Enhanced Face/Audio Emotion Recognition: Video and Instance Level Classification using ConvNets and Restricted Boltzmann Machines

Juan M. Mayor Torres
University of Trento
Signals and Interactive Systems Lab
Trento, Italy
juan.mayortorrres@unitn.it

Evgeny A. Stepanov
University of Trento
Signals and Interactive Systems Lab
Trento, Italy
evgeny.stepanov@unitn.it

ABSTRACT

Face-based and audio-based emotion recognition modalities have been studied profusely obtaining successful classification rates for arousal/valence levels and multiple emotion categories settings. However, recent studies only focus their attention on classifying discrete emotion categories with a single image representation and/or a single set of audio feature descriptors. Face-based emotion recognition systems use a single image channel representations such as principal-components-analysis whitening, isotropic smoothing, or ZCA whitening. Similarly, audio emotion recognition systems use a standardized set of audio descriptors, including only averaged Mel-Frequency Cepstral coefficients. Both approaches imply the inclusion of decision-fusion modalities to compensate the limited feature separability and achieve high classification rates. In this paper, we propose two new methodologies for enhancing face-based and audio-based emotion recognition based on a single classifier decision and using the EU Emotion Stimulus dataset: (1) A combination of a Convolutional Neural Networks for frame-level feature extraction with a k-Nearest Neighbors classifier for the subsequent frame-level aggregation and video-level classification, and (2) a shallow Restricted Boltzmann Machine network for arousal/valence classification.

KEYWORDS

ConvNets, RBM, Face emotion, Audio Emotion, EU Emotion Stimulus

ACM Reference format:
Juan M. Mayor Torres and Evgeny A. Stepanov. 2017. Enhanced Face/Audio Emotion Recognition: Video and Instance Level Classification using ConvNets and Restricted Boltzmann Machines. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.

1 INTRODUCTION

Emotions and complex mental states have been studied through different image-based and audio-based modalities achieving important milestones in emotion recognition state-of-the-art [19].

Facial features are considered to be an informative channel for the recognition of Ekman's six basic emotions [6] (*angry, afraid, happy, sad, disgusted,* and *surprised* with their corresponding appraisal mechanisms); thus, they are considered an important reference point for multiple other facial feature-based emotion recognition systems [15]. These particular studies include facial features such as Facial Action Coding System (FACS) [4], non-linear image kernel models [25], and Convolution Neural Networks (ConvNets) probabilities [7, 14].

Likewise, recent studies such as *Emonets* [8] and [5] (winner of the ICMI 2013 emotion recognition challenge) use ConvNets to discriminate sets of isotropic image's patches from affective faces; thus extending the classification paradigm for frame-level and video-level annotation. However, studies that involve ConvNets and multiple emotion categories tend to use a single face representation (e.g. PCA, isotropic smoothing) limiting the feature space in the frame aggregation and the subsequent video-level classification. On the other hand, audio features such as low-level audio descriptors (LLD) are considered a primary score for understanding the emotion implied in affective dyadic interactions [3]. Fundamental frequency (F_0), Mel-Frequency Cepstral Coefficients (MFCC), Jitter and the percentages of pauses are considered important features for describing prosodical cues and the related affective utterance [24]. However, the multilingual variability and the considerable limitation for extracting significant features from unimodal audio instances/utterances are considered critical drawbacks for the classification of multiple affective categories.

Despite the emotion recognition systems have included facial and audio features in an separated or a synchronized modalities increasing classification rates, some image normalization models such as ZCA whitening and PCA are not included for face-based modalities, and low-level feature sets and descriptors such as Zero-Crossing Rate or the quartile and interquartile ranges of the low-level descriptors for audio-based modalities; thus constraining their evaluation to small set of emotion categories and limiting the inter-classes separability. These drawbacks imply the combination of multiple classifiers' decision (e.g. DBN, ConvNet, RBM, Auto-encoder) to achieve a considerable emotion recognition rate [12, 23].

In this paper, we propose two pipelines for face and audio emotion classification using a single classifier decision and without including any type of decision-fusion modality: (1) A *face-based* emotion recognition using the facial contours detected by *dlib* for face alignment, and a combination of the *ConvNets* described by [9] and a subsequent frame aggregation (video-level) classification using k-Nearest Neighbor (kNN) or two-layer Restricted Boltzmann

WI '17, Leipzig, Germany
© 2017 ACM. 978-1-4503-4951-2/17/08…$15.00
DOI: 10.1145/3106426.3109423

Machine (RBM) classifier. (2) An *audio-based* emotion recognition using low-level descriptors extracted from multiple speech signal excerpts to feed the two-layer RBM network classifier. We use the video and audio data from EU-emotion Stimulus dataset – a set of emotion stimuli designed for understanding complex emotions in Asperger Condition (ASC) populations. In this study we report the performances of kNN and RBM classifiers predicting the six basic Ekman's emotions and *neutral* for face-based classification, and the performances of kNN and RBM classifiers for separate high/low levels of arousal and valence classes respectively.

2 EU-EMOTION STIMULUS DATASET

The EU-emotion Stimulus (EESS) dataset presented by O'Reilly et. al [13] is an induced stimuli dataset developed for understand complex emotion elicitation and the subsequent neurophysiological responses in ASC patients. EESS was collected as a part of the ASC-Inclusion project www.ascinclusion.eu. EESS consists of a set facial expressions, voices, and body gestures that are annotated using 20 different emotion/mental state labels such as *afraid, angry, ashamed, bored, disappointed, disgusted, excited, frustrated, happy, hurt, interested, jealous, joking, kind, proud, sad, sneaky, surprised, unfriendly,* and *worried* plus an extra *neutral* category.

Emotions were enacted by 17 different actors of 5 different ethnicities. Each actor only portrayed 10 different emotions: two subsets of 3 basic emotions (Ekman's emotions) and 7 additional complex emotions that were evenly assigned to the actors. On the other hand, the vocal emotional expressions were represented by prepared scripts assigned for each particular emotion individually. Each actor used different scripts to portray low and high intensity for the 6 basic emotions *angry, afraid, happy, sad, disgusted,* and *surprised* for both video and audio.

The data was annotated using 14 online surveys representing a total of 1431 responses. Only six out of these 14 surveys were used for annotation, each micro-task consisted of 15-30 stimuli. Annotators gave responses for 3 different tasks: *recognition, emotion impression (valence and arousal)*, and *intensity*. Final results for emotion recognition were expressed in chance-corrected scores, a 63% represent the overall chance-corrected data for the 20 emotion/mental states plus neutral.

In our experiments, we use the EESS UK English subset of data and only the six Ekman's basic emotion labels. Table 1 gives the number of instances, video frames and time-length for each modality – face and vocal expressions – for each EESS emotion/mental state included in this study.

3 METHODOLOGY

Table 1 shows the facial and the vocal expression contents of EESS: the number of videos (# Vid.), the corresponding number of frames for all videos (# Fra.), and the total time-length of the videos (T_{vid} [s]) per emotion. The number of instances/utterances (# Inst.), the corresponding 25ms speech segments contained in all the instances (# Segm.), and the total time-length of the utterances (T [s]). In summary, the total number of video frames and speech segments are 26638 for faces and 14859 for audio respectively, however, in our study we only include the six basic emotion categories such as *angry, afraid, happy, sad, disgusted, surprised* and *neutral* for being consistent with the baseline. A total of 63 videos composed of 9073 frames, and 193 instances composed of 5697 audio segments composed our evaluation subset, grouping the audio segments in high/low arousal valence levels for audio-based as we explain below. We exclude all the frames and audio instances that belong to low intensity emotion categories.

Table 1: EESS contents for facial expressions and voices. Contents for all the 21 emotions/mental states are specified as well as the number of frames and audio segments for face-based and audio-based pipelines respectively. (Disapp.) is the abbreviation of Disappointed, and (Unfriend.) the abbreviation of Unfriendly

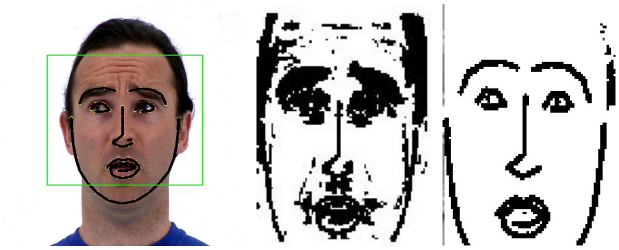
Emotion/ Mental States	**Facial Expressions**			**Voices**		
	# Vid.	**# Fra.**	T_{vid} **[s]**	**# Inst.**	**# Segm.**	T **[s]**
Afraid	7	1255	46	31	510	12.7
Angry	7	1016	38.3	35	1032	25.8
Ashamed	8	660	26.4	26	496	12.4
Bored	8	1458	52.6	31	800	20
Disapp.	6	522	20	21	758	18.9
Disgusted	9	1345	49.1	29	575	14.3
Excited	7	1051	38.8	31	665	16.6
Frustrated	7	1159	43.1	26	840	21
Happy	8	1166	43.3	38	1031	25.7
Hurt	6	422	16.9	22	603	15.1
Interested	6	643	25.2	32	931	23.3
Jealous	7	1382	50.3	21	441	11.1
Joking	7	1141	42	28	620	15.5
Kind	9	1394	52.2	31	724	18.1
Proud	8	1076	61.3	29	332	8.3
Sad	8	1254	40.9	31	824	20.6
Sneaky	7	1053	46.8	27	502	12.5
Surprised	8	884	39.5	24	598	14.9
Unfriend.	8	1005	34.5	24	747	18.6
Worried	8	1687	38.5	20	703	17.6
Neutral	16	2153	81.9	29	1127	28.2
TOTAL	**165**	**26638**	**887.9**	**586**	**14859**	**371.5**

In the following subsections we explain the details of the face-based and audio-based emotion recognition systems, evaluating our performances with a 5-fold cross-validation and comparing our pipelines to the top current state-of-the-art systems.

3.1 Face-based

RGB images were extracted from each EESS video with an initial resolution of 1440×1080. Each EESS video was resampled to 30 fps using ffmpeg, and each of these frames contains the figure of a single actor standing in front of camera and portraying the corresponding emotion.Therefore, we applied a haarcascade alt-tree classifier through openCV to detect frontal faces [11] that guarantees a minimum of 40% of continuous face detection rate in our tests.

Figure 1a shows the RGB image with the face inclosed in a green rectangle. Table 1 shows the number of valid frames with detected faces per emotion. Our experiments for face-based emotion recognition are (1) an evaluation of the ConvNets model described by Kahou et. al [9] using a ZCA whitening normalization (Figure 1b), and a KNN and a shallow RBM classifier evaluation for the frame-aggregation task, (2) a similar analysis but using the dlib landmarks (Figure 1c) as input, and (3) a baseline replication of by Kahou et. al [9] using the isotropic smoothing extracted from INline image toolbox [22], and the radial basis function (RBF) SVM for the frame-aggregation and subsequent video-classification.



