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.

Table 1
Overview of the SEMAINE database as used for the 2011 Audio/Visual Emotion Challenge [6].

	Train	Develop	Test	Total
# Sessions	31	32	32	95
# Frames	501 277	449 074	407 772	1 358 123
# Words	20 183	16 311	13 856	50 350
Avg. word duration [ms]	262	276	249	263

the distinctions between everyday emotion categories. Arousal is the individual's global feeling of dynamism or lethargy and subsumes mental activity as well as physical, preparedness to act as well as overt activity. Expectation also subsumes various concepts that can be separated as expecting, anticipating, being taken unaware. Power subsumes two related concepts, power and control. Valence subsumes whether the person rated feels positive or negative about the things, people, or situations at the focus of his/her emotional state. Fig. 1 shows example screenshots for low and high arousal, expectation, power, and valence. In Fig. 2, a series of word-level screenshots of a user and the corresponding valence annotation can be seen. A detailed description on the annotation process can be found in [6].

The word timings were obtained by running an HMM-based speech recogniser in forced alignment mode on the manual transcripts of the interactions. The recognizer used tied-state cross-word triphone left-right (linear) HMM models with 3 emitting states and 16 Gaussian mixture components per state.

3. Feature extraction

3.1. Audio feature extraction

Our acoustic feature extraction approach is based on a large set of low-level descriptors and derivatives of LLD combined with suited statistical functionals to capture speech dynamics within a word. All features and functionals are computed using our on-line audio analysis toolkit openSMILE [28]. The audio feature set identical to the 2011 Audio/Visual Emotion Challenge baseline acoustic feature set applied in [6] and consists of 1941 features, composed of 25 energy and spectral related low-level descriptors × 42 functionals, 6 voicing related LLD × 32 functionals, 25 delta coefficients of the energy/spectral LLD × 23 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced/unvoiced durational features. Details on the LLD and functionals are given in Tables 2 and 3, respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition. The functional set has been based on similar sets, such as the one used for the Interspeech 2011 Speaker State Challenge [37], but has been manually reduced to avoid LLD/functional combinations that produce values which are constant, contain very little information and/or a high amount of noise. One example for a LLD/functional combination that contains no information is 'minimum pitch' which is always zero.

3.2. Linguistic and non-linguistic feature extraction

Linguistic features are extracted using the SEMAINE 3.0 ASR system [4]. It applies openSMILE as front-end to extract 13 Mel-Frequency Cepstral Coefficients (MFCC) together with first and second order temporal derivatives every 10 ms (window size 25 ms). The HMM back-end is based on the open-source Julius decoder [38]. Both, a back-off bigram language model and tied-state triphone acoustic models were trained on the COSINE corpus [39], the SAL database [40], and the training set of the SEMAINE database [14]. All of these corpora contain spontaneous, conversational, and partly emotional speech. The phoneme HMMs consist of three states with 16 Gaussian mixtures per state. Models for non-linguistic vocalizations (laughing, breathing and sighing) consist of nine emitting states.

Typically, one (key)word is detected for every audio chunk (which correspond to single words), however the recognizer is not restricted to detect exactly one word, thus insertions and deletions are possible.

Fig. 1. Examples for low and high arousal, expectation, power, and valence.

Fig. 2. Series of word-level screenshots of a user together with the corresponding valence annotation.

From the detected sequence of words a bag-of-words vector is computed. The general procedure is as follows:

- a word list (also including non-linguistic vocalizations) is built from all the recognized words in the training and development set,
- words that occur less than 10 times in the union of training and development set are removed from the word list,
- the dimensionality of the bag-of-words vector equals the size of the remaining word list (141 words),
- for the current chunk a bag-of-words vector is built by setting each element corresponding to a detected word to the word confidence score; all other elements in the vector are set to zero; if the recognizer output for one word is empty, all elements of the vector are set to zero.

3.3. Visual feature extraction

Generally, a large variety of purely visual emotion recognition systems has been presented in recent years, including combinations of Local Binary Patterns and Support Vector Machines [41], methods based on deformed grids and SVMs [42], Haar-like features modeled via AdaBoost [43], approaches using Gabor filters and non-negative matrix factorization [44], and variable-intensity models [45]. An overview of different visual emotion recognition techniques is given in [46]. Further audiovisual approaches have been presented in [47–51].

For the Video Sub-Challenge of the 2011 Audio/Visual Emotion Challenge, a variety of purely vision-based emotion recognition approaches were presented. In Table 4, we give a brief overview over these methods. Cruz et al. [52] proposed an approach that registers the face with an avatar and subsequently computes Local Phase Quantization (LPQ) features. Ramirez et al. [33] extract high-level features such as eye gaze direction, smile intensity, and head tilt. Glodek et al. [35] use Gabor filters to extract video features. Similarly, Dahmane et al. [53] employ Gabor filter energy to extract visual features.

In this study, we focus on the extraction of facial movement features as input for context-sensitive LSTM-based emotion classification from video. The face is detected with a Viola–Jones detector [54] and tracked with a camshift tracker that is based on a probability image built from a color histogram of the facial pixels. Subsequently, the face is cut out and rotated so that it is upright, before the optical flow with respect to the previous frame is computed. The optical flow field is subdivided into 49 subregions and the average in x- and y-direction is computed. Compared to [29], our method is faster and also extracts head tilt in addition to facial movement features. Furthermore, unlike the Audio/Visual Emotion Challenge baseline video feature extractor [6] which is based on dense local appearance descriptors, our approach does not rely on correct eye detection. Note that the video-based methods presented as Video Sub-Challenge contributions (see Table 4 [52,33,35,53]) do not compensate for head tilt. In our approach, this problem is addressed with ellipse fitting and subsequent tilt rectification.

3.3.1. Baseline video feature extractor

The baseline video feature extractor for the 2011 Audio/Visual Emotion Challenge [6] works as follows: First, the face position is detected by a Viola Jones face detector which computes a squared window containing the face. To refine the detected face region, the location of the right eye (p_r^x, p_r^y) and the position of the left eye (p_l^x, p_l^y) are detected. To this aim, a Haar-cascade object detector is applied. Once the two eyes are detected, the image can be rotated by angle α so that the eyes lie on a horizontal line. The image is scaled so that the distance between the eyes is exactly 100 pixels. Subsequently, a squared face region of 200×200 pixels is cropped out so that the middle of the right eye is at $(p_r^x, p_r^y) = (80, 60)$.

Uniform Local Binary Patterns (LBP) [55] are used as dense local appearance descriptors. Consisting of eight binary comparisons per

Table 2
31 acoustic low-level descriptors (LLD).

Energy & spectral (25)
Loudness (auditory model based), zero crossing rate,
Energy in bands from 250–650 Hz, 1 kHz–4 kHz,
25%, 50%, 75%, and 90% spectral roll-off points,
Spectral flux, entropy, spectral variance, skewness, kurtosis,
Psychoacousitc sharpness, harmonicity, MFCC 1–10
Voicing related (6)
F_0 (Sub-harmonic summation (SHS) followed by Viterbi smoothing),
Probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"),
Logarithmic Harmonics-to-Noise Ratio (logHNR)

Table 3
Set of all 42 functionals.

Statistical functionals (23)
(Positive[b]) arithmetic mean, root quadratic mean, standard deviation, flatness,
Skewness, kurtosis, quartiles, and inter-quartile ranges,
1%, 99% percentile, percentile range 1%–99%
Percentage of frames contour is above: min + 25%, 50%, and 90% of the range
Percentage of frames contour is rising,
Max, mean, min segment length[c], standard deviation of segment length[c]
Regression functionals[a] (4)
Linear regression slope, and approximation error (linear),
Quadratic regression coefficient a, and approx. error (linear)
Local minima/maxima related functionals[a] (9)
Mean and standard deviation of rising and falling slopes (minimum to maximum),
Mean and standard deviation of inter maxima distances
Amplitude mean of maxima
Amplitude mean of minima
Amplitude range of maxima
Other[a, c] (6)
Linear prediction gain, linear prediction coefficients 1–5

[a] Not applied to delta coefficient contours.
[b] For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied.
[c] Not applied to voicing related LLD.

Table 4
Approaches proposed for the Video Sub-Challenge of the 2011 Audio/Visual Emotion Challenge.

Reference	Method
[52]	Viola–Jones + LPQ
[33]	Omron OKAO Vision software (high-level features)
[35]	Gabor filters
[53]	Gabor filter energy
Our approach	Viola–Jones + segmented optical flow and head tilt

pixel, they are fast to compute. By employing uniform LBPs instead of full LBPs and aggregating the LBP operator responses in histograms taken over regions of the face, the dimensionality of the features is rather low (59 dimensions per image block). The registered face region is divided into 10×10 blocks, resulting in 5900 features. The baseline method also uses head tilt α and the distance between the eyes in the original image $d = ||p_r\ p_l||^2$. Here, the variables P_r and P_l denote the position of the right and left eyes.

3.3.2. Proposed visual feature extraction method

In order to compute the visual low-level features applied in our proposed LSTM-based audiovisual emotion recognition framework we go through the steps depicted in the block diagram in Fig. 3. Note that we only use data from the frontal view color camera. In Block 1 the face is detected by a Viola Jones face detector [54] (currently one of the best face detectors [56]). From the detected face a histogram is built for tracking (Block 2 in Fig. 3). The face that has been detected in the first frame is cut out and transformed into the hue-saturation-value (HSV) color space and the entries of the histogram M are computed:

$$M(h,s,v) = \sum_{x,y} \begin{cases} 1 & \text{if } T_H(x,y) = h \cap T_S(x,y) = s \\ & \quad \cap T_V(x,y) = v \\ 0 & \text{else} \end{cases} \quad (1)$$

where T is the detected face region that is taken as template. The indices h, s, and v denote hue, saturation, and value, respectively. Each of the three components of the HSV color model has 20 bins in the histogram. For each pixel $I(x,y)$ in the current image the probability of a facial pixel can be approximated by

$$P_f(x,y) = \frac{M(I_H(x,y), I_S(x,y), I_V(x,y))}{N}, \quad (2)$$

with N being the number of template pixels that have been used to create the histogram. The face is considered as detected when there is a sufficiently large amount of facial pixels in the upper half of the image. Subsequently, the face is tracked with a camshift tracker [57] which takes the probability image as input. The location, the size, and the orientation of the face are computed according to [57]. One advantage of the camshift tracker is that it is comparatively robust which is important for a reliable facial movement feature extraction. Furthermore, it operates fast and also computes the tilt of the head, as can be seen in Fig. 3.

Subsequently, the face is cut out and the tilt is undone (Block 3). The face in the up-right pose is compared to the previous frame. Note that we use the tilt θ itself as one facial low-level feature. In Block 4, 98 facial movement features are extracted as follows. The optical flow between the rectified face and the face of the previous frame is computed. As an example, Fig. 4 depicts a subject that opens its mouth. In this case the y-values of the rectangles of the lip region are high.

The cut out face is then subdivided into $7 \times 7 = 49$ rectangles. For each of these rectangles the average movement in x- and y-direction is computed. These movements are further features in addition to the tilt θ, so that we extract a total of 99 visual low-level features per frame. The computation of the low-level features takes 50 ms per frame for a C++ implementation on a 2.4 GHz Intel i5 processor with 4 GB RAM.

In order to map the sequence of frame-based video features to a single vector describing the word-unit, statistical functionals are applied to the frame-based video features and their first order delta coefficients. This step is conceptually the same as for the audio features, except that different functionals are used, considering the different properties of the video features. Note that words shorter than 250 ms are expanded to 250 ms which means that the time windows containing very short words can contain (fractions of) other words. The following functionals are applied to frame-based video features: arithmetic mean (for delta coefficients: arithmetic mean of absolute values), standard deviation, 5% percentile, 95% percentile, and range of 5% and 95% percentile. Fewer functionals as for audio features are used to ensure a similar dimensionality of the video feature vector and the audio feature vector. The resulting per-word video feature vector has $5 \times 2 \times 99 = 990$ features.

Fig. 5 shows the importance of the subregions of the face for the video-based discrimination between high and low arousal, expectation, power, and valence. Importance was evaluated employing the ranking-based information gain attribute evaluation algorithm implemented in the Weka toolkit [58]. As input for the ranking algorithm, we used all 990 features extracted from each instance in the training set together with the ground truth annotation of the respective emotional dimension. In Fig. 5, the shading of the facial regions indicates the importance of the features corresponding to the respective region. As expected, the small remaining background parts are less important than the subregions containing facial information. Within the face, the eye regions contain slightly more information. Overall, we

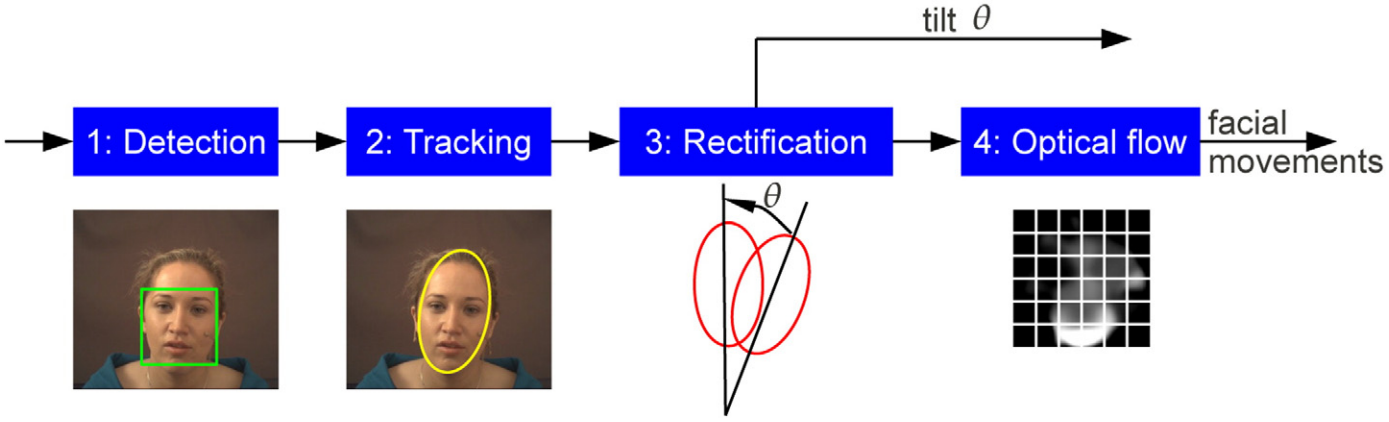

Fig. 3. Basic steps for the computation of the low-level visual features.

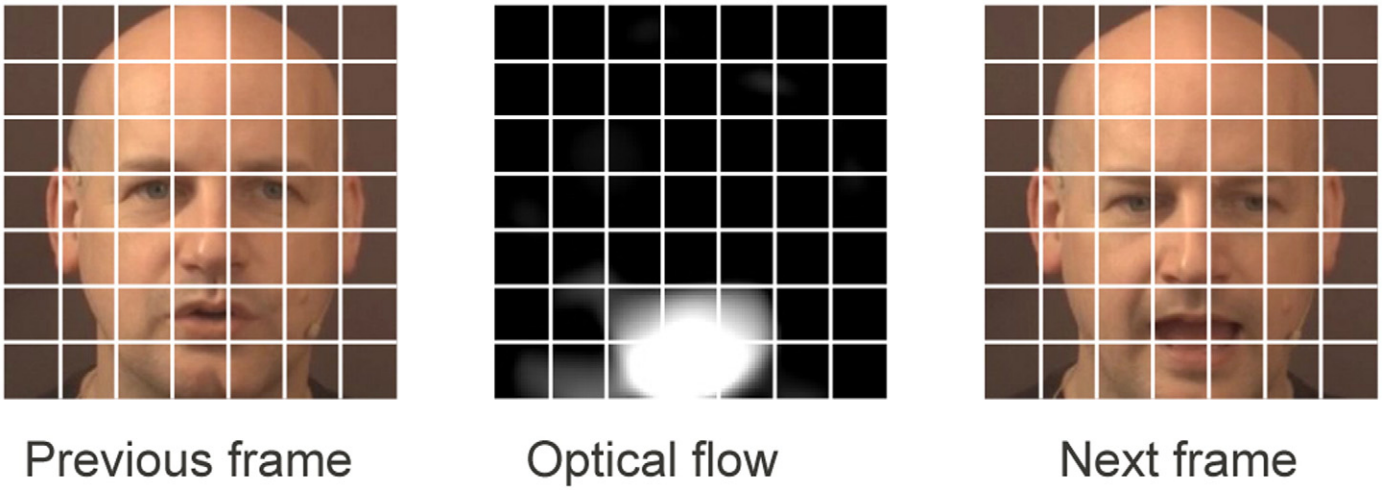

Fig. 4. Example for optical flow computation: between the frames there is a substantial change in the mouth region.

observe that relevant information about a subject's emotional state can be found in multiple regions of the face and not just in the upper or lower face, corresponding to the eye and mouth region, respectively.

4. Classification

Widely used classifiers operating on static word- or turn-level feature vectors are, e.g., Support Vector Machines or Multilayer Perceptrons. Context-dependent classification frameworks are mostly based on Hidden Markov Models, as for example the approach proposed by the winners of the Audio Sub-Challenge of the 2011 Audio/Visual Emotion Challenge [59]. To exploit context between successive speech segments for improved audiovisual emotion recognition, this study considers *recurrent* neural network architectures which take into account past observations by cyclic connections in the network's hidden layer. For off-line sequence labeling problems, also *future* context can be modeled via *bidirectional* RNNs (BRNN). Bidirectional networks have access to both, past and future observations by applying two hidden layers, one for forward processing and one for backward processing. These two hidden layers are connected to the same output layer (see [60] for details). For emotion recognition BRNNs can be employed whenever the real-time constraint can be relaxed, i.e., when focusing on off-line processing or when a short latency is tolerable, so that the system can be operated with a look-ahead buffer.

In our experiments we investigate a more advanced technique for neural network based context modeling. It is based on the Long Short-Term Memory principle originally introduced in [18] and improved in [61]. LSTM networks use so-called memory blocks instead of conventional hidden cells which allows them to access and model a self-learned amount of long-range temporal context. Each memory block consists of one or more memory cells and multiplicative input, output, and forget gates. The cell input is scaled by the activation of the input gate, the output by the activation of the output gate, and the previous cell value by the activation of the forget gate. Thus, the network can perform read, write, and reset operations, and – unlike traditional RNNs which are affected by the *vanishing gradient problem* – has access to an arbitrary amount of context information. Fig. 6 shows the basic architecture of an LSTM memory block with one memory cell.

The initial version of the LSTM architecture proposed in [18] contained only input and output gates to enable an architecture that can store and access activations via gate activations. Later, in [61], the authors found that for long sequences it is beneficial to allow the memory cells to reset themselves whenever the network needs to *forget* past inputs. This led to the inclusion of the so-called *forget gates*. In our experiments we exclusively consider the enhanced LSTM version including forget gates. The overall effect of the gate units is that the LSTM memory cells can store and access information over long periods of time and thus avoid the vanishing gradient problem. For instance, as long as the input gate remains closed (corresponding to an input gate activation close to zero), the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate. Consequently, the number of memory blocks in the network

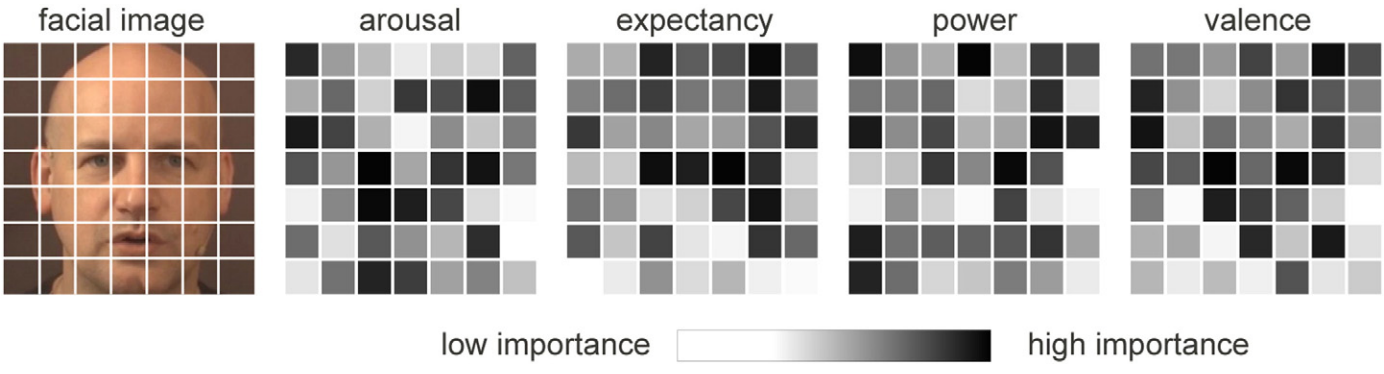

Fig. 5. Importance of facial regions for video feature extraction according to the ranking-based information gain attribute evaluation algorithm implemented in the Weka toolkit [58]. Information gain is evaluated for each emotional dimension. The shading of the facial regions indicates the importance of the features corresponding to the respective region.

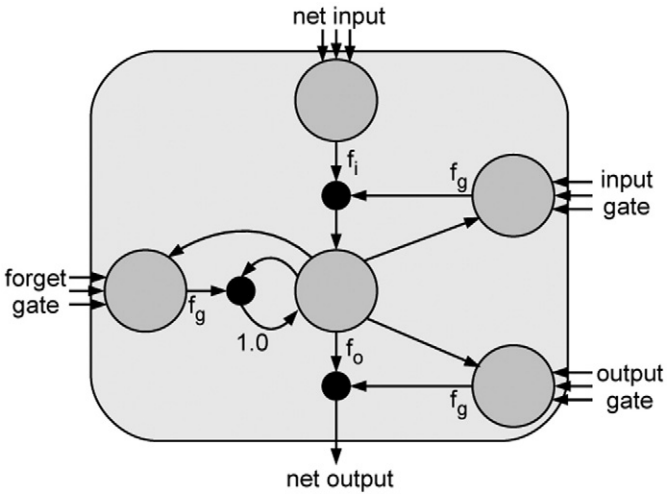

Fig. 6. LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively; f_i, f_g, and f_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state.

specifies how many different (weighted) inputs can be stored at each timestep to influence future output activations.

LSTM networks have shown remarkable performance in a variety of pattern recognition tasks, including phoneme classification [20], handwriting recognition [62], keyword spotting [63], affective computing [17], and driver distraction detection [64]. Further details on the LSTM technique and on its bidirectional extension (BLSTM) can be found in [62]. In [65], a methodology for determining the *amount* of temporal context information relevant for emotion recognition is proposed and evaluated.

Fig. 7 shows the overall system architecture of our LSTM-based audio-visual emotion recognition framework applying early (i.e., feature-level) fusion. The openSMILE audio feature extractor provides framewise MFCC features for the speech recognition module as well as statistical functionals of acoustic features for the LSTM network. In addition to audio features, the network also processes the linguistic feature vector provided by the ASR system and video features computed by the facial feature extractor to generate the current emotion prediction.

5. Experiments and results

5.1. Audio/visual emotion challenge

All experiments are carried out on the Audiovisual Sub-Challenge task as described in Section 2. To gain first insights concerning the optimal combination of modalities (i.e., acoustic, linguistic, and visual features) and the number of training epochs needed for LSTM network training, we performed initial experiments using the training set for network training and the development set for testing, before we focused on the actual challenge task which consists in training on the union of the training and the development set and testing on the test set. The task is to discriminate between high and low AROUSAL, EXPECTATION, POWER, and VALENCE. As the class distribution in the training set is relatively well balanced, the official challenge measure is weighted accuracy, i.e., the recognition rates of the individual classes weighted by the class distribution. However, since the instances of the development and test sets are partly unbalanced with respect to the class distributions, we also report unweighted accuracies (equivalent to unweighted average recall). This imbalance holds in particular for the Audio and Audio-Visual Sub-Challenge as they consider word-level modeling rather than frame-based recognition.

5.2. Experimental settings

We investigate the performance of both, bidirectional LSTMs and unidirectional LSTM networks for fully incremental on-line audiovisual affect recognition. Separate networks were trained for each emotional dimension. The following modality combinations were considered: acoustic features only, video features only, acoustic and linguistic features (including non-linguistic vocalizations), acoustic and video features, as well as acoustic, (non-)linguistic, and video features.

As in our previous studies on affective computing (e.g., [13]), all LSTM networks consist of 128 memory blocks. Each memory block contains one memory cell. The number of input nodes corresponds to the number of different features per speech segment and the number of output nodes corresponds to the number of target classes, i.e., we used two output nodes representing high and low AROUSAL, EXPECTATION, POWER, and VALENCE, respectively. A common method to improve generalization and to prevent over-fitting of neural networks to the training data is to add a small amount of noise to the inputs at each training epoch. In our experiments, we added zero mean Gaussian noise with standard deviation 0.6 to the inputs during training. All networks were trained using a learning rate of 10^{-5}. As for standard feedforward neural networks, the learning rate for LSTM network training defines how 'aggressively' the network weights are updated in the direction of the negative error gradient during application of the gradient decent algorithm. The bidirectional networks consist of two hidden layers (one for forward and one for backward processing) with 128 memory blocks per input direction. Parameters such as learning rate and the number of memory blocks were configured according to our experience with similar recognition tasks [17,13]. To validate whether better recognition performance can be obtained when changing the number of memory blocks, we evaluated hidden layer sizes of between 80 and 160 memory blocks on the development set. Yet, for none of the modality combinations a modified

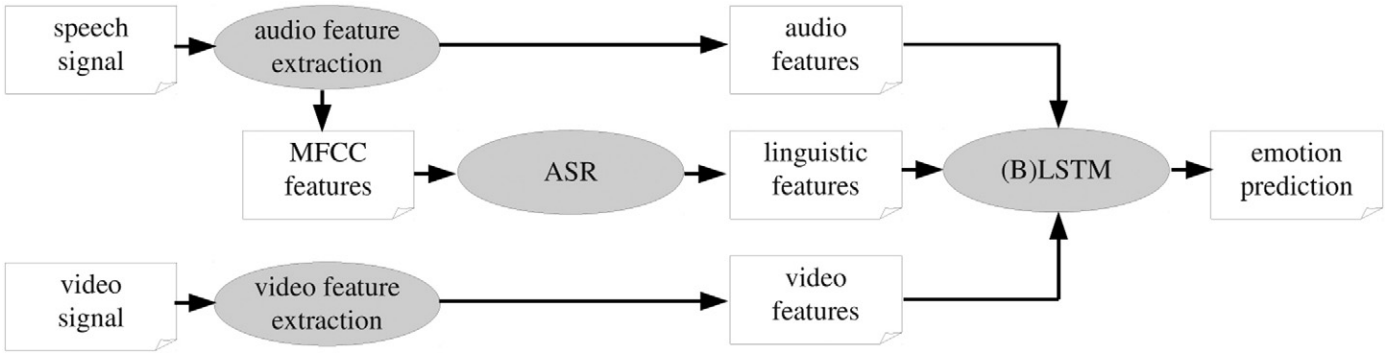

Fig. 7. System architecture for early fusion of acoustic, linguistic, and video features.

hidden layer size could significantly outperform networks using the default setting of 128 memory blocks. The resulting number of variables that need to be estimated during network training is equivalent to the number of weights in the network, e.g., an LSTM network that processes the full feature set consisting of acoustic, linguistic, and video information has 2 094 210 weights.

As abort criterion for training we periodically evaluated the classification performance on the development set and used the network which achieved the best results on the development set. The number of training epochs needed until the best performance was reached was around 30 epochs for recognition of EXPECTATION, POWER, and VALENCE, and 60 epochs for AROUSAL classification. All input features were mean- and variance normalized with means and variances computed from the training set.

Alternatively to early fusion of modalities on the feature level, we also consider a simple late fusion (LF) technique which consists in training separate networks for each modality and summing up the output activations of the respective networks before deciding about the estimated class that can be inferred from the highest (overall) output activation.

5.3. Results and discussion

Table 5 shows both, weighted accuracies (WA) and unweighted accuracies (UA) obtained when training on the training set of the 2011 Audio/Visual Emotion Challenge and testing on the development set. Results are shown for BLSTMs, LSTM networks, and for the SVM approach applied in [6]. We consider various modality combinations as mentioned in Section 2. Note that the results for SVMs processing audio and video data are missing as they are not reported in [6]. For the Audiovisual Sub-Challenge, the development set data has been used in [6] only to train the fusion engine – this, however, is not necessary in our experiments since we focus on early fusion and on a simple late fusion scheme that does not require training.

The performance difference between unidirectional and bidirectional LSTM networks is comparatively small. In some cases (e.g., classification of AROUSAL using acoustic and linguistic features), LSTM networks perform even slightly, but not significantly better than BLSTM nets. This means that modeling only past context does not necessarily downgrade recognition results compared to bidirectional modeling, which is important for incremental on-line applications in which future context is not available due to real-time constraints. The performance of the different feature groups (acoustic, linguistic, video) heavily depends on the considered emotional dimension. For AROUSAL, the best WA of 68.5% is obtained for acoustic features only, which is in line with previous studies showing that audio is the most important modality for assessing AROUSAL [13]. However, the classification of EXPECTATION seems to benefit from including visual information as the best WA (67.6%) is reached for LSTM networks applying late fusion of audio and video modalities. Similar to AROUSAL, POWER is best classified via speech-based features. Bidirectional networks for classifying extscpower cannot be enhanced by linguistic features, however, for unidirectional modeling WA significantly increases from 64.7% to 66.2% when using linguistics in addition to audio features. For VALENCE, the inclusion of video information helps, leading to a WA of 69.8% when using BLSTM networks and audiovisual data. The effectiveness of the emotion recognition approaches using only video information also depends on the emotional dimension. For AROUSAL and EXPECTATION, BLSTM modeling of facial movement features prevails, while for POWER and VALENCE, we observe slightly, but not significantly better results for SVM-based classification of local appearance descriptors as proposed in [6] and for unidirectional LSTM modeling. On average the best performance on the development set is obtained for bidirectional processing and acoustic and visual features (mean WA of 66.7%). Yet, in this case there is no significant difference between bi- and unidirectional processing, as LSTM networks achieve almost the same WA on average (66.5%). For each emotional dimension, context modeling via LSTM increases accuracies compared to the static SVM-based technique applied in [6]. Furthermore, late fusion tends to prevail over early fusion.

To investigate whether a smaller feature space leads to better recognition performance, we repeated all evaluations on the development set applying a Correlation based Feature Subset Selection (CFS) [66] for each modality combination. The corresponding results can be seen in Table 6. For most settings, CFS does not significantly improve the average weighted accuracy. However, for recognition based on video only, CFS leads to a remarkable performance gain, increasing the average WA from 60.4% to 65.8% for unidirectional LSTM networks.

The results for the official Audiovisual Sub-Challenge test set can be seen in Table 7. Networks were trained on the training and development set. According to optimizations on the development set, the number of training epochs was 60 for networks classifying AROUSAL and 30 for all other networks. Networks processing video data only are based on a video feature set reduced via CFS, whereas for all other networks, we did not apply CFS. All network parameters (number of memory blocks, learning rate, etc.) were identical to the previous set of experiments on the development set. We compare BLSTM and LSTM modeling to all

Table 5
Development set of the Audiovisual Sub-Challenge; no feature selection: weighted accuracies (WA) and unweighted accuracies (UA) for the discrimination of high and low AROUSAL, EXPECTATION, POWER, and tscvalence using acoustic (A), linguistic (L), and video (V) features combined with different classifiers. LF: late fusion; the best weighted accuracies for each emotional dimension are highlighted.

		AROUSAL		EXPECTATION		POWER		VALENCE		Mean
Classifier	Features	WA	UA	WA	UA	WA	UA	WA	UA	WA
BLSTM	A	**68.5**	69.3	64.3	53.5	66.1	53.3	66.3	56.1	66.3
BLSTM	A+L	67.8	69.0	64.8	52.0	65.5	53.9	66.3	56.2	66.1
LSTM	A	**68.5**	68.6	66.1	55.9	64.7	56.1	65.6	55.2	66.2
LSTM	A+L	68.2	68.8	65.2	51.9	**66.2**	55.0	63.8	55.9	65.9
SVM [6]	A	63.7	64.0	63.2	52.7	65.6	55.8	58.1	52.9	62.7
BLSTM	V	62.3	62.9	62.3	51.8	55.2	53.0	63.3	60.5	60.8
LSTM	V	60.3	61.3	60.4	57.7	57.0	50.4	64.0	57.9	60.4
SVM [6]	V	60.2	57.9	58.3	56.7	56.0	52.8	63.6	60.9	59.5
BLSTM	A+V	67.7	68.0	63.1	53.4	60.6	55.0	67.2	61.8	64.7
BLSTM	A+L+V	66.9	67.0	66.2	57.3	63.4	52.3	65.9	61.5	65.6
LSTM	A+V	68.0	67.5	65.7	57.7	63.8	54.7	65.5	59.5	65.8
LSTM	A+L+V	67.4	66.8	65.3	56.7	61.7	54.2	67.6	62.8	65.5
BLSTM (LF)	A+V	67.9	69.3	65.0	53.2	64.0	55.5	**69.8**	61.3	**66.7**
BLSTM (LF)	A+L+V	67.0	68.6	65.7	51.6	63.6	55.7	**69.8**	61.2	66.5
LSTM (LF)	A+V	62.6	64.3	**67.6**	57.6	65.1	56.0	68.2	57.7	65.9
LSTM (LF)	A+L+V	66.3	67.4	63.9	58.1	66.0	53.9	66.4	58.2	65.7

Table 6
Development set of the Audiovisual Sub-Challenge; CFS feature selection: weighted accuracies (WA) and unweighted accuracies (UA) for the discrimination of high and low AROUSAL, EXPECTATION, POWER, and xtscvalence using acoustic (A), linguistic (L), and video (V) features combined with different classifiers. LF: late fusion; the best weighted accuracies for each emotional dimension are highlighted.

		AROUSAL		EXPECTATION		POWER		VALENCE		Mean
Classifier	Features	WA	UA	WA	UA	WA	UA	WA	UA	WA
BLSTM	A	71.3	70.2	66.2	51.0	66.0	56.4	65.9	60.6	67.4
BLSTM	A+L	**73.7**	74.4	66.1	53.1	64.6	55.7	65.8	57.2	67.6
LSTM	A	70.4	69.8	**67.7**	54.6	64.9	58.8	63.1	55.3	66.5
LSTM	A+L	71.9	71.1	63.1	55.5	66.6	56.3	64.7	56.9	66.6
BLSTM	V	59.8	58.8	66.2	50.1	64.1	57.5	63.3	56.0	63.4
LSTM	V	62.7	61.5	66.0	50.1	**70.2**	62.4	64.3	52.7	65.8
BLSTM	A+V	67.8	69.5	64.3	52.3	60.1	57.0	64.7	58.8	64.2
BLSTM	A+L+V	69.9	70.7	63.3	50.4	61.9	56.1	61.4	55.9	64.1
LSTM	A+V	69.7	70.8	64.5	52.0	63.5	56.8	62.4	53.0	65.0
LSTM	A+L+V	70.4	71.3	65.7	53.3	63.5	55.9	62.9	53.2	65.6
BLSTM (LF)	A+V	68.5	67.5	66.7	50.4	64.2	52.7	**69.1**	60.6	67.1
BLSTM (LF)	A+L+V	72.3	72.3	66.6	50.9	64.4	54.0	67.9	58.5	**67.8**
LSTM (LF)	A+V	65.7	63.7	67.4	52.1	68.0	58.6	66.8	54.8	67.0
LSTM (LF)	A+L+V	64.8	63.5	67.1	54.9	68.1	57.3	65.7	56.4	66.4

Table 7
Test set of the Audiovisual Sub-Challenge: weighted accuracies (WA) and unweighted accuracies (UA) for the discrimination of high and low AROUSAL, EXPECTATION, POWER, and VALENCE using acoustic (A), linguistic (L), and video (V) features combined with different classifiers. LF: late fusion; the best weighted accuracies for each emotional dimension are highlighted.

Classifier	Features	AROUSAL		EXPECTATION		POWER		VALENCE		Mean
		WA	UA	WA	UA	WA	UA	WA	UA	WA
BLSTM	A	69.2	69.1	63.1	54.6	59.6	52.9	68.7	57.4	**65.2**
LSTM	A	**71.2**	71.2	57.6	48.7	57.4	50.4	68.7	59.5	63.7
SVM [30]	A	59.8	59.7	63.6	50.0	57.9	48.4	**70.2**	54.9	62.9
ELM-NN [31]	A	52.0	52.3	63.7	50.1	62.2	50.7	69.1	50.0	61.8
AdaBoost [32]	A	57.6	57.5	62.2	49.6	54.2	47.9	60.3	47.6	58.6
LDCRF [33]	A	60.9	60.4	53.2	44.1	56.8	45.7	60.9	45.8	57.9
GMM [34]	A	55.3	55.2	56.1	50.7	49.1	45.3	50.9	48.4	52.9
BLSTM	V	43.1	42.9	**68.6**	62.0	44.8	41.0	51.7	52.4	52.1
LSTM	V	48.6	48.7	65.6	60.2	37.6	35.8	60.8	52.2	53.1
SVM [52]	V	47.8	47.4	62.0	54.8	57.9	47.4	69.6	50.2	59.3
LDCRF [33]	V	53.2	53.1	46.8	43.2	57.3	50.5	59.3	50.7	54.1
BLSTM	A + V	58.3	58.1	64.1	59.5	46.9	45.4	51.1	45.4	55.1
BLSTM	A + L + V	58.8	58.6	60.8	54.8	46.9	44.0	57.1	50.2	55.9
LSTM	A + V	56.3	56.2	61.6	54.1	46.7	45.8	61.2	53.9	56.5
LSTM	A + L + V	57.9	57.8	64.0	58.6	47.6	44.8	55.7	47.9	56.3
SVM [6]	A + V	67.2	67.2	36.3	48.5	62.2	50.0	66.0	49.2	57.9
LDCRF [33]	A + V	65.6	65.3	53.4	49.2	**62.9**	58.3	59.5	49.6	60.3
MLP [35]	A + V	54.1	54.3	58.5	57.8	42.7	40.0	44.8	35.9	50.0
BLSTM (LF)	A + V	69.5	69.4	63.6	54.5	55.8	49.3	69.6	59.2	64.6
BLSTM (LF)	A + L + V	63.3	63.2	62.9	53.0	53.1	48.4	57.6	46.9	59.2
LSTM (LF)	A + V	67.8	67.8	58.3	48.9	57.5	50.3	68.7	59.3	63.1
LSTM (LF)	A + L + V	70.3	70.3	62.5	52.2	56.0	50.4	69.2	58.5	64.5

other approaches proposed for the Audiovisual Sub-Challenge, including Support Vector Machines [6,30], extreme learning machine based feedforward neural networks [31], AdaBoost [32], Latent-Dynamic Conditional Random Fields [33], Gaussian Mixture Models [34], and a combined system consisting of MLPs and HMMs [35]. Note, however, that these classification techniques do not necessarily use the same set of audio (and video) features, thus, Table 7 compares the overall approaches of different research groups rather than the effectiveness of the different classifiers. Similar to our experiments on the development set, audio features lead to the best result for AROUSAL classification. When applying LSTM modeling we reach a WA of 71.2% which is the best result reported for this task so far. Also for BLSTM-based classification of EXPECTATION using facial movement features, the obtained WA of 68.6% is higher than what is reported for other techniques. For POWER, we were not able to outperform audiovisual classification with Latent-Dynamic Conditional Random Fields as proposed in [33]. For VALENCE, the audio features used in [30] lead to the highest accuracy (70.2%). When computing the average WA, we find that a remarkable average performance can be obtained for systems exclusively processing audio data (for an overview over the statistical significance of the performance difference between the audio-based approaches, see Table 8). This suggests that even though video information helps for some emotional dimensions (such as EXPECTATION), on average acoustic features contribute the most to the assessment of affective states in the SEMAINE scenario. Interestingly, in our evaluations on the test set, the performance gap between early and late fusion of modalities via LSTM networks is significantly more pronounced than in our initial experiments on the development set. The average WA values we obtain for BLSTMs (65.2%) and LSTMs (63.7%) processing acoustic features prevail over all other approaches applied for this task by the challenge participants. Thus, we can conclude that the LSTM architecture is well suited for modeling affect in human conversations and that the exploitation of long-range temporal context not only helps humans to judge a conversational partner's emotional state but also increases the accuracy of automatic affect sensing in human–computer interaction.

Table 8
Statistical significance of the average performance difference between the audio-based classification approaches denoted in the column and the approaches in the table header (evaluations on test set of the Audiovisual Sub-Challenge); '−': not significant; 'o' significant at 0.1 level; '+': significant at 0.05 level; '++': significant at 0.001 level. Significance levels are computed according to the z-test described in [67].

	LSTM	SVM	ELM-NN	AdaBoost	LDCRF	GMM [34]
BLSTM	o	+	++	++	++	++
LSTM		−	+	++	++	++
SVM [30]			−	++	++	++
ELM-NN [31]				+	++	++
AdaBoost [32]					−	++
LDCRF [33]						++

6. Conclusion and future work

In this article, we proposed an automatic emotion recognition framework exploiting acoustic, linguistic, and visual information in affective interactions. We aimed to improve recognition performance by modeling the temporal continuity of human affect via a suited machine learning technique. As previous studies report excellent results for speech-based emotion recognition using Long Short-Term Memory neural networks [17], we built a system based on LSTM long-range temporal context modeling in order to discriminate between high and low levels of AROUSAL, EXPECTATION, POWER, and VALENCE using statistical functionals of a large set of acoustic low-level descriptors, linguistic information (including non-linguistic vocalizations), and facial movement features. To get an impression of the effectiveness of context-sensitive LSTM-based audiovisual emotion recognition compared to other recently published approaches, we train and evaluate our system on data sets defined in the 2011 Audio/Visual Emotion Challenge and strictly adhere to the challenge conditions. For the emotional dimensions AROUSAL and EXPECTATION, our framework leads to the best accuracies reported so far (71.2% and 68.6%, respectively). Averaged over all four emotional dimensions, we obtain a (weighted) accuracy of 65.2% via bidirectional LSTM modeling of acoustic features, which is higher than all other average accuracies reported for this task in literature up to now. The absolute values of the reported accuracies seem low in comparison to easier scenarios, such as the discrimination of acted, prototypical emotions. However, the considered scenario reflects realistic conditions in natural interactions and thus highlights the need for further research in the area of affective computing in order to get closer to the human performance in judging emotions.

Our future research in the area of video feature extraction will include the application of multi-camera input to be more robust to head rotations. We plan to combine the facial movements of the 2D camera sequences to predict 3D movement. For this purpose we intend to employ a deformable 3D model. Further, we want to evaluate advanced network architectures in the back-end, such as bottleneck networks [68]. Another possibility to increase recognition performance is to allow asynchronities between audio and video, e.g., by applying hybrid fusion techniques like asynchronous HMMs [69] or multi-dimensional dynamic time warping [48]. Finally, it would be interesting to fuse the results of all challenge participants to make use of the potentially complementary information generated by the individual techniques. To obtain the best possible recognition performance, future studies should also investigate which feature-classifier combinations lead to the best results, e.g., by combining the proposed LSTM framework with other audio or video features proposed for the 2011 Audio/Visual Emotion Challenge.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (1) (2001) 32–80.

[2] T. Ziemke, R. Lowe, On the role of emotion in embodied cognitive architectures: from organisms to robots, Cogn. Comput. 1 (1) (2009) 104–117.

[3] L. Devillers, C. Vaudable, C. Chastagnol, Real-life emotion-related states detection in call centers: a cross-corpora study, Proc. of Interspeech, Makuhari, Japan, 2010, pp. 2350–2353.

[4] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, M. Wöllmer, Building autonomous sensitive artificial listeners, IEEE Trans. Affective Comput. doi:10.1109/T–AFFC.2011.34.

[5] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, Being bored? Recognising natural interest by extensive audiovisual integration for real-life application, Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior, 27 (12, 2009, pp. 1760–1774.

[6] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, AVEC – the First International Audio/Visual Emotion Challenge, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 415–424.

[7] B. Schuller, M. Wimmer, L. Mösenlechner, D. Arsic, G. Rigoll, Brute-forcing Hierarchical Functionals for Paralinguistics: A Waste of Feature Space? Proc. of ICASSP, Las Vegas, NV, 2008, pp. 4501–4504.

[8] S. Steidl, B. Schuller, A. Batliner, D. Seppi, The Hinterland of Emotions: Facing the Open-microphone Challenge, Proc. of ACII, Amsterdam, The Netherlands, 2009, pp. 690–697.

[9] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey, Proc. of IEEE Conference on Face and Gesture Recognition, Santa Barbara, CA, USA, 2011, pp. 827–834.

[10] M. Grimm, K. Kroschel, S. Narayanan, Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech, Proc. of ICASSP, Honolulu, Hawaii, 2007, pp. 1085–1088.

[11] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie, On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum Using Acoustic and Linguistic Cues, Journal on Multimodal User Interfaces (JMUI), Special Issue on Real-time Affect Analysis and Interpretation: Closing the Loop in Virtual Agents, 3, 2009, pp. 7–19.

[12] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic Emotion Recognition: A Benchmark Comparison of Performances, Proc. of ASRU, Merano, Italy, 2009, pp. 552–557.

[13] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive Multimodal Emotion Recognition from Speech and Facial Expression Using Bidirectional LSTM Modeling, Proc. of Interspeech, Makuhari, Japan, 2010, pp. 2362–2365.

[14] G. McKeown, M.F. Valstar, M. Pantic, R. Cowie, The SEMAINE Corpus of Emotionally Coloured Character Interactions, Proc. of ICME, 2010, pp. 1–6.

[15] B. Schuller, G. Rigoll, M. Lang, Hidden Markov Model-based Speech Emotion Recognition, Proc. of ICASSP, Hong Kong, China, 2003, pp. 1–4.

[16] L.F. Barrett, E.A. Kensinger, Context is routinely encoded during emotion perception, Psychol. Sci. 21 (2010) 595–599.

[17] M. Wöllmer, B. Schuller, F. Eyben, G. Rigoll, Combining Long Short-Term Memory and Dynamic Bayesian Networks for incremental emotion-sensitive artificial listening, IEEE J. Sel. Top. Sign. Proces. 4 (5) (2010) 867–881.

[18] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (1997) 1735–1780.

[19] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, in: C. Kremer, J.F. Kolen (Eds.), A Field Guide to Dynamical Recurrent Neural Networks, IEEE Press, 2001, pp. 1–15.

[20] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (5–6) (2005) 602–610.

[21] S. Fernandez, A. Graves, J. Schmidhuber, An Application of Recurrent Neural Networks to Discriminative Keyword Spotting, Proc. of ICANN, Porto, Portugal, 2007, pp. 220–229.

[22] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, G. Rigoll, Robust Discriminative Keyword Spotting for Emotionally Colored Spontaneous Speech Using Bidirectional LSTM Networks, Proc. of ICASSP, Taipei, Taiwan, 2009, pp. 3949–3952.

[23] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, G. Rigoll, Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework, Cogn. Comput. 2 (3) (2010) 180–190.

[24] M. Wöllmer, F. Eyben, B. Schuller, G. Rigoll, A Multi-stream ASR Framework for BLSTM Modeling of Conversational Speech, Proc. of ICASSP, Prague, Czech Republic, 2011, pp. 4860–4863.

[25] M. Wöllmer, B. Schuller, G. Rigoll, A Novel Bottleneck-BLSTM Front-end for Feature-level Context Modeling in Conversational Speech Recognition, Proc. of ASRU, Waikoloa, Big Island, Hawaii, 2011, pp. 36–41.

[26] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning Emotion Classes – Towards Continuous Emotion Recognition with Modelling of Long-range Dependencies, Proc. of Interspeech, Brisbane, Australia, 2008, pp. 597–600.

[27] M.A. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, IEEE Trans. Affect. Comput. 2 (2011) 92–105.

[28] F. Eyben, M. Wöllmer, B. Schuller, openSMILE — The Munich Versatile and Fast Open-source Audio Feature Extractor, Proc. of ACM Multimedia, Firenze, Italy, 2010, pp. 1459–1462.

[29] D. Arsic, B. Hörnler, B. Schuller, G. Rigoll, A Hierarchical Approach for Visual Suspicious Behavior Detection in Aircrafts, Proceedings of the 16th international conference on Digital Signal Processing, 2009, pp. 639–645.

[30] A. Sayedelahl, P. Fewzee, M. Kamel, F. Karray, Audio-based Emotion Recognition from Natural Conversations Based on Co-Occurrence Matrix and Frequency Domain Energy Distribution Features, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 407–414.

[31] L. Cen, Z.L. Yu, M.H. Dong, Speech Emotion Recognition System based on L1 Regularized Linear Regression and Decision Fusion, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 332–340.

[32] S. Pan, J. Tao, Y. Li, The CASIA Audio Emotion Recognition Method for Audio/Visual Emotion Challenge 2011, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 388–395.

[33] G. Ramirez, T. Baltrusaitis, L.P. Morency, Modeling Latent Discriminative Dynamic of Multi-dimensional Affective Signals, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 396–406.

[34] J.C. Kim, H. Rao, M.A. Clements, Investigating the Use of Formant Based Features for Detection of Affective Dimensions in Speech, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 369–377.

[35] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, F. Schwenker, Multiple Classifier Systems for the Classification Of Audio-Visual Emotional States, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 359–368.

[36] J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, P. Ellsworth, The world of emotions is not two-dimensional, Psychol. Sci. 18 (2) (2007) 1050–1057.

[37] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, The Interspeech 2011 Speaker State Challenge, Proc. of Interspeech 2011, Florence, Italy, 2011, pp. 3201–3204.

[38] A. Lee, T. Kawahara, Recent Development of Open-source Speech Recognition Engine Julius, Proc. of APSIPA ASC, Sapporo, Japan, 2009, pp. 131–137.

[39] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, J. Bilmes, The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments, Comput. Speech Lang. 26 (1) (2011) 52–66.

[40] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis, The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data, Affective Computing and Intelligent Interaction, vol. 4738/2007, Springer, 2007, pp. 488–500.

[41] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vision Comput. 27 (6) (2009) 803–816.

[42] I. Kotsia, I. Pitas, Facial expression recognition in image sequences using geometric deformation features and support vector machines, IEEE Trans. Image Process. 16 (1) (2007) 172–187.

[43] P. Yang, Q. Liu, D.N. Metaxas, Boosting encoded dynamic features for facial expression recognition, Pattern Recognit. Lett. 30 (2) (2009) 132–139.

[44] I. Kotsia, I. Buciu, I. Pitas, An analysis of facial expression recognition under partial facial image occlusion, Image Vision Comput 26 (7) (2008) 1052–1067.

[45] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, Y. Sato, Pose-invariant facial expression recognition using variable-intensity templates, Int. J. Comput. Vis. 83 (2) (2009) 178–194.

[46] Y. Tian, T. Kanade, J.F. Cohn, Handbook of Face Recognition, Ch. Facial Expression Analysis, Springer, London, 2011, pp. 487–519.

[47] F. Eyben, M. Wöllmer, M. Valstar, H. Gunes, B. Schuller, M. Pantic, String-based audiovisual fusion of behavioural events for the assessment of dimensional affect, Proc. of FG, 2011, pp. 322–329.
[48] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, G. Rigoll, A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams, Neurocomputing 73 (1–3) (2009) 366–380.
[49] A. Metallinou, S. Lee, S. Narayanan, Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice, International Symposium on Multimedia, Los Alamitos, CA, USA, 2008, pp. 250–257.
[50] Z. Zeng, J. Tu, B. Pianfetti, T.S. Huang, Audio-visual affective expression recognition through multistream fused HMM, IEEE Trans. Multimedia 10 (4) (2008) 570–577.
[51] O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, F. Lepore, Audio-visual integration of emotion expression, Brain Res. 1242 (2008) 126–135.
[52] A. Cruz, B. Bhanu, S. Yang, A psychologically-inspired match-score fusion model for video-based facial expression recognition, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 341–350.
[53] M. Dahmane, J. Meunier, Continuous emotion recognition using Gabor energy filters, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 351–358.
[54] P.A. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[55] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[56] D. Greig, Video object detection speedup using staggered sampling, IEEE Workshop on Applications of Computer Vision (WACV), 2009, pp. 23–29.
[57] G.R. Bradski, Computer vision face tracking for use in a perceptual user interface, Tech. Rep. Q2, Intel Technol. J. (1998) 1–15.
[58] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (2009) 10–18.
[59] H. Meng, N. Bianchi-Berthouze, Naturalistic affective expression classification by a multi-stage approach based on Hidden Markov Models, Proc. of First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with ACII, Memphis, Tennessee, USA, 2011, pp. 378–387.
[60] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (1997) 2673–2681.
[61] F. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with LSTM, Neural Comput. 12 (10) (2000) 2451–2471.
[62] A. Graves, Supervised sequence labelling with recurrent neural networks, Ph.D. thesis, Technische Universität München (2008).
[63] M. Wöllmer, B. Schuller, A. Batliner, S. Steidl, D. Seppi, Tandem decoding of children's speech for keyword detection in a child–robot interaction scenario, IEEE Trans. Audio Speech Lang. Process. 7 (4) (2011) 1–26.
[64] M. Wöllmer, C. Blaschke, T. Schindl, B. Schuller, B. Färber, S. Mayer, B. Trefflich, On-line driver distraction detection using Long Short-Term Memory, IEEE Trans. Intell. Transp. Syst. 12 (2) (2011) 574–582.
[65] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, S. Narayanan, Analyzing the Memory of BLSTM Neural Networks for Enhanced Emotion Classification in Dyadic Spoken Interactions, Proc. of ICASSP, Kyoto, Japan, 2012.
[66] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition Morgan Kaufmann, San Francisco, 2005.
[67] G.W. Snedecor, W.G. Cochran, Statistical methods, 8th ed. Iowa State University Press, 1989.
[68] F. Grezl, M. Karafiat, K. Stanislav, J. Cernocky, Probabilistic and Bottle-neck Features for LVCSR of Meetings, Proc. of ICASSP, Honolulu, Hawaii, 2007, pp. 757–760.
[69] S. Bengio, An asynchronous Hidden Markov Model for audio-visual speech recognition, Adv. NIPS 15 (2003) 1–8.