Emotion Recognition in the Wild: Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion

Fabien Ringeval, Shahin Amiriparian, Florian Eyben, Klaus Scherer[*] ,Björn Schuller [† *]
Machine Intelligence & Signal Processing Group
Technische Universität Müchen
Munich, Germany
fabien.ringeval@tum.de

ABSTRACT

In this paper, we investigate the relevance of using voice and lip activity to improve performance of audiovisual emotion recognition in unconstrained settings, as part of the 2014 Emotion Recognition in the Wild Challenge (EmotiW14). Indeed, the dataset provided by the organisers contains movie excerpts with highly challenging variability in terms of audiovisual content; e. g., speech and/or face of the subject expressing the emotion can be absent in the data. We therefore propose to tackle this issue by incorporating both voice and lip activity as additional features in a decision-level fusion. Results obtained on the blind test set show that the decision-level fusion can improve the best mono-modal approach, and that the addition of both voice and lip activity in the feature set leads to the best performance ($UAR = 35.27\%$), with an absolute improvement of 5.36% over the baseline.

Categories and Subject Descriptors

H.5.1 [**Information systems**]: Information systems applications—*Multimedia information systems*

Keywords

Emotion Recognition; Multimedia; Voice Activity Detection; Lip Activity Detection; Decision-Level fusion

1. INTRODUCTION

Automatic Emotion Recognition has become a major field of research in the last decade. Early research focused on theoretical definitions of emotion [5], and automatic recognition on prototypically, acted databases [4, 22]. Recently, more naturalistic data, e. g., [7, 16, 17], as well as other paralinguistic phenomena besides emotion, such as social signals and autism [20], physical and cognitive load [19], have been addressed. Emotion recognition on real-life data suffers from two issues: first, the variance of emotions expressed is very high in relation to the data available (sparseness), second, state-of-the-art methods are largely affected by additive and convolutive background noise [10]. While the second issue can be eased by multimodal approaches [7, 21] or noise robustness counter measures [10], the first issue remains.

To overcome the data-sparseness, larger and more realistic multimodal emotion datasets are required, such as the one collected for the SEMAINE project [16], or the RECOLA dataset [17]. An endless resource of acted, but realistic emotional portrayals seems to be available in TV series and movies. The first database building on this kind of data was introduced in [13]. It contains excerpts from the Vera am Mittag TV show. Recently, facial expressions from TV shows and movies were used for emotion analysis, namely the Static Facial Expressions in the Wild database (SFEW) [8] and the Acted Facial Expressions in the Wild (AFEW) database [9]. Last year, the first Emotion in the Wild (EmotiW 2013) challenge [7] provided an audiovisual dataset (AFEW + audio tracks) and a platform for researchers to create, extend, and validate their methods on real-world movie data.

The AFEW database contains video clips collected by searching closed caption keywords for emotion related content. The labels obtained in this way were validated by human annotators in order to cope with incorrect or unrelated captions [7]. For the second EmotiW challenge [6], an updated version of this database is used, namely version 4.0. The given training set has 578 video clips extracted from movies labelled with six emotional expressions (Angry, Disgust, Fear, Happy, Sad and Surprise) and a neutral state. In the development/validation set, 383 video samples with corresponding labels are contained. 407 video clips without labels are released as blind evaluation/test set.

This paper describes our submission to the 2nd EmotiW challenge (EmotiW 2014) and is organised as follows: In Section 2 we introduce our developed system composed of four components, voice and lip activity detection, feature extraction and multimodal classification. The results obtained on the development and the evaluation set are discussed in Section 3. Section 4 highlights the differences between our results and the baseline results of the challenge, summarises the paper and discusses the direction of future work.

[*]The author is affiliated with the Swiss Center for Affective Sciences at the University of Geneva, Geneva, Switzerland.

[†]The author is also affiliated with the Department of Computing, Machine Learning Group at the Imperial College London, London, UK.

ICMI'14, November 12–16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2663204.2666271.

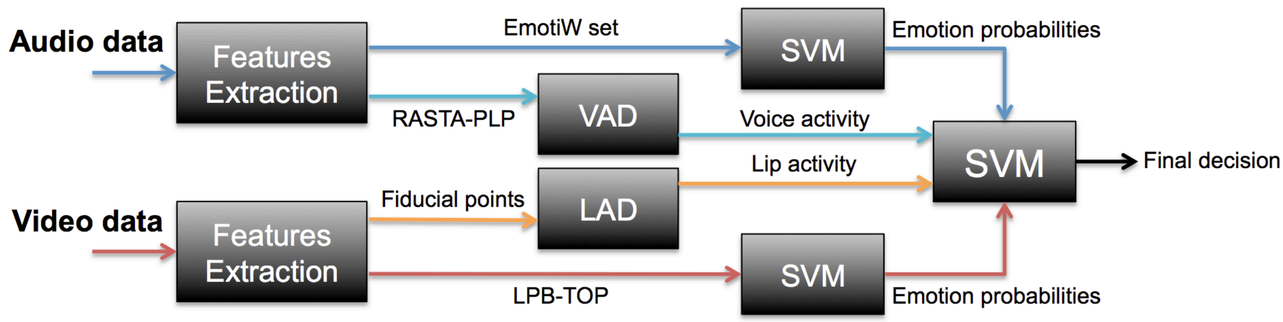

Figure 1: Flowchart of the emotion recognition system: mono-modal SVM based emotion recognition + decision-level fusion with both voice and lip activity as additional features.