(a) Face Detected using openCVs **(b) ZCA_{white} image** **(c) dlib face contour**

Figure 1: Afraid face in 1a color RGB image, 1b ZCA_{white} representation, and 1c dlib landmark contours representation

3.1.1 ZCA Whitening. ZCA whitening is a linear transformation derived by a PCA whitening process plus a zero mean subtraction per image or a set of random cropped images for feeding deep neural network schemes [2]. First, the image is cropped around the green rectangle and it is transformed to 100×100 gray-scale using pixel-averaging. Subsequently, we subtract the amplitude mean for each image row and calculate the image covariance matrix described by $\Sigma = XX^T$ with X as the gray-scale image. To implement the PCA we developed a single value decomposition for non-singular Σ and then with these eigenvectors matrix S we calculate the ZCA_{white} following Equation 1

$$ZCA_{white} = \lambda tr \left(\frac{1}{\sqrt{(tr(S) + \epsilon)}} \right) \Sigma \lambda^T X \tag{1}$$

We set a fudge factor $\epsilon = 0.1$ for avoiding an extra blurring level in the resulting image, thus obtaining a similar representation shown in Figure 1b. λ is the rotation for the covariance matrix singular value decomposition. The resulting faces were giving as input to the ConvNets using an a-priori cropping and flipping as we explain below.

3.1.2 dlib Face contour landmarks. 100×100 Color images were either processed using an openCV library extension dlib [16] for efficient face contours detection and head pose alignment. Dlib library uses a Histogram of Oriented Gradients (HOG) and a Linear SVM classifier to infer the position of 68 x-y landmark points distributed among the allocated face [10].

These landmark points are concatenated and overlayed in the color image as Figure 1a shows. To extract and preserve only the black contour from the subsequent gray-scale image we applied a amplitude mask saturating any pixel that has a value above 2, thus obtaining a representation shown in Figure 1c. In this particular case, we use the dlib contours for two purposes: First, we use the landmarks to align the faces using the eyes center similar to [27]. ZCA_{white} and isotropic smoothing images were both aligned before the frame-level training. Second, we use the representation from Figure 1c as an input for the ConvNets normalizing the images from uint8 to double. Previous implementations [26] have proposed the evaluation and detection of face landmarks using ConvNets a-priori, in our particular case we use dlib landmarks as auxiliary features in the analysis in order to explore how the frame-level and video-level accuracies are correlated with the dlib landmarks as features.

3.1.3 Frame-level Classification: ConvNets. Each image representation ZCA, dlib, and isotropic are resized from 100×100 to 48×48 for being consistent with the input dimensionality expected by the Kahou et. al [9] ConvNets. The ConvNets architecture is shown in Figure 2. There the input image is decomposed from a 48×48 to a 5×5 resolution.

We follow the model and the parameters established by *Emonets* [8] with some changes. The size of each convolutional filter and the max/average pooling was set in 5×5. The first section of the ConvNets is composed of 64 filters except for the last section of convolution and pooling that is connected with the *softmax* layer composed of 128 filters. This ConvNets was trained with a learning rate of 0.0003 for modifying filters weights and biases, and a weight decay per epoch equal to 0.004. All the convolutional+pooling blocks have ReLU activation function layers as outputs.

The frame-level training task starts with the cropping and horizontal flipping of all the training face images. Specifically, the 48×48 input image is cropped in a flipped 40×40 random patch per epoch being consistent with the same random patch among the training batches.

Each fold in the cross-validation (5-Fold) includes only the frames that belong to a particular EESS video, in this experiment we did not include any extra dataset for training, thus each test set is composed of the frames of a particular video without any overlap with the other folds belonging to the training set; thus assuring the consistency with the EESS video annotation.

We set the maximum number of training epochs to 200 for each cross-validation, however we set an early-stopping when the training error in the frame-level is below 5% and the video-level accuracy for the test is set above 50%. For the frame-aggregation (refer section 3.1.4) we group the probabilities in output of the *softmax layer* per epoch.

The frame-level performances reported below such as top5err, top1err, F1, and accuracy are evaluated per training epoch and averaged for all the 5 folds.

3.1.4 Frame-Aggregation: Video-Classification. The frame-aggregation is similar to the process established by Kahou et. al [9]. For each training epoch the probabilities calculated by the *softmax* layer were grouped in 10 bins for the training and the test set.

These probability groups are sorted depending on the order of the frames in the corresponding video. Thus, per video and per batch we average the seven corresponding probabilities scores obtaining 70 features per each video at the end of the ConvNets.

Our main contribution in the frame-aggregation is changing the radial kernel SVM described in the baseline system to a kNN or a

