Fusion of Audio- and Visual Cues for Real-Life Emotional Human Robot Interaction

Ahmad Rabie
ahmad.rabie@hs-ruhrwest.de

Uwe Handmann
uwe.handmann@hs-ruhrwest.de

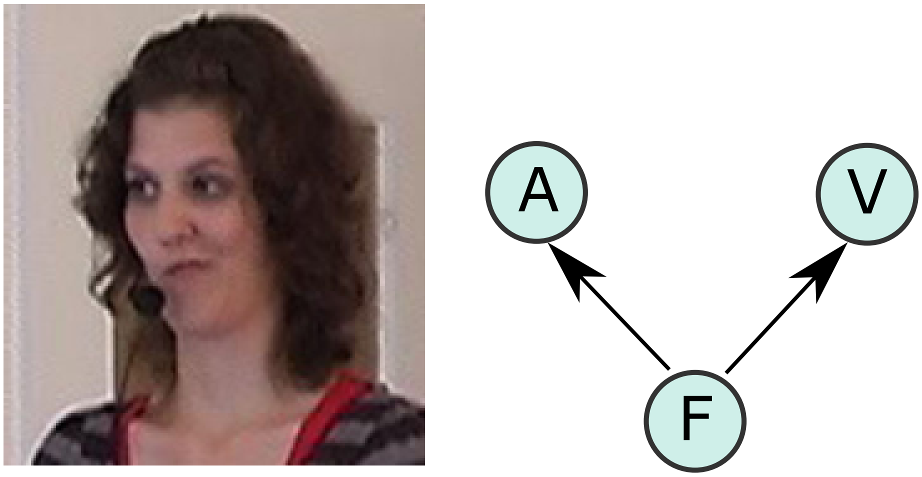
Institute of Informatics
University of Applied Sciences; HRW
Mülheim & Bottrop, Germany

Abstract. Recognition of emotions from multimodal cues is of basic interest for the design of many adaptive interfaces in human-machine interaction (HMI) in general and human-robot interaction (HRI) in particular. It provides a means to incorporate non-verbal feedback in the course of interaction. Humans express their emotional and affective state rather unconsciously exploiting their different natural communication modalities such as body language, facial expression and prosodic intonation. In order to achieve applicability in realistic HRI settings, we develop person-independent affective models. In this paper, we present a study on multimodal recognition of emotions from such auditive and visual cues for interaction interfaces. We recognize six classes of basic emotions plus the neutral one of talking persons. The focus hereby lies on the simultaneous online visual and accoustic analysis of speaking faces. A probabilistic decision level fusion scheme based on Bayesian networks is applied to draw benefit of the complementary information from both – the acoustic and the visual – cues. We compare the performance of our state of the art recognition systems for separate modalities to the improved results after applying our fusion scheme on both DaFEx database and a real-life data that captured directly from robot. We furthermore discuss the results with regard to the theoretical background and future applications.

1 Introduction

Recognizing emotions is widely accepted as one relevant step towards more natural interaction in human-robot and, more general, human-machine interaction. The new scientific understanding of emotions on the one hand, and the rapid evolution of computing system skills on the other, provided inspiration to numerous researchers to build machines that will have the ability to recognize, express, model, and communicate emotions.

In order to exploit emotional cues also in technical interfaces, the recognition of dedicated emotions is in particular necessary. Indeed, humans articulate emotions using different modalities in parallel (cf. Fig. 1(a) for an example from a human-robot interaction study). On the one hand, the different modalities



(a) Emotion articulation in an HRI task

(b) The Bayesian network structure for decision-level fusion

Fig. 1.

transport a significant amount of redundancy, allowing a more robust perception of one's emotion. On the other hand, dedicated emotions might be easier to read from one cue than another. Consequently, the recognition of emotions from multi-modal cues already has a certain tradition in research with a focus on visual and auditory cues, due to their relevance in human-human interaction. In order to apply emotion recognition in real-world interactive systems, a multi-modal, online system is proposed that analyzes users' faces and voices in order to classify emotions. As we are interested in the online analysis of verbal interactions, the paper focuses on the multi-modal analysis of talking interlocutors which is different from most approaches which focus on non-talking faces.

This paper first briefly introduces the relevance and related work on multi-modal recognition of emotion on natural interaction in Sec. 2. Afterwards, our systems for visual (Sec. 3.1) and acoustic (Sec.3.2) recognition are introduced. We present a comprehensive study using the DaFEx database [1] on recognizing the six basic Ekmanian emotions (anger, disgust, happiness, fear, sadness, and surprise)[2] plus a neutral class and discuss the results with regard to a fusion scheme and its accordance to theoretical models. An evaluating study of unimodal- and bimodal recogniton of five basic emotions (anger, fear, neutral, sadness, and suprise) from data captured directly from robot's camera and microphon demonstrates the ability of our bimodal system to be applied for natural, unrestricted and life-like human-robot interaction in Sec. 4.

2 Emotions in Natural Interaction

Though the study of emotion has a long tradition in psychology, approaches to the automatic recognition of emotions have emerged only in the last decade when the necessity to deal with the affective state of a user has become obvious for efficient and user-friendly human-computer interaction. For example, in tutoring systems or computer games, knowing about the user's feeling of boredom, frustration or happiness can increase learning success or fun in the game. In human-robot interaction, affective reactions of the robot, following the recognition of the user's emotional state, can make the interaction more natural and human-like.

Possible modalities to exploit for automatic recognition are language (acoustic and linguistic information), facial expressions, body gestures, bio signals (e. g. heart rate, skin conductance), or behavioral patterns (such as mouse clicks). Though one modality alone can already give information on the affective state of a user, humans always exploit all available modalities, and if an automatic systems attempts to reach human performance, the need for multi-modality is obvious. Thereby not only consent results of different modalities lead to more confident decisions, but also conflicting results can be helpful [3], e. g. to detect pretended or masked emotions, or to find out more reliable modalities for certain emotions. The most obvious modalities in human-human conversation, and also in human-robot conversation which we aim to enhance, are speech and facial expressions. Most related work has focused on the offline analysis of actors [4,5] or spontaneous emotions databases [6]. [3] present a framework for the fusion of multiple modalities for emotion recognition, however, without evaluation.

The novel aspect of our work is the using of technology that is fully capable of online recognition of emotion for natural human robot interaction. We presented an offline analysis of an actors database as previous work. With this analysis we first wanted to find a suitable fusion scheme motivated by the uni-modal results of which emotions are better recognised by which modality. This fusion scheme is then straight-forward applied in real-time scenario of human robot interaction.

3 Bi-Modal Emotion Recognition

Theories of modality fusion in human perception do not agree on how information from different modalities should be integrated. For example, the Fuzzy Logical Model of Perception (FLMP) [7] states that stimuli from different modalities should be treated as independent sources of information and be combined regardless of the kind of information they contain. This view is not undisputed (i.e. [8]) and it has been argued that the FLMP does not work well when confronted with conflicting information from different modalities [9]. Perceptual results suggest that, at least for the case of emotion recognition, the modalities should be weighted according to which information that they convey best [10]: the visual modality primarily transmits valence (positive or negative value) whereas the auditory channel mainly contains information about activation.

In our work we challenge this approach by analysing the auditory and visual stimuli with respect to their general discriminative power in recognizing emotions. Note that in our work we focus on interactive scenarios and are thus targeting at systems that are able to work online. The approaches we present in this paper are, therefore, not only being tested offline on existing databases but have proven their applicability in robotic applications in real world settings [11,12]. This is in contrast to other work (e.g. [4,5]), which has focussed on offline emotion recognition only. The following three sections will provide a brief introduction on the respective unimodal analysis techniques as well on the proposed probabilistic decesion level fusion.

3.1 Visual Facial Expression Recognition

In order to recognize basic emotion visually, we take a closer look into the interlocutor's face. The basic technique applied here are Active Appearance models (AAMs) first introduced by Cootes et al [13]. The generative AAM approach uses statistical models of shape and texture to describe and synthesize face images.

An AAM, that is built from training set, can describe and generate both shape and texture using a single appearance parameter vector, which is used as feature vector for the classification. The "active" component of an AAM is a search algorithm that computes the appearance parameter vector for a yet unseen face iteratively, starting from an initial estimation of its shape. The AAM fitting algorithm is part of the integrated vision system [11] that consists of three basic components. Face pose and basic facial features (BFFs), such as nose, mouth and eyes, are recognized by the face detection module [14]. This face detetion in particular allows to apply the AAM approach in real-world enviroments as it has proven to be robust enough for face identification in human robot interaction in natural environments [15]. The coordinates representing these features are conveyed to the facial feature extraction module. Here, the BFFs are used to initialize the iterative AAM fitting algorithm. After the features are extracted the resulting parameter vector for every image frame is passed to a classifier which categorizes it in one of the six basic emotions in addition to the neutral one. Besides the feature vector, AAM fitting also returns a reconstruction error that is applied as a confidence measure to reason about the quality of the fitting and also to reject prior false positives resulting from face detection. As classifier a one-against-all Support Vector Machine is applied. The whole system is applicable in soft real-time, running at a rate of approximate (5) Hz on recent PC hardware.

3.2 Emotion Recognition From Speech

For the recognition of emotions from speech, EmoVoice, a framework that features offline analysis of available emotional speech databases, as well as online analysis of emotional speech for applications, is used [16]. The approach taken there is purely based on acoustic features, that is no word information is used. As a first step in feature extraction, a large vector of statistical features based on prosodic and acoustic properties of the speech signal was calculated for each utterance in the DaFEx database. From this large vector of over 1400 features the most relevant ones were selected by correlation-based feature subset selection [17]. This selection is necessary to increase performance as well as speed of classification. By this way, 71 features related to pitch, energy, MFCCs, to linear regression and range of the frequency spectrum of short-term signal segments, to the speech proportion and to the length of voiced and unvoiced parts in an utterance, and the number of glottal pulses remained. The full procedure of extracting features is described in [18,16]. For classification, again support vector machines were used, but with a linear kernel. The feature selection is typically done offline, but the feature extraction and classification can be done in real-time. Utterances as classification units, which are normally not available in online applications, can be replaced by an on-the-fly segmentation into parts with voice activity.

3.3 Probabilistic Decision Level Fusion

As affective states in interaction are usually conveyed on different cues at the same time, we agree with other works summarized in [19] that a fusion of visual and acoustic recognition yields significant performance gains. Hence, we followed the idea of an online integration scheme based on the prior offline analysis of recognition results on a database. In current classification fusion research, usually two types of multi-modal fusion strategies are applied, namely feature level fusion and decision level fusion. Both types combine different modalities of data to achive better recognition performance. In the former one, the feature spaces of all modalities are merged into one feature space, which is then conveyed to a single classifier. While in the latter type the classification is performed on each modality separately, then the results of each modality are fused to a final class-prediction accuracy. Due to the inherently different nature of our visual and accoustic cues, we decided for a decision-level fusion scheme. But instead of applying majority voting or other simple fusion techniques, we explicitly take the performance of each individual classifier into account and weight it according to their respective discrimination power.

The proposed probablistic approach for this fusion are Bayesian networks with a rather simple structure depicted in Fig. 1(b). Based on the classification results of the individual visual and acoustic classifiers, we feed these into the Baysian network as *evidences* of the observable nodes (A and V, respectively). By Bayesian inference the posteriori probabilities of the unobservable affective fusion (F) node are computed and taken as final result.

The required probability tables of the Bayesian network are obtained from a performance evaluation of the individual classifiers in an offline training phase based on ground-truth annotated databases. Therefore, confusion matrices of each classifier are turned into probability tables modeling the dependent observation probabilities of the model according to the arrows in Fig. 1(b). In the notion of Zeng *et al.* [19], our fusion scheme is referred to as model-level instead of decision-level fusion, as it takes the respective classification performance models into account.

4 Evaluation on Real-Life Data

As a previous work toward a bimodal system with online ability all systems are evaluated on the DaFEx database [1], which consists of 1008 short video clips of eight Italian actors (4 male and 4 female). Each clip comprises then deliberate presentation of one of the six Ekman's basic emotions plus the neutral one and lasts between 4 and 27 sec. The DaFEx database is divided into six blocks, in two of them namely block 3 and block 6 the actors present facial expression without speaking, in the remainder the actors speak during their emotional performance. Each actor in each of these block performs the seven emotion three times with different intensities (high, medium, and low).

The subset of DaFEx was chosen that contained only videos where the actors were speaking namely (block 1, 2, 4, and 5). Due to the small sample size, the same actors were used for training and test; but it shall be noted that both recognizers apply person independent models. However, the same leave-one-out cross-validation is used for the different modalities. Training is done on three blocks and evaluation of the performance of each uni-modals is performed on the one remaining test block. The probability tables for the Bayesian fusion model are obtained from validation of the performance on the three training blocks. The fusion performance is tested again on the test block. In cross-validation, all permutation of blocks are applied to training and test respectively.

	Ang	Dis	Fea	Hap	Neu	Sad	Sur	Total
Vis	94.44	73.61	58.33	80.55	79.16	72.22	62.91	**74.46**
Aco	68.05	51.38	48.61	50.00	87.49	69.44	58.33	**61.90**
Bimodal	81.94	87.50	52.78	86.11	86.11	74.99	77.77	**78.17**

Table 1. Recognition rates achieved by each unimodal und the bimodal for each individual emotion.

Table 1 depicts the achieved significant overall improvement of the proposed fusion scheme applying our simple Bayesian networks model proposed in Sec. 3.3. The fused system has the advantage over the vision- and audio-based unimodal of about 4% and 16% points, respectively. The 2^{nd} and the 7^{th} columns (Dis, Sur respectively) reveal a high accuracy of the fusion model for recognizing disgust and surprise respectively in contrast to the stand alone uni-modal models, indicating that both cues obviously comprise complementary information that facilitate eased discrimination in the joint analysis. In contrast, from column (Fea) it is noticeable that both uni-modal cues comprise only redundant information so that the fusion yields no improvement with regard to discrimination ability for the recognition of fear. Overall our system achieves good results on the DaFEx database, which are comparable with those reported for human observers [1]. However, the results achieved by the bimodal system emphasize putting forward the fusion scheme of Sec. 3.3 toward efficient recognition of emotion for HRI [20].

As we are striving in this work to give the robot a bimodal emotion recognition ability that is based on analyzing facial expressions and speech information, the systems are afresh evaluated on data set with subjects in a real-life conditions. Four subjects have participated in this test (one female and three males). The whole procedure is divided into training and test phases. For one subject both phases were conducted in the same day; for two others the test was is conducted in the following day, while for the fourth subject the time interval was two days.

In the training phase the subjects are asked to display facial expressions of five emotion classes: anger, happiness, neutral, sadness, and surprise with and without speaking. The average amount of data captured from each subject for each facial expression class was 246 images. To create conditions of real-life human-robot interaction as much as possible, the subjects are allowed to move arbitrarily in front of the camera. During this phase a person-independent AAM, which is built from a subset of the DaFEx database of talking and non-talking subjects, is used to extract the emotion-related facial features. These features are then conveyed to train a person-dependent SVM.

	Anger	Happiness	Neutral	Sadness	Surprise
Anger	**57.72**	00.60	12.19	28.54	00.95
Happiness	02.96	**67.46**	21.00	07.15	01.42
Neutral	05.21	00.00	**64.36**	30.42	00.00
Sadness	02.98	00.00	17.32	**79.18**	00.53
Surprise	05.57	00.88	31.35	10.55	**51.64**
Total	**64.07**				

Table 2. Confusion matrix obtained by using the facial-expression-based system in the test session of displaying emotions deliberatively; rows represent the ground truth.

In the test phase the subjects are asked to display facial expressions and utter a few sentences (in general five) expressing as much an emotions as possible [1]. The above-mentioned AAM is used to extract facial features, which are labeled with the proper emotional class by the above-trained SVM. In this session a person-independent speech-based emotion recognizer is utilized to categorize each utterance into the proper emotional class. An average of 145.25 images from each subject for each emotion are used as test data. The validation matrix for the fusion scheme of each subject was an averaged confusion matrix (CPT), which is obtained from the performance of both individual systems on the three remaining subjects.

Table 2 illustrates the result obtained by using only the facial-expression-based emotion analysis system to recognize emotions that are deliberatively displayed by the subjects. As depicted in the table, the most negative emotion – sadness – and the most positive emotion – happiness – are recognized the best. Neutral also has a relatively high recognition rate, which can serve to distinguish between emotional and non-emotional states of the interactant. The mutual confusion between sadness and neutral indicates the similarity between them when the distinguishing is based only on analyzing the associated facial

[1] The sentences were emotional words free

expressions. The fact that surprise is a transient state, difficult to hold, which changes rapidly into another one (in our test it changed generally into the neutral state), could be the reason for the relatively high confusion of surprise with neutral.

	Anger	Happiness	Neutral	Sadness	Surprise
Anger	**75.00**	00.00	06.25	18.75	00.00
Happiness	25.00	**43.75**	25.00	06.25	00.00
Neutral	20.00	00.00	**50.00**	30.00	00.00
Sadness	22.36	11.11	06.25	**60.28**	00.00
Surprise	16.67	00.00	12.50	22.92	**47.92**
Total	**55.39**				

Table 3. Confusion matrix obtained by using the facial-expression-based system in the test session of expressing emotions via facial expressions and speech tone simultaneously; rows represent the ground truth.

The results obtained by analyzing facial expressions during speech are illustrated in the table 3. The results present the recognition rates after applying majority voting for each utterance that doesn't include a pause longer than 200 ms. As in the evaluation with the database (offline evaluation), facial-expression-based analysis of emotion delivered lower recognition rates when the subjects were engaged in conversational sessions; 64.07% for the former and 55.39% for the latter. The higher recognition rate of anger during speech compared to anger displayed deliberatively could be because majority voting over the time of each sentence is applied in the former, while the recognition rate of the latter is computed for the entire video sequence.

	Anger	Happiness	Neutral	Sadness	Surprise	Total
Vis	75.00	43.75	50.00	60.28	47.92	55.39
Aco	33.04	15.42	36.25	23.06	10.42	23.63
Audio-Visual	75.00	50.00	68.75	49.03	47.92	58.14

Table 4. The performance of each stand-alone unimodal systems against the one of the bimodal system. All results are obtained from a test in a real-life condition.

Table 4 illustrates the results obtained from both the stand-alone and bimodal systems. The low rates delivered by the speech-based emotion analysis system - the first raw - could be because a person-independent classifier is used, which is trained on a speech-based emotion database that does not include the subjects participating in the evaluation procedure. Nevertheless, it can be seen that the whole performance of the bimodal system has an advantage over both facial-expression- and speech-information-based systems, which satisfy the goal of the fusion scheme proposed previously. However, when the performance of each channel on each emotion is considered it is notable that the recognition rate of happiness and neutral is enhanced when the bimodal system is employed, which indicates that the cues of both modalities comprise complementary information for these two emotions. In contrast, from the first and fifth rows, it is noticeable that both unimodal cues comprise only redundant information so that combining both modalities yields no improvement with regard to discrimination ability for the recognition of anger and surprise. Furthermore, the fourth column indicates that both modalities deliver conflicting information, which causes sadness to be recognized even less than the stand-alone facial-expression-based modality.

The comparison between the performance of all of the systems in the cases of offline (DaFEx database) and online (data captured in real-life conditions) evaluation shows better performance of the systems in the former case, especially of the speech-information-based system. These performance differences were greatly expected because (I) the speech-information-based system in the former was trained using data from the same subjects who had participated in the evaluation test, (II) the facial-expression-based system of the former case was trained and tested on a relatively constrained set of data (the actors displayed almost a frontal-view facial expression with constrained head movements while they were sitting in front of the camera), and (III) the degraded performance of both unimodal systems will consequentially lead to a degraded performance of the bimodal system.

5 Conclusion and Outlook

In this paper we presented our approaches on single cue analysis and multi-cue probabilistic decision-level fusion for emotion recognition. As we strive to recognize the basic emotions in real interaction, we presented a person-independent model and restricted ourselves to the challenge of talking persons in this database-based study.

The results indicate that the performance of each modality is highly varying with the respective emotion class which is in line with hypotheses of modality fusion in human perception [10,7]. Based on these results we put forward our fusion scheme where each modality is weighted according to its discriminative power for a specific emotion by applying Bayesian networks trained according to the performance of the individual classifiers. Towards our goal of real-life Human-Robot Interaction our system presents an advanced improvement not only due its reasonable accuracy in emotion recognition but also due its applicability as an online system in less constrained environments and without any further prior processing [4,6].

References

1. Battocchi, A., Pianesi, F., Goren-Bar, D.: A first evaluation study of a database of kinetic facial expressions (dafex). In: Proc. Int. Conf. Multimodal Interfaces, ACM Press (2005) 214–221
2. Ekman, P., Friesen, W.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions. Prentice Hall (1975)

3. Paleari, M., Lisetti, C.L.: Toward multimodal fusion of affective cues. In: Proc. ACM int. workshop on Human-centered multimedia, New York, NY, USA, ACM (2006) 99–108
4. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proc. Int. Conf. Multimodal Interfaces. (2004)
5. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proc. Int. Conf. Multimodal Interfaces, New York, NY, USA, ACM (2006) 146–154
6. Zeng, Z., Hu, Y., Fu, Y., Huang, T.S., Roisman, G.I., Wen, Z.: Audio-visual emotion recognition in adult attachment interview. In: Proc. Int. Conf. on Multimodal Interfaces, New York, NY, USA, ACM (2006) 139–145
7. Massaro, D.W., Egan, P.B.: Perceiving affect from the voice and the face. Psychonomoic Bulletin and Review (**3**) 215–221
8. de Gelder, B., Vroomen, J.: Bimodal emotion perception: integration across separate modalities, cross-modal perceptula grouping or perception of multimodal events? Cognition and Emotion **14** (2000) 321–324
9. Schwartz, J.L.: Why the FLMP should not be applied to McGurk data .. or how to better compare models in the bazesian framework. In: Proc. Int. Conf. Audio-Visual Speech Processing. (2003) 77–82
10. Fagel, S.: Emotional mcgurk effect. In: Proc. Int. Conf. on Speech Prosody, Dresden, Germany (2006)
11. Rabie, A., Lang, C., Hanheide, M., Castrillon-Santana, M., Sagerer, G.: Automatic initialization for facial analysis in interactive robotics (2008)
12. Hegel, F., Spexard, T., Vogt, T., Horstmann, G., Wrede, B.: Playing a different imitation game: Interaction with an empathic android robot. In: Proc. Int. Conf. Humanoid Robots. (2006) 56–61
13. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. PAMI **23** (2001) 681–685
14. Castrillón, M., Déniz, O., Guerra, C., Hernández, M.: Encara2: Real-time detection of multiple faces at different resolutions in video streams. Journal of Visual Communication and Image Representation **18** (2007) 130–140
15. Hanheide, M., Wrede, S., Lang, C., Sagerer, G.: Who am i talking with? a face memory for social robots (2008)
16. Vogt, T., André, E., Bee, N.: Emovoice — A framework for online recognition of emotions from voice. In: Proc. Workshop on Perception and Interactive Technologies for Speech-Based Systems, Irsee, Germany (2008)
17. Hall, M.A.: Correlation-based feature subset selection for machine learning. Master's thesis, University of Waikato, New Zealand (1998)
18. Vogt, T., André, E.: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proc. of IEEE Int. Conf. on Multimedia & Expo, Amsterdam, The Netherlands (2005)
19. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transaction on Pattern Analysis and Macine Intellegence **31** (2009) 39–58
20. Rabie, A., Vogt, T., Hanheide, M., Wrede, B.: Evaluation and discussion of multimodal emotion recognition. In: ICCEE. (2009)