2. SYSTEM

An overview of the system developed for the EmotiW 2014 challenge is shown in Fig. 1. Mono-modal emotion recognition is first performed separately on audio features and video features by using a supervised SVM learning. Outputs of these two systems (i. e., emotion probabilities) are then merged with the estimated mean voice and lip activity and a second SVM is used to predict the final emotion decision.

2.1 Voice Activity Detection

We used a voice activity detector to estimate the probability ρ_a of having speech in the audio instances of the dataset; we thus make the hypothesis that the emotion labelling procedure did not take into account the music content of the audiovisual excerpt, but only the spoken content for the audio modality. Because the probability of speech ρ_a was used for the emotion decision-level fusion, cf. Fig. 1, it was estimated only on the development and test dataset.

As the data provided for the EmotiW Challenge contain a high level of background noise and music, we employ our robust voice activity detector based on LSTM-RNN as introduced in [11], using topology $N1$. The input frontend to the neural network extracts RASTA-PLP [15] coefficients 1–18 and their first order delta regression coefficients. Our openSMILE toolkit [12] is used to extract the RASTA-RLP features. The output activation of the LSTM-RNN represents voice activity as values from approx. -1 to +1. This output was normalised into probabilities according to the minimum and maximum values observed on the validation set. Fig. 2 shows the histogram of the voice activity probability ρ_a estimated on all audio frames (frame rate is 10 ms) from both validation and test set. Frames for which the voice probability ρ_a is superior or equal to 0.5 can be interpreted as containing speech. Therefore, 38.48% of the overall audio frames doesn't contain speech.

2.2 Lip Activity Detection

Because speech can be present in the audiovisual instances while not being produced by the person seen in the video, we computed the lip activity from the video data[1]. We thus make the hypothesis that the emotion labelling was performed on the speech produced by the person seen in the video regarding the audio modality.

[1]In some rare cases (e. g., when depicting surprise), it is possible that the person seen in the movie doesn't talk while having the mouth open.

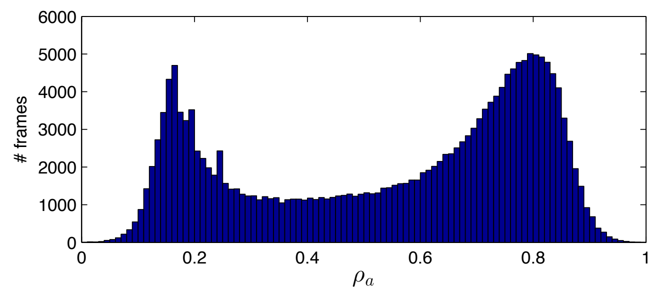

Figure 2: Histogram of voice activity ρ_a computed at the frame level (10 ms) on the development and test datasets; values of ρ_a that are superior or equal to 0.5 mean that speech is present in the corresponding frames.

The probability of lip activity ρ_l was estimated from the fiducial points provided by the organisers of the challenge, using a technique similar to the classic adaptive appearance model (AAM) methodology [1]. The fiducial points are estimated according to a given head pose, which the angle θ ranges from $-90°$ to $+90°$, with a step of 15°. The number of fiducial points modelling the face varies according to the head pose: there are 49 points for $\theta \in F : [-45°, +45°]$ and 39 points for $\theta \in NF$: $[-90°, -60°] \cup [+60°, +90°]$. Because it appeared that the alignment of the fiducial points on the face did not perform well with $\theta \in NF$, we only considered the frames where $\theta \in F$; Table 1 provides the association between the fiducial points and their corresponding region of the face; Fig. 3 shows the detected face and the fiducial points on a video frame of an instance labeled as happy in the training partition. As we noticed that errors in the detection of the fiducial points can appear with $\theta \in F$, we used the following list of checks:

1. the mean horizontal coordinate of the points modelling the nose has to be located in the middle of those computed on the points modelling the left and the right eye
2. the mean vertical coordinate of the points modelling the eyebrow has to be located above the mean vertical coordinate of the points modelling the corresponding eye, i. e., left and right, respectively

Table 1: Correspondances between the 49 detected fiducial points with the 7 modelled regions of the face when the absolute angle $|\theta|$ of the estimated head pose is inferior or equal to 45°

Indice of fiducial points	Face region
1–5	left eyebrow
6–10	right eyebrow
11–19	nose
20–25	left eye
26–31	right eye
32–43	outer mouth
44–49	inner mouth

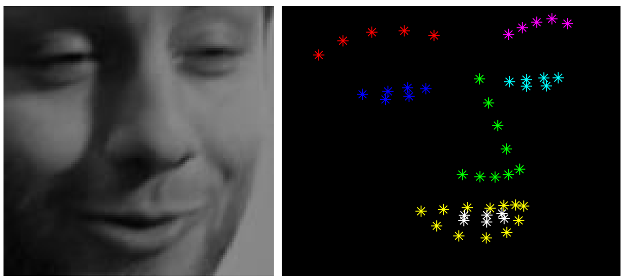

Figure 3: Detected face and fiducial points on a video frame of an instance labelled as happy (file: 000201320.avi, frame: 34; $\theta = 0°$, $\rho_l = 0.07$).

3. horizontal and vertical extremum of the points modelling the inner region of the mouth have to be bounded by the horizontal and vertical extremum of the points modelling the outer region of the mouth, respectively

On a total of 54.7 k frames available on the validation and test partitions, 4.94 % did not contain fiducial points (i. e., failure in the detection, e. g., not or partially visible face, too low level of luminosity), 5.07 % were obtained with a value of $|\theta| > 45°$ and 4.83 % were rejected by our check list.

To compute the probability of lip activity ρ_l, we first calculated the area A_{im} of the polygon formed by the points modelling the inner mouth region; the coordinates of the fiducial points were normalised in $[0, 1]$ to remove the influence of having various sizes of the face model over the instances. Because the area of the inner mouth region A_{im} depends on the angle of the head pose, we normalised this value with the maximum value observed on each head pose in absolute (accordingly), i. e., $|\theta| = [0°, 15°, 30°, 45°]$, to obtain the probability of lip activity ρ_l; the minimum value of A_{im} was always found to be equal to 0 for all angles $\theta \in F$ of head pose.

Fig. 4 shows the histogram of the lip activity probability ρ_l estimated on both validation and test set at the frame level (frame rate is 40 ms). This histogram shows that there is a high number of frames for which the probability of lip activity is close to 0, i. e., having the mouth closed; 22.07 % of the processed video frames present a value of $\rho_l < 0.1$. This percentage is lower than the one obtained on speech (38.48 % of frames doesn't contain speech), because there is the case where the mouth can be opened without producing sound.

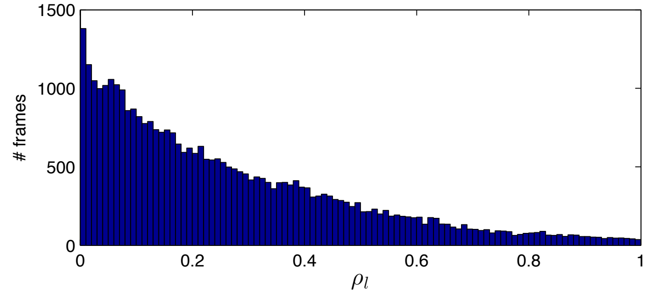

Figure 4: Histogram of the probability of lip activity ρ_l computed at the frame level (40 ms) on the development and test datasets; the lip activity ρ_l is computed by the area formed by the points modelling the inner region of the mouth, normalised by the maximum value observed on each head pose in absolute (accordingly), i. e., $|\theta| = [0°, 15°, 30°, 45°]$.

2.3 Features Extraction

We describe below the two feature sets we used for extracting information from the audio and video data, respectively.

2.3.1 *Audio Features*

In contrast to large scale brute-force feature sets, which have been successfully applied to many speech and music classification tasks, e. g., [20, 23], smaller, expert-knowledge based feature sets have shown high robustness for emotion recognition [3]. In this light, we assembled a small acoustic feature set for the EmotiW14 Challenge, using our openSMILE toolkit [12].

The set contains 102 parameters in total. The parameters are based on the following Low-Level Descriptors (LLD): Fundamental Frequency (F_0) represented on a logarithmic scale as well as on a linear scale, Loudness (computed as the sum of the intensities in 26 Mel-frequency scale auditory spectrum bands (20–8000 Hz)), Mel-Frequency Cepstral Coefficients (MFCC) 1–4, Jitter, Shimmer, Harmonics-to-Noise Ratio (HNR), Formants 1–3 (frequency, bandwidth and log of amplitude relative to F_0), spectral slopes and spectral flux. As functionals to summarise the descriptors over an analysis segment, mean and standard deviation are applied to all LLD, and to loudness and F_0 additionally the following functionals are applied: percentiles (20, 50, 80) and the range of percentiles 20–80, as well as the means and standard deviations of the slopes of rising and falling contour parts, and of the length of voiced and unvoiced segments. Further, the equivalent sound level (i. e., the average RMS energy converted to dB) is included.

2.3.2 *Video Features*

We used the feature set provided by the organisers as visual descriptors of the face contained in the video frames [7]. The system requires first to detect the face in the video frames; the MoPs framework was used for detection [26][2] and the Intraface tracker was employed for tracking [24][3]. Features of the face were then extracted with the Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) method [25]. The set contains 2832 parameters in total.

[2] `http://www.ics.uci.edu/~xzhu/face/`

[3] `http://www.humansensing.cs.cmu.edu/intraface/`

Table 2: Matrix of confusion obtained on the 7 classes of the EmotiW14 dataset with a emotion perception test; training+validation+test datasets; last value in bold is the % UAR.