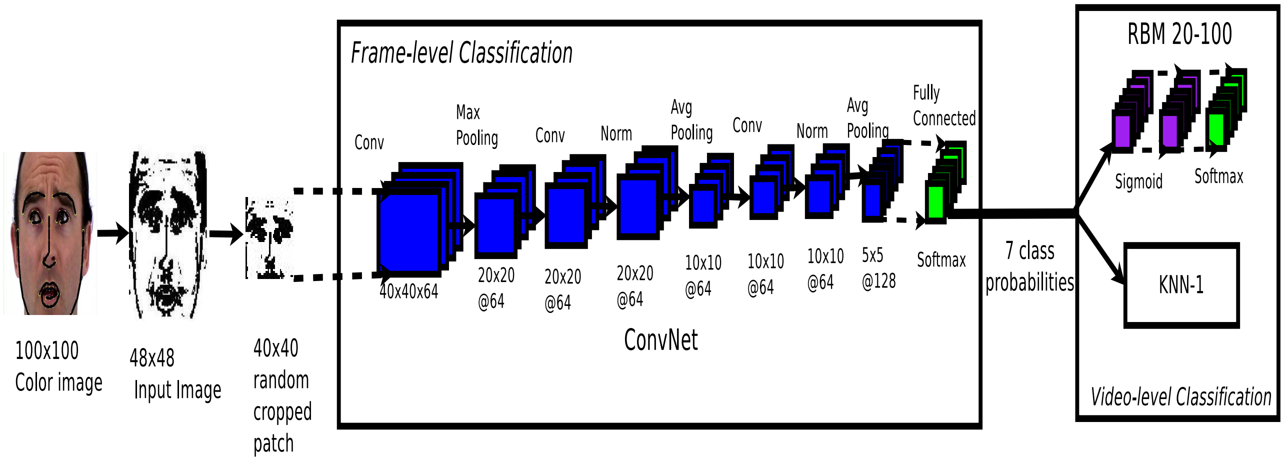


Figure 2: Block diagram of the face-based emotion recognition pipeline. Input image is transformed first to gray-scale representation to reduce RGB channels from 3 to 1. Subsequently, each detected face-image was resized to 48x48 and normalized using ZCA whitening. The ZCA_{white} image is cropped in 40x40 random patches per each training epoch and flipped horizontally before it is given as an input in the Kahou et. al [9] ConvNet. Thus, before video-level classification we extract the seven probabilities scores assigned for each emotion category + neutral, averaging the scores in 10 groups of frames per video.

two-layer RBM network classifiers. Specifically, we set the number of nearest-neighbors to 1; thus referring to this particular classifier as KNN-1.

On the other hand, the RBM is a 20-100 fully-connected greedy network with *sigmoid* activation functions. We train the classifier following [1] with 10 epochs of Contrastive Divergence (CD-1) pretraining and 200 iterations of fine-tunning. For the pre-training we use a fixed learning rate of 0.1, and for the fine-tunning we start the initial learning rate in 0.2 using a rate decay of 0.001 per training epoch.

The video-level predicted labels are obtained through a *softmax* output layer in the case of the RBM, and the inferred index group for the KNN-1. Figure 2 shows the video-level classification block at the end of block diagram.

3.2 Audio-based

Audio-based pipeline is composed of the following steps: Each .mp3 file in the EESS repository were transformed to .wav using *ffmpeg*. The sampling frequency and the bits per sample were preserved for each file being 44 kHz and 16 respectively. Figure 3 shows the complete block diagram of the audio-based emotion recognition pipeline.

3.2.1 Audio Descriptors: Feature Sets. We use openSMILE for extracting and concatenate four different features set per voice instance: (1) The first feature set is calculated from *prosodicAcf.conf* file. This feature set is composed of the voice probability, the F_0 value calculated from cepstrum, and the loudness value estimation. (2) A set extracted using the *MFCC12_E_D_A.conf* file. This set is composed of 36 MFCC features such as 12 general MFCC coefficients calculated from 12 Mel-bands, other 12 delta coefficients, and other 12 calculated by the acceleration method. (3) The Interspeech 2009 Emotion Challenge feature set [18] described in the *IS09_emotion.conf*. This set is composed of 384 features including basic stats (e.g. max, min, skewness, stddev, kurtosis) and contour slopes from 12 smoothed low level descriptors calculated for each instance/utterance. (4) The acoustic descriptors from the *emobase.conf* set including some features such as amplitude max/min, range, arithmetic mean, linear regression coefficients, linear and quadratic errors, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges. We exclude the features that are in common between the sets.

The final feature set is composed of 1230 features as a result of concatenating all the features described above. The final feature vector is obtained averaging the low-level features calculated among the 25ms audio segments (10ms overlap) indicated in the Table 1 per instance.

To evaluate our tests in the instance level classification we used the baseline proposed by Sagha et. al [17], specifically the EESS english modality in which only the Interspeech 2009 Emotion Challenge feature set was included.

3.2.2 Instance Level Classification: RBM/kNN. In this evaluation we work consistently with audio instance/utterance annotation. 193 voice instances were labeled with Ekman's basic emotions as we explained above. We assign these seven labels among the high/low arousal/valence levels following the baseline Sagha et. al [17] and the re-mapping of the circumplex axis denoted as the Geneva wheel [20].

Sagha et. al assigned EESS basic emotions labels such as *angry, afraid, disgusted* and *sad* to the negative (low) valence level, and *happy, surprised,* and *neutral* to positive (high) valence. Likewise, *afraid, sad, neutral* and *disgusted* were assigned to negative (low) arousal level, *angry, happy,* and *surprised* were assigned to positive (high) arousal.

Figure 3 shows the two classifiers for the audio instance-level classification: A 50-50 two-layer RBM with sigmoidal activation functions and a *softmax* output layer. This RBM classifier is trained

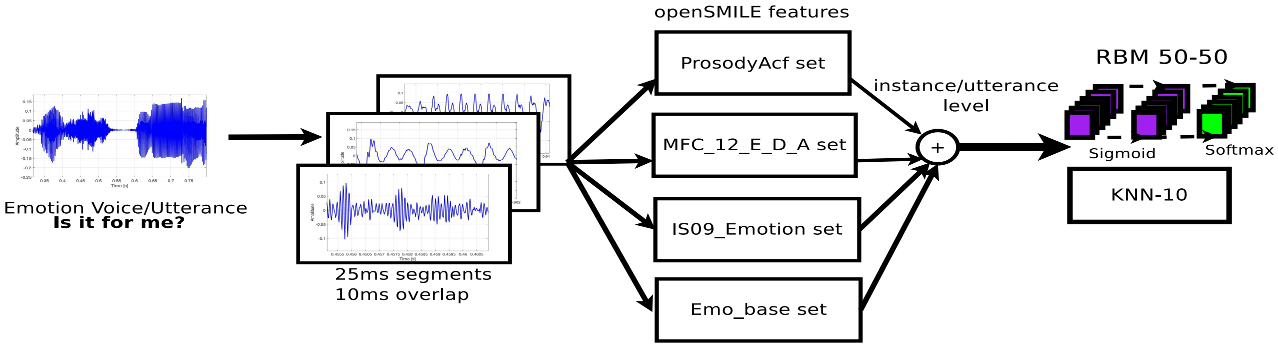


Figure 3: Block diagram of the audio-based emotion recognition pipeline. Input instance/utterance is segmented in 25ms length 10ms overlap. For each audio segment we extracted the *prosodicAcf, MFCC12_E_D_A, IS09_emotion,* and *emobase* features set, averaging the low-level descriptors per instance and feeding the RBM 50-50 and KNN-10 classifiers for the instance-level classification.

with 10 CD-1 pre-training epochs and a fixed learning rate of 0.01, and a 1200 fine-tunning iterations with a initial learning rate of 0.1 and a rate decay of 0.05 per epoch. We normalize the instances for the RBM classifier training based on Equation 2 in which X_k is the instance/example k_{th}.

$$X_k^{norm} = \frac{X_k - min(X_k)}{max(X_k) - min(X_k)} \tag{2}$$

On the other hand, we set a kNN classifier with 10 nearest neighbors (KNN-10). The baseline classifier is a SVM linear kernel as Sagha et. al [17] describe. We evaluate the performance of these classifiers in a 5-Fold cross-validation modality.

4 RESULTS

The face-based and audio-based emotion recognition pipelines are evaluated using a 5-Fold cross-validation. As we mentioned above, both the frame-level classification and video-level classification for the face-based pipeline are evaluated using the 5-fold cross-validation at the end of the training or when the early-stopping criterion is met.

4.1 Face-based

Average frame-level and video-level performances are shown in Tables 3 and 2 respectively. The performances for the frame-level are evaluating using the Accuracy (Acc.) to avoid the incidence of 0 true positives in precision and recall scores for certain classes such as *surprised* and *disgusted* with few face detected frames. Accuracy values are calculated for video-level classification. Precision (Pr), Recall (Re), and F1 scores are reported in the confusion matrices of Figures 6 and 7.

Changing the input representations such as ZCA_{white} and dlib shows an increased accuracy in comparison to the isotropic smoothing baseline in the video-level classification. Likewise, changing the classifier from RBF SVM to RBM 20-100 and/or KNN-1 shows an increase in accuracy; thus supporting that changing the separability of image features is an important step for obtaining a better emotion recognition rate in a multi-class data.

Specifically, the KNN-1 classifier obtains the best video-level performance for both representations – ZCA_{white} and dlib. Some differences are observed for the dlib performances in Figures 7a, 7b, and 7c. The confusion matrices for dlib representation including the KNN-1 results are more unbalanced in comparison to ZCA_{white} (see Figures 6a, 6b, and 6c). A clear evidence of this measure is given by the increased difference between the Precision and Recall values in dlib confusion matrices in comparison to the ZCA_{white}. This suggests that the multiple PCA eigenvalues extracted from a grayscale image as a whole are more discriminant than the equivalent representation from the face contour landmarks only. These results can be also supported with the considerable differences between Precision and Recall values for dlib confusion matrices that are not presented in the ZCA_{white} case.

Surprisingly, the dlib representation yields slightly better performance for the frame-level classification in comparison to ZCA_{white}, as can be observed in Table 3. We do not find any significant differences between the new representation and the baseline; however, for all the cases RBM 20-100 and KNN-1 are outperforming the baseline methodology.

Top1errors and top5errors for frame-level classification are reported in Figure 4. Baseline system errors show increased values in comparison with dlib and ZCA_{white} errors, being consistent with confusion matrices shown in Figures 6 and 7. Confusion matrices for baseline methodology using the isotropic smoothing are shown in Figures 5a, 5b, and 5c.

Table 2: Video-level average accuracies for the 5-fold cross-validation. Accuracies are calculated with a sum off all the member of diagonal over the sum of all the members of matrix for each fold. Results in italics represent the Kahou et. al baseline [9]

Video-Level	ZCA_{white}	dlib	Isotropic
SVM RBF	0.482±0.022	0.511±0.022	*0.309±0.172*
RBM 20-100	0.534±0.022	0.534±0.032	0.334±0.125
KNN-1	**0.541±0.137**	**0.537±0.153**	0.473±0.051

