Multimodal Fusion Based on Information Gain for Emotion Recognition in the Wild

Esam Ghaleb, Mirela Popa, Enrique Hortal and Stylianos Asteriadis
Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, Netherlands
{esam.ghaleb,mirela.popa,enrique.hortal,stelios.asteriadis}@maastrichtuniversity.nl

***Abstract*—In this paper we present a novel approach towards multi-modal emotion recognition on a challenging dataset AFEW'16, composed of video clips labeled with the six basic emotions plus the neutral state. After a preprocessing stage, we employ different feature extraction techniques (CNN, DSIFT on face and facial ROI, geometric and audio based) and encoded frame-based features using Fisher vector representations. Next, we leverage the properties of each modality using different fusion schemes. Apart from the early-level fusion and the decision level fusion approaches, we propose a hierarchical decision level method based on information gain principles and we optimize its parameters using genetic algorithms. The experimental results prove the suitability of our method, as we obtain 53.06% validation accuracy, surpassing by 14% the baseline of 38.81% on a challenging dataset, suitable for emotion recognition in the wild.**

Keywords*—Emotion Recognition; Multimodal Fusion; Information Gain; Genetic Algorithm.

I. Introduction

The recent technological advancements brought interactivity between people and digital devices to a completely different level, making computers and mobile phones an important part of our daily lives. A natural way in which people communicate with each other is based on emotions. Therefore, there is an increased interest in the human computer interaction (HCI) field towards enhancing digital devices with emotion recognition abilities for obtaining a more natural HCI experience. Emotions can be expressed using both verbal and non-verbal cues such as facial expressions, gestures or the tone of the voice. Facial expressions represent one of the most significant cues for recognizing emotions, due to their universality proven by Ekman [1] who found that six basic emotions (happiness, fear, sadness, disgust, surprise and anger) are the same across cultures. The applications of an automatic facial recognition system go beyond HCI, being useful also in website customization, gaming industry, humanoid robots, as well as in improving online education systems. Due to the potential applications of such a system, there have been done many research studies in classifying faces in still images [2] or in video sequences [3] into one of the six basic emotions.

Data mining algorithms employed for recognizing the six basic emotions have been rather successful on posed datasets gathered in controlled environments such as the Cohn-Kanade [4], the JAFFE [5], the CMU Pose Illumination and Expression (PIE) [6] or the MMI database [7]. While recently, efforts were devoted to more challenging datasets, captured in uncontrolled spontaneous conditions such as the Acted Faces in the Wild (AEFW) dataset [8], containing video clips of unconstrained facial expressions, with varied head poses, occlusions and challenging illumination conditions. The palette of feature extraction techniques employed for facial expressions recognition contains appearance based methods (Gabor filters [9], LBP [10], SIFT [11]), geometric features [12] and also unsupervised feature learning methods such as the recently adapted CNN models [13]. On top of the extracted features, multiple classification algorithms are used, varying from SVM with different kernel methods [14], neural networks [3], Boltzaman machines [15] to deep architectures [16]. Apart from the visual modality, audio-based emotion recognition is also promising and features such as prosody, jitter, or the fundamental frequency proved to be useful [17].

Furthermore, studies in multi-modal emotion recognition showed the benefits of fusing visual and acoustic information [18], due to the complementarity of the two modalities. Therefore, in this paper we propose a multi-modal framework for emotion recognition from video sequences, by taking advantage of both visual and audio features. Moreover, one of the main contributions of this paper consists of proposing a hierarchical fusion approach, which combines feature level and decision level fusion in an efficient manner, using information gain principles, which is depicted in Figure 1. The proposed fusion framework is general enough to be useful also for other tasks such as behaviour or object recognition, as long as there are available different types of features which are complementary.

In our approach, we take advantage of different feature extraction algorithms, extracted from the audio channel and also from the entire face or from salient facial regions of interest (ROI), (e.g. eyes, nose, mouth, forehead and chin), such as dense scale invariant feature transformation (DSIFT), geometric features, and a pre-trained CNN model for face recognition provided by the Visual Geometry Group (VGG-face) [16], denoted as a set of M features on the 2-nd layer of Figure 1. Each of these features are useful, while one constraint in fusing them is given by their different underlying probabilities and ranges. Another contribution of this paper refers to encoding the different features using Fisher Vector [19] representations, which are useful at projecting all types of features in the same space and also as it facilitates the analysis of videos with different lengths, while efficiently capturing the facial dynamics (the 3-rd layer in Figure 1). Next, we use an efficient algorithm for feature-level fusion, which finds the

best types of features to be fused in a hierarchical manner, based on minimizing the KullbackLeibler (KL) divergence [20] between the probability distribution function (PDF) of true labels and the PDF of predicted labels, obtained after employing a classification algorithm. For example, at a first stage the mouth region features and audio features are fused in a new feature vector and also DSIFT features and geometric features are fused in another one.

Then, at the next stage (the 5-th layer in Figure 1), the two new obtained feature vectors are fused using a decision-level fusion algorithm which optimizes the weights of each modality using a Genetic Algorithm (GA).

The proposed framework is useful, as, instead of fusing all features at an early stage as described by [21] or at the end of the pipeline as proposed by [22], it searches for the best combinations at different processing stages for finding complementary modalities. Furthermore, the use of a genetic algorithm facilitates finding the optimum weights for the decision level fusion. We evaluated our proposed approach on the challenging AEFW dataset [23] and compared it with a deep learning architecture.

The remaining of this paper is organized as follows. In Section 2, related work is presented showing the popular trends in multimodal emotion recognition. Section 3 explains the pre-processing stage and the feature extraction methods from both video and audio modalities. Our proposed framework towards emotion recognition, based on multi-modal fusion of vision and audio modalities is introduced in Section 4, highlighting different fusion schemes. Next, the experimental results are presented in details in Section 5, while the paper ends with conclusions and directions for future work.

II. RELATED WORK

In human emotion recognition, the goal is to predict high-level affective content from low-level human-centered signals such as video, audio, and body posture. The description of the high level affective content can be roughly divided into two subcategories: discrete emotions and continuous emotions in the valence and arousal space. According to the discrete category, there exist six basic emotions (sadness, happiness, fear, anger, surprise, and disgust) proposed by the physiologist Paul Ekman in [1] which are universally recognized and shared across cultures. While most of the previous works on facial expression recognition are based on this type of categorization [24], [14], [3], there have been various studies, which addressed the dimensional space of human emotions such as [25], [18], where they used a 2D valence-arousal emotion model.

A. Feature Extraction

Previous work on facial emotion recognition mostly use hand crafted features [24], [25]. The pipeline of these studies starts by performing face detection and is followed by extracting facial features such as LBP[10], Gabor wavelets[9], and SIFT[11]. In addition, these features were extended to capture the spatio-temporal space such as BoW on SIFT features in [26] and LBP-TOP [27]. With the recent improvements in neural networks, deep architectures have become popular and effective for extracting high level features from data and specifically from facial images [28], [16]. Recently, deep learning approaches for feature extraction have surpassed the traditional ones and emerged in an enormous impact and improvement in many pattern recognition and classification tasks. In computer vision, Convolutional Neural Network (CNN) is a well-known deep learning architecture for feature extraction from images. In our work, we benefit from this model as well, by using the state-of-art VGG-Face face representation which proved to be discriminative and efficient in face recognition [16]. In [3], [14], CNN features were extracted by fine-tuning pre-trained models for facial emotion recognition. In [29], a hybrid neural network is presented, that combines Recurrent Neural Networks (RNN) with CNN to encode facial motion throughout video frames.

B. Multimodal Learning

Data perception can be achieved through different and complementary modalities such as audio, video, and skeleton joints. The joint analysis of the sensory inputs leads to an improved recognition of the environment, since it enhances the understanding of an event through different channels. However, each modality, has its own feature distribution and statistical properties, and different sensory data have high non-linear relationships. Therefore, concatenating the input features of the different channels is not efficient. There have been many studies that try to optimize a framework and benefit from data of various modalities in order to obtain free modality shared description to represent the correlation between different modalities. Multimodal learning has been applied for several tasks which involve various data sensors, such as person identification, emotion recognition [30], multimedia retrieval [15], and gesture and action recognition [31]. In [32] a deep learning method for audiovisual speech recognition was proposed, where the authors used different settings and scenarios in order to find a framework that would obtain a shared representation for both modalities. One of the main constraints of the scheme presented in that paper is the complexity of the applied architecture. In addition, when analysing the performance of the late and early fusion, the results show the inefficiency of the deep learning based multimodal learning. This can be traced to the fact that deep learning requires much more data to learn a shared representation than other models.

In our study we employ a Fisher vector representation for encoding low-level features of different modalities. It functions as a higher layer representation of those features, as it projects them into the same space, where they share common statistical and distribution properties. Comparing to deep learning, Fisher vector representation has advantages such as its compactness and efficiency, while it can be computed using a small number of parameters (GMM parameters) [33].

C. Multimodal Emotion Recognition

Similarly, there have been various studies that cover multimodal learning for emotion recognition. For example, in [30], multimodal deep learning was applied to learn a shared representation for audiovisual emotion recognition. Other studies exploited late level fusion. In [3], separate methods for each modality were developed, (e.g. CNN for facial images and Restricted Boltzmann Machines (RBM) for audio information),

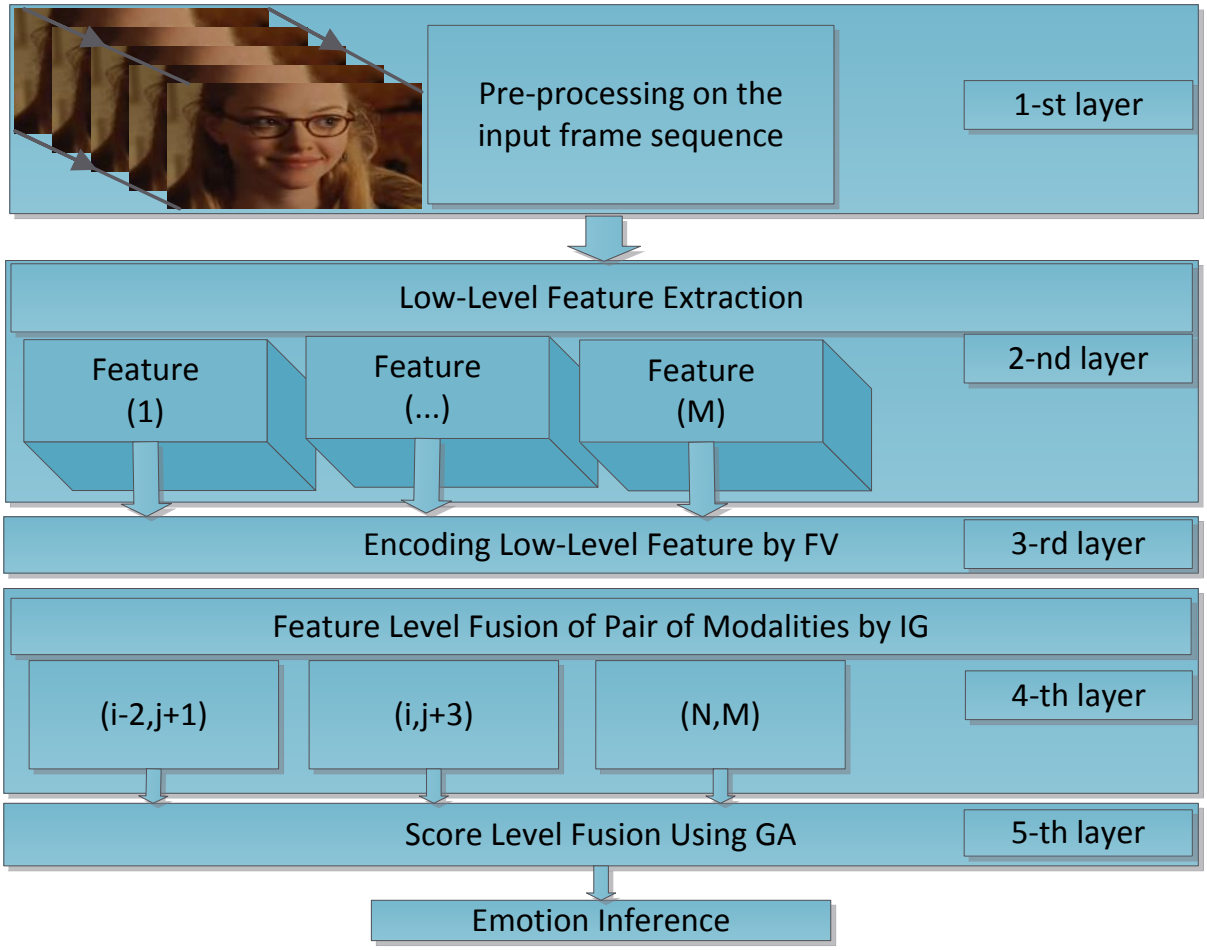

Fig. 1. Hierarchical multimodal fusion framework based on feature level and score level fusion. The proposed scheme starts with a pre-processing layer for face and facial landmark detection, and face alignment. This layer is followed by extracting a set of M low-level features (e.g. DSIFT, CNN, and geometric features). The third and fourth layers include high level representation of features by Fisher Vector encoding (FV) and selecting pair of modalities based on information gain principles (IG). In the fourth layer are included examples of the selected pair modalities having indices $(i, j) \in M$. The last layer of the framework depicts score level fusion optimized using a Genetic Algorithm (GA).

followed by a combination of the score of each modality in late fusion by grid search. Similarly in [14], they benefited from kernel methods for video feature representation and the fusion of the different modalities was achieved in a probabilistic manner at a late stage. In [34], different classifiers were trained for each modality, and then combined using a late fusion approach based on a genetic algorithm.

In our work, we target the task of emotion recognition in the wild using both schemes of multimodal learning: early and late fusion. In the early level fusion phase, we first project modality features into a common space that shares similar properties using Fisher vector encoding, and then decide the best combination of the modalities by employing information gain principles for modality selection. In late fusion, we benefit from the resulted modalities of early fusion, and take into consideration the performance of each modality prior to Fisher vector encoding to spot the complementary modalities for better emotion prediction.

III. Proposed Framework

In this section, we first explain the preprocessing phase of facial images and how we obtain a face track from a video. Then we present the low-level feature extraction methods implemented in our framework for different modalities: audio and visual (geometric and appearance features). Finally, we describe feature encoding and representation by means of Fisher vectors for video modeling and projecting features into the same space.

A. *Preprocessing*

Facial Landmark Detection: Succeeding the step of face detection, we detect 49 landmarks and track them in each

Fig. 2. Cropped and aligned faces from the AFEW dataset.

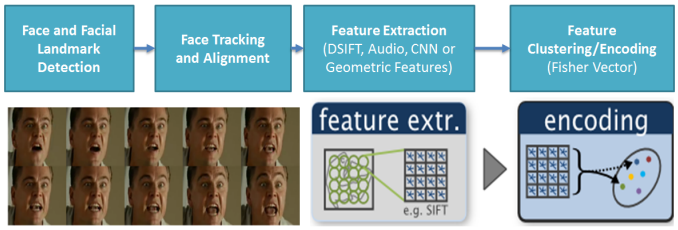

Fig. 3. Face pre-processing and Feature Extraction and Encoding.

frame of a video, using the Supervised Descent Method (SDM) [35]. SDM is a successful face shape regression technique, which begins with an initial S_0 face shape and progressively predicts the final shape of the facial landmarks in an iterative way. Comparing to other techniques, this method provides robust and accurate landmark positions in challenging conditions, such as varying illumination and pose, and low quality images. In addition, it gives a reliable and robust tracking of facial landmarks in the wild, in real-time.

Face Alignment: Face alignment is an essential step in facial emotion recognition. It is the process of registering faces with respect to facial landmarks (e.g. eyes, nose, mouth, and chin) of the canonical frame. This process fixes the landmark positions in aligned images and it is carried out by similarity transformation. In our work, we use facial landmarks provided by SDM landmark detector and perform a similarity transformation that aligns faces to the fixed canonical frame based on eye centers positions. In addition, facial images are cropped and re-sized to a fixed resolution: 224×224. Examples of aligned and cropped faces are depicted in Fig. 2, while Fig. 3 presents examples of tracked facial images from the AFEW dataset.

B. Low-Level Feature Extraction

Emotion recognition relies on representative data along with accurate and discriminative descriptors. This type of information contributes in enhanced recognition and classification accuracy. Accordingly, in this paper, we extract a set of low-level descriptors for the visual and audio modalities. Then, we use Fisher vectors for video modeling and projecting them into the same space. In the reminder of this subsection, we outline

Fig. 4. Illustration of the six salient facial regions of interest (ROI): left eye, right eye, forehead, mouth, nose and the region between eyes.

the low-level features used in our work: DSIFT, handcrafted geometric, CNN and audio features.

1) DSIFT Features: Dense Scale Invariant Feature Transform (DSIFT) has been widely used for image representation in the last decade, in many computer vision recognition tasks [36], [37]. In DSIFT, instead of sparsely detecting and selecting the facial key-points, we compute the DSIFT histogram densely over a given image, using a certain scale factor and step size. This has an advantage since it does not rely on facial landmark detection. We divide the facial images into a grid of overlapping blocks with a step size equal to 1. Specifically, the block size is 24×24. Later, we compute a DSIFT histogram for each block. This step is repeated in 5 scales, with a scale factor equal to $\sqrt{2}$.

In this work, we compute DSIFT with two approaches: (i) DSIFT on the entire facial image; and (ii) DSIFT on six distinct facial regions of interest (ROI): left eye, right eye, forehead, mouth, nose and the region between eyes. These six facial ROI are illustrated in Fig. 4. We extract and crop the ROI using the facial landmarks provided by [35]. Then, DSIFT features are extracted from each region separately. In the reminder of this paper, we refer to the DSIFT extracted from the entire facial image as DSIFT, while we call the DSIFT computed on ROI as ROI DSIFT.

2) CNN Features: Our CNN face representation is based on the VGG-face model [16], which is a 16-layer convolutional neural network (CNN) model trained with 2.6M facial images of 2.6K people for face recognition in the wild. We use this model for feature extraction by employing the output of the sixth Fully Connected layer (FC6) as the facial signature. This layer outputs a 4096 dimensional feature vector.

3) Geometric Features: Another feature representation deals with the shape and location of the facial landmarks (e.g. mouth, eyes, eyebrows, nose, and chin). Different facial expressions correspond to different shape deformations of the facial landmarks. The location of the fiducial points was chosen according to the facial model proposed by [35] and is shown in Fig. 5. These landmarks are transformed and fitted with the same alignment used for face registration.

An alternative is to use the fiducial points coordinates as features in the classification process, but this representation achieves poor performance, as it is not able to capture the

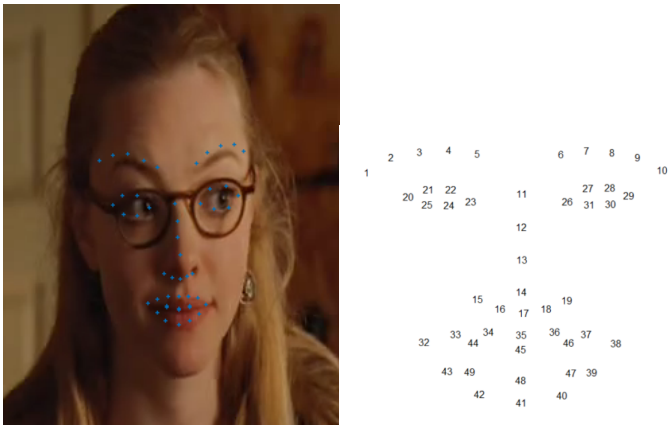

Fig. 5. Facial landmarks provided by [35].

TABLE I. AUDIO FEATURES: LOW LEVEL DESCRIPTORS

Low Level Descriptors (LLD)	Audio Features
Energy/Spectral LLD	PCM Loudness
	MFCC [0-14]
	log Mel Frequency Band [0-7]
	Line Spectral Pairs (LSP) frequency [0-7]
	F0
	F0 Envelope
Voicing related LLD	Voicing Prob.
	Jitter Local
	Jitter consecutive frame pairs
	Shimmer Local

variations between different individuals. For increasing the discriminative power of the feature set, we compute geometric features, which may be represented by segments, perimeters, or areas of the figures formed by the fiducial points.

Following the works in [38] and [39] we obtain a set of features including: Euclidean distances, angles and curvatures between fitted facial landmarks, followed by applying a normalization step. For example, the set of extracted features include but are not limited to: mouth and eyes aspect ratios, lower and upper lips and mouth corners' angles, nose tip-mouth corner angles, eyebrow slope, mouth corner and mouth bottom angles, and the curvature of lower-outer and lower-inner lips.

4) Audio Features: We utilize the speech analysis openSMILE toolkit [40] for audio features extraction. This popular and widely used library extracts features that capture both voice quality and prosodic characteristics of a speaker. We follow the audio feature extraction as explained in [41]. The set of audio features used in this study consists of: 34 energy & spectral related low-level descriptors (LLD)×21 functionals, 4 voicing related LLD×19 functionals, 34 delta coefficients of energy & spectral LLD×21 functionals, 4 delta coefficients of the voicing related LLD×19 functionals and 2 voiced/unvoiced durational features. The details for the LLD are included in Table I. The functionals computed on the LLD include: arithmetic mean, standard deviation, skewness, kurtosis, quartiles, quartile ranges, percentile 1%, 99%, percentile range, position max./min, up-level time 75/90, linear regression coefficient, and linear regression error (quadratic/absolute).

C. Feature Encoding and Video Modeling

Video Modeling: In this work, we adopt the usage of Fisher vectors for encoding and clustering different low-level features for each modality. The features are not only pooled from one still image, instead they are pooled from all the frames across a face track. As suggested in [37], we use video-pooling, where we compute a single fisher vector over the whole face track by pooling together low-level features (e.g. DSIFT, or CNN features) from all facial images in a track. This kind of representation has many advantages comparing to still image based representation for various reasons: (i) it encodes the spatio-temporal information in a face track, (ii) it captures the motion of the face over time which leads to a better description of the different low-level features; and (iii) it dramatically reduces the dimensionality of data by producing a single discriminative descriptor for a video.

Fisher Vector Representation: The pipeline for Fisher vector encoding typically starts with extracting a set of features (e.g. DSIFT, geometric features etc.), and then aggregates the large set of feature vectors across all frames in a track into a high dimensional Fisher vector which is better suited for linear classification. This is achieved by fitting a parametric generative model such as Gaussian Mixture Models (GMM) to the features. GMM can be referred to as a *probabilistic visual vocabulary*. The next step consists of encoding the gradient of the local descriptors log-likelihood with respect to the GMM parameters. The GMM parameters are estimated on a large set of local descriptors using the Expectation Maximization (EM) algorithm to optimize the log-likelihood.

In Fisher vector computation, the covariance of the GMM is assumed to be diagonal and only the derivatives with respect to Gaussian mean and covariance are considered. This leads to a vectorial representation that obtains the average first and second order difference between the features and each of the GMM centers:

$$\Phi(k)^{(1)} = \frac{1}{N\sqrt{w_k}} \sum_{p=1}^{N} \alpha_p(k)\left(\frac{X_p - \mu_k}{\sigma_k}\right) \quad (1)$$

$$\Phi(k)^{(2)} = \frac{1}{N\sqrt{2w_k}} \sum_{p=1}^{N} \alpha_p(k)\left(\frac{(X_p - \mu_k)^2}{\sigma_k} - 1\right) \quad (2)$$

Where w_k, μ_k, σ_k are the GMMs weights, means and diagonal covariance, respectively. These parameters are computed on each of the low-level features of the training set. $\alpha_p(k)$ is the soft assignment of the p-th feature x_p to the k-th GMM component. Fisher vectors dimensionality is $2Kd$ which depends on the number of the GMM components (K), and the dimensionality of the employed set of features. Then, a Fisher vector ϕ is computed by stacking the differences (the assignment of the low-level features to the first and second differences of GMM centers): $\phi = [\Phi(1)^{(1)}, \Phi(1)^{(2)}, ..., \Phi(K)^{(1)}, \Phi(K)^{(2)}]$.

A Fisher Vector representation has many advantages: (i) it is a generic representation which combines the benefits of generative and discriminative approaches, (ii) it can be computed using a small number of parameters (GMM parameters), (iii) more importantly, it is efficient and it shows a significant benefit when used in combination with linear classifiers such as linear-SVM [33].

IV. Multimodal Fusion

In this section, we present the two fusion approaches employed in the proposed framework, feature level fusion based on information gain and score level fusion, improved by means of a genetic algorithm. We propose a framework for multimodal emotion recognition, which combines different modalities in a hierarchical and collaborative fashion, using both early and late fusion schemes. These two techniques aim to maximize the benefit of different modalities in emotion recognition. In the rest of this section, first, we introduce our approach by explaining how information gain and Fisher vector representation are involved in early level fusion. Then, we describe our method of collaborative late level fusion that captures the performance of each modality per emotion to enhance the final decision making.

A. Feature Level Fusion

In our study, we apply various feature extraction and representation techniques for different modalities. Accordingly, their data comes from diverse input channels. Therefore, each modality has its own distinct feature distribution properties. However, a multimodal fusion and feature learning method can be used to capture the correlations between these modalities in real word data, by employing a feature level representation. As a result, similarity in the representation space, must reflect the similarity in corresponding concepts. For example, speech and facial images are correlated in the real world when people express their emotions. People often tend to speak loudly when they are angry, or they use a certain tone of voice accompanied by a facial expression to indicate their affective states. We use Fisher vector encoding to map the extracted features into a common space, for achieving a higher layer feature description, which shares similar statistical and discriminative properties. Thus, different modalities are projected into one domain by means of fisher vectors, enabling and supporting feature concatenation. The newly obtained fisher vector based representation is independent of the input modality, opposite to the low-level features which are modality-dependent. Furthermore, the proposed data representation is useful at capturing the non-linear relationships between the different modalities employed in our work.

B. Feature Level Fusion Based on Information Gain Principles

For optimizing the feature level fusion of different modalities and selecting the best combination among the possible ones, we used measures from information theory, as the Kullback-Leibler (KL) divergence [20], which is useful at measuring the distance between two probability distributions (PDF). In our framework we aim to minimize the distance between the PDF of the true labels, denoted with Y and the PDF of the predicted labels for each modality ($X_k, k \in \{1,\ldots,n_{mod}\}$), obtained using a classification algorithm on top of the modality k features.

$$KL(X_k||Y) = \sum_{i=1}^{N_{lab}} X_k(i) log \frac{X_k(i)}{Y(i)} \tag{3}$$

where N_{lab} is the number of labels, X_k is the PDF of predicted labels for the k modality, and n_{mod} is the number of input modalities denoted by different types of features, both visual and audio. By minimizing the KL divergence, we aim to obtain a PDF as close as possible to the ground truth PDF, increasing in this way the performance accuracy of our emotion recognition framework. As the KL divergence is not symmetric, we employ in our work the symmetric version [42], for obtaining a general framework, which is not affected by the order of the modalities in the fusion process:

$$I(X_k,Y) = \frac{KL(X_k||Y) + KL(Y||X_k)}{2} \tag{4}$$

Furthermore, the set of modalities which are fused at the feature-level are selected by minimizing the KL divergence between the PDF of the true labels and the PDF of the predicted labels using a set of fused modalities, achieving in this way a result as close as possible to the expected one:

$$\underset{k,j}{\operatorname{argmin}} I(\{X_k,X_j\},Y), k,j \in \{1,\ldots,n_{mod}\}, k<j \tag{5}$$

C. Score Level Fusion

In our work, we observed that emotional states are more dominant depending on the existing modalities, e.g. some of them are visual prevailing, while others are stronger displayed through the audio modality. As modalities can be complementary to each other and display varying performance characteristics across emotions, we take advantage of this aspect for predicting emotional states in a collaborative manner at the decision level. We apply this scheme in two stages, first we learn classifiers for each single modality separately, and then we combine the scores of specific modalities at the decision level. In the first stage, each modality classifier is regarded as an expert model due to its distinctive performance in emotion prediction. In this phase, we take advantage of the best fused modalities obtained using the information gain principles presented in section IV-B. In addition, we also use particular classification techniques for each modality or fused feature vector before feeding it into the decision level algorithm, as different classifiers are better fitted for specific modalities[34]. Then, we apply a weighting scheme that takes into consideration the performance of each modality with respect to each affective state. The final decision is obtained using a weighted-sum of the prediction given by each modality. For optimizing our results, we employ a genetic algorithm (GA) for assigning weights to each modality score for each affective state.

For the aforementioned reasons, we applied a re-weighting per modality and per emotion as a hyper-parameter search over the model prediction scores for each emotion. This optimized search algorithm adjusted the parameters to produce a collaborative and complementary scheme. Accordingly, GA learns the weights of the final decision for the modalities combination and their predictions. The search space S of GA depends on the number of modalities fed into it: n_{lab}, and the number of predictions n_{lab} for each modality which is fixed to 7 in our case (the number of basic emotions and the neutral state). Therefore, the search space matrix S has $[n_{mod} \times n_{lab}]$ dimensions.

Prior to learning the weighting scheme of the selected modalities, we considered lower and upper bounds constraints to avoid over-fitting the given modalities by GA. We use the following constraints to regularize the learning during the weight parameters search:

$$0 \leqslant S(k,i) \leqslant 1, \& k \in \{1,\ldots,n_{mod}\}, i \in \{1,\ldots,n_{emo}\} \quad (6)$$

V. Evaluation and Results

In this section, we first introduce the dataset chosen for our experiments, then we present an extensive study and evaluation of the modalities employed in our proposed framework. We first evaluate each modality separately to assess their discriminative properties and to estimate their efficiency. Then we apply feature level fusion on all modalities. Finally, we apply the proposed hierarchical scheme for selecting modality combination based on information gain principles and genetic algorithm optimization.

A. Dataset

Acted Faces Emotion In The Wild (AFEW): There are several facial expressions datasets gathered in controlled environments, which mainly contain still images or videos of frontal faces. Furthermore, the facial expressions are posed, limiting the capacity of the data to reflect real-world challenging conditions. Therefore, we chose to base our work on the AFEW dataset for several reasons: (i) AFEW is a challenging dataset with occlusions, varying illumination and head poses, which meets real-world conditions; (ii) it provides baseline results and an evaluation protocol which is useful to evaluate the efficiency of our scheme and (iii) it is currently studied by the research community, as it was the subject of several competitions over the last few years.

The Acted Faces Emotion in the Wild dataset is divided into three subsets: Train (773 samples), Validation (383 samples) and Test (593 samples), while only the Train and Validation sets are publicly available. It has both audio and video modalities. In this dataset, the task is to classify a sample audio-video clip into one of the seven categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. The baseline results are based on Mixture of Parts and INTRAFACE for prepossessing and facial alignment and LBP-TOP and SVM for feature extraction and classification, achieving 38.8% accuracy for the validation set and 40.47% for the test set.

The dataset has in the wild settings, containing wide pose, expression and illumination variation, which reflect the real-world challenging conditions. Fig. 6 illustrates examples of still images and a face track, where the various challenging illumination and pose conditions can be noticed.

Video Modeling: As we described in section III-C, we use video-pooling, where the low-level features are pooled from all the frames across a face track in each video of the AFEW train and validation sets. Then, we compute a single Fisher vector over the whole face track by aggregating and encoding low-level features (e.g. DSIFT or CNN features) of all frames.

Fig. 6. Example of still images of affective states and a face track from the AFEW dataset.

TABLE II. Performance of individual modalities on AFEW validation set using linear SVM classifier

Modalities	Features	Accuracy
Visual	FV on DSIFT	39.4%
	FV on ROI DSIFT	39.2%
	FV on CNN features	40.0%
	FV on hand crafted geometric features	32.8%
Audio	FV on audio features	36.4%
	Raw audio features without FV	30.8%

B. Evaluations Metrics

In our experiments, we take into consideration several evaluation criteria: (i) Accuracy, which is the number of correctly classified video samples; (ii) Confusion Matrix between the ground truth and the predicted emotion labels and (iii) Symmetric KL-Divergence, where we aim to minimize the symmetric KL-divergence between the predicted labels and the true labels. We train our proposed approach on the train set and test it on the validation set.

C. Unimodal Experiments

Firstly, we apply the evaluation metrics for each representation separately on the AFEW validation set. These experiments aim to show the baseline performance of different features for both visual and audio modalities, which are presented in Table II. The best results are obtained for the visual modality, where CNN appearance based features are slightly better than the baseline results. Another interesting finding is represented by obtaining an improved accuracy of audio features when encoded with Fisher vectors in comparison to the raw audio features. As shown in last two raws of table II, this gain in the performance by almost 6% is significant.

D. Multimodal Emotion Prediction

Feature Level Fusion: was introduced in sections III-C and IV-A. We first encode the low-level features of audio and visual modalities using a Fisher vector representation. To such an extent, we obtain a general representation of each modality that shares similar distribution properties. Next we concatenate the Fisher vectors of pair modalities and then perform the classification task using linear Support Vector Machines (linear-SVM).

In case of concatenating the Fisher vectors of all modalities, the accuracy on the AFEW validation set is 45.6%. In addition, using IG principles, based on minimizing the symmetric KL-divergence between the predicted labels of concatenated modalities and the ground truth labels of the test samples, we selected the best combination of features to concatenate,

Fig. 7. Confusion matrix of the AFEW validation set for the IG based feature level fusion.

followed by the emotion prediction task. This leads to an overall accuracy of 47.5%, for a reduced set of modalities composed of (CNN, geometric, DSIFT and audio). Fig. 7 shows the confusion matrix of affective states corresponding to this approach. However, fusing all the modalities into one feature vector is less efficient for the classification task and also slower in comparison with to the following scheme of score level fusion, which is based on the fusion of the best pair modalities.

Dataset Influence: in the AFEW dataset, there are a number of videos for which it is very hard to decide their emotion label only from the visual information. For example, we noticed that, facial expressions, in many videos, labeled as surprise have been classified as an angry emotion by several human annotators, observation also supported by [43]. Therefore, we need more contextual and complementary information to enhance the accuracy and to correctly classify these ambiguous videos. Thus, the audio modality represents one way to boost up the performance of the classification task by adding contextual information. In addition, we observed that in both fusion schemes, employing different features and modalities led to a better accuracy. In Table II, the performance of separate modalities is reported, (e.g. face and audio accuracies are 39.4% and 36.4% respectively). Accordingly, table III illustrates the performance for feature level fusion, where we notice that the fusion of visual and audio modalities increases the performance to 43.3%, mainly due to the complementarity of the two channels.

Furthermore, as we aim to investigate the advantages of a hierarchical fusion scheme, we apply the information gain theory based on minimizing the KL-divergence for deciding the best pair of modalities to be combined in feature level fusion. In Table III, the three best pairs of modalities are included, obtained by concatenating the FV of the following features: (i) ROI DSIFT and geometric features, (ii) audio and DSIFT features, and (iii) geometric and DSIFT features. The KL-divergence and the obtained accuracy on the AFEW validation set, are shown in Table III. We notice the increase in the performance over the unimodal results in Table II in all

TABLE III. PERFORMANCE OF FEATURE LEVEL FUSION ON CONCATENATED PAIR MODALITIES OF AFEW VALIDATION SET.

Fusion	Modalities	Sym-KLDV	Accuracy
FLF	ROI DSIFT and Geometric	0.2622	43.6%
FLF	Audio and DSIFT	0.2626	43.33%
FLF	Geometric and DSIFT	0.3244	40.6%

TABLE IV. SCORE LEVEL FUSION (SLF) OF PAIR MODALITIES IN TABLE III ON AFEW VALIDATION SET.

Score Level Fusion	Accuracy
Genetic Algorithm Based Fusion	48.9%
Performance Based Weights Fusion	44.4%

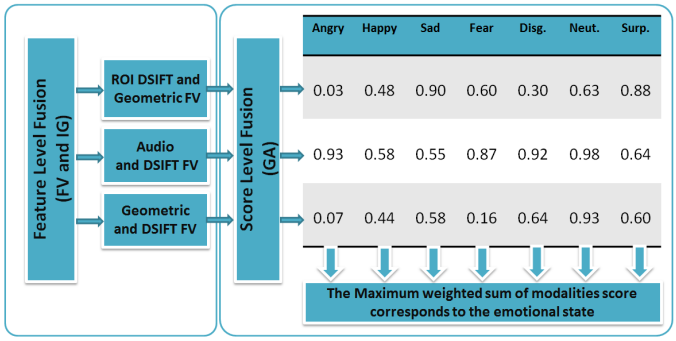

Fig. 8. The resulted modalities and features from feature level fusion by FV and IG, and the weights per-modality and per emotion obtained by score level fusion using GA.

cases, which proves the benefits of both feature level fusion and of the Fisher vector representation.

Score Level Fusion: Following the feature-level fusion step, we fed the obtained pair modalities predicted scores into late level fusion, and searched for the best weights to fuse them. As emotions are more dominant depending on the audio or visual modalities, score level fusion aims to breakdown the fusion into this level, where we assign weights per-modalities and per-emotion. For achieving this purpose, we employ two approaches: (i) firstly we use as weights of each modality the diagonal elements of the confusion matrix; (ii) the second technique uses GA for searching the best weights to fuse the given modalities.

In the first case of using the performance based weights, the overall accuracy is 44.4%. However, in the second case, we apply a genetic algorithm as an optimization search algorithm, using 5-fold cross validation. Fig. 8 depicts the score level fusion approach together with the weights per-modality and per-emotion in the best performing case. The GA-optimized search resulted in an enhanced performance with an average accuracy of 48.9%. The results obtained in both cases are shown in Table IV. In comparison to the feature level fusion and the performance based weights late fusion, the genetic algorithm outperformed both approaches, leading to a better fusion model.

In addition, the best weights of among the 5 folds gave an even better performance, obtaining an accuracy of 53.06%. Fig. 9 contains the normalized confusion matrix for the validation set obtained using the best weights of the score level fusion. When compared with the confusion matrix for the best feature-

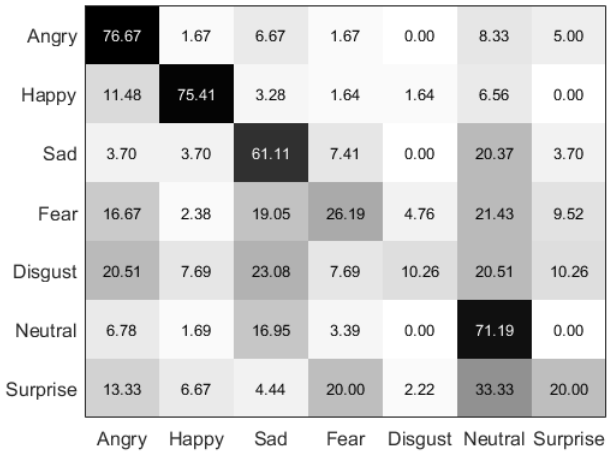

Fig. 9. Confusion matrix of the AFEW validation set for the best score level fusion.

TABLE V. PERFORMANCE OF DIFFERENT METHODS ON AFEW VALIDATION SET

Approach	Accuracy
Baseline AFEW [48]	38.8%
Gideon, et al. [44]	43.86%
Chen, et al. [45]	50.65%
Ding, et al.[46]	51.20%
Ours	**53.06%**
Barga, et al.[47]	**59.42%**

level fusion, we notice a substantial improved performance for several emotions (angry, sad and neutral).

Therefore, we can notice the advