Use of Nonverbal Communication in Dialog System

JANA KLECKOVA, PAVEL KRAL, JANA KRUTISOVA
Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia in Pilsen,
Univerzitni 22, Pilsen, 306 14
CZECH REPUBLIC
http://www.kiv.zcu.cz

Abstract: - Understanding human emotions and their nonverbal messages is one of the most necessary and important skills for making the next generation of human-computer interfaces (HCI) easier, more natural and effective. Indeed, the first step toward an automatic emotion sensitive human-computer system having the ability to automatically detect users' nonverbal signals is the development of an accurate and real-time automatic NVC analyzer. Such an analyzer must deal mainly with users' facial expressions and paralanguage. The main goal of this paper is to compare different methods to combine the results of both classifiers. A prototype of the dialog system was developed in the Department of Computer Science . The proposed system is fully automatic, user-independent and real-time working. Several experiments show that the speech recognition quality is increased by using nonverbal information.

Key-Words: - Dialog system, nonverbal communication, prosody, face detection and localization, feature extraction and classification.

1 Introduction

Nonverbal communication (NVC) consists of several categories like facial expressions, gaze, proxemics, paralanguage, adornment, etc. In fact, each of these NVC categories function in a distinctive way, and quite different stories can be told about each of them. Therefore, dealing with a system for automatic processing of all of these categories, this work will deal "only" with the automatic recognition of facial expressions, which are an important aspect of our communication and according to many experts, the most important category of nonverbal communication. They regulate social behavior, signal communicative intent, and are related to speech production.

Paralinguistic information is defined as the information that is not inferable from the written counterpart but is added deliberately by the speaker to modify or supplement the linguistic information [9]. A written sentence can be uttered in various ways to different intentions, attitudes, and speaking styles which are under the conscious control of the speaker. Non-linguistic information concerns such factors as the age, gender, physical and emotional states of the speaker, etc. These factors are not directly related to the linguistic and paralinguistic contents of the utterances and, in general, cannot be controlled by the speaker, though it is possible for a speaker to control the way of speaking to intentionally convey an emotion, or to simulate an emotion, as it is done by actors. Understanding human emotions and their nonverbal messages is one of the most necessary and important skills for making the next generation of human-computer interfaces (HCI) easier, more natural and effective. Indeed, the first step toward an automatic emotion sensitive human-computer system having the ability to automatically detect users' nonverbal signals is the development of an accurate and real-time automatic NVC analyzer. Such an analyzer must deal mainly with users' facial expressions and paralanguage. The part of this problem leads to the consideration of the nonverbal communication in developed dialog system [9].

2 Nonverbal communication

Nonverbal communication has many functions in the communication process. By virtue of nonverbal communication, we simply express our emotions. In many cases we are able to exhibit our feelings by facial expression and gestures much more quickly than by using words. It regulates relationships and may support or replace verbal communication. On the other hand nonverbal communication has its disadvantages and seamy sides too. Difficulties may arise if communicators are unaware of the types of messages they are sending, and how the receiver is interpreting those messages. No dictionary can accurately classify nonverbal signals. Their meanings vary not only by culture and situation, but also by the degree of intention of their use. Many of them are ambiguous and could cause misunderstandings. Effective communication is the combined harmony of verbal and nonverbal actions. We can divide the functions of communication into four parts, see Fig.1. The first is called expressive function and it demonstrates outward our inner emotional state. The second is signal function intended for sending simple or complex information. The third is the descriptive function serving for sending more complex and complicated information.

