Knowledge-Based Systems 71 (2014) 34–40

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Predicting emotions in facial expressions from the annotations in naturally occurring first encounters

Costanza Navarretta *

University of Copenhagen, Njalsgade 140, Build. 25, 4. 2300 Copenhagen S, Denmark

ARTICLE INFO

Article history:
Received 24 November 2013
Received in revised form 21 April 2014
Accepted 23 April 2014
Available online 4 May 2014

Keywords:
Multimodal corpus
Multimodal communication
Emotion
Machine learning
Feedback
Turn management
Annotation

ABSTRACT

This paper deals with the automatic identification of emotions from the manual annotations of the shape and functions of facial expressions in a Danish corpus of video recorded naturally occurring first encounters. More specifically, a support vector classified is trained on the corpus annotations to identify emotions in facial expressions. In the classification experiments, we test to what extent emotions expressed in naturally-occurring conversations can be identified automatically by a classifier trained on the manual annotations of the shape of facial expressions and co-occurring speech tokens. We also investigate the relation between emotions and the communicative functions of facial expressions. Both emotion labels and their values in a three dimensional space are identified. The three dimensions are Pleasure, Arousal and Dominance.

The results of our experiments indicate that the classifiers perform well in identifying emotions from the coarse-grained descriptions of facial expressions and co-occurring speech. The communicative functions of facial expressions also contribute to emotion identification. The results are promising because the emotion label list comprises fine grained emotions and affective states in naturally occurring conversations, while the shape features of facial expressions are very coarse grained. The classification results also assess that the annotation scheme combining a discrete and a dimensional description, and the manual annotations produced according to it are reliable and can be used to model and test emotional behaviours in emotional cognitive infocommunicative systems.

1. Introduction

This paper deals with the emotions, attitudes and affective states (emotions henceforth) shown by facial expressions in a corpus of video recorded naturally occurring dyadic first encounters. The emotions which people show in everyday conversations are often not so strong or easily identifiable as the basic universal emotions which many researchers have worked with [14,42]. Emotions often reveal what people really feel about the communicative situation, the interlocutor and the content of the ongoing conversation. Furthermore, they play an important role not only in cognitive processes [21] but also in people's social life [13]. Thus, it is important to include the identification and generation of emotions in implementations of plausible cognitive infocommunicative devices.

Many factors influence communication, such as the cultural and communicative settings, the relationship and role of the participants, their age and number. All these aspects and their relation to communicative multimodal behaviours must be determined in order to build human-like behavioural models and test affective communicative systems. Our point of departure is computational linguistics, and we account for emotions in everyday human–human communication, that is in intra-cognitive communication [2].

Our study concerns emotions in a corpus of video recorded Danish conversations between two subjects who meet each other for the first time. The corpus contains manual annotations of the shape and function of communicative body behaviours as well as descriptions of emotions in facial expressions. The basic emotions Ekman and Friesen [14], Ekman [42] which have been studied in most research are only a subset of the emotions which are expressed in conversations. These can often be subtle reactions to the ongoing conversation situation, its content and the interlocutors' behaviours.

Moreover, differing from most emotion studies which are based on acted data, such as [5,15,7,9], we work with non-acted conversations and with the emotions expressed by the participants. Thus, it is not possible to verify whether the identified emotions are in fact the emotions felt by the participants. However, this is what

* Tel.: +45 35329079.
E-mail address: costanza@hum.ku.dk

http://dx.doi.org/10.1016/j.knosys.2014.04.034
0950-7051/

happens in the real world where each person has to process and interpret the emotions expressed spontaneously or displayed by the interlocutors. Inter-annotator agreement tests were therefore performed to ensure that the emotion labels annotated have an acceptable reliability. Agreement for 28 emotion labels was over 0.6 in terms of kappa score Cohen [10], a result which is better than those reported in the literature for similar tasks, inter alia Martin et al. [24]. Furthermore, all the annotations in the corpus were checked by one or two coders. A more detailed discussion and evaluation of the annotation scheme and procedure were provided in [30] where a first analysis of the emotions annotated in half of the corpus was also presented.

In the present work, we describe experiments in which classifiers are trained on the manual annotations of the shape and functions of facial expressions to identify emotions expressed in the corpus. We also wanted to investigate which information types contribute mostly to classification and, thus, determine possible relations between shape and functional features of facial expressions, co-speech and emotions.

The results of our experiments are quite promising and show that emotions expressed in natural occurring conversations can be identified on the basis of even coarse-grained descriptions of the shape of facial expressions and co-speech. Thus, it is possible to model subtle emotions expressed by people in conversations and use these models for identifying user emotional behaviours and generating plausible emotional cognitive infocommunicative devices.

The paper is organised as follows. In Section 2, we discuss related work, and in Section 3 we present the data. Section 4 accounts for the machine learning experiments on the annotations, whose results are discussed in Section 5. Finally, in Section 6 we conclude and present future work.

