Video-Based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion

Heysem Kaya[a,*], Furkan Gürpınar[b], Albert Ali Salah[b]

[a]*Department of Computer Engineering, Çorlu Faculty of Engineering*
Namık Kemal University, 59860, Çorlu, Tekirdağ, TURKEY
[b]*Department of Computer Engineering*
Boğaziçi University, 34342, Bebek, İstanbul, TURKEY

Abstract

Multimodal recognition of affective states is a difficult problem, unless the recording conditions are carefully controlled. For recognition "in the wild", large variances in face pose and illumination, cluttered backgrounds, occlusions, audio and video noise, as well as issues with subtle cues of expression are some of the issues to target. In this paper, we describe a multimodal approach for video-based emotion recognition in the wild. We propose using summarizing functionals of complementary visual descriptors for video modeling. These features include deep convolutional neural network (CNN) based features obtained via transfer learning, for which we illustrate the importance of flexible registration and fine-tuning. Our approach combines audio and visual features with least squares regression based classifiers and weighted score level fusion. We report state-of-the-art results on the EmotiW Challenge for "in the wild" facial expression recognition. Our approach scales to other problems, and ranked top in the ChaLearn-LAP First Impressions Challenge 2016 from video clips collected in the wild.

Keywords: EmotiW, emotion recognition in the wild, multimodal fusion, convolutional neural networks, kernel extreme learning machine, partial

*Corresponding author. This is the uncorrected author proof of the accepted paper. Please cite this paper as: "Kaya, H., F. Gurpinar, A.A. Salah, Video-Based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion, Image and Vision Computing, 2017".

Email addresses: hkaya@nku.edu.tr (Heysem Kaya), furkan.gurpinar@boun.edu.tr (Furkan Gürpınar), salah@boun.edu.tr (Albert Ali Salah)

1. Introduction

Audio- and video-based emotion recognition in the wild is challenging, because of noise, large idiosyncratic variances, and sensor-related differences. This paper describes a multimodal approach for audio-visual emotional expression recognition. Our approach processes complementary features with summarizing functionals, classifies these descriptors with a set of least squares based classifiers, and combines their output with decision level fusion.

The Emotion Recognition in the Wild (EmotiW) Challenge provides out of laboratory data -Acted Facial Expressions in the Wild (AFEW)-, collected from videos that mimic real life, and poses a very difficult and realistic problem [5, 6, 7].

This paper extends our contribution to the EmotiW 2015 Challenge [22], which was ranked the second in the official competition, in terms of the introduced approach, level of detail, and amount of experimental validation. [22] proposed the combination of multiple visual features with audio over summarizing functionals. Here, we extend this framework by employing Deep Convolutional Neural Network (CNN) features, as well as an investigation of suitable transfer learning strategies that can be used with CNNs. We show that incorporating multiple registration schemes into the model helps transfer learning. We also illustrate the success of the approach on two additional in the wild datasets, namely, the ChaLearn-LAP 2016 First Impressions Challenge database, and the EmotiW 2016 corpus, which is an extension of the 2015 corpus. Our system achieves the best official result in the ChaLearn-LAP First Impressions Challenge. We further validate our visual features using two widely used lab-controlled corpora, namely the Extended Cohn-Kanade dataset (CK+) [33] and MMI [50].

The remainder of this paper is organized as follows. In the next section we discuss related work on video-based facial expression recognition in naturalistic settings. Section 3 describes the proposed approach. In Section 4 we briefly introduce the corpora and baseline feature sets. In Section 5, we give experimental results. Finally, Section 6 concludes the paper and summarizes our findings.

2. Related Work

Facial affective displays in real life involve subtle changes, in contrast to the exaggerated displays in posed expressions [60]. Therefore, facial expression recognition "in the wild" poses much greater challenges, and the recognition accuracies are invariably much lower. Furthermore, controlled illumination and pose settings in the lab are not available in the wild, which adds purely physical challenges to the problem, and causes issues starting from the detection phase [59].

We focus here only on video-based approaches, which implies both dynamic and multimodal information. Multimodal approaches to emotional expression recognition leverage both paralinguistic audio cues, as well as the synchronization between modalities to improve robustness. Early multimodal approaches focused on coarse affective states (e.g. positive and negative states), because data collection and annotation for natural or spontaneous scenarios was difficult (see [60] for a comprehensive survey of earlier approaches, and [55] for available databases).

Together with developments of multimodal expression research in more natural conditions, it became obvious that the temporal dynamics of expressions contained rich information [61]. Krumhuber et al. previously used a 2-person trust game setting to illustrate that facial dynamics significantly affect social judgments [27]. Research on smile dynamics illustrated fake and genuine enjoyment smiles have quite distinctive dynamics, and that humans are sensitive to such cues [9]. However, automatic classification of subtle distinctions requires controlled and high-quality data, which cannot be assumed for naturalistic settings. The Emotion Recognition In The Wild (EmotiW) Challenge, which started in 2013, initiated an effort to overcome challenges of data collection, annotation, and standardized testing for multimodal and dynamic emotion recognition in the wild. The challenge used the AFEW corpus, which was collected from movies with close-to-real-world conditions [7]. It was quickly understood that for image-based processing, it was particularly important to get good face detection and alignment, as well as rejection of non-face images [29]. In the top performing system of the first challenge, visual bag of words features, gist features, paralinguistic audio features and Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) features obtained from aligned faces were each processed by RBF kernels, and combined with a multi kernel support vector machine [45]. The contribution of audio was very small (its normalized weight was 1.8, as opposed to 50.1

for visual HOG features). The weight for features extracted over the entire scene was even smaller than the audio.

In the 2013 EmotiW Challenge, Liu et al. treated images from each video clip as a set and represented them as points in a Grassmanian manifold [29]. Audio was separately modeled and fused linearly with video-based classifiers. One year later, they extended their approach by representing each video clip using three kinds of image set models (i.e. linear subspace, covariance matrix, and Gaussian distribution) respectively. As features, they used HOG, dense SIFT, as well as features extracted from the 9-layer deep convolutional neural network (CNN) pre-trained with ImageNet [26]. This system achieved the best test set accuracy of 50.37%, and represented a significant improvement over the systems submitted to the 2013 challenge. The first runner-up system proposed a hierarchical voting classifier, and showed that the addition of audio features had a small, but persistent impact (from 44.72% to 47.17%) [47]. Multiple kernel learning and SVM were popular in the submissions, but the top system used a partial least squares based classifier.

In both years, surprise and disgust were the most difficult classes to recognize, whereas happy, angry, and neutral had relatively high accuracies. By 2015, it was clear that deep neural networks harbored great potential for describing features for non-controlled settings. Their main advantage was a resistance to alignment issues, as well as to noise. In their submission to the 2015 EmotiW Challenge, Ebrahimi Kahou et al. used a Recurrent Neural Network (RNN) combined with a CNN to model the expression dynamics [11]. Their results suggested that temporal integration was better than averaging per-frame decisions. They also combined feature level and decision level fusion, noting that such powerful and complex models were prone to overfitting the training set, as evidenced by the great discrepancy between training and validation set accuracies (98.3% vs. 26.6% for activity modality). Their solution was to adopt early stopping during training. The usage of rectified linear units and dropout strategies are also commonly employed to control overlearning in CNNs [26]. These approaches will directly benefit from increases in the amount of training data. The top performing system used a CNN model as well, but it fused the CNN model output with one audio and three linear SVMs trained with AU-aware facial feature relations on two face scales [56].

The number of examples in the challenge database are increased to 1645 videos (AFEW 5.0) during the years, but the database still has some shortcomings. Disgust, sadness and surprise were found to be really difficult to

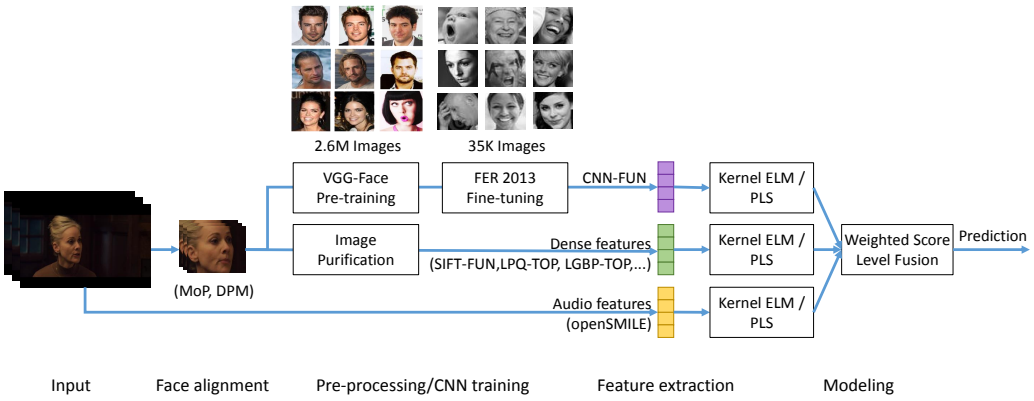

Figure 1: Overview of the proposed approach.

reliably classify on this database. There are few other initiatives to collect and evaluate in the wild data. The AM-FED corpus contains 242 videos of people watching commercials on a computer [35]. The RECOLA database presents 27 videos that are continuously annotated in time and space for arousal and valence dimensions [42]. A very recent database is collected by Zafeiriou et al. from YouTube, containing around 500 videos and annotated with regards to valence and arousal [58].

3. Proposed Method

The proposed approach is illustrated in Figure 1, and contains the detection of the face, its alignment (registration) with a fixed model or with a set of facial landmarks, and subsequent feature extraction and classification. We use two separate alignment options, which greatly improves processing in the CNN. The CNN pre-training and fine tuning stages are carried out prior to processing of the target emotional video corpus. Audio is separately processed and fused at the decision level. We describe each phase in the pipeline separately.

3.1. Preprocessing and Image Purification

Facial registration is one of the most important steps in face image processing. To guard against registration errors, we applied a purification step in processing faces of the EmotiW dataset. To deal with rotated faces and

false positives in face detection, we use a principal component analysis (PCA) based method to automatically remove false detections, as shown to be effective in [24, 29, 47]. The idea is to measure the mean reconstruction error per image, after projecting the images to the PCA space and back. We discard the frames with a high reconstruction error, as these are probably poorly detected or poorly aligned images. We remove videos that have less than three valid images from the validation set. For the sequestered test set, all instances are retained. After purification, images are resized to 64×64 pixels. We manually remove poorly aligned images from the training set, which improves the quality of training.

3.2. Visual Descriptors

We extract and compare Scale Invariant Feature Transform (SIFT) [32], Histogram of Oriented Gradients (HOG) [3], Local Phase Quantization (LPQ) [17, 19], Local Binary Patterns (LBP) [37] and its Gabor extension (LGBP), as well as Deep Convolutional Neural Network (CNN) based visual descriptors. For LPQ, LBP, and LGBP, the Three Orthogonal Planes (TOP) extension is popularly used in video modeling [62]. This extension applies the relevant descriptor on XY, XT and YT planes (where T represents time) independently, and concatenates the resulting histograms. Also in our implementation, we divide the video into two equal length volumes over the time axis and extract spatio-temporal TOP features from each volume to further enhance temporal modeling. In the following, we provide brief explanations of LPQ, LBP and LGBP descriptors.

3.2.1. Local Phase Quantization

The LPQ features are computed by taking 2-D Discrete Fourier Transform (DFT) of M-by-M neighborhoods of each pixel in the gray scale image. 2D-DFT is computed at four frequencies $\{[a,0]^T, [0,a]^T, [a,a]^T, [a,-a]^T\}$ with $a = 1/M$, which correspond to four of eight neighboring frequency bins centered at the pixel of interest. The real and imaginary parts of resulting four complex numbers are separately quantized using a threshold of zero, which gives an eight bit string. This string is then converted into an integer value in the range of [0-255]. The pixel based values are finally converted into a histogram of 256 bins. In order to partly provide structural information of facial features, the face is divided into non-overlapping regions and an LPQ histogram is computed per region and concatenated for the final image representation [17]. LPQ variants are successfully used in audio-visual emotion

recognition problems [20] and served as baseline in the Audio-Visual Emotion Challenge (AVEC) 2013 [52].

3.2.2. Local Binary Patterns

After face alignment and conversion to gray scale, LBP computation amounts to finding the sign of difference with respect to a central pixel in a neighborhood, transforming the binary pattern into an integer and finally converting the patterns into a histogram. Uniform LBP clusters 256 patterns into 59 bins, and takes into account occurrence statistics of common patterns [37]. As in LPQ, the face is divided into non-overlapping regions and an LBP histogram is computed per region.

3.2.3. Local Gabor Binary Patterns

In LGBP, the images are convolved with a set of 2D complex Gabor filters to obtain Gabor-pictures, then LBP is applied to each Gabor-picture (or Gabor-video). A 2D complex Gabor filter is the convolution of a 2D sinusoid (carrier) having phase P, spatial frequencies u_0 and v_0 with a 2D Gaussian kernel (envelope) having amplitude K, orientation ρ, and spatial scales a and b. For simplicity, and in line with [1], we take $a = b = \sigma, u_0 = v_0 = \phi$ and $K = 1$ to obtain:

$$G(x, y) = e^{-\pi\sigma^2((x-x_0)^2_\rho+(y-y_0)^2_\rho)}e^{j(2\pi\phi(x+y)+P)}, \tag{1}$$

where the subscript ρ stands for a clockwise rotation operation around reference point (x_0, y_0) such that:

$$(x - x_0)_\rho = (x - x_0)cos\rho + (y - y_0)sin\rho \tag{2}$$
$$(y - y_0)_\rho = -(x - x_0)sin\rho + (y - y_0)cos\rho. \tag{3}$$

Note that the effect of the phase is canceled out, since only the magnitude response of the filter is used for the descriptor.

