Image and Vision Computing 31 (2013) 153–163

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework☆

Martin Wöllmer *, Moritz Kaiser, Florian Eyben, Björn Schuller, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, Theresienstr. 90, 80333 München, Germany

ARTICLE INFO

Article history:
Received 31 October 2011
Received in revised form 13 February 2012
Accepted 7 March 2012

Keywords:
Emotion recognition
Long Short-Term Memory
Facial movement features
Context modeling

ABSTRACT

Automatically recognizing human emotions from spontaneous and non-prototypical real-life data is currently one of the most challenging tasks in the field of affective computing. This article presents our recent advances in assessing dimensional representations of emotion, such as arousal, expectation, power, and valence, in an audiovisual human–computer interaction scenario. Building on previous studies which demonstrate that long-range context modeling tends to increase accuracies of emotion recognition, we propose a fully automatic audiovisual recognition approach based on Long Short-Term Memory (LSTM) modeling of word-level audio and video features. LSTM networks are able to incorporate knowledge about how emotions typically evolve over time so that the inferred emotion estimates are produced under consideration of an optimal amount of context. Extensive evaluations on the Audiovisual Sub-Challenge of the 2011 Audio/Visual Emotion Challenge show how acoustic, linguistic, and visual features contribute to the recognition of different affective dimensions as annotated in the SEMAINE database. We apply the same acoustic features as used in the challenge baseline system whereas visual features are computed via a novel facial movement feature extractor. Comparing our results with the recognition scores of all Audiovisual Sub-Challenge participants, we find that the proposed LSTM-based technique leads to the best average recognition performance that has been reported for this task so far.

1. Introduction

As speech recognition systems have matured over the last decades, automatic emotion recognition (AER) can be seen as going one step further in the design of natural, intuitive, and human-like computer interfaces. Multimodal human-machine communication systems that not only take into account *what* the user says but also *how* the user communicates, are usually perceived as more natural and more enjoyable to use [1]. Examples for successful applications of socially competent human–computer interaction via automatic emotion recognition can be found in the areas of human–robot communication, call center dialog systems, computer games, and conversational agents [2–4]. Since most of today's computer systems are equipped with microphones and cameras, audio and video are the most important non-obtrusive modalities based on which affect recognition can be performed. Audio and video channels can provide complementary information and tend to improve recognition performance if they are used in a combined multimodal setup [5]. This led to a large number of studies investigating audiovisual emotion recognition (e.g., [6]).

The accuracy of automatic emotion recognition heavily depends on the considered scenario: Acted, prototypical emotions recorded in a laboratory environment typically lead to high recognition rates that can compete with human performance in classifying these affective states [7]. These conditions, however, do not reflect real-life scenarios in which non-prototypical spontaneous emotions have to be modeled in an *open-microphone* setting [8], i.e., the system has to listen and observe (time-) continuously. Such challenges demand for 'second generation' AER systems that focus on realistic data and are able to account for the complexity, subtlety, continuity, and dynamics of human emotions [9]. Currently, we are observing a shift from modeling prototypical emotional categories such as *anger* or *happiness* to viewing human affect in a continuous orthogonal way by defining *emotional dimensions* including for example *arousal* and *valence*. This allows researchers to model emotions either in a fully value-continuous way (e.g., via regression approaches as in [10,11]) or by using discretized emotional dimensions, for example for the discrimination of high vs. low arousal or positive vs. negative valence [12,13]. Systems applying the latter approach have the advantage of detecting a defined set of user states which can be easily used as input for automatic dialog managers that have to decide for an appropriate system response given a certain affective state of the user [4].

The 2011 Audio/Visual Emotion Challenge [6] focuses on exactly these kinds of discretized emotional dimensions. More specifically, this challenge was organized to provide research teams with unified training, development and test data sets that can be used to compare individual approaches applying a defined test scenario and defined performance measures. The task was to classify two levels of *arousal*,

☆ This paper has been recommended for acceptance by Hatice Gunes and Bjoern Schuller.

* Corresponding author. Tel.: +49 89 289 28550; fax: +49 89 289 28535.
E-mail address: woellmer@tum.de (M. Wöllmer).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2012.03.001

expectation, *power*, and *valence* from audiovisual data as contained in the SEMAINE database [14]. Compared to rather 'friendly' test conditions as considered in the early days of emotion recognition research [15], this scenario is exceedingly challenging and typically leads to results from below chance-level accuracies to around 70% accuracy for a two-class task.

One approach towards reaching acceptable recognition performance even in challenging conditions is the modeling of contextual information. Even for humans it can be difficult to judge a person's emotional state from a short isolated utterance. Thus, modern AER is influenced by the growing awareness that context plays an important role in expressing and perceiving emotions [16]. Human emotions tend to evolve slowly over time which motivates the introduction of some form of context-sensitivity in emotion classification frameworks. Up to now, most recognition systems only consider feature-level context *within* a spoken utterance or video segment, e.g., via the Markov assumption when applying Hidden Markov Models (HMM) foremotion recognition from framewise low-level features. Yet, recent studies show that also higher-level context modeling between successive utterances increases the accuracy of AER systems [17].