Partition	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	% Recall
Angry	120	68	6	2	13	1	4	56.08
Disgust	3	49	0	5	37	27	11	37.12
Fear	4	5	101	0	21	5	23	63.52
Happy	1	3	1	217	21	2	7	86.11
Neutral	4	11	10	8	224	17	10	78.87
Sad	2	20	34	4	28	105	5	53.03
Surprise	1	8	17	5	22	4	72	55.81
% Precision	88.89	29.88	59.76	90.04	61.20	65.22	54.55	**61.51**

2.4 Machine Learning

Mono-modal emotion recognition of the 7 classes was performed by using a SVM classifier with the SMO training algorithm. For transparency and reproducibility, we used the implementation provided in the Weka data mining software [14] – version 3-6-10. Two types of features normalisation were used: either a normalisation between $[0-1]$ or a standardisation, i. e., subtracting the mean and dividing by the standard deviation. We used either a linear (the degree of exponent varied between 1 and 3 with a step of 1) or a gaussian kernel (the gamma coefficient varied on a logarithmic scale with 10 values between 10^{-5} and 0.5), and optimised the complexity parameter on a logarithmic scale (10 values between 10^{-3} and 1). The SVM was configured to provide as output the probability of each class by using logistic regression models. Performance was optimised on the validation set using the training set for learning the models, and the unweighted average recall (UAR), i. e., the mean value of the recall of each class in percentage, was used as metric; chance score is therefore equal to $1/7 = 14.29\,\%$.

For the multi-modal emotion recognition, we used as features the sets of probabilities obtained on the audio and video features, respectively. Additionally, we added the probability of voice and lip activity (the mean value was computed for each instance) to this feature vector. Finally, another SVM was used on this feature set, with the same set of parameters and configurations as used for mono-modal emotion recognition. Considering the limited number of instances available for each class on the validation partition (3 classes contain less than 50 instances, the other 4 classes contain less than 65 instances), we optimised the performance on this partition with a 2-fold cross validation.

3. RESULTS

We first present below the results obtained in a perception test of the full dataset; results obtained on the test partition were submitted as non-candidates for the challenge. The performance obtained by our system on the audio and video modalities are then described, followed by those obtained with the multi-modal fusion. A discussion of the results over each emotion class is provided at the end of this section.

3.1 Human perception test

We performed a perception test on all the audiovisual data provided for the EmotiW14 challenge to estimate the performance of human labelling in an emotion recognition task. One author of this paper labelled all instances of the corpus in a randomised order. Table 2 shows the confusion matrix obtained with this perception test; results obtained on the training, validation and test partitions were summed. The obtained performance ($\%\,UAR = 61.51$) is quite low and shows that the emotion classes contained in the EmotiW14 dataset are not easy to identify even for a human. In particular, 'Disgust' was the worst recognised emotion and the most confused, whereas 'Happy' was the best recognised emotion and the less confused. In comparison, the performance obtained by human labelling of audiovisual data on the GEMEP corpus - acted data from professional actors - is higher ($\%\,UAR = 76.00$) [2]; for comparison, we considered the 6 following classes: 'Joy (elation)', 'Hot anger (rage)', 'Panic fear', 'Sadness (depression)', 'Disgust' and 'Surprise', cf. Table 5 in [2]. This difference in accuracy of emotion perception is probably due to the fact that: (1) the audiovisual data present in the EmotiW14 dataset are highly compressed, which reduces the quality of the stimulus (2) the emotions portrayed in the GEMEP database are more prototypical than those used in EmotiW14, e. g., 'Angry' vs. 'Rage', 'Fear' vs. 'Panic fear', 'Sadness' vs. 'Depression' and (3) EmotiW14 contains an additional neutral case that is not present in GEMEP.

3.2 Audio features

Results obtained with the audio features (EmotiW set; 102 parameters, cf. section 2.3.1) on the validation partition are displayed in Fig. 5. The best performance ($\%\,UAR = 32.82$) is obtained with a gaussian kernel ($\gamma = 0.5$) and the lowest tested value of complexity, i. e., $C = 10^{-3}$, combined with a standardisation of the features. The absolute improvement over the baseline ($\%\,UAR = 23.74$) is 9.08 %; the baseline feature set includes 1582 parameters, which were proposed for the INTERSPEECH 2010 Paralinguistic challenge [18]. We can notice from Fig. 5 that performance decreases when the complexity parameter of the SVM increases, which let us suppose that there are many outliers in the audio data, since the system performs better when these outliers are not considered for computing the decision frontier. In the average, the RBF kernel performed better than the linear kernel, and the standardisation better than the normalisation (extremums are more sensitive to noise than mean and variance). Predictions on the test partition were submitted with the linear ($\%\,UAR = 32.16$) and the gaussian ($\%\,UAR = 31.61$) kernels, and we obtained again a better performance than the baseline ($\%\,UAR = 23.82$); absolute improvement with the linear kernel is 8.34%. Therefore, a smaller, expert-knowledge based acoustic feature set shows higher robustness for emotion recognition than a large scale brute-force feature set, as found in [3].

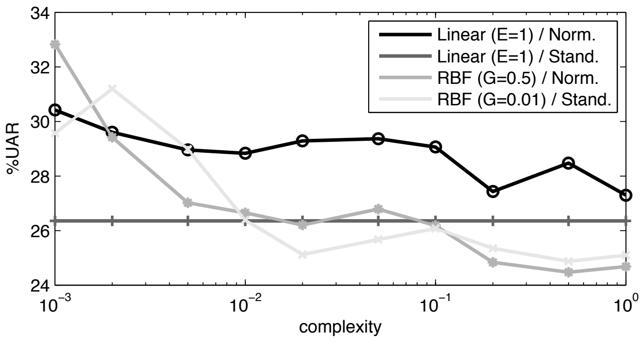

Figure 5: Emotion recognition performance on the EmotiW14 dataset (7 classes) with audio features (EmotiW set; 102 parameters) and SVM classifier, for different types of kernel (linear, RBF), normalisation procedures (norm: normalisation between $[0-1]$, stand.: standardisation to zero mean and unit variance) and values of complexity.

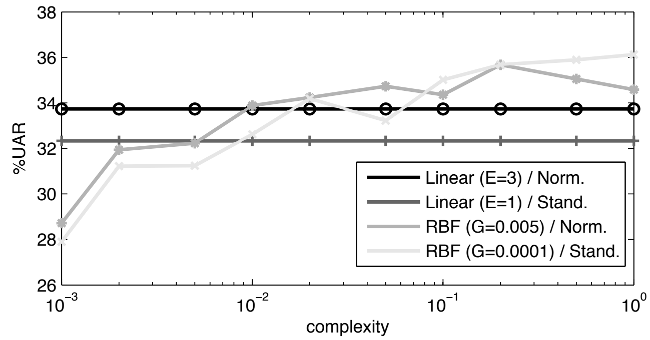

Figure 6: Emotion recognition performance on the EmotiW14 dataset (7 classes) with video features (baseline set; 2832 parameters) and SVM classifier, for different types of kernel (linear, RBF), normalisation procedures (norm: normalisation between $[0-1]$, stand.: standardisation to zero mean and unit variance) and values of complexity.

3.3 Video features

Results obtained with the video features (baseline set; 2832 parameters, cf. section 2.3.2) are displayed in Fig. 6. The best performance ($\%UAR = 36.13$) is obtained with a gaussian kernel ($\gamma = 10^{-4}$) and the highest tested value of complexity, i.e., $C = 1$, combined with a standardisation of the features. The absolute improvement obtained over the baseline ($\%UAR = 31.49$) is 4.64 %, which is smaller than the one achieved with audio features (9.08%), but we used as video data the baseline feature set and only tuned the parameters of the SVM and the normalisation technique. Whereas the performance doesn't vary over the different values of complexity for the linear kernel (same support vectors were probably obtained for the different values of complexity), the performance increased with the value of complexity with the gaussian kernel; the RBF based projection of the features helped to find a better discriminant space. Predictions on the test partition were submitted with the linear ($\%UAR = 29.07$) and the gaussian ($\%UAR = 32.33$) kernels, and we obtained a better performance than the baseline ($\%UAR = 29.91$) only with the RBF kernel; absolute improvement is 2.42 %. Video features thus did perform better than the audio features for the emotion recognition of the 7 classes of the EmotiW14 dataset, on both validation and test partitions. However, the difference in performance between audio and video features is rather small, especially on the test partition, which thus lets room for improvement by using a multi-modal approach.

3.4 Decision-level fusion

Predictions obtained with the audio and video features were learned by another SVM in order to perform a decision-level fusion. We used the same ensemble of machine learning settings for kernel and complexity as those employed for the mono-modal emotion recognition. Multi-modal predictions were performed separately regarding the type of kernel used for the audiovisual modalities, i.e., either linear or RBF kernel. Results obtained on the validation partition (2-fold cross validation) are depicted in Fig. 7. The best performance ($\%UAR = 37.78$) was obtained on the mono-modal predictions made with a linear kernel, with the following configuration: linear kernel (2^{nd} order), normalisation of features in $[0-1]$ and with the lowest value of complexity, i.e., $C = 10^{-3}$. The absolute improvement over our best mono-modal approach ($\%UAR = 36.13$) is 1.65 %, and up to 12.41 % over the baseline ($\%UAR = 25.37$), which was obtained with a feature-level fusion – performance has dropped compared to the mono-modal baseline. Predictions on the test partition were submitted with the linear ($\%UAR = 30.46$) and the RBF ($\%UAR = 33.58$) based decision-level fusion. Whereas the linear kernel performed best on the validation partition for the decision-level fusion, the RBF kernel provided the best performance on the test partition. A small improvement can be observed compared to our best mono-modal result ($\%UAR = 32.33$). Therefore, a decision-level fusion appears more appropriate than a feature-level fusion, for the multi-modal emotion recognition of the EmotiW14 dataset, since we improved the performance whereas a drop was observed on the baseline.