For LPQ, LBP, and LGBP, the Three Orthogonal Plane (TOP) variants are implemented. This extends these image-based features into the spatiotemporal domain. The LGBP-TOP feature was used as a baseline feature in the AVEC 2014 challenge [51].

3.2.4. CNN Features

Because of the recent success of deep convolutional neural network (CNN) approaches, we integrate pre-trained CNN models into our method. The

main disadvantage of CNNs is that they require very large amounts of data in order to avoid over-fitting. One way of getting around this problem is using pre-trained CNN models for visual feature extraction and use transfer learning to adapt the models to the particular application setting [25, 57, 36]. The winners of EmotiW 2013 [21] and 2014 [31] video challenges employed pre-trained CNN models, while Yao et al. trained a 5-hidden-layer CNN directly from the video frames for the EmotiW 2015 challenge [56]. The best fusion system reported in [56] had a test set accuracy of 53.80%, while our official challenge submission, which did not employ any CNN features, was the first runner-up with 53.62% accuracy [22]. In this paper, we improve our fusion framework with CNN features, similar to the transfer learning approach taken in [36].

To use CNN features, we initialize a pre-trained CNN model [38] and then use the FER 2013 dataset [15] for fine-tuning. In [36] the authors exploited pre-trained models based on the ImageNet dataset, followed by two stages of fine-tuning, on FER 2013 and Static Facial Expression Sub-Challenge (SFEW) Dataset from the EmotiW 2015 Challenge, respectively [8]. In our approach, we start from the VGG-Face model that is trained for face recognition [38], and then apply the FER 2013 emotion corpus for fine tuning. Our preliminary experiments suggested that VGG-Face is more suitable for this task compared to ImageNet, which is developed for object recognition.

For feature extraction from the fine-tuned CNN model, we first rescale the input images to 224×224 pixels, then normalize them by subtracting the average image of the VGG-Face network. We use the activations of the top-level convolution layer as image features for each video frame. These frame-level features are summarized over the entire video, using functional encoding as explained in the next section.

For the fine-tuning of the VGG-Face network for the emotion recognition task, we investigated various options in our preliminary analysis. We found that combining weight decay and dropout for regularization gives the best results on the FER validation set. We carry out a multi-stage fine-tuning. In the first stage, we fine-tune on the FER public test set, and run weight updates for five epochs. In the second stage, we update the upper layers (higher than layer 27) using the FER private test set, and update for another ten epochs. We finally use the AFEW validation set for a third fine-tuning stage.

During the training of the deep networks, we oversample the training images by rotating them around their center by a random angle between

$-15°$ and $15°$, and by circularly shifting the images in the horizontal and vertical directions by an amount no more than 20% of the image size. This approach helps our network to be more robust against alignment errors. In Figure 2, we show the training curves of two stages of fine-tuning of the network with the FER dataset, where we set the learning rate and weight decay to 0.0005, momentum to 0.9, and dropout probability to 0.8 [25]. We observe that the validation set error is lower than that of the training set, which suggests that overlearning is not an issue here. Note that these curves represent the error on the FER training and validation partitions, and not the challenge corpus.

3.3. Video Modeling

In this study, we enhance the diversity of learners by incorporating functionals on frame-level features and by Fisher vector encoding of low-level descriptors.

3.3.1. Functionals

In the state-of-the-art acoustic feature extraction pipeline, it is common to use a large set of summarizing functionals over the low level descriptor contours. In video modeling, on the other hand, this approach is rarely taken, and only simple functionals such as mean and standard deviation are used. In this work, in addition to the mean and standard deviation, we use three functionals based on polynomials, fit to each descriptor contour. The first is curvature, which is the leading coefficient of the second order polynomial. The other two are the slope and the offset, respectively, computed from the first order polynomial. As a consequence, when we enable functional-based encoding, the video feature vector dimensionality is five times the frame-level dimensionality.

3.3.2. Fisher Vector Encoding

The Fisher vector (FV) provides a supra-frame encoding of the local descriptors, quantifying the gradient of the parameters of the background model with respect to the data. Given a probability model parametrized with θ, the expected Fisher information matrix $F(\theta)$ is the expectation of the second derivative of the log likelihood with respect to θ:

$$F(\theta) = -E[\frac{\partial^2 \log p(\mathcal{X}|\theta)}{\partial \theta^2}]. \tag{4}$$

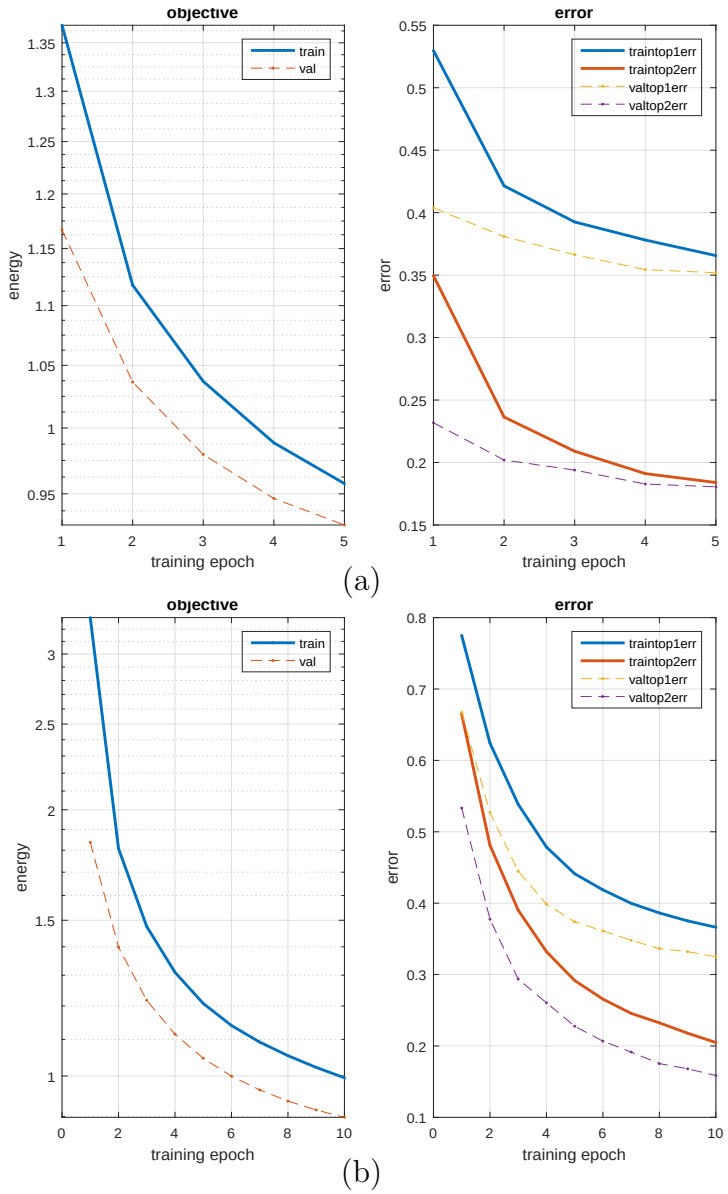

Figure 2: Fine-tuning VGG-Face with dropout on FER-2013 (a) public (b) private test set. The left column shows the softmax loss, whereas the right column shows the top-1 and top-2 classification errors.