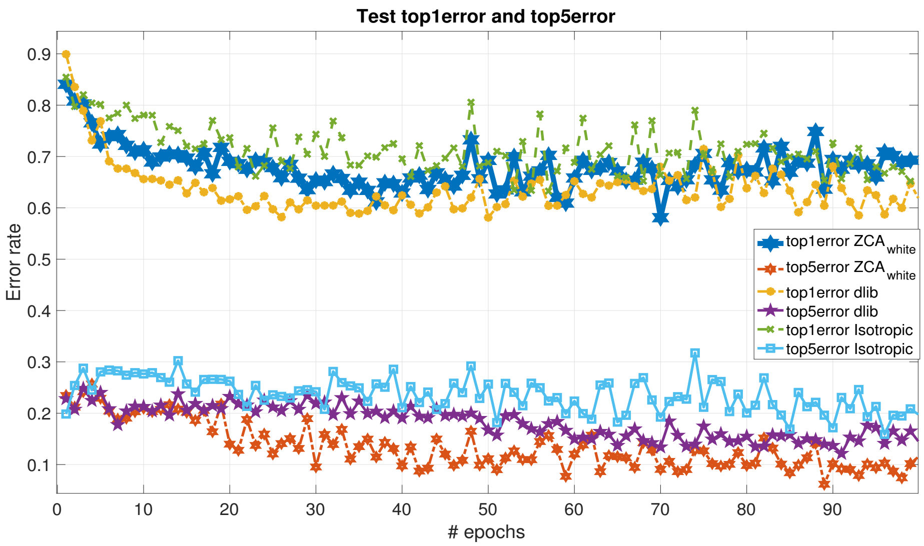
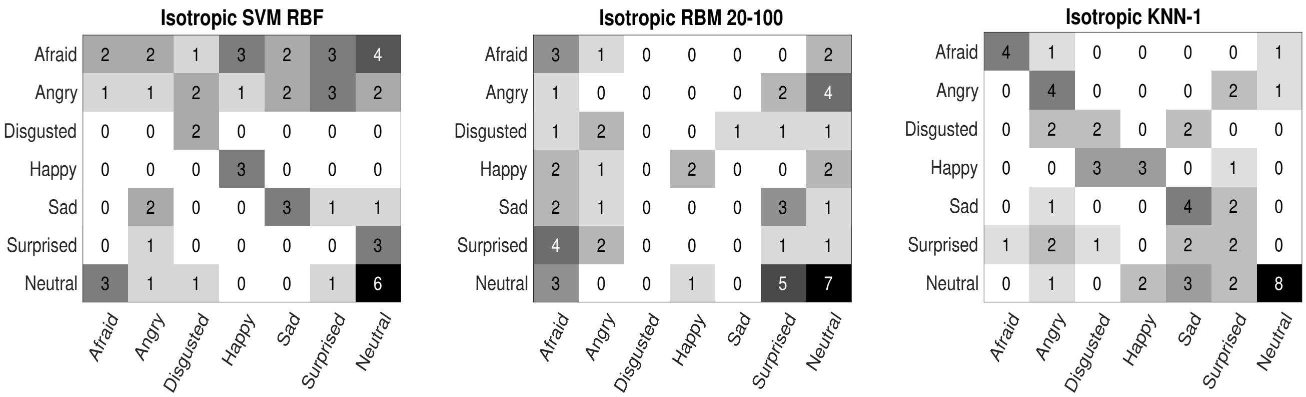


Figure 4: Top1errors and top5errors for Isotropic (baseline), ZCA_{white}, and dlib contours representations in frame-level classification modality.



Isotropic SVM RBF

	Afraid	Angry	Disgusted	Happy	Sad	Surprised	Neutral
Afraid	2	2	1	3	2	3	4
Angry	1	1	2	1	2	3	2
Disgusted	0	0	2	0	0	0	0
Happy	0	0	0	3	0	0	0
Sad	0	2	0	0	3	1	1
Surprised	0	1	0	0	0	0	3
Neutral	3	1	1	0	0	1	6

(a) SVM RBF Acc=0.3756, Pr=0.3812, Re=0.3731, F1=0.3792

Isotropic RBM 20-100

	Afraid	Angry	Disgusted	Happy	Sad	Surprised	Neutral
Afraid	3	1	0	0	0	0	2
Angry	1	0	0	0	0	2	4
Disgusted	1	2	0	0	1	1	1
Happy	2	1	0	2	0	0	2
Sad	2	1	0	0	0	3	1
Surprised	4	2	0	0	0	1	1
Neutral	3	0	0	1	0	5	7

(b) RBM, Acc=0.4012,Pr=0.4020,Re=0.3914,F1=0.3999

Isotropic KNN-1

	Afraid	Angry	Disgusted	Happy	Sad	Surprised	Neutral
Afraid	4	1	0	0	0	0	1
Angry	0	4	0	0	0	2	1
Disgusted	0	2	2	0	2	0	0
Happy	0	0	3	3	0	1	0
Sad	0	1	0	0	4	2	0
Surprised	1	2	1	0	2	2	0
Neutral	0	1	0	2	3	2	8

(c) KNN-1, Acc=0.4211, Pr=0.4001, Re=0.4131 F1=0.4045

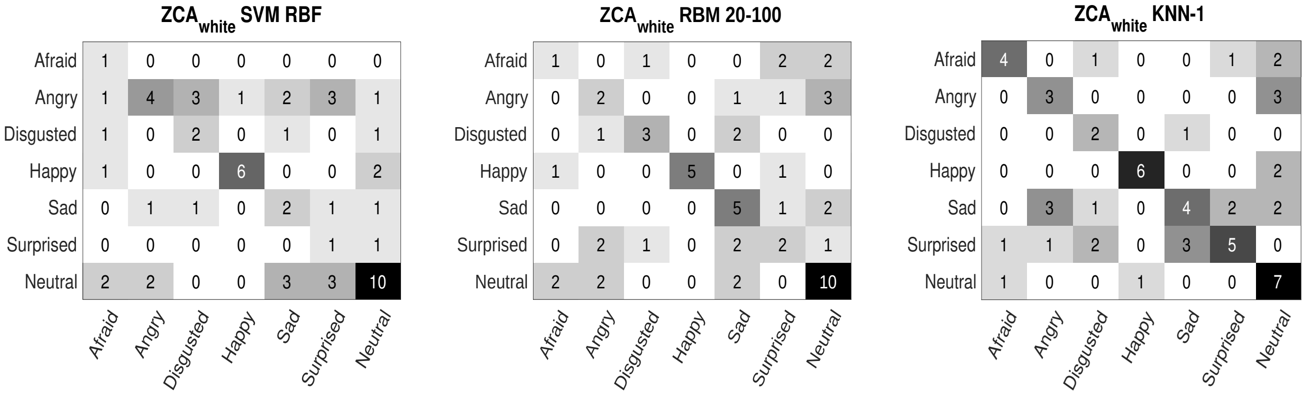
Figure 5: Confusion matrices for Isotropic smoothing input representation and video-level classification: Figure 6a shows the confusion matrix for the SVM RBF classifier similar to baseline Kahou et. al [9], Figure 6b shows the confusion matrix for RBM 20-100 classifier, and 6c the confusion matrix for KNN-1. The numbers labeled for each cell of these confusion matrices represent the number of videos that are predicted as the emotion category in the y-axis giving by the real emotion category in the x-axis.

Table 3: Frame-level average accuracies for the 5-fold cross-validation calculated after the *softmax* layer in the output of the ConvNet. A *softmaxloss* operator was used to infer the final label per frame. Results with italics font represent the Kahou et. al baseline [9]

Frame-Level	ZCA_{white}	dlib	Isotropic
Acc	0.401±0.158	**0.431±0.095**	*0.312±0.131*

4.2 Audio-based

The average audio-based classification results are shown in Table 4. We report Acc, Pr, Re, F1 score values for all the classifiers explained above: The linear SVM proposed by the baseline, the two layer RBM 50-50, and the KNN-10. Results show that using a type of classifier such as RBM 50-50 and/or KNN-10 is more accurate than using linear kernel SVM. For both arousal and valence high/low levels classifications, the RBM 50-50 shows an increased and significantly better performances (Pr, Re, and Acc) in comparison to the baseline and KNN-10 classifier. As we expected, valence high/low level

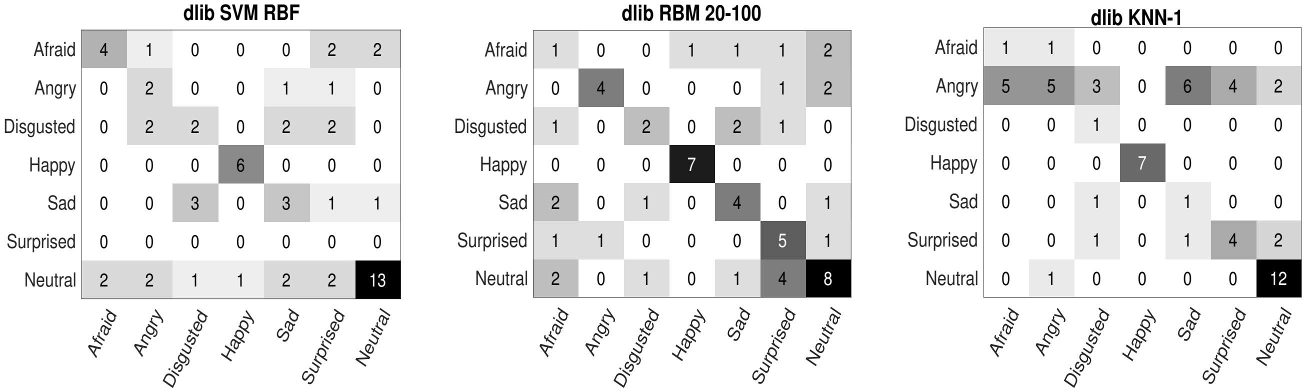


(a) SVM RBF Acc=0.4310, Pr=0.4110, Re=0.3946, F1=0.4026

(b) RBM, Acc=0.4828,Pr=0.4524,Re=0.4848,F1=0.4680

(c) KNN-1, Acc=*0.5345*, Pr=*0.5636*, Re=*0.5497*, F1=*0.5565*

Figure 6: Confusion matrices for ZCA_{white} input representation and video-level classification: Figure 6a shows the confusion matrix for the SVM RBF classifier, Figure 6b shows the confusion matrix for RBM 20-100 classifier, and 6c the confusion matrix for KNN-1.



(a) SVM Acc=0.5101, Pr=0.5234, Re=0.4758, F1=0.4958

(b) RBM, Acc=0.5345, Pr=0.5437, Re=0.5281, F1=0.5358

(c) KNN-1, Acc=*0.5345*, Pr=*0.6604*, Re=*0.4889*, F1=*0.5610*

Figure 7: Confusion matrices for dlib contour input representation and video-level classification: Figure 7a shows the confusion matrix for the SVM RBF classifier, Figure 7b shows the confusion matrix for RBM 20-100 classifier, and 7c the confusion matrix for KNN-1.

Table 4: Audio-based: instance level classification average results for the 5-Fold cross-validation. Values with (*) are $p < 0.05$ inter-classifier

Instance/Level	Arousal				Valence			
	Pr	Re	F1	Acc	Pr	Re	F1	Acc
Sagha et. al [17]	0.526±0.065	0.527±0.066	0.526±0.067	0.525±0.073	0.597±0.033	0.596±0.035	0.597±0.042	0.585±0.043
KNN-10	0.627±0.056	0.615±0.052	0.619±0.055	0.631±0.054	0.688±0.048*	0.698±0.045*	0.693±0.051*	0.697±0.052
RBM 50-50	**0.622±0.091***	**0.651±0.098***	**0.635±0.092***	**0.640±0.093**	**0.741± 0.022***	**0.735±0.034***	**0.738±0.045***	**0.742±0.052***

classification yields better performances, which is consistent with other studies on emotion instance/utterance levels classification [21].

5 CONCLUSIONS

In this face-based and audio-based emotion recognition experiments we identified that the ConvNets represent a fundamental and primary type of discriminative approach for classifying affective faces. The inclusion of multiple input representations such as ZCA_{white} and dlib, and adding kNN and RBM classifiers in the frame-aggregation task offers a robust enough system for affective faces recognition without combining other systems' decisions and extra multimodal approaches. For more efficiency in terms of the classification, the kNN shows an increased performance independently of the type of input representation. In terms of voice expressions, the inclusion of multiple standardized datasets yields a more accurate pipeline than using linear classifiers and short length datasets for identifying arousal/valence levels. RBM and kNN classifiers are again more indicated for this task than the SVM.

In the near future the inclusion of context-aware datasets such as EU-emotion Stimulus represent an open door for new implementations of ConvNets fed by pictorial and/or audio stimuli in a combination with cognitive and/or neurophysiological responses.

REFERENCES

[1] Yoshua Bengio et al. 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning* 2, 1 (2009), 1–127.

[2] Adam Coates and Andrew Y Ng. 2011. Selecting receptive fields in deep networks. In *Advances in Neural Information Processing Systems*. 2528–2536.

[3] Morena Danieli, Giuseppe Riccardi, and Firoj Alam. 2015. Emotion unfolding and affective scenes: A case study in spoken conversations. In *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*. ACM, 5–11.

[4] Sidney K D'Mello, Scotty D Craig, and Art C Graesser. 2009. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology* 4, 3-4 (2009), 165–187.

[5] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 467–474.

[6] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[7] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[8] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10, 2 (2016), 99–111.

[9] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. 2013. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 543–550.

[10] Mahir Faik Karaaba, Olarik Surinta, LRB Schomaker, and Marco A Wiering. 2016. Robust face identification with small sample sizes using bag of words and histogram of oriented gradients. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 582–589.

[11] Shengcai Liao, Anil K Jain, and Stan Z Li. 2016. A fast and accurate unconstrained face detector. *IEEE transactions on pattern analysis and machine intelligence* 38, 2 (2016), 211–223.

[12] Heather O'leary, Juan M. Mayor Torres, Walter E. Kaufmann, and Mustafa Sahin. 2017. Classification of Respiratory Disturbances in Rett Syndrome patients using Restricted Boltzmann Machine. In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE.

[13] Helen O'Reilly, Delia Pigat, Shimrit Fridenson, Steve Berggren, Shahar Tal, Ofer Golan, Sven Bölte, Simon Baron-Cohen, and Daniel Lundqvist. 2016. The EU-emotion stimulus set: a validation study. *Behavior research methods* 48, 2 (2016), 567–576.

[14] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 439–448.

[15] Peter Robinson and Tadas Baltrušaitis. 2015. Empirical analysis of continuous affect. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 288–294.

[16] Anwar Saeed, Ayoub Al-Hamadi, and Heiko Neumann. 2017. Facial point localization via neural networks in a cascade regression framework. *Multimedia Tools and Applications* (2017), 1–23.

[17] Hesam Sagha, Pavel Matejka, Maryna Gavryukova, Filip Povolny, Erik Marchi, and Björn Schuller. 2016. Enhancing multilingual recognition of emotion in speech by language identification. *Interspeech 2016* (2016), 2949–2953.

[18] Björn W Schuller, Stefan Steidl, Anton Batliner, et al. 2009. The INTERSPEECH 2009 emotion challenge.. In *Interspeech*, Vol. 2009. 312–315.

[19] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. 2005. Multimodal approaches for emotion recognition: a survey. In *Electronic Imaging 2005*. International Society for Optics and Photonics, 56–67.

[20] Ingo Siegert, Ronald Böck, Bogdan Vlasenko, David Philippou-Hübner, and Andreas Wendemuth. 2011. Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self assessment manikins. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 1–6.

[21] Mohammad Soleymani, Guillaume Chanel, Joep JM Kierkels, and Thierry Pun. 2008. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. Ieee, 228–235.

[22] Vitomir Štruc and N Pavešic. 2011. Photometric normalization techniques for illumination invariance. *Advances in Face Image Analysis: Techniques and Technologies* (2011), 279–300.

[23] Bo Sun, Liandong Li, Xuewen Wu, Tian Zuo, Ying Chen, Guoyan Zhou, Jun He, and Xiaoming Zhu. 2016. Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild. *Journal on Multimodal User Interfaces* 10, 2 (2016), 125–137.

[24] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Felix Weninger, Björn Schuller, and Gerhard Rigoll. 2012. Fully automatic audiovisual emotion recognition: Voice, words, and the face. In *Speech Communication; 10. ITG Symposium; Proceedings of*. VDE, 1–4.

[25] Jingjie Yan, Wenming Zheng, Qinyu Xu, Guanming Lu, Haibo Li, and Bei Wang. 2016. Sparse Kernel Reduced-Rank Regression for Bimodal Emotion Recognition From Facial Expression and Speech. *IEEE Transactions on Multimedia* 18, 7 (2016), 1319–1329.

[26] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence* 38, 5 (2016), 918–930.

[27] Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2879–2886.