Emotion Recognition Based on Joint Visual and Audio Cues

Nicu Sebe[1], Ira Cohen[2], Theo Gevers[1], Thomas S. Huang[3]

[1]University of Amsterdam, The Netherlands
[2]HP Labs, USA
[3]University of Illinois at Urbana-Champaign, USA

Abstract

Recent technological advances have enabled human users to interact with computers in ways previously unimaginable. Beyond the confines of the keyboard and mouse, new modalities for human-computer interaction such as voice, gesture, and force-feedback are emerging. However, one necessary ingredient for natural interaction is still missing - emotions. This paper describes the problem of bimodal emotion recognition and advocates the use of probabilistic graphical models when fusing the different modalities. We test our audio-visual emotion recognition approach on 38 subjects with 11 HCI-related affect states. The experimental results show that the average person-dependent emotion recognition accuracy is greatly improved when both visual and audio information are used in classification.

1. Introduction

In many important HCI applications such as computer aided tutoring and learning, it is highly desirable (even mandatory) that the response of the computer takes into account the emotional or cognitive state of the human user [7]. Computers today can recognize much of what is said, and to some extent, who said it but, they are having difficulties when it comes to how things are said, the affective channel of information. Affective communication explicitly considers how emotions can be recognized and expressed during human-computer interaction. Addressing the problem of affective communication, in [1] were identified three key points to be considered when developing systems that capture affective information: embodiment (experiencing physical reality), dynamics (mapping experience and emotional state with its label), and adaptive interaction (conveying emotive response, responding to a recognized emotional state).

Mehrabian [8] indicated that when judging someone's affective state, people mainly rely on facial expressions and vocal intonations. Thus, affect recognition should be performed in a multimodal framework [15]. In this paper, our main goal is to combine cues from facial expression and vocal information so that the affective state of a person can be inferred more accurately.

In this paper, when performing emotion recognition we have a more general human-computer interaction application in mind. As a consequence, besides the 6 universal emotions (happy, surprise, angry, disgust, fear, and sad) we take into consideration other affective states that indicate user's cognitive/motivational states: interest, boredom, confusion, and frustration. We test our bimodal affect recognition approach on 38 subjects with these 11 affective states (neutral state is also considered). The experimental results show that the average person-dependent emotion recognition accuracy is greatly improved when both visual and audio information are used in classification.

2. Related Work

So far, the studies in facial expression recognition and vocal affect recognition have been done largely independent of each other. Most current works in facial expression recognition use still photographs or video sequences where the subject exhibits only facial expressions without speaking any words. Similarly, the works on vocal emotion detection used only the audio information. There are situations where people would speak and exhibit facial expressions at the same time. For example, "he said hello with a smile." Pure facial expression recognizers may fail because the mouth movements may not fit the description of a pure "smile." For computers to be able to recognize emotional expression in practical scenarios, these cases must be handled.

According to [9], only a few reports [2][4][10][13][14] of bimodal affect recognition are found. While the recent research and technology advances make multimodal analysis of human emotions feasible, progress in this direction is only in its infancy. Compared to the previous reports of bimodal affect recognition listed in [9] and [11], the contributions of this paper are as follows:

1) 11 affective states are analyzed, including 4 HCI-related affective states (confusion, interest, boredom, frustration). The other works (excepting [14]) only analyzed 5-6 basic emotions.
2) 38 subjects are tested. The numbers of subjects in [2][4][13] are at most five. Thus, the generality of their algorithms is not guaranteed.
3) Bayesian networks are applied for bimodal fusion. The authors of [2][4] applied rule-based methods for combining two modalities while in [13] a weighted intensity summation is used. It is not clear whether their rules or methods are suitable for more subjects.

4) Integrate a variable into the Bayesian network that indicates whether the person is speaking or not. In [14] a smoothing method is applied to reduce the detrimental influence of speech on the information provided by facial expression; [2][4][13] ignored this problem.

3. Feature extraction

The face tracker we use is based on a system developed by Tao and Huang [12] and described in detail in [3]. A snap shot of the system with the face tracking and the recognition result is shown in Figure 1.

Figure 1. A snap shot of our facial expression recognition system. On the right side is a wireframe model overlayed on a face being tracked. On the left side the correct expression, Angry, is detected.

This face tracker uses a model-based approach where an explicit 3D wireframe model of the face is constructed. In the first frame of the image sequence, landmark facial features such as the eye and mouth corners are selected interactively. A face model consisting of 16 surface patches embedded in Bezier volumes is then warped to fit the selected facial features. The surface patches defined this way are guaranteed to be continuous and smooth. The shape of the mesh can be changed by changing the locations of the control points in the Bezier volume.