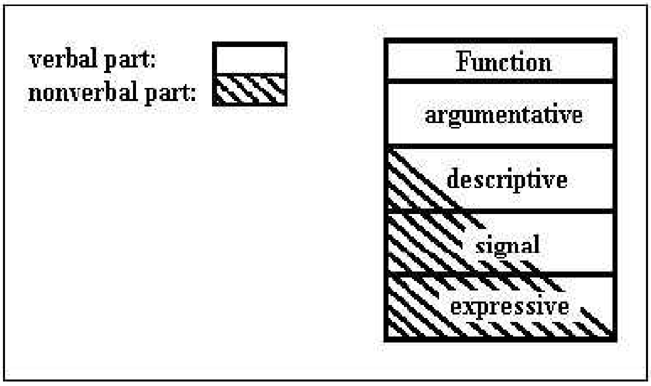

Fig.1 Ratio of verbal and nonverbal parts

2.1 Facial expression

Facial expression carries most of our nonverbal meanings and often is considered as the most important category of nonverbal communication (by many experts 55-85 percent of NVC is exchanged by them). Although the human face is capable of creating 250,000 expressions, less than 100 sets of them constitute meaningful symbols. Three main categories of conversational signals have been identified: syntactic display - used to stress words, or clauses (raising or lowering eyebrows can be used to emphasize a word or clause), speaker displays: illustrate the ideas conveyed ("I don't know" can be expressed by the corners of the mouth being pulled up or down), and listener comment display - used in response to an utterance (incredulity can be expressed with a longer duration of eyebrow raising). Facial expressions are generated by contractions of facial muscles, which result in temporally features such as eye lids, eye brows, lips, skin texture, etc., and are often revealed by wrinkles and bulges. In general, facial actions last 250 ms to 5 seconds. In order to accurately measure and describe facial expressions, we need to know the location of facial actions, their intensity and their dynamics. It should be noted that measuring facial expressions and their intensity is a very difficult task, due to many problems such as their variation and appearance from one individual to another based on their age, ethnicity, sex, facial hair, cosmetic products, etc.

2.2 Prosody

Following the conclusions of previous studies [8,9], only the two most important prosodic attributes are considered: F0 and energy. The F0 curve is computed with the autocorrelation function. The F0 and energy values are computed on every overlapping speech window [10]. The F0 curve is completed by linear interpolation on the unvoiced parts of the signal. Then, each sentence is decomposed into 20 segments and the average values of F0 and energy are computed within each segment. This number is chosen experimentally [2]. We thus obtain 20 values of F0 and 20 values of energy per sentence. Let us call F the set of prosodic features for one sentence. We test two classifiers: a multi-layer perceptron (MLP) that computes P (CjF) and a Gaussian mixture model (GMM) that models P (FjC). Both classifiers error rates are reported in the following experiments, but as they give comparable results, the combination with the lexical classifier uses only the prosodic GMM.

3 Automatic Nonverbal Communication Analyzer

As noted previously, nonverbal communication plays a crucial role in man-to-man interaction. Also, it is clear that correct and effective recognition of spontaneous speech without the use of nonverbal communication (facial expression, paralanguage, etc.) is impossible. Likewise, being aware of how the user is receiving a piece of information provided would be very valuable. Being able to know when the user needs more feedback, by observing cues about his emotional state has many advantages. The human sensory system uses multimodal analysis of multiple communication channels (HCI). This human ability could be our goal for developing an emotion-sensitive and intelligent multimodal HCI. For developing such a system, we need an automatic multimodal nonverbal communication analyzer (ANCA). The performance of a multimodal system (e.g., our nonverbal communication analyzer) does not depend only on the number and types of integrated modalities (sight, sound, touch, etc.). Another important issue is how the data carried by multiple channels should be fused to achieve high performance in recognizing NVC. In general, fusion of multimodal information has always been a topic of discussion, because the technique of how these modalities are fused plays an essential role.

3.1 Automatic Bimodal Audiovisual ANCA for Speech Recognition

An automatic nonverbal communication analyzer needs to mainly deal with users' facial expressions and users' paralanguage. An automatic bimodal audiovisual ANCA needs for capturing nonverbal signals in speech and in facial expressions at least one microphone (audio processing) and one camera (visual processing).

3.1.1 Audio Input

The audio input carries various kinds of information, and considering only the verbal part without regarding the manner in which it was spoken and analyzing facial expressions, might lead to overlooking important information of utterance or even misunderstand it. According to our experiments, the accuracy of automated speech recognition, which is about 80 to 90 percent for neutrally spoken words, tends to drop to 50 to 60 percent if it concerns emotional speech.

Type of utterance	Utterance number
Statements (S)	566
Orders (O)	125
Yes/no questions (Q[y/n])	282
Other questions (Q)	1200
Total	2173

Tab.1 Experimental corpus – type of utterance

3.1.1 Visual Input

The facial expression recognition problem can be divided into the following three partial problems: face detection; facial feature extraction; facial expression classification. In despite of significant advances of computer vision in recent years, developing robust and accurate facial expression recognition in an automatic way and in real-time is still very problematic and at present belongs to one of the greatest dreams and most active areas in the computer vision. All existing systems have their advantages and disadvantages. Many of them are not real-time or if are real-time, they are not fully automatic. In addition, the systems, which are real-time and fully automatic are very limited or are not accurate. Another large problem is that there is a strong need to generate an equivalent to the FERET (Face Recognition Technology) database of facial images for face recognition specifically for facial expression recognition. Approaches for extraction and representation of facial features can be categorized according to several factors. In general, we can distinguish them as local or holistic. Local approaches deal with features, which are prone to change with facial expressions. Holistic approaches deal with the face as a whole.

4 System ARFE (Automatic Recognition of Facial Expression)

ARFE is based on state of the art approaches, is a multi-user system and has two working modes: static images and dynamic images (photo and video) mode. This is possible due to the fact, that each frame is processed and classified separately. The system automatically detects frontal faces in complex backgrounds and makes classification for each found face. The only requirements of the system are frontal faces, a good illumination condition and acceptable light direction. In other words, faces should not contain shadows and must be well lighted. After a face was detected in an arbitrary image, which could be a digitized video signal or a digitized image, the face finder returns the coordinates of a square box around the face. The feature selection step results in a choice of the subset of those features, which best describe our classification classes. The classification is performed also by Adaboost, which combines several weak classifiers (in our case a naive Bayesian classifier) to get a strong and accurate one. For searching the face of an unknown size in the whole input image, we can move the search window across the image several times at different scales, and check each location using our resultant classifier. It is interesting, that most of features selected by Adaboost are from the first and last scale of Gabor filters.

5 Experimental Implementation

The dataset consists of 30 adult volunteers. None of the subjects wore eyeglasses. Some of the subjects had hair covering their foreheads, no subject wore caps, or had makeup on their brows, eyelids or lips. The subjects included both male (70%) and female (30%). The important condition was maximum illumination with a minimum of facial shadows. The primary idea was to ask each volunteer to look at some examples of all 6+1 facial expressions (happy, fear, anger, disgust, sadness, surprise and neutral) and try to copy them. In fact, the gathered training set provides only three acceptable facial expressions: neutral, happy and surprise. Unfortunately, also not all of the surprise expressions are really perfect surprise expressions. The final training set contains 900 images, 30 images per expression from 30 volunteers.

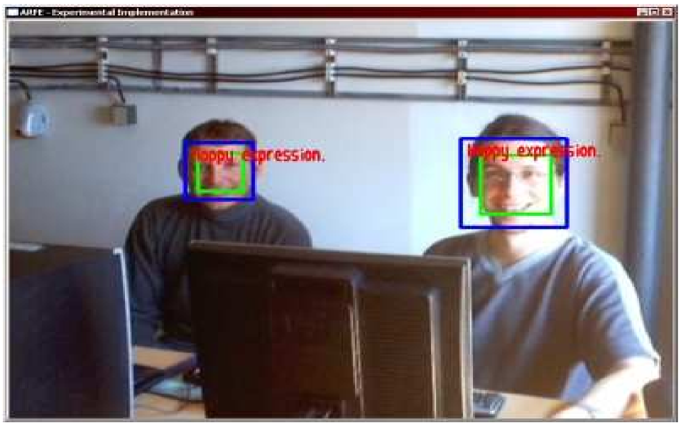

Fig.10 The facial hair causes an incorrect classification results

Corpus, which contains some human-human dialogs, is used to validate the proposed methods. It was created at the University of West Bohemia mainly by members of the Department of Computer Science and Engineering. For the next experiments, it has been labelled manually with the following set of dialog acts: statements, orders, yes/no questions and other questions.

ACC in [%]					
Classifier	S	O	Q[y/n]	Q	Total
First word (lexic)	88.5	90.4	92.9	94.2	92.3
GMM (prosody)	47.7	43.2	40.8	44.3	44.7
MLP (prosody)	38.7	49.6	52.6	34.0	43.5

Tab. 2 Recognition score

The middle part of table 2 shows the recognition accuracy with the prosodic GMM and MLP. The best recognition accuracy is obtained with the 3-mixtures GMM. It is difficult to use more Gaussians, because of the lack of training data, mainly for class O. The best MLP topology uses three layers: 40 inputs, 18 neurons in hidden layer and 4 outputs. The global accuracies of the GMM and MLP classifiers are comparable. Notice that these recognition scores are much lower than the one obtained with the lexical features, but our objective is to show that prosody may nevertheless bring some relevant clues that are not related to the words sequence.

6 Conclusion

In this work, we have studied and compared different methods to combine information of a facial expression, and lexical and prosodic information in the context of automatic dialog act recognition. We have presented an automatic user-independent real-time facial expression recognition system, called ARFE. The most important feature and advantage of ARFE in comparison with many other existing systems is that ARFE is fully automatic. It is a necessary condition for using it in the dialog system. This system provides an excellent face detection system. But unfortunately, the face localization performed by this system is not accurate enough for an automatic facial expression recognition system and without any doubt is in need of an improvement. The lexical knowledge source is by far the most important one, because on a Czech corpus that simulates the first application it already gives recognition accuracy about 92 %. However, we showed that it is possible to improve this baseline result by combining this lexical classifier with a prosodic one with a MLP. Then, a statistically significant 2 % absolute improvement can be achieved, which is actually very close to the potential improvement derived from the correlation matrix between both classifiers. This confirms that prosodic clues are complementary to the lexical ones, as it has been already suggested in other studies such as [2]. All the other combination schemes, and in particular the unsupervised ones, do not reach the level of the lexical classifier alone. This shows the importance to fine-tune the combiner on a development corpus in our experimental set-up. This can be explained by significant difference in the performances of both classifiers, and also by a small number of combined experts. The first perspective of this work will consist to use an automatic speech recognizer, such as the one described in [6], instead of the manual word transcriptions used in our experiments.

References:

[1] H. Wright, Automatic Utterance Type Detection Using Suprasegmental Features. In ICSLP'98, Sydney, 1998, p. 1403.

[2] E. Shriberg et al., Can prosody aid the automatic classification of dialog acts in conversational speech? *In Language and Speech,* 1998, vol. 41, pp. 439–487.

[3] Chengjun Liu and Harry Wechsler, Independent Component Analysis of Gabor Features for Face Recognition. *IEEE Trans. Neural Networks*, Vol. 14, no. 4, 2003, pp. 919-928

[4] Maja Pantic and Leon J.M. Rothkrantz, Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, 2000

[5] Paul Viola and Michael J. Jones, *Robust Real-time Object Detection.* Cambridge Research Laboratory, February 2001.

[6] K. Ekstein and T. Pavelka, Lingvo/laser: Prototyping concept of dialogue information system with spreading knowledge. *In NLUCS'04, Porto, Portugal,* April 2004, pp. 159–168.

[7] Nikolaus Voß and Barbel Mertsching. Design and Implementation of an Accelerated Gabor Filter *Bank Using Parallel Hardware.* Springer-Verlag Berlin Heidelberg 2001.

[8] Jana Kleckova, Jana Krutisova, Vaclav Matousek, Important Prosody Characteristics for Spontaneous Speech Recognition. *In: Proceedings of the 9th International Conference on Neural Information ICONIP 2002*, 2002, pp. 824 -830

[9] J. Kleckova and V. Matousek, *Using Prosodic Characteristics in Czech Dialog System.* In Interact'97, 1997.

[10] Jana Kleckova, Developing the Database of the Spontaneous Speech Prosody Characteristics. *In: Proceedings of the Int. Conference EUROSPEECH 99,* Vol. 2, 1999, pp. 731-734