The idea in FV in relation to $F(\theta)$ is taking the derivative of the model parameters and normalizing them with respect to the diagonal of $F(\theta)$ [39]. To make the computation feasible, a closed form approximation to the diagonal of $F(\theta)$ is proposed [39]. As a probability density model $p(\theta)$, Gaussian Mixture Models (GMM) with diagonal covariances are used. A K-component GMM is parametrized as $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$, where the parameters correspond to zeroth (mixture proportion), first (mean), and second order (covariance) statistics, respectively. It has been shown that using the zeroth order statistics is equivalent to the Bag of Words (BoW) model [46], however in FV, they were found to have a negligible effect on performance [39]. Therefore, only gradients of $\{\mu_k, \Sigma_k\}_{k=1}^{K}$ are used, giving a $2 \times d \times K$ dimensional super vector, where d is the descriptor dimensionality.

In order to efficiently learn a Background Probability Model (BPM) using GMM with diagonal covariances, we first need to decorrelate the data gathered from all instances. Principal Component Analysis (PCA) is applied on the data for this purpose. Once the parameters of PCA projection and GMM are learned, we use all descriptors from each video without sub-sampling to represent them as a Fisher Vector.

3.4. Model Learning

To learn a classification model, we employ kernel extreme learning machine (ELM) [18] and Partial Least Squares (PLS) regression due to their fast and accurate learning capabilities.

ELM proposes a multilayer perceptron architecture, but unsupervised, even random generation of the hidden node output matrix $\mathbf{H} \in \mathbb{R}^{N \times h}$, where N and h denote the number of instances and the hidden neurons, respectively. The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T} \in \mathbb{R}^{N \times L}$, where L is the number of classes. $\mathbf{T}$ is composed of continuous annotations in case of regression, therefore is a vector. In the case of L-class classification, $\mathbf{T}$ is represented in one vs. all coding:

$$\mathbf{T}_{n,l} = \begin{cases} +1 & \text{if } y^n = l, \\ -1 & \text{if } y^n \neq l. \end{cases} \tag{5}$$

The second level weights $\beta \in \mathbb{R}^{h \times L}$ are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. The output weights can be learned via:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{6}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [40] that gives the minimum L_2 norm solution to $||\mathbf{H}\beta - \mathbf{T}||$, simultaneously minimizing the norm of $||\beta||$. To increase the robustness and generalization capability, the optimization problem of ELM is reformulated using a regularization coefficient on the residual error $||\mathbf{H}\beta - \mathbf{T}||$. The learning rule of this alternative ELM is related to Least Square SVMs (LSSVM) via the following output weight learning formulation:

$$\beta = \mathbf{H}^T(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T}, \tag{7}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix, and C, which is used to regularize the linear kernel $\mathbf{H}\mathbf{H}^T$, corresponds to the complexity parameter of LSSVM [48]. This formulation is further simplified by noting that the hidden layer matrix need not be generated explicitly given a kernel $\mathbf{K}$, which can be seen identical to Kernel Regularized Least Squares [18, 41]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}. \tag{8}$$

The second approach we use for classification is partial least squares (PLS) regression. PLS regression between two sets of variables $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N \times p}$ is based on decomposing the matrices as $\mathbf{X} = \mathbf{U}_x\mathbf{V}_x + r_x$, $\mathbf{Y} = \mathbf{U}_y\mathbf{V}_y + r_y$, where $\mathbf{U}$ denotes the latent factors, $\mathbf{V}$ denotes the loadings and r stands for the residuals. The decomposition is done by finding projection weights $\mathbf{W}_x, \mathbf{W}_y$ that jointly maximize the covariance of corresponding columns of $\mathbf{U}_x = \mathbf{X}\mathbf{W}_x$ and $\mathbf{U}_y = \mathbf{Y}\mathbf{W}_y$. For further details of PLS regression, the reader is referred to [54]. When PLS is applied to the classification problem in a one-versus-all setting, it learns the regression function between the feature matrix $\mathbf{X}$ and the binary label vector $\mathbf{Y}$, and the class giving the highest regression score is taken as prediction. The number of latent factors is a hyper-parameter to tune via cross-validation.

3.5. Fusion

We contrast two fusion schemes, namely early (feature level) and late (decision level) fusion. As discussed in Section 2, we expect late fusion to produce better results, and we investigate two score fusion alternatives. The first is simple weighted fusion (SF) of scores, where the classifier confidence scores S^A and S^B are fused using weight $\gamma \in [0, 1]$:

$$S^{fusion} = \gamma * S^A + (1 - \gamma) * S^B. \tag{9}$$

The fusion parameter is optimized on the development set, as it is the case with other hyper-parameters.

Secondly, we apply weighted score fusion (WF) for each model and class. Let M and L denote the number of models and classes, respectively. The optimal fusion weights $W_{i,j}^{fusion} \in [0,1], 1 \leq i \leq M, 1 \leq j \leq L, \sum_{i=1}^{M} W_{i,j} = 1$, are searched over a pool of randomly generated matrices. This scheme considers the individual confusion matrices of each sub-system, and was shown to be successful on previous EmotiW challenges, [21, 24].

In the next section, we briefly introduce the challenge data [7], the baseline features, and the experimental protocols.

4. EmotiW 2015/2016 Corpora and Experimental Protocols

In line with the previous two challenges of the series, EmotiW 2015/2106 challenges present video clips collected from movies representing close-to-real-world conditions [4, 7, 8]. The challenge datasets are partitioned into training, development, and test sets. The EmotiW 2016 corpus shares the same validation set with the 2015 corpus. The training and test sets of the 2016 Challenge are obtained by adding new clips to corresponding sets of the 2015 Challenge. The distribution of instances over partitions based on modality (i.e. audio, video) is given in Table 1. Rows three and four report the videos with at least two detected face frames when using mixture of parts (MoP) [14] or deformable parts model (DPM) [34] alignment methods (see Section 5.1.1).

Table 1: Instance distribution over partitions. -MoP: MoP-based alignment, -DPM: Deformable Parts Model based alignment.

	AFEW 2015			**AFEW 2016**		
#	Train	Val	Test	Train	Val	Test
Clips	711	383	539	773	383	593
Audio	711	383	539	773	383	593
Video-MoP [14]	698	371	539	756	371	593
Video-DPM [34]	708	380	528	761	380	586

The baseline video features consist of Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) compacted via uniform LBP [37]. We use the baseline audio features, which are extracted via openSMILE tool [13]

using INTERSPEECH 2010 Paralinguistic Challenge baseline set [44]. The 1 582-dimensional feature set covers a range of popular low level descriptors such as Fundamental Frequency (F0), Mel Frequency Cepstral Coefficients (MFCC [0-14]), and Line Spectral Pairs Frequency [0-7], mapped to a fixed length feature vector by means of functionals such as arithmetic mean and extrema.

We use three pre-processing alternatives for any given feature channel: no-normalization (used for histogram-type descriptors), min-max normalization into [0,1] range, and z-normalization (i.e. subtracting the mean and dividing by the standard deviation). The audio and video features are kernelized by means of Linear and Radial Basis Function (RBF) kernels, prior to model learning by classifiers. We optimize the model hyper-parameters on the challenge validation set. These are RBF kernel parameter γ, regularization parameter τ for ELM and number of latent components for PLS. Similarly, the number of PCA eigenvectors p and GMM components K used in the FV encoding are optimized using cross-validation.

5. Experimental Results

We report results from six different experimental settings, four of which deal with in the wild conditions. We first provide experimental results on the EmotiW 2015 Challenge corpus, and then apply the best performing system on the EmotiW 2016 Challenge corpus. As an additional check for the effectiveness of visual features, we report results on lab-controlled data, namely, Extended Cohn Canade (CK+) and MMI corpora. We then adapt the proposed system to two datasets from different (but related) problems, namely, the RECOLA database for continuous valence/arousal prediction and ChaLearn-LAP for first impression estimation.

5.1. EmotiW 2015 - Video Based Emotion Recognition Challenge

We first summarize the accuracy obtained on the validation set with different visual features and classifiers in Table 2. Both ELM and PLS classifiers are used with linear and RBF kernels. Analyzing these unimodal systems, we observe that i) the best accuracies of individual features range from 42% to 45% and ii) only models trained on HOG-FUN, CNN-FUN and LGBP-TOP exceed 44% accuracy on the validation set. The audio features alone do not reach the accuracy levels of the visual features. The CNN-FUN results

reported here is obtained after fine tuning on the FER dataset, which gives around 10% improvement.

For multimodal evaluation, we investigate feature- and decision-level fusion. While feature-level fusion improves the results, we obtain better results with simple weighted score fusion compared to feature-level fusion (see Table 3). Once the unimodal classifiers are optimized, late fusion is less costly and reaches a higher accuracy. Consequently, we focus on late fusion and further apply random weighted score fusion in our final system.

Table 2: Validation set accuracies of visual feature types with MoP alignment. Best results for each feature type are shown in **bold**. Accuracy of audio features included on the last row for comparison.

	Lin		**RBF**	
Feature	**ELM**	**PLS**	**ELM**	**PLS**
CNN-FUN	43.40	41.78	**44.47**	43.13
HOG-FUN	39.02	**44.99**	41.46	42.01
SIFT-FUN	40.92	**43.63**	42.28	41.19
SIFT-FV	40.11	**41.73**	40.38	40.38
LBP-TOP	41.19	41.46	**42.01**	40.65
LGBP-TOP	43.63	43.90	**44.44**	44.17
LPQ-TOP	40.65	41.19	40.65	**42.01**
Audio	**36.59**	**36.29**	35.51	34.73

Table 3: Best validation set results of pairwise feature fusion/simple weighted score fusion on MoP aligned images.

	HOG-FUN	LBP-TOP	LGBP-TOP	SIFT-FV
CNN-FUN	43.40/46.61	45.01/46.34	**47.71**/46.88	44.74/45.26
HOG-FUN	-	45.82/47.70	44.47/47.15	43.67/**48.24**
LBP-TOP		-	42.86/44.44	44.74/45.26
LGBP-TOP			-	43.67/45.80

5.1.1. Deep Transfer Learning

The baseline of EmotiW uses mixture of parts (MoP) alignment for faces. We hypothesize that selecting a suitable alignment and cropping will improve

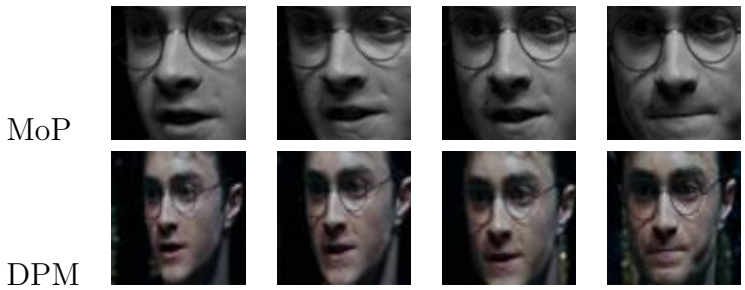

Figure 3: Example training video aligned using two different methods.

the processing of CNN features. Subsequently, we employ deformable parts model (DPM) based face detection [34] on AFEW videos. As can be seen in Table 1, the number of training and validation set videos having more than two valid facial frames are higher with DPM compared to MoP. Inspired by [43], we run the DPM face detector on rotated version of the original image between -45° and 45° in 5° increments, in order to deal with in-plane rotations. We then take the output with the maximum face score. Figure 3 shows examples of MoP- and DPM-based alignment.

The DPM alignment is coarser, but gives better results with the CNN model, as the images are more similar to the imaging conditions for pre-training and fine-tuning of the CNN. While this is an expected result, the magnitude of the difference is surprising, and illustrates the superiority of CNN for dealing with scale changes and for using appearance dynamics around the borders of the face. Other visual descriptors work better with the finer alignment of MoP.

We contrast two pre-trained CNN models in Table 4. Based on our preliminary studies, we report only results with features extracted from the 33rd layer of the CNNs. Both of these pre-trained models gave better results compared to a CNN pre-trained on ImageNet during our preliminary studies.

We first note that when using the raw VGG-Face model, employing DPM-based alignment gives a dramatic improvement over MoP-based alignment (34.1% $\rightarrow$ 44.15%). This result verifies our hypothesis that the alignment quality strongly affects the performance. More importantly, we observe a marked improvement due to transfer learning on the FER dataset (44.15% $\rightarrow$ 50.80%).

The results reported in Table 4 are obtained using min-max normalization. We seek to further improve the performance by applying the cas-

Table 4: Comparison of validation set accuracies (%) of CNN features over CNN model, fine tuning and classifier/kernel alternatives.