Once the model is constructed and fitted, head motion and local deformations of the facial features such as the eyebrows, eyelids, and mouth can be tracked. First the 2D image motions are measured using template matching between frames at different resolutions. Image templates from the previous frame and from the very first frame are both used for more robust tracking. The measured 2D image motions are modeled as projections of the true 3D motions onto the image plane. From the 2D motions of many points on the mesh, the 3D motion can be estimated by solving an overdetermined system of equations of the projective motions in the least squared sense.

The recovered motions are represented in terms of magnitudes of some predefined motion of various facial features. Each feature motion corresponds to a simple deformation on the face, defined in terms of the Bezier volume control parameters. We refer to these motions vectors as Motion-Units (MU's). Note that they are similar but not equivalent to Ekman's AU's [5] and are numeric in nature, representing not only the activation of a facial region, but also the direction and intensity of the motion.

As audio features we use three kinds of prosody features for affect recognition (also used in [14]): the logarithm of energy, the syllable rate, and two pitch candidates together with their corresponding scores. The log energy is computed by $E = \log \sum_{i=1}^{N} x_i^2$ where N is the frame length and x_i is the i^{th} signal in that frame. For pitch extraction, an autocorrelation based pitch detector is used to extract two candidates of pitch frequency. The autocorrelation function is the correlation of a waveform with itself by delaying some time lag. The mathematical definition of the autocorrelation function is the following: $Xor_p = \log \sum_{i=1}^{N} x_{i+p} x_i$ where x_{i+p} is the $(i+p)^{th}$ signal in that frame. The autocorrelation of periodic signal is also periodic. As the signal lags to the length of one period, the autocorrelation increases to the maximum; the first peak indicates the period of the signal. Finally, the pitch is detected by

$$P_1 = \arg\max_{P_{\min} \le p \le P_{\max}} Xor_p$$

where $P_{\min}$ is the possible minimum pitch and $P_{\max}$ is the possible maximum pitch. The search range for pitch is set to be 50~1000Hz. In addition to the pitch with the maximum autocorrelation score, the pitch P_2 with the second maximum of autocorrelation score is also chosen as a pitch candidate. Also, the autocorrelation scores are treated as features to detect whether the frame is a vowel. The syllable rate is computed by: *#syllables/duration* where *duration* is the segment duration (0.5 s). To detect the numbers of syllables in the segments (*#syllables)*, a threshold-based speech detection method is used to detect the syllables in the signal. In detail, the frame is considered as speech if the following condition is satisfied:

$$E > 50 \wedge Xor_{P_1} > 0.5 \wedge Xor_{P_2} > 0.5 \wedge \left| \frac{P_{i1}}{P_{i2}} - 2 \right| < 0.2$$

where E is the log energy of one frame, Xor_{P1} and Xor_{P2} are the autocorrelation scores of the two pitch candidates, and P_{i1} and P_{i2} are the larger and the smaller values of P_1 and P_2, respectively. After speech detection, we can count the number of speech segments and compute the syllable rate as one dimension of prosody features. The prosody modality in our experiment can output 92 frames per second in real-time conditions.

The features extracted are used as inputs to a classifying stage described in the next section.

4. Bayesian Networks and Fusion

A typical issue of multimodal data processing so far is that the multisensory data are typically processed separately and only combined at the end. Yet this is almost certainly incorrect; people display audio and visual communicative signals in a complementary and redundant manner. Chen and Huang [2] have shown this experimentally. In order to accomplish a human-like multimodal analysis of multiple

input signals acquired by different sensors, the signals cannot be considered mutually independent and cannot be combined in a context-free manner at the end of the intended analysis but, on the contrary, the input data should be processed in a joint feature space and according to a context-dependent model [15]. In practice, however, besides the problems of context sensing and developing context-dependent models for combining multisensory information, one should cope with the size of the required joint feature space, which can suffer from large dimensionality, different feature formats, and timing. Our approach to achieve the target tightly coupled multisensory data fusion is to develop context-dependent versions of a suitable method such as the Bayesian inference method [3].

If we consider the state of the art in audio and visual signal processing, noisy and partial input data should also be expected. A multimodal system should be able to deal with these imperfect data and generate its conclusion so that the certainty associated with it varies in accordance to the input data. Probabilistic graphical models, such as hidden Markov models (including their hierarchical variants), Bayesian networks, and dynamic Bayesian networks are very well suited for fusing such different sources of information. These models can handle noisy features, temporal information, and missing values of features all by probabilistic inference. Hierarchical HMM-based systems [3] have been shown to work well for facial expression recognition. Dynamic Bayesian networks and HMM variants have been shown to fuse various sources of information in recognizing user intent, office activity recognition, and event detection in video using both audio and visual information [6]. The success of these research efforts has shown that fusing audio and video for detection of discrete events using probabilistic graphical models is possible. Therefore, in this work we propose the Bayesian network topology for recognizing emotions from audio and facial expressions presented in Figure 2.

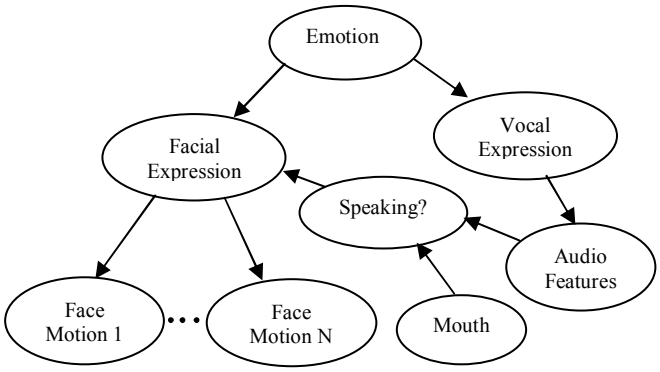

Figure 2. Bayesian network topology for bimodal emotion expression recognition.

The network topology combines the two modalities in a probabilistic manner. The top node is the class variable (recognized emotional expression). It is affected by the recognized facial expressions, the recognized vocal expressions, and by the context in which the system operates (if that is available). Vocal emotions are recognized from audio features extracted from the person's audio track. Facial expressions are recognized by facial features tracked using video, but the recognition is also affected by a variable that indicates whether the person is speaking or not. Recognizing whether a person is speaking uses both visual cues (mouth motion) and audio features. The parameters of the proposed network are learned from data. By using this framework, inferring the human emotional expression can be performed even when some pieces of information are missing, e.g., when audio is too noisy, or the face tracking loses the face.

5. Experiments

In the previous reports [2][4][13], the datasets used were so small that the generality of their methods is not guaranteed. In addition, they only detected 5-6 basic emotional states that are not directly related to human computer interaction. However, as was noticed in [14], the subjects facing a computer tutor seldom express these basic emotions except the neutral state. Actually, detecting some special affects, including interest, boredom, confusion, and frustration, is very important for the system to interact naturally with their users. These affects indicate the cognitive/motivational states of the subjects' learning. They provide information about whether the subject is engaged or whether the subject is having difficulties during the learning activities.

In our experiments we used a large-scale database [14] that is more related to the human-computer interaction; 11 affect categories were used which include 7 basic affects (i.e. happiness, sadness, fear, surprise, anger, disgust, and neutral), and 4 HCI-related affects (i.e. interest, boredom, confusion, and frustration). We tested our methods on 38 subjects (24 females and 14 males).

The subjects consist of mostly graduate and undergraduate students in various fields. Some staff and faculty members also volunteered to participate. Although the subjects displayed affect expression on request, minimal instruction was given to the subjects. In particular, no instruction on how to portray the affects was given. They were simply asked to display facial expressions and speak appropriate sentences. Each subject was required to pose a pure facial expression without speech three times, followed by a facial expression with speech three times, and then a pure facial expression three more times.

In the dataset in which subjects were facing the camera, we found that appropriate sentences made subjects more natural, and this way it was easier for them to express affects than without speech. In particular, without appropriate sentences, some subjects found it difficult to display subtle differences among the 4 HCI-related affects (confusion, interest, boredom, and frustration). On the other hand, speaking reduces the distinctiveness of the different facial expression. As one can see from Figure 3, when the "speaking" variable is not used in the Bayesian network ("No speaking" results) the performance is substantially reduced.

		Detected										
		Neutral	Happy	Surprise	Anger	Disg	Fear	Sad	Frust	Puzz	Inter	Bore
Desired	Neutral	**98.34**	0.43	0.27	0.00	0.00	0.00	0.54	0.00	0.00	0.42	0.00
	Happy	3.47	**92.05**	0.68	0.48	0.58	0.28	1.16	0.00	0.00	0.82	0.48
	Surprise	4.40	0.26	**89.30**	1.07	0.00	1.52	1.76	0.00	1.06	0.00	0.63
	Anger	1.15	0.24	0.47	**94.63**	1.42	0.23	1.16	0.05	0.42	0.20	0.03
	Disgust	2.27	0.81	0.77	3.38	**88.42**	0.81	0.77	0.99	0.50	0.55	0.73
	Fear	1.61	0.20	0.20	4.13	1.51	**90.02**	0.81	0.00	0.91	0.21	0.30
	Sad	4.30	0.46	0.46	3.07	2.67	0.00	**85.04**	1.46	1.31	0.31	0.92
	Frustration	4.21	0.17	1.51	3.31	0.37	2.57	2.21	**83.56**	1.00	0.37	0.72
	Puzzlement	1.98	0.21	0.63	2.82	1.05	1.14	0.21	0.73	**89.01**	0.97	1.25
	Iinterest	2.14	0.36	1.42	3.20	1.78	1.07	1.42	0.36	1.06	**85.05**	2.14
	Boredom	2.02	0.19	0.16	1.12	0.53	0.66	2.62	0.86	0.91	0.43	**91.50**

Table 1. Confusion-matrix for similarity classification based on the combined featural and configural information.

For every subject, the data was divided into two parts. One half of every subject's frames are selected as training data, and the other half as testing data. Our experimental results (Figure 3) show that the average recognition accuracy is about 56% for the face-only classifier, about 45% for the prosody-only classifier, but around 90% for the bimodal classifier. The testing time was about 7.8 seconds for each of the three methods and for every affect. Also, our results are comparable with the ones of [14] where the authors report average bimodal results of 89%.

The male-female average confusion matrix for the bimodal classifier is presented in Table 1. The analysis of the confusion matrix shows that the neutral state has the highest recognition accuracy, and is the state with which the other affective states are most confused. One reason behind this result is that every affective expression in the dataset started from the neutral state, and ended at the neutral state. Thus, the first and last few frames of each affect are very close to neutral states. In addition, anger, boredom, and happiness have high recognition accuracies while sadness and interest have low accuracies. Contrary to the neutral state, the frustrated state is of lowest recognition accuracy, and is the state with which other affects are not likely to be confused.

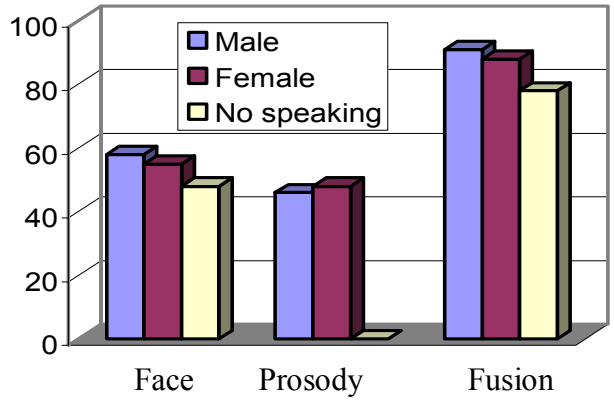

Figure 3. Average affect classification accuracies

References

[1] Bianchi-Berthouze, N. and Lisetti, C., Modeling multimodal expression of user's affective subjective experience, *User Modeling and User-adapted Interaction,* 12:49-84, 2002.

[2] Chen, L. and Huang, T., Emotional expressions in audio-visual human computer interaction, in Proc. *ICME,* 423-426, 2000.

[3] Cohen, I., Sebe, N., Garg, A., Chen, L., and Huang, T., Facial expression recognition from video sequences: Temporal and static modeling, *CVIU*, 91(1-2):160-187, 2003.

[4] De Silva, L.C. and Ng, P.C., Bimodal emotion recognition, in Proc. *Face and Gesture Recognition Conf.,* 332-335, 2000.

[5] Ekman, P. and Friesen, W., *Facial Action Coding System: Investigator's Guide,* Consulting Psychologist Press, 1978.

[6] Garg, A., Pavlovic, V., and Rehg, J., Boosted learning in dynamic Bayesian networks for multimodal speaker detection, *Proceedings of the IEEE,* 91(9):1355-1369, 2003.

[7] Goleman D., *Emotional Intelligence,* Bantam Books, 1995

[8] Mehrabian, A., Communication without words, *Psychology Today,* 2(4):53-56, 1968.

[9] Pantic, M. and Rothkrantz, L., Toward an affect-sensitive multimodal human-computer interaction, *Proceedings of the IEEE,* 91(9):1370-1390, 2003.

[10] Song, M., Bu. J., Chen. C., and Li, N., Audio-visual based emotion recognition: A new approach, in Proc. *CVPR,* 1020-1025, 2004.

[11] Sebe, N, Cohen, I., and Huang, T., Multimodal emotion recognition, *Handbook of Pattern Recognition and Computer Vision,* World Scientific, 2005.

[12] Tao, H. and Huang, T., Connected vibrations: A modal analysis approach to non-rigid motion tracking, in Proc. *CVPR,* 735-740, 1998.

[13] Yoshitomi, Y., Kim, S., Kawano, T., and Kitazoe, T., Effect of sensor fusion for recognition of emotional states using voice, face image, and thermal image of face, in Proc. *Int'l Workshop on Robot-Human Interaction,* 178-183, 2000.

[14] Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T., Roth, D., and Levinson, S., Bimodal HCI-related affect recognition, in Proc. *ICMI,* 2004.

[15] Pantic, M., Sebe, N., Cohn, J., and Huang, T., Affective multimodal human-computer interaction, in Proc. *ACM Multimedia, 2005.*