Among various classification frameworks that are able to exploit turn-level context, so-called Long Short-Term Memory (LSTM) networks [18] tend to be best suited for long-range context modeling in emotion recognition. Unlike conventional recurrent neural network (RNN) architectures, LSTM is able to incorporate an arbitrary, self-learned amount of context into the decoding process. They were shown to prevail over standard RNNs for recognition tasks that presume the ability to learn long-range temporal dependencies between network input activations as they overcome the *vanishing gradient problem* (see [19]). First attempts to use LSTM for speech processing concentrate on phoneme recognition and keyword spotting [20–23]. These studies show that modeling not only past but also future context via *bidirectional* Long Short-Term Memory (BLSTM) networks can further enhance context-sensitive sequence processing. Recent publications reveal that also continuous automatic speech recognition (ASR) benefits from LSTM modeling [24,25]. First experiments on (unidirectional) LSTM-based continuous emotion recognition from speech can be found in [26]. This study focuses on the recognition of both, continuous and discretized levels of arousal and valence, showing that LSTM architectures outperform Support Vector Machines (SVM), Support Vector Regression (SVR), and Conditional Random Fields (CRF). Further gains in speech-based affective computing could be obtained with combined acoustic-linguistic modeling, improved LSTM architectures including so-called *forget gates* (see Section fsec:classification), and bidirectional processing [17].

Apart from preliminary experiments using facial marker information as additional input modality [13] and a recent study on subject dependent recognition of arousal and valence [27], LSTM architectures have hardly been applied for audiovisual emotion recognition. In this article, we propose an LSTM-based emotion classification framework which exploits acoustic, linguistic, and visual information. Focusing on the Audiovisual Sub-Challenge of the 2011 Audio/Visual Emotion Challenge, we investigate which modalities contribute to the discrimination between high and low levels of arousal, expectation, power, and valence. Furthermore, we analyze which emotional dimensions benefit the most from unidirectional and bidirectional Long Short-Term Memory modeling. By comparing our results with all other contributions to the Audiovisual Sub-Challenge task, we provide an overview over recent approaches towards audiovisual emotion recognition as well as over their strengths and weaknesses with respect to the modeling of the different emotional dimensions.

The audio feature extraction front-end applied in our study is based on our open-source toolkit openSMILE [28] which is able to extract large sets of prosodic, spectral, and voice quality low-level descriptors (LLD) combined with various statistical functionals in real-time. Linguistic features, including non-linguistic vocalizations such as *laughing*, *breathing*, and *sighing* are extracted with an ASR engine optimized for real-time emotional speech recognition. Our method to compute low-level facial movement features was inspired by [29] and requires only one monocular camera. The computation time per frame is about 50 ms, i.e., almost real-time.

We evaluate our audiovisual LSTM technique on both, the development set and the official test set of the Audiovisual Sub-Challenge. This allows us to compare our results to various other methods proposed for this task so far, including Support Vector Machines [6,30], extreme learning machine based feedforward neural networks (ELM-NN) [31], AdaBoost [32], Latent-Dynamic Conditional Random Fields (LDCRF) [33], Gaussian Mixture Models (GMM) [34], and a combined system consisting of Multilayer Perceptrons (MLP) and HMMs [35].

The article is structured as follows: Section 2 provides an overview of the SEMAINE database and the challenge task, Section 3 details our methods for acoustic, linguistic, and visual feature extraction, Section 4 reviews the principle of Long Short-Term Memory, and Section 5 contains our experimental results.

2. The SEMAINE database

The freely available audiovisual SEMAINE corpus[1] [14] was recorded to study natural social signals that occur in conversations between humans and artificially intelligent agents. It has been used as training material for the development of the SEMAINE system [4] – an emotionally sensitive multimodal conversational agent.

The scenario used during the creation of the database is called the Sensitive Artificial Listener (SAL). It involves a user interacting with emotionally stereotyped characters whose responses are stock phrases keyed to the user's emotional state rather than the content of what he/she says. For the recordings, the participants are asked to talk in turn to four characters. These characters are Prudence, who is sensible; Poppy, who is happy; Spike, who is angry; and Obadiah, who is sad and depressive.

The data used for the 2011 Audio/Visual Emotion Challenge[2] is based on the 'Solid-SAL' part of the SEMAINE database, i.e., the users do not speak with artificial agents but instead with human operators who pretend to be the agents (Wizard-of-Oz setting). Further details on the interaction scenario can be found in [6].

Video was recorded at 49.979 frames per second at a spatial resolution of 780×580 pixels and 8 bits per sample, while audio was recorded at 48 kHz with 24 bits per sample. Both, the user and the operator were recorded from a frontal view by both a greyscale camera and a color camera. In addition, the user is recorded by a greyscale camera positioned on one side of the user to capture a profile view of the whole scene, including their face and body. Audio and video signals were synchronized with an accuracy of 25 μs.

The 24 recordings considered in the Audio/Visual Emotion Challenge consisted of three to four character conversation sessions each and were split into three speaker independent partitions: a training, development, and test partition each consisting of eight recordings. As the number of character conversations varies between recordings, the number of sessions is different per set: The training partition contains 31 sessions, while the development and test partitions contain 32 sessions. Table 1 shows the distribution of data in sessions, video frames, and words for each partition.

In our experiments we exclusively focus on the *Audiovisual Sub-Challenge* of the emotion challenge. Thus, our test set consists only of the sessions that are intended for this sub-challenge, meaning only 10 out of the 32 test sessions.

For the challenge, the originally continuous affective dimensions *arousal*, *expectation*, *power*, and *valence* were redefined as binary classification tasks by testing at every frame whether they are above or below average. As argued in [36], these four dimensions account for most of

[1] www.semaine-db.eu.
[2] www.avec2011-db.sspnet.eu.