2. Related work

In this section, we discuss related work on the identification, annotation and classification of emotions.

Early studies on the identification of emotions in facial expressions have focused on six or eight so called universal emotions, see inter alia [14,42]. Cowie and Douglas-Cowie [11] apply machine learning to a large speech database to identify emotions expressed vocally. Scherer [45] presents acoustic parameters which are relevant for recognising and synthesising a large number of emotions.

Pantic and Rothkrantz [40] and Ioannou et al. [17] address the automatic identification of emotions from still pictures of facial expressions while Black and Yacoob [3] extend recognition to sequences of images. More recently, researchers have used more modalities in order to identify emotions automatically. For example, speech and facial expressions are taken into account in [6,5,15], while Castellano et al. [8] also include body movements. Most of these studies deal with a limited number of acted basic emotions.

Differing from most of these studies, we deal with all types of emotions which occur in a corpus of naturally occurring first encounters. Although we only deal with facial expressions and co-speech in this study, the whole context was taken into account by the coders when annotating emotions.

Various models for describing emotions have been proposed and applied. They can mainly be distinguished in the following types: categorical, dimensional and combined. Categorical models consist of lists of discrete emotion labels [14,36,1], while dimensional models describe emotions by their position on the models' dimensional axes. The number and type of axes vary. Examples of dimensional models are those proposed in [48,44]. In particular, Russell and Mehrabian [44] account for emotions placing them in a three dimensional space model. The three dimensions are Pleasure, Arousal and Dominance (PAD) which have a positive and negative pole. The Pleasure dimension describes positive versus negative affective state, and the Arousal dimension indicates high versus low level of physical activation and/or mental alertness. Finally, the Dominance dimension expresses whether the subject has the feeling of having control and influence over the situation and other people versus the feeling of being controlled and influenced by others or by the situation. The emotions which are accounted for by Russel and Mehrabian do not only describe occasional passionate states, but all emotional states which people experience in their everyday life. We follow Russel and Mehrabian's approach to emotions as well as their definition of the PAD dimensions.

Examples of combined discrete and dimensional models are in [46,24]. Scherer [46] distributes emotion labels in a semantic grid with respect to more dimensions, while Martin et al. [24] propose and test a complex annotation scheme for annotating emotions in video clips of French TV interviews (EmoTv corpus). The scheme describes emotions by eighteen labels and four abstract dimensions: activation, valence, intensity, control. Both a major and a minor emotion label can be assigned and non-basic emotional patterns can be coded. These specify, inter alia, whether emotions are acted to mask the real emotion or they are related directly to a cause. The test of the annotations of 40 coders showed that it is extremely difficult for human coders to identify emotions and especially masked acted emotions. Since the identification of emotions is complex also for humans, Kipp and Martin [20] introduce a simpler annotation scheme than that proposed by Martin et al. [24]. They apply this scheme to annotate the emotions expressed by the hand gestures of the protagonists in two film versions of Arthur Miller's *Death of a Salesman*. The scheme is a simplification of Russel and Mehrabian's PAD model in which the three dimensions are bipolar (±), but are connected to an intensity dimension with three values: *low*, *neutral* and *high*. Kipp and Martin [20] analyse the annotated data and find a correlation between types of hand movement and PAD coding.

In our annotation scheme, we built upon Kipp and Martin [20]'s approach. More precisely, we combined a simplification of their dimensional annotation with the emotion label list proposed in the MUMIN model [1]. The MUMIN list is open-ended because emotions, affective states and attitudes are related to the communicative setting and, thus, can change from a setting to the other and cannot be predicted in advance. In our simplified version of Kipp and Martin [20]'s approach, only the bipolar values of the three PAD dimensions are coded. The correspondence between PAD values and emotions labels were determined in advance by three coders who also decided when to add emotion labels to the open-ended list. The combined annotation strategy was applied in the annotation of emotions in the Danish NOMCO corpus. The coders developed an annotation manual in which the correspondence between emotion labels and PAD values was described by placing the labels in a PAD space where 10 degrees of intensity in each dimension were distinguished. The use of the combined model resulted in improved inter-coder agreement scores with respect to the use of emotion labels or PAD values [30]. Navarretta [30] reports that inter-coder agreement in terms of Cohen's kappa [10] on 25 emotion labels was 0.61 while it was between 0.72 and 0.82 for the PAD values.

Multimodal data are complex, and their annotation often consists of many features which are correlated at different levels. Therefore, it is useful to apply machine learning algorithms to multimodal annotations in order to determine the correlation between the various features. For example, Louwerse et al. [23,22] apply a classifier on annotated English map-task dialogues in order to identify the relation between facial expressions, gaze and speech. Similarly, supervised learning algorithms have been used