		ELM		PLS	
CNN Model		Linear	RBF	Linear	RBF
VGG-Face [38]	Original	39.89	42.29	41.76	44.15
	Fine-tuned	48.94	48.67	49.20	**50.80**
VGG-M-2048 [2]	Original	36.68	36.15	35.88	35.09
	Fine-tuned	41.42	42.74	37.47	39.58

caded normalization strategies proposed in [23]. The combination of Z-normalization, Power normalization, and instance level L_2 normalization (in this order) reaches a validation set accuracy of 51.6% with Linear Kernel ELM. This is the highest validation set accuracy obtained from single-modality, single-feature type models on this challenge corpus. The scores and parameters of this model are used in the subsequent fusion experiments.

5.1.2. Weighted Score Fusion and Test Set Results

The results reported here adhere to the EmotiW 2015 - Video based Emotion Recognition Sub-Challenge protocol. The test set labels are sequestered, and only limited submissions are allowed. In total, seven submissions were evaluated on the test set during the challenge, and eight systems afterwards, including probes for unimodal predictions.

Our first test set submission consisted of weighted fusion of audio, LBP-TOP and LGBP-TOP. This setup was based on [24], and provided us with a competitive benchmark. This system reached a test set accuracy of 50.28%. During our analyses, we observe that fusing the best scores from PLS and ELM might result in lowered performance due to varying score ranges of these classifiers. For these reasons, we opt to fuse the best models from MoP-aligned images with just the PLS classifier. Table 5 summarizes our best test set results. Our top test set result on the challenge (53.62%) was obtained using weighted score fusion of audio model with five visual models. In this paper, we report 54.55% accuracy using deep transfer learning with appropriate strategies and suitably preprocessed images. The CNN-FUN feature with DPM alignment gives the best unimodal performance on the test set (51.39%) using the challenge protocol, which is dramatically higher compared to the same feature type extracted from images aligned with MoP (42.86%).

Table 5: Validation and test set accuracies of the submitted systems. First part: top two systems submitted and evaluated in the official challenge [22] without CNN features. Second part: Systems using CNN features. WF: weighted fusion, FF: feature level fusion, MoP: MoP-based alignment, DPM: DPM-based alignment.

System	Val	Test
WF(Audio, LBP-TOP$_{MoP}$, LGBP-TOP$_{MoP}$)	50.14%	50.28%
WF (Audio, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$, SIFT-FV$_{MoP}$, LBP-TOP$_{MoP}$, LPQ-TOP$_{MoP}$)	**52.30%**	**53.62%**
CNN-FUN$_{MoP}$	44.47%	42.86%
CNN-FUN$_{DPM}$	51.60%	51.39%
WF(Audio, CNN-FUN$_{MoP}$, SIFT-FV$_{MoP}$ LBP-TOP$_{MoP}$, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$)	53.70%	51.76%
WF(Audio, CNN-FUN$_{DPM}$, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$)	**57.02%**	**54.55%**

The confusion matrices of best test set submissions with and without CNN features are shown in Figure 4. We observe that recall of Fear, Happy and Sad classes improve in the latter, increasing the Unweighted Average Recall (UAR) performance from 45.51% to 47.77%, along with an increase in accuracy.

Comparison of the results with the top performing systems of the official challenge is shown in Table 6. The results show the difficulty of working "in the wild". The winner and the second runner up systems used CNN features.

Table 6: Comparison of our approach with the top three systems of the EmotiW 2015 Challenge.

Work	**Val**	**Test**
Baseline	36.08	39.33
Yao et al. [56]	49.09	53.80
Kaya et al. [22]	52.30	53.62
Kahou et al. [11]	-	52.88
This paper	57.02	**54.55**

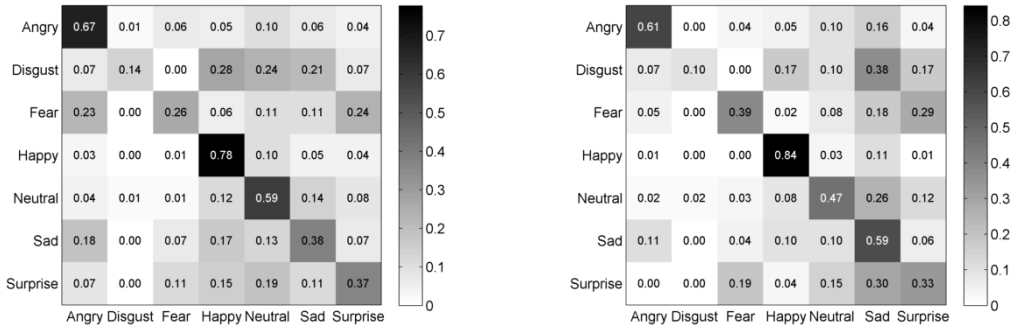

Figure 4: Test set confusion matrices (a): without CNN (53.62%), (b) with CNN (54.55%).

5.2. EmotiW 2016 - Video Based Emotion Recognition Challenge

After comprehensive experiments on the EmotiW 2015 Challenge corpus, we use our optimized system and the set of learned models on the 2016 Challenge corpus. The instance distributions were shown in Table 1. To see the generalization capability of the proposed system, we first employ the models and the fusion matrix learned on the 2015 Challenge data. We then use the same processing pipeline to learn new models and obtain a new fusion matrix for the 2016 Challenge data. The validation and test set results of these two alternative systems are shown in Table 7.

Interestingly, the models trained on the 2015 corpus show better generalization on the 2016 test set, reaching 52.11% accuracy. One reason may be that the 2016 training set contains additional samples from a particular ethnicity to improve ethnicity balance, but the test set does not reflect this change. The confusion matrix for this submission is shown in Figure 5. We observe the highest recalls for "Happy" and "Anger" classes (high arousal classes), followed by "Sadness" and "Neutral" (low arousal classes). The lowest recall is for the "Disgust" class, in line with the results obtained on previous EmotiW challenges. We also observe that a high proportion of "Disgust," "Surprise," and "Neutral" instances are incorrectly classified as "Sad". Apart from class imbalance, these confusions can be attributed to the subtleness of expressions for corresponding classes.

The fusion weights of the system achieving the highest test set accuracy are shown in Figure 5. Here, we observe that the audio modality has the highest weights for positive arousal, negative valence classes ("Anger" and "Fear"). "Disgust" samples can only be identified via subtle visual cues.