The values of voice and lip activity (cf. section 2.1 and 2.2, respectively) were added to the feature vector used to perform the multi-modal decision-level fusion, i.e., emotion predictions obtained with the audio and video features. The goal is to provide the system some knowledge regarding the occurrence of speech in the audiovisual data, because some instances of the dataset doesn't contain speech, cf. Fig. 2 and 4. Results obtained on the validation partition (2-fold cross validation) are depicted in Fig. 8. The best performance ($\%UAR = 38.78$) was obtained on the mono-modal predictions made with a linear kernel, with the following configuration: RBF kernel ($\gamma = 5.10^{-3}$), normalisation of features in $[0-1]$ and with the complexity equal to $C = 0.2$. The absolute improvement over the previous approach, i.e., without adding the voice and lip activity, is quite small but still observable: 1.00 %. An improvement over the decision-level fusion was also observed on the test partition with the mono-modal predictions made with the RBF kernel: we obtained a performance of $\%UAR = 35.27$, for an absolute improvement of 1.69 %. These results show that the use of voice and lip activity as additional features helps to improve the multi-modal emotion recognition of the EmotiW14 dataset.

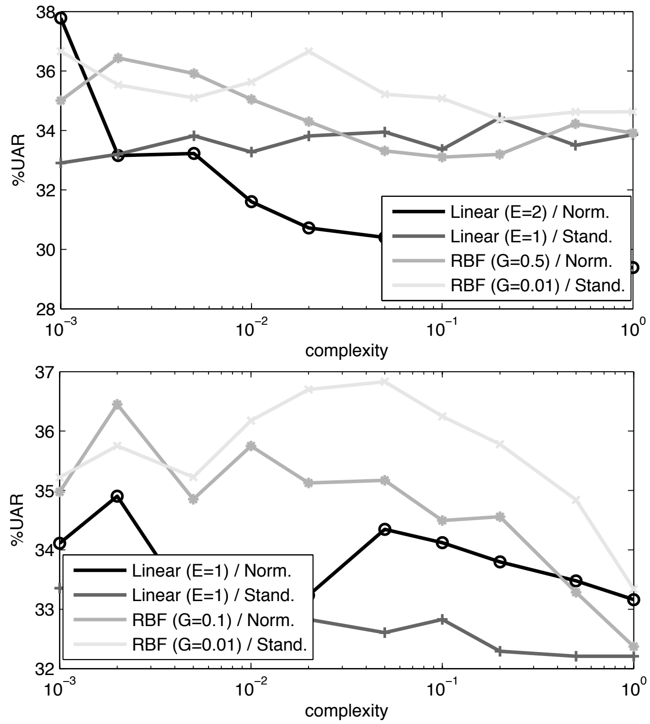

Figure 7: Emotion recognition performance on the EmotiW14 dataset (7 classes) with decision-level fusion of audio and video predictions (top: estimated with a linear kernel, bottom: estimated with a gaussian kernel), for different types of kernel (linear, RBF), normalisation procedures (norm: normalisation in $[0-1]$, stand.: standardisation to zero mean and unit variance) and values of complexity.

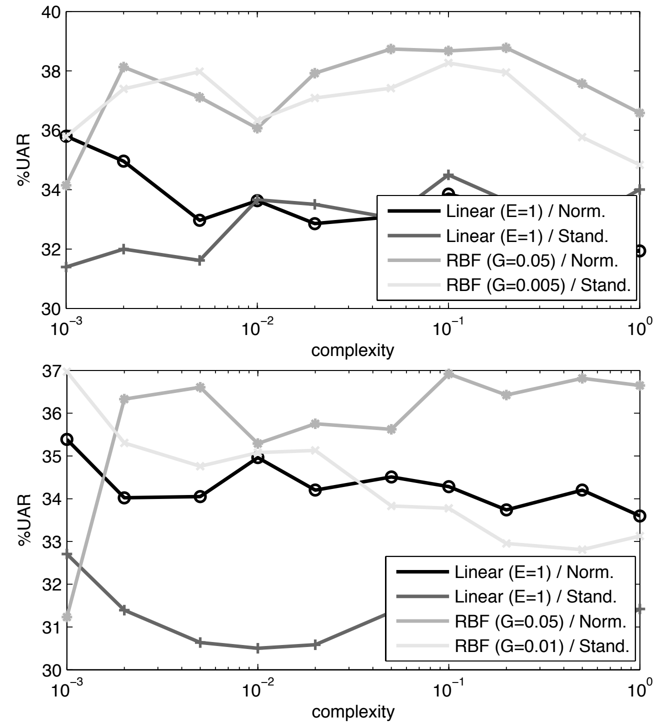

Figure 8: Emotion recognition performance on the EmotiW14 dataset (7 classes) with decision-level fusion of audio and video predictions (top: estimated with a linear kernel, bottom: estimated with a gaussian kernel) combined with voice and lip activity, for different types of kernel (linear, RBF), normalisation procedures (norm: normalisation in $[0-1]$, stand.: standardisation to zero mean and unit variance) and values of complexity.

A summary of the best performance obtained on the automatic emotion recognition on the validation and test partitions of the EmotiW14 dataset for the different studied approaches, i. e., audio, video, decision-level fusion and with voice and lip activity, is given in Table 3.

3.5 Performance over the emotion classes

A detailed description of the automatic emotion recognition performance is given for each emotion class in Table 4; performance of human perception test and the baseline system are included as well. Interestingly, the automatic recognition system based on audio features performed better than human labelling for the emotion 'Angry' on the test partition; the multi-modal decision-level fusion with VAD and LAD also performed better on the validation partition. The audio modality provided the best performance on test for both 'Angry' and 'Neutral' classes, video modality performed best for 'Disgust' and 'Happy', whereas the multi-modal system, i. e., audio and video combined, performed best for 'Fear' and 'Sad'; performance was the same on 'Surprise' for both video and audiovisual based recognition systems. The analysis of the GEMEP corpus also shows that 'Fear', 'Sad' and 'Surprise' are best recognised with audio-visual data in comparison to mono-modal data, cf. Table 5 in [2].

4. CONCLUSIONS

We investigated the relevance of using voice and lip activity to improve performance of audiovisual emotion recognition in unconstrained settings, as part of the 2014 Emotion Recognition in the Wild Challenge (EmotiW14). A small, expert-knowledge based acoustic feature set (EmotiW: 102 parameters) was used for emotion recognition on audio data, and it showed higher robustness than the large scale brute-force feature set (INTERSPEECH 2010 Paralinguistic Challenge: 1582 parameters); the absolute improvement was equal to 9.08 % on the validation set and 8.34 % on the test set. Regarding video features, we used the baseline set (LBP-TOP method; 2832 parameters) proposed by the organisers. A tuning of the parameters of the SVM allowed to improve the baseline system with an absolute improvement of 4.64 % on the validation set and 2.42 % on the test set. Whereas the performance dropped with the multi-modal baseline system (feature-level fusion) compared to the mono-modal baseline system, our decision-level fusion achieved an absolute improvement of 1.65 % on the validation set and 1.25 % on the test set compared to our best mono-modal performance; a decision-level fusion appears thus more suitable than a feature-level fusion, for the multi-modal emotion recognition of the EmotiW14 dataset. Finally, the addition of both voice and lip activity as features in the multi-modal

Table 4: Emotion recognition performance (%*recall*) obtained on the 7 classes of the EmotiW14 dataset with perception test (human labelling), mono-modal and multi-modal (i. e., decision-level fusion) systems – best configurations of SVM are retained here; UAR: unweighted average recall; the baseline corresponds to the best system used by the organiser (i. e., using only video features); best automatic recognition performance obtained on the test set are in bold.

Partition	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	%UAR
Perception test								
Validation	68.75	27.50	58.70	92.06	93.65	54.10	52.17	63.85
Test	46.55	46.15	78.26	91.35	64.96	62.26	65.39	64.99
Baseline								
Validation	50.00	25.00	15.22	57.14	34.92	16.39	21.74	31.49
Test	36.21	34.62	26.09	**41.98**	40.17	22.64	7.69	29.91
Audio								
Validation	57.81	12.50	43.48	30.16	65.08	16.39	4.35	32.82
Test	**60.35**	3.85	41.30	27.16	**64.10**	24.53	3.85	32.16
Video								
Validation	54.24	28.21	22.73	61.91	45.90	20.34	19.56	36.13
Test	37.93	**38.46**	23.91	35.80	42.74	32.08	**15.38**	32.33
Audio+Video								
Validation	60.93	12.50	34.78	55.56	39.68	26.23	34.78	37.78
Test	50.00	7.69	**52.17**	40.74	42.74	30.19	11.54	33.58
Audio+Video+VAD+LAD								
Validation	70.31	0.0	34.78	52.38	74.60	18.03	8.69	36.97
Test	53.45	7.69	50.00	40.74	41.88	**37.74**	**15.38**	**35.27**

Table 3: Performance on the automatic emotion recognition of the EmotiW14 dataset (7 classes) using different approaches: audio features, video features, audiovisual decision-level based fusion and audiovisual decision-level based fusion with voice (VAD) and lip (LAD) activity; baseline on audio was obtained with a linear kernel; baseline on video was obtained with a gaussian kernel; baseline on audio+video was obtained with a feature-level fusion (RBF kernel).

% UAR	Validation	Test
Audio		
Baseline	23.74	23.82
Linear kernel	30.42	**32.16**
Gaussian kernel	**32.82**	31.61
Video		
Baseline	31.49	29.91
Linear kernel	33.74	29.07
Gaussian kernel	**36.13**	**32.33**
Audio+Video		
Baseline	25.37	23.02
Linear kernel	**37.78**	30.46
Gaussian kernel	36.83	**33.58**
Audio+Video+VAD+LAD		
Linear kernel	**38.78**	31.13
Gaussian kernel	36.97	**35.27**

decision-level fusion allowed to obtain the best overall performance on both validation ($\%UAR = 38.78$) and test ($\%UAR = 35.27$) sets. Therefore, the add of knowledge of the occurrence of speech in the audiovisual data helps the system to know which modality to trust.

More complex machine learning algorithms than SVM, such as those exploiting non-linear dependencies (e. g., deep neural networks - DNN), could probably provide some additional improvement in the automatic emotion recognition of the highly challenging EmotiW14 dataset. However, we do not believe that recurrent based architectures, such as LSTM or BLSTM, would help further, because the duration of the instances is quite small (maximum duration is 5.4 s). Finally, some improvement could probably be obtained by tuning a bit more the feature sets. Regarding audio data, finely tuned features selection could be performed from a large brute-force feature set, whereas for video data, geometric based information could be added to the feature set, since the baseline set contains only appearance based information.

5. ACKNOWLEDGMENTS

This research was supported by an ERC Advanced Grant in the European Community's 7th Framework Programme under grant agreement 230331-PROPEREMO (Production and perception of emotion: an affective sciences approach) to Klaus Scherer, and by the National Center of Competence in Research (NCCR) Affective Sciences financed by the Swiss National Science Foundation (51NF40-104897) and hosted by the University of Geneva.

6. REFERENCES

[1] A. Abel, A. Hussain, Q. D. Nguyen, F. Ringeval, M. Chetouani, and M. Milgram. Maximising Audiovisual Correlation with Automatic Lip Tracking and Vowel Based Segmentations. *Biometric ID Management and Multimodal Communication, joint COST 2101 and 2102 International Conference, Madrid, Spain, September 16-18 2009, Lecture Notes in Computer Science*, 5707:65–72, 2009.

[2] T. Bänziger, M. Mortillaro, and K. Scherer. Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception. *Emotion*, 12(5):1161–1179, 2012.

[3] D. Bone, C.-C. Lee, and S. S. Narayanan. Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features. *IEEE Transactions on Affective Computing (TAC)*, 5(2):201–213, April-June 2014.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Proc. of INTERSPEECH 2005*, pages 1517–1520, Lisbon, Portugal, 2005. ISCA.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, January 2001.

[6] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion Recognition in the Wild Challenge 2014: Baseline, Data and Protocol. In *Proc. of ICMI 2014*, Istanbul, Turkey, November 2014. ACM.

[7] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proc. of ICMI 2013*, pages 509–516, Sydney, Australia, December 2013. ACM.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark. In *Proc. of ICCV Workshops 2011*, pages 2106–2112, Barcelona, Spain, November 2011. IEEE.

[9] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *MultiMedia, IEEE*, 19(3):34–41, July-September 2012.

[10] F. Eyben, F. Weninger, and B. Schuller. Affect Recognition in Real-Life Acoustic Conditions – A New Perspective on Feature Selection. In *Proc. of INTERSPEECH 2013*, pages 2044–2048, Lyon, France, August 2013. ISCA.

[11] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies. In *Proc. of ICASSP 2013*, pages 483–487, Vancouver, Canada, May 2013. IEEE.

[12] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of ACM Multimedia 2010*, pages 1459–1462, Florence, Italy, 2010. ACM.

[13] M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German audio-visual emotional speech database. In *Proc. of ICME 2008*, pages 865–868, Hannover, Germany, June 2008. IEEE.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, June 2009.

[15] H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America (JASA)*, 87(4):1738–1752, November 1990.

[16] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing (TAC)*, 3(1):5–17, January-March 2012.

[17] F. Ringeval, A. S.onderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proc. of EmoSPACE 2013, held in conjunction with FG 2013*, Shanghai, China, April 2013. IEEE.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. of INTERSPEECH 2010*, pages 2794–2797, Makuhari, Japan, 2010. ISCA.

[19] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *Proc. of INTERSPEECH 2014*, Singapore, Singapore, September 2014. ISCA.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. of INTERSPEECH 2013*, pages 148–152, Lyon, France, August 2013. ISCA.

[21] B. Schuller, M. Valstar, R. Cowie, and M. Pantic. AVEC 2012: The Continuous Audio/Visual Emotion Challenge – An Introduction. In *Proc. of ICMI 2012*, pages 361–362, Santa Monica, CA, USA, October 2012. ACM.

[22] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *Proc. of ASRU 2009*, pages 552–557, Merano, Italy, December 2009. IEEE.

[23] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Emotion Science*, 4(Article ID 292):1–12, May 2013.

[24] X. Xiong and F. De la Torre. Supervised Descent Method and its Applications to Face Alignment. In *Proc. of CVPR 2013*, pages 532–539, Portland, OR, USA, 2012. IEEE.

[25] G. Zhao and M. Pietikainen. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):915–928, June 2007.

[26] X. Zhu and D. Ramanan. Face Detection, Pose Estimation, and Landmark Localization in the Wild. In *Proc. of CVPR 2012*, pages 2879–2886, Providence, RI, USA, 2012. IEEE.