynamic or sequence-based* classification, respectively (Petridis et al., 2009). Commonly used static classifiers are support vector machines, neural networks, and decision trees. Dynamic Bayesian networks, hidden Markov models, and their variations (e.g., coupled hidden Markov models) constitute the well-known dynamic classifiers.

Researchers claim that in the static classification case, dynamic properties of human affective behavior should be captured by the features, while in dynamic classification, they are dealt with by the classifier. Vogt et al. (2008) argue that in speech-based emotion recognition, most works use different feature representation for static and dynamic classification; therefore, it is not possible to clearly attribute the higher recognition accuracy to either classification technique (dynamic versus static). A number of researchers reported that dynamic classifiers are better suited for person-dependent facial expression recognition (e.g., Cohen et al., 2003), which is likely to be the case for affect recognition from other modalities. This was attributed to the fact that dynamic classifiers are more sensitive to both differences in terms of appearance change and differences in temporal patterns among individuals. Static classifiers were reported as being more reliable when the frames represent the apex of an expression (Cohen et al., 2003). Other researchers reported that the frame-based classification outperforms the sequence-based classification in the task of temporal segment detection from face and body displays (e.g., Gunes & Piccardi, 2009). Overall, the usefulness of static versus dynamic classification depends on the feature representation (frame- versus window-based feature representation) and the task at hand (Petridis et al., 2009).

DISCRETE VERSUS CONTINUOUS RECOGNITION

Traditionally, research in the field of automatic affect recognition has focused on recognizing discrete, basic emotional states from posed data acquired in laboratory settings. However, these models are deemed unrealistic, as they are unable to capture the nonbasic and subtle affective states exhibited by humans in everyday interactions. Therefore researchers have started adopting a dimensional description of human emotion, where an emotional state is characterized in terms of a number of latent dimensions (Gunes & Schuller, 2013; Gunes et al., 2011, Kleinsmith et al., 2011). Two dimensions are deemed sufficient for capturing most of the affective variability: valence and arousal, signifying respectively how negative/positive and active/inactive an emotional state is. Other dimensions have also been proposed (Fontaine et al., 2007).

Dimensional quantized (discrete) classification of affect is usually done by reducing the prediction problem to a two/three/four-class classification problem (e.g., positive versus negative or active versus passive classification (Nicolaou et al., 2010, 2011a). The choice of classifier depends on the context and the application. Classification methods used for discrete affect detection and recognition include, among others, support vector machines (SVMs), multilayer perceptron networks, k-nearest neighbor classifiers, naïve Bayes classifiers, radial basis function networks, linear discriminant analysis, conditional random fields, hidden Markov models (HMMs), and variations of these (e.g., coupled HMMs or asynchronous HMMs) (Nicolaou et al., 2010). Various frameworks that combine the benefits of multiple classifiers have also been proposed (e.g., a multilayer hybrid framework for classification (Nicolaou et al., 2011b; Meng et al., 2013).

Continuous affect measurements should be able to produce continuous values for the target dimensions. Some of the classification schemes that have been explored for this task are support vector regression, relevance vector machines, and long short-term recurrent neural networks (e.g., Nicolaou et al., 2011a, 2012). Overall, for automatic affect analysis of continuous input, there is no agreement on how

to model dimensional affect space (continuous versus quantized) and which classifier is better suited for automatic multimodal analysis of continuous affective input.

The two emerging trends in continuous affect prediction are the so-called output-associative prediction (e.g., Nicolaou et al., 2012) and the design of emotion-specific classification schemes (e.g., Nicolaou et al., 2011b). Output-associative prediction exploits the correlations between the dimensions and learns dependencies among the predicted values. Creating emotion-specific schemes for continuous prediction of emotions is relatively new and needs to be investigated further.

Stream Fusion

EARLY FEATURE-LEVEL FUSION

In automatic affect prediction, feature-level fusion is obtained by concatenating all the features from multiple cues into one feature vector, which is then fed into a machine learning model (e.g., Nicolaou et al., 2011a). If the frame rate of the audio stream differs from that of the video stream (e.g., 50 Hz versus 25 fps), some form of adaptation is needed during feature-level fusion (e.g., Nicolaou et al., 2011a; Petridis et al., 2009). Feature-level fusion becomes more challenging as the number of features increases and when the features are of very different nature. Synchronization then becomes of utmost importance.

LATE SEMANTIC FUSION

The most straightforward approach whereby to tackle modality fusion is at the decision level, since feature and time dependence are abstracted. Each classifier processes its own data stream and the multiple sets of outputs are combined at a later stage to produce the final hypothesis. Decision-level fusion can be obtained at the *soft level* (a measure of confidence is associated with the decision) or at the *hard level* (the combining mechanism operates on single hypothesis decisions). There has been some work on combining classifiers and providing theoretical justification for using simple operators such as majority vote, sum, product, maximum/minimum/median, and adaptation of weights.

Explicit fusion of multimodal data refers to first automatically detecting behavioral cues that are known to convey important affective information (e.g., head nods, smiles, pauses) and then fusing explicitly only these higher-level cues. A representative example of explicit fusion is the work of Eyben et al. (2011), who proposed a string-based approach for fusing the behavioral events from visual and audio modalities (i.e., facial action units, head nods and shakes, and verbal and nonverbal vocal cues) to predict human affect in a continuous dimensional space in terms of arousal, expectation, intensity, power, and valence dimensions. A number of approaches have also been reported for explicit synchronization purposes of multiple streams. For instance, Gunes et al. (2009) identified the neutral-onset-apex-offset-neutral phases of face and body expressions recorded via separate cameras and synchronized the information from face and body streams at the phase level (i.e., by detecting the apex phase of face and body expressions stream).

HYBRID FUSION

Since humans display multimodal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e., loss of mutual correlation between the modalities). *Model-level fusion* has been adopted to mitigate the issues pertinent to feature- and decision-level fusion by exploiting the correlations between the modalities while relaxing the requirement of synchronization. By doing this, model-level fusion has the potential of capturing correlations and structures embedded in the continuous output of the classifiers or regressors from different sets of cues. It may use Bayesian networks, multistream fused HMM, tripled HMM, neural networks, etc. (see Zeng et al., 2009 for details on these).

Overall, finding the best way to fuse the modalities for automatic emotion prediction remains an open issue in the field. An emerging trend in affective data fusion is called output-associative fusion (e.g., Nicolaou et al., 2011a). This fusion method capitalizes on the fact that the emotion dimensions (valence and arousal) are correlated. In order to exploit these correlations and patterns, the output-associative fusion framework aims to learn the dependencies that exist among the predicted dimensional values.

Multimodal Affect Detection for (Soft) Real-Time HCI and HRI: Methodological Considerations and Case Studies

The timely analysis and interpretation of a user's affective state is of primary importance for HCI and HRI in real-world settings. For example, it is vital for embodied agents and robots to establish an affective loop with the user through the generation

of a response that is appropriate to the way the user is feeling. Despite the large body of existing literature on affect recognition, examples of automatic affect recognition systems for integration in HCI and HRI frameworks are still not numerous. Further, not many system prototypes have been designed which can work in real environments in the long term. The next sections present two case studies where a multimodal affect recognition system has been successfully applied to real-world HCI and HRI scenarios.

Multimodal Affect Detection for a Sensitive Artificial Listener—Results and Lessons Learned from the SEMAINE Project

The SEMAINE system is a pioneering effort in creating dynamic, expressive, and adaptive virtual agents by analyzing the multimodal nonverbal communicative behavior of the human user in soft real-time. The system aims to engage the user in a dialog and create an emotional workout by paying attention to the user's nonverbal expressions, and reacting accordingly. It focuses on the *soft skills* that humans naturally use to keep a conversation alive (e.g., backchannel feedback such as nodding and smiling). The SEMAINE system avoids task-oriented dialogue, instead, it models the type of interaction sometimes found at parties: you listen to someone you want to chat with, and without really understanding much of what the other person is saying, you exhibit all the signs that are needed for him or her to continue talking to you. The SAL characters can speak to engage the user in a simple dialogue as well as show nonverbal listener signals (Figure 17.1). The approach has been test run using various "Wizard of Oz" setups that have allowed the fine tuning of the scripts used by the various characters in order to react to the emotional state of the user in plausible ways despite the lack of language understanding. The SEMAINE system has been demonstrated at]: International Conference on Affective Computing and Intelligent Interaction (ACII 2009) (Schröder et al., 2009) and IEEE International Conference on Automatic Face and Gesture Recognition FG'11 (Schröder et al., 2011).

AUDIOVISUAL AFFECT RECOGNITION IN A REAL-LIFE SYSTEM

In a real-life system, such as the SEMAINE system, affective data can be thought of as uninterrupted streams originating from a variety of sensors (cameras, microphones, etc.); to achieve optimal affect prediction, prior to recognition, or simultaneously with this, there is a requirement to segment the data and to determine analysis duration (Gunes et al., 2011) or the unit of analysis (Schuller et al., 2011b). Segmenting multimodal data in a meaningful way is directly related to the level at which the detection results should be accurate and that at which the detection results should be analyzed and outputted (frame, millisecond, second, or minute level). The current solution is to employ various window sizes depending on the modality. The achievement of real-time affect prediction requires a small window size to be used for analysis (i.e., a few seconds), but obtaining a reliable prediction accuracy requires longer-term monitoring. Overall, the challenge for future research is to find an appropriate unit of analysis which is sensitive to the context at hand. Another issue is that research on affect analysis and affect generation (synthesis) appear to be detached from each other even in multiparty and multidisciplinary projects such as SEMAINE (Schröder et al., 2011). Investigation of how to interrelate these in earlier stages will provide valuable insight into the realization of affect-sensitive systems that are able to interpret multimodal and continuous input and respond appropriately.

Fig. 17.1 A user conversing with one of the SAL characters (i.e., Poppy).

THE AUDIOVISUAL EMOTION CHALLENGES

The AVEC series of two consecutive public challenges on audiovisual emotion recognition is the first of its kind for multimodal affect detection. It is based on data collected in the SEMAINE project, and offers a test bed for uni- and multimodal emotion recognition including acoustic, linguistic, and video cues. Four affect dimensions are to be assessed: arousal, expectation, power, and valence. While the annotation of the data was done in a continuous manner both in time and values, the first challenge, as held in 2011 (Schuller et al., 2011), required participants to solve a two-class problem with respect to above or below the average value per dimension. In addition, the video stream was chunked in two ways over time: per frame for the video only task and per word for the audio and audiovisual tasks. In the second round held in 2012 (Schuller et al., 2012), this was changed to a continuous regression-type measurement in value either at the frame or word level. The 2011 installation thus used three different test partitions for three subchallenges, providing files containing either audio only (as test partition for the audio subchallenge), or video only (as test partition for the video subchallenge), or both (for the audiovisual subchallenge) to ensure that only this modality was used for result assessment. In the 2012 installation, the same test partition was used no matter which modality was exploited for the best result. Instead, two types of subchallenges focused either on fully continuous (i.e., frame-level emotion assessment) or word-level assessment. This means that emotion needed to be recognized either for every frame or per word (i.e., over a larger frame that lasted as long as each spoken word). Only parts where audio was actually present were used.

Besides the audiovisual data, 1,941 (2011)/1,841 (2012) precomputed audio features brute-forced by functional application to LLDs (see Feature Representation—Frame- Versus Window-Based, p. 248) and 5,908 video features are given for optional usage and baselines. These features and the data are freely available to experiment with; however, the labels of the test partition remain with the organizers and results can be acquired by submission of predictions on these instances.

Various classification methods have been applied to the 2011 audiovisual task of AVEC: support vector machines, extreme learning machine-based feed forward neural networks, AdaBoost, Gaussian mixture models, and a combined system consisting of MLPs and HMMs. At the time of the challenge, latent-dynamic conditional random fields led to the best result of 60.3% weighted accuracy on average over the four dimensions (Ramirez et al., 2011). Later, the best audiovisual result to date was reached by long short-term recurrent neural networks (Wöllmer et al., 2012) with 64.6% weighted accuracy for late fusion. The authors used the baseline acoustic feature set and optical flow video features after rectifying the tracked facial region.

The 2012 event highlights the particular challenge of fully continuous emotion assessment: 0.456 as cross-correlation coefficient was reached by the winning team (Nicolle et al.) as averaged over the four affective dimensions. In the case of the word-level subchallenge, this measure exceeded only 0.28.

Multimodal Affect Recognition for a Robotic Companion—Results and Lessons Learned from the LIREC Project

The EU FP7 LIREC (LIving with Robots and intEractive Companions) project (2008–2012) explored long-term social relationships with socially intelligent robotic companions. Within LIREC, MyFriend is an HRI scenario that showcases an iCat robot acting as a game companion for young children (Figure 17.2). The robot plays chess with children, provides affective feedback based on the moves on an electronic chessboard, interacts with them by displaying facial expressions and verbal utterances, and reacts empathically based on the valence of the affects the children experience throughout the game (Castellano et al., 2013).

CONTEXT-SENSITIVE AFFECT RECOGNITION IN REAL-WORLD HRI SETTINGS

The robotic game companion is built on a novel platform for affect sensitive, adaptive HRI. The platform integrates an array of sensors in a modular client-server architecture that includes a vision module, a game engine, an affect recognition module, an empathic behavior generation engine coupled with an action selection and an appraisal mechanism, and the iCat robot module. After every move made by the user, the user's affective state is inferred by the affect recognition module based on behavioral indicators provided by the vision module (i.e., probability of smile, eye gaze) and contextual indicators (i.e., game-related features) extracted by the game engine (Castellano et al., 2012).

The affect recognition module consists of an SVM-based valence detector. It continuously receives synchronized features from the vision module and the game engine and, as output, provides

Fig. 17.2 A user interacting with iCat in a primary school.

probability values for the valence of the user's affect. At any time during the interaction, the iCat module can send a request to the affect recognition module to evaluate the affective state experienced by the user in the previous N seconds of the game/interaction. Information about the user's affective state is then used by the robot to select and generate an empathic intervention, such as providing encouraging comments or suggesting a good move. The valence detector was trained using the Inter-ACT corpus, an affective and contextually rich multimodal video corpus including spontaneous expressions of children playing chess with the iCat robot in a primary school and a chess club (Castellano et al., 2010b). Nonverbal behaviors (i.e., smiles and eye gaze) and game-based features (i.e., state of the game and game evolution) were automatically extracted and synchronized before being combined in a joint feature space for training the valence detector. An SVM classifier with radial basis function (RBF) kernel achieved a recognition performance of 63% in a three-class valence classification problem (three labels: positive, negative, or neutral) (Castellano et al., 2013).

Results from studies integrating the platform for affect-sensitive, adaptive human-robot interaction in the robotic companion and carried out in a semicontrolled environment in a primary school showed that affect sensing and empathic interventions lead to increased engagement with the robot and an increased perception of friendship from the robot (Leite et al., 2012a) as compared with neutral behavior. Affect sensing and empathic interventions also allowed the establishment of interactions that were more engaging and more successful over extended periods of time, which is an important requirement for companionship (Leite et al., 2012b).

Important challenges for future research in the domain of affect recognition for social robots include the design of systems that can adapt to specific users, successfully encode relationships between contextual features and between context and other modalities, and are highly robust (e.g., they are capable of performing successful continuous affect prediction over extended periods of time).

Conclusions and Future Directions

This chapter provided an introduction to multimodal affect recognition for naturalistic HCI and HRI. We showed how the latest trends in multimodal affect recognition research are opening up new opportunities for real-time interactions with embodied agents and robots in real-world settings. Particularly, we identified key challenges in the design of a multimodal affect recognition system for naturalistic HCI and HRI. These include:

1. *The collection and annotation of data containing spontaneous affective expressions.* Real-world HCI and HRI require affect recognition systems trained with corpora and databases that contain spontaneous and subtle rather than acted and prototypical affective expressions.

2. *The choice of appropriate methods for feature representation and feature selection in a multimodal context.*

Different modalities tend to operate on different time levels and may be dependent one another; additionally, some of them may be more important than others for the purpose of affect prediction in a specific application scenario. Hence there is the need to choose appropriate methods for feature representation and feature selection in a multimodal context.

3. *The design of affect recognition systems sensitive to context.*

Context can be used as an additional modality to improve the performance of an affect recognition system. A key challenge here is how to model and encode relationships between different types of context and between context and other modalities.

4. *The design of classification schemes that take into account the dynamic nature of affect and the relationship between different feature sets.*

Continuous affect prediction has been shown to be successful in addressing the dynamic nature of affective states; novel methods for multimodal fusion need to take into consideration the underlying relationships and correlations between feature sets in different modalities and affect dimensions.

While several issues in multimodal affect recognition require further investigation, we have shown how initial attempts at addressing these challenges can lead to the successful integration of multimodal affect recognition systems in HCI and HRI frameworks.

Acknowledgement

The works of the authors are partially supported by the following grants: G. Castellano by the European Commission (EC) via by the EU FP7 ICT-317923 project EMOTE (EMbOdied-perceptive Tutors for Empathy-based learning), H. Gunes by the EPSRC EP/L00416X/1 Digital Personhood project 'Being There' (Humans and Robots in Public Space). The authors acknowledge that they are solely responsible for the content of this publication. It does not represent the opinion of the EC and/or EPSRC and the EC and/or EPSRC is not responsible for any use that might be made of data appearing therein

References

Bickmore, T., & Picard, R. (2005). Establishing and maintaining long-term human—computer relationships. *ACM Transactions on Computer-Human Interaction* 12(2), 293–327.

Breazeal, C. (2009). Role of expressive behaviour for robots that learn from people. *Philosophical Transactions of the Royal Society B, 364*, 3527–3538.

Calvo, R. A., & D'Mello, S. K. (2010). Affect Detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2010a). Affect recognition for interactive companions: Challenges and design in real-world scenarios. *Journal on Multimodal User Interfaces*, 3(1–2), 89–98.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2010b). Inter-ACT: An affective and contextually rich multimodal video corpus for studying interaction with robots. In *Proceedings of the ACM international conference on multimedia 2010* (pp. 1031–1034). New York: Association for Computing Machinery.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2012). Detecting engagement in HRI: An exploration of social and task-based context. In *Proceedings of the IEEE/ASE international conference on social computing (SocialCom'12)*IEEE Press.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2013). Multimodal affect modelling and recognition for empathic robot companions. *International Journal of Humanoid Robotics*, 10(1), 1–23

Cohen, I., Sebe N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1–2), 160–187.

D'Mello, S. K., & Kory, J. (2012). Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies. In L. P. Morency et al. (Eds.), *Proceedings of the 14th ACM international conference on multimodal interaction* (pp. 31–38). New York: Association for Computing Machinery.

Ennis, C., Peters, C., & O' Sullivan, C. (2011). Perceptual effects of scene context and viewpoint for virtual pedestrian crowds, *ACM Transactions on Applied Perception*, 8(2), 10:1–10:22.

Eyben, F., Woellmer, M., Valstar, M. F., Gunes, H., Schuller, B., & Pantic, M. (2011). String-based audiovisual fusion of behavioural events for the assessment of dimensional affect.

Proceedings of IEEE conference on automatic face and gesture recognition (pp. 322–329), IEEE Press

Fontaine, J., Scherer, K., Roesch, E., & Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12).

Grandjean, D., & Peters, C. (2011). Novelty processing and emotion: conceptual developments, empirical findings and virtual environments. In P. Petta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-oriented systems: The humaine handbook* (pp. 441–458). New York: Springer.

Gunes, H., & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2), 120–136.

Gunes, H., Schuller, B., Pantic, M., & Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of IEEE FG 2011* (pp. 827–834).

Gunes, H., & Piccardi, M. (2009). 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.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the ACM international conference on multimedia 2005* (pp. 677–682).

Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics Part B., 41*, 1027–1038.

Leite, I., Castellano, G., Pereira, A., Martinho, C. & Paiva, A. (2012a). Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In *Proceedings of the ACM/IEEE international conference on human-robot interaction (HRI'12)*. ACM Press

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2012b). Long-term Interactions with empathic robots: Evaluating perceived support in children. In *Proceedings of the international conference on social robotics*. Springer-Verlag, 298–307.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. In *IEEE International conference on automatic face and gesture recognition (FG2011)* (pp. 57–64).

Martinez, H. P., & Yannakakis, G. N. (2011). Mining multimodal sequential patterns: A case study on affect recognition. In *Proceedings of the 13th international conference on multimodal interaction (ICMI'11)*. New York: Association for Computing Machinery. (pp. 3–10)

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroeder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing, 3*, 5–17.

McKeown, G., Curran, W., McLoughlin, C., Griffin, H., & Bianchi-Berthouze, N. (2013). Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. In *Proceedings of 2nd international workshop on emotion representation, analysis and synthesis in continuous time and space (EmoSPACE), in conjunction with the IEEE conference on automatic face and gesture recognition.*

Meng, H., & Bianchi-Berthouze, N. (2013). Affective state level recognition in naturalistic facial and vocal expressions, *IEEE Transactions on Systems, Man, and Cybernetics Part B*, in press.

Meng, H., & Bianchi-Berthouze, N. (2011). Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In *Proceedings of the 4th international conference on affective computing and intelligent interaction* (pp. 378–387). New York: Springer.

Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing (IMAVIS)* (Special Issue on Affect Analysis in Continuous Input), 31(2), 137–152.

Morency, L.-P., de Kok, I., & Gratch, J. (2008). Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI'08)* (pp. 181–188), ACM Press.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2012). Output-associative RVM regression for dimensional and continuous emotion prediction. *Image and Vision Computing Journal*(Invited Paper for the Special Issue on Best of 2011 Automatic Face and Gesture Recognition), 30 (3), 186–196.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011a). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* (Special Issue on Affect Based Human Behavior Understanding), 2(2), 92–105.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011b). A multi-layer hybrid framework for dimensional emotion classification. In *Proceedings of ACM multimedia* (pp. 933–936).

Nicolaou, M. A., Gunes, H., & Pantic, M. (2010). Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Proceedings of international conference on pattern recognition* (pp. 3695–3699).

Nicolle, J., Rapp, V., Bailly, K., Prevost, L. & Chetouani, M. (2012). Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on multimodal interaction* (pp. 501–508). New York: Association for Computing Machinery.

Peters, C., Asteriadis, S., & Karpouzis, K. (2010). Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces*, 3(1–2), 119–130.

Petridis, S., Gunes, H., Kaltwang, S., & Pantic, M. (2009), Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. *Proceedings of ACM international conference on multimodal interfaces* (pp. 23–30).

Ramirez, G., Baltrusaitis, T., & Morency, L. P. (2011). Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Proceedings first international audio/visual emotion challenge and workshop, AVEC 2011* (held in conjunction with the international HUMAINE association conference on affective computing and intelligent interaction 2011, ACII 2011) (Vol. II, pp. 396–406). New York: Springer.

Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T., & Rigoll, G. (2008). Detection of security related affect and behaviour in passenger transport. In *Proceedings INTERSPEECH 2008, 9th annual conference of the international speech communication association* (pp. 265–268), ISCA/ASSTA, ISCA.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011). AVEC 2011—The first international audio/visual emotion challenge. In *Proceedings first international audio/visual emotion challenge and workshop, AVEC 2011* (held in conjunction with the international

HUMAINE association conference on affective computing and intelligent interaction 2011, ACII 2011) (Vol. II, pp. 415–424). New York: Springer.

Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2012). AVEC 2012—The continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on multimodal interaction* (pp. 449–456). New York: Association for Computing Machinery.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D.,...Wöllmer, M. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2), 165–183.

Schröder, M., Bevacqua, E., Eyben, F., Gunes, H., Heylen, D., ter Maat, M.,...Wöllmer, M. (2009). A demonstration of audiovisual sensitive artificial listeners. In *Proceedings of IEEE conference on affective computing and intelligent interaction* (pp. 263–264).

Schröder, M., Pammi, S., Gunes, H., Pantic, M., Valstar, M., Cowie, R.,...de Sevin, E. (2011). Come and have an emotional workout with sensitive artificial listeners! In *Proceedings of IEEE conference on automatic face and gesture recognition* (pp. 646).

Vogt, T., Andre, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In *LNCS 4868*, (pp. 75–91).

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B. & Rigoll, G. (2012). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* (special issue on affect analysis in continuous input), 31(2), 153–163.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39–58.