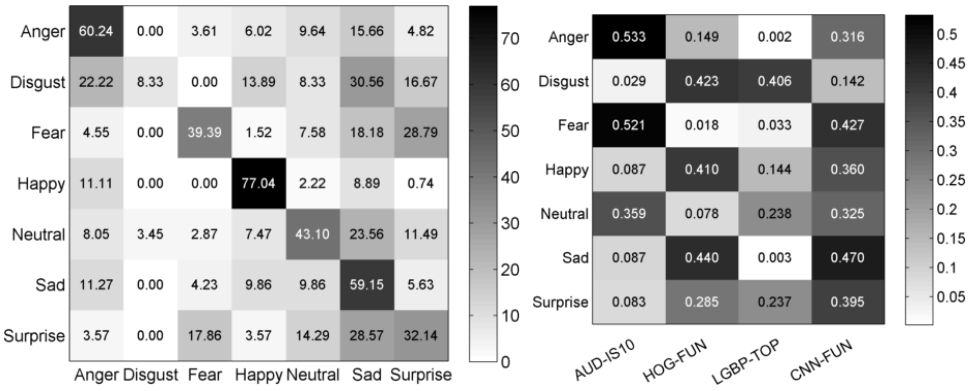

Figure 5: Left: Confusion matrix of the best test set submission in EmotiW 2016 (52.11% accuracy). Right: Fusion weights used in the best test set submissions on two challenge corpora (54.55% and 52.11% accuracy on EmotiW 2015 and 2016 Challenges, respectively)

Regarding the sub-systems, the model trained with CNN-FUN features has the highest average weight (0.348), followed by HOG-FUN (0.258) and audio (0.243), respectively. This finding shows the importance of CNN features for the final predictions.

Table 7: Accuracy of four submitted systems on the EmotiW 2016 test set. The corpus used for model training/system development is given in square brackets.

System	Val	Test
Baseline - LBP-TOP	38.81	40.47
1 - WF (Audio, CNN-FUN$_{DPM}$, LGBP-TOP$_{MoP}$ [2016])	57.70	46.21
2 - WF (Audio, CNN-FUN$_{DPM}$, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$ [2015])	57.02	**52.11**
3 - WF (Audio, CNN-FUN$_{DPM}$, LGBP-TOP$_{MoP}$, HOG-FUN$_{MoP}$ [2016])	55.35	48.40
4 - Equal Weight Score Fusion of Systems 2 and 3	57.18	51.77

5.3. Experiments on Extended Cohn Kanade and MMI Corpora

We validate the proposed set of visual features using two frequently used controlled emotional expression datasets, namely the Extended Cohn-Kanade dataset (CK+) [33] and MMI [50]. The alignment of both datasets are carried out by DPM-based face detection. We re-scale cropped images to 224×224 pixels, which is compatible with the CNN model.

CK+ [33] features 593 emotional portrayals from neutral face (onset) to apex, out of which 327 videos are evaluated that contain one of seven basic emotions (Anger, Contempt, Disgust, Fear, Happiness, Sadness, and

Surprise). Here, we use leave-one-subject-out (LOSO) protocol with 118-folds as in [33].

MMI [50] is also portrayed, however, it poses a higher challenge compared to CK+ due to occlusions and pose variations. Out of 326 videos, 213 are annotated with six basic emotions. In our experiments, we use this subset with basic emotion annotation, excluding fully profile videos and implementing a 10-fold subject independent cross-validation for comparability with the literature [28, 30, 53].

A comparison with the some state-of-the-art studies are given in Table 8, and illustrates that the proposed approach performs equally well in controlled settings.

Table 8: Comparison of our proposed approach with state-of-the-art results on CK+ and MMI.

	CK+ Dataset		MMI Dataset	
Work	Mean Acc.	Overall Acc.	Mean Acc.	Overall Acc.
Wang et al. [53]	86.30	88.80	59.70	60.50
Liu et al. [28]	87.90	92.40	62.20	63.40
Liu et al. [30]	94.82	96.64	**71.38**	**74.63**
This paper	**98.31**	**98.47**	70.31	72.46

5.4. Experiments on the RECOLA Corpus

To illustrate the performance of proposed CNN features, we apply the corresponding pipeline to the RECOLA multi-modal emotion corpus, which is composed of 27 videos, continuously annotated for valence and arousal dimensions [42]. This dataset is used both in 2015 and 2016 editions of the AVEC challenge series. The challenge setting evenly partitions the videos into training, development, and test sets. Videos are 300 seconds long, and ground-truth annotations are given with a rate of 25Hz [49]. The challenge baseline benefits from the fusion of audio, video, and physiological modalities, such as Electro-Cardiogram and Skin Conductance Rate. Eight classifiers are trained on different modalities, and fused.

We introduce some changes to test our approach on this dataset. Since the task is continuous prediction, the summarizing functionals cannot be applied over the whole video. We use a symmetric window for each frame, or for a subset of frames, and select specific window sizes for each modality and

affective dimension [49]. The baseline system applies a post-processing that includes shifting predictions to compensate for annotation delay, as well as centering and scaling [49]. We include these steps in our model. The baseline system also estimates missing frames for each modality using other modalities. Because we lack data from the physiological modalities, we could not apply this estimation, and subsequently our reported accuracy is impaired by missing frames. The challenge performance measure is the Concordance Correlation Coefficient (CCC), which takes into account the difference between the prediction and ground-truth means.

The summary of experiments with the baseline and our proposed features, as well as their weighted fusion results are given in Table 9. We observe that substituting the baseline appearance feature (LGBP-TOP) with CNN dramatically improves the performance for valence prediction (0.481 $\rightarrow$ 0.550). When CNN is fused with the three baseline sub-systems, the CCC measure is further improved. The contribution of CNN features for arousal is lower.

Table 9: Results for baseline (BAS) and proposed CNN features on the RECOLA corpus

CCC		Arousal		Valence	
#	System	Dev	Test	Dev	Test
1	AUDIO_{BAS}	0.796	**0.648**	0.455	0.375
2	GEO_{BAS}	0.379	0.272	0.612	0.507
3	LGBP-TOP_{BAS}	0.483	0.343	0.474	0.486
4	CNN_{DPM}	0.503	0.301	0.543	0.505
5	WF(1,2,3)	0.797	0.451	0.660	0.481
6	WF(1,2,4)	0.804	0.461	0.677	0.550
7	WF(1,2,3,4)	**0.809**	0.465	**0.690**	**0.566**

5.5. Experiments on ChaLearn-LAP First Impressions Challenge

We have used the proposed system on the ChaLearn-LAP First Impressions Challenge at ICPR 2016 [12]. In this challenge, in the wild videos of people are evaluated to guess the first impressions given to other people, as measured by a Big-5 personality traits questionnaire. 6000 videos were used for training, 2000 for validation, and 2000 videos with sequestered labels were used for testing the systems. Our submission to the challenge followed the same pipeline, with an additional (but very low-impact) channel for deep scene features [16]. Also, the output layer estimates continuous personality

trait estimations, instead of discrete emotional labels. The rationale behind applying a similar transfer learning approach is to benefit from influences of emotional facial expressions, as well as ambient cues on first impressions. Our system received the first place in the challenge, reaching an accuracy of 0.913 averaged over five personality dimensions, illustrating that the processing pipeline we proposed in this paper can generalize to other problems with minimal adaptation. We refer the reader to [16] for the details.

6. Conclusions

In this study, we proposed a system for multimodal expression recognition in the wild. Our findings confirm those of [10], who reported that multimodality brings in diminishing returns for natural (or seminatural) data. For the arousal predictions on the RECOLA corpus, the multimodal system performs poorer than the best unimodal system.

On the EmotiW challenges, we saw that the recall of *disgust* and *fear* classes was very low. In audio-only models, it is possible to obtain higher recall for the *fear* class. On the other hand, *disgust* can be distinguished (up to a point) with visual cues, especially with dynamics of facial structure. In the wild studies exacerbate the difficulty of catching subtle cues, as the noise-related variation can be higher than variation due to these cues. Consequently, some classes are affected more by these conditions.

We have illustrated the impact of flexible alignment (and fine-tuning) on transfer learning for CNN models. The pre-trained CNN models extract rich features, and serve to complement even within-modality systems. We have illustrated that CNN features can be enhanced by summarizing functionals, which is an approach frequently followed for the audio domain.

Class imbalance mitigates the learning issues for the automatic systems, but we note that the data used for the EmotiW Challenge is also challenging for human annotators, and the gap between human and computer performance is ever decreasing.

References

[1] T. R. Almaev and M. F. Valstar. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 356–361. IEEE, 2013.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR*, volume 1, pages 886–893, 2005.

[4] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 427–432, 2016.

[5] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction*, pages 461–466, 2014.

[6] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 509–516, 2013.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, July 2012.

[8] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proc. ACM ICMI*, ICMI '15, pages 423–426, New York, NY, USA, 2015. ACM.

[9] H. Dibeklioğlu, A. A. Salah, and T. Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *Proc. ECCV*, pages 525–538. Springer, 2012.

[10] S. D'Mello and J. Kory. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proc. ACM ICMI*, pages 31–38. ACM, 2012.

[11] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proc. ACM ICMI*, pages 467–474. ACM, 2015.

[12] H. J. Escalante, V. Ponce-López, J. Wan, M. Riegler, B. Chen, A. Clapes, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, H. Müller, and M. Larson. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *ICPR Contest proceedings*, 2016.

[13] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. of the Intl. Conf. on Multimedia*, pages 1459–1462. ACM, 2010.

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[15] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[16] F. Gürpınar, H. Kaya, and A. A. Salah. Multimodal fusion of audio, scene, and face features for first impression estimation. In *Proc. ICPR*, 2016.

[17] J. Heikkilä, V. Ojansivu, and E. Rahtu. Improved blur insensitivity for decorrelated local phase quantization. In *20th International Conference on Pattern Recognition (ICPR '10)*, pages 818–821, 2010.

[18] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *SIEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):513–529, 2012.

[19] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Cybernetics, IEEE Transactions on*, 44(2):161–174, 2014.

[20] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proc. IEEE FG*, pages 314–321. IEEE, 2011.

[21] S. E. Kahou, C. Pal, X. Bouthillier, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proc. ACM ICMI*, ICMI '13, pages 543–550, 2013.

[22] H. Kaya, F. Gürpınar, S. Afshar, and A. A. Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *Proc. ACM ICMI*, ICMI '15, pages 459–466, New York, NY, USA, 2015. ACM.

[23] H. Kaya, A. A. Karpov, and A. A. Salah. Fisher vectors with cascaded normalization for paralinguistic analysis. In *INTERSPEECH*, pages 909–913, Dresden, Germany, 2015.

[24] H. Kaya and A. A. Salah. Combining modality-specific extreme learning machines for emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 10(2):139–149, 2016.

[25] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2):173–189, 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[27] E. Krumhuber, A. S. Manstead, D. Cosker, D. Marshall, P. L. Rosin, and A. Kappas. Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, 7(4):730–735, 2007.

[28] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Proc. ACCV*, pages 143–157. Springer, 2014.

[29] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen. Partial least squares regression on Grassmannian manifold for emotion recognition. In *Proc. ACM ICMI*, pages 525–530. ACM, 2013.

[30] M. Liu, R. Wang, S. Li, Z. Huang, S. Shan, and X. Chen. Video modeling and learning on Riemannian manifold for emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 10(2):113–124, 2016.

[31] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *Proc. ACM ICMI*, ICMI '14, pages 494–501, New York, NY, USA, 2014. ACM.

[32] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[33] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proc. IEEE CVPR Workshops*, pages 94–101. IEEE, 2010.

[34] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Proc. ECCV*, pages 720–735. Springer International Publishing, 2014.

[35] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proc. IEEE CVPR Workshops*, pages 881–888, 2013.

[36] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proc. ACM ICMI*, pages 443–449. ACM, 2015.

[37] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[38] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[39] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, Minnesota, USA, 2007.

[40] C. R. Rao and S. K. Mitra. *Generalized inverse of matrices and its applications*, volume 7. Wiley New York, 1971.

[41] R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. *NATO Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.

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

[43] R. Rothe, R. Timofte, and L. Gool. DEX: Deep EXpectation of apparent age from a single image. In *Proc. IEEE CVPR Workshops*, pages 10–15, 2015.

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

[45] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proc. ACM ICMI*, pages 517–524. ACM, 2013.

[46] J. Sivic and A. Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, 2009.

[47] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In *Proc. ACM ICMI*, pages 481–486. ACM, 2014.

[48] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[49] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of AVEC'16, co-located with ACM MM 2016*, Amsterdam, The Netherlands, October 2016. ACM.

[50] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION: Corpora for Research on Emotion and Affect*, page 65, 2010.

[51] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *Proc. of the 4rd ACM Intl. Workshop on Audio/Visual Emotion Challenge*, AVEC '14, 2014.

[52] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013–The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In *Proc. of the 3rd ACM Intl. Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 3–10, 2013.

[53] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proc. IEEE CVPR*, pages 3422–3429, 2013.

[54] H. Wold. Partial least squares. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, pages 581–591. Wiley New York, 1985.

[55] C.-H. Wu, J.-C. Lin, and W.-L. Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3:e12, 2014.

[56] A. Yao, J. Shao, N. Ma, and Y. Chen. Capturing AU-aware facial features and their latent relations for emotion recognition in the wild. In *Proc. ACM ICMI*, pages 451–458. ACM, 2015.

[57] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proc. ACM ICMI*, pages 435–442. ACM, 2015.

[58] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect "in-the-wild". In *Proc. IEEE CVPR Workshops*, pages 36–47, 2016.

[59] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

[60] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[61] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.

[62] 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*, 29(6):915–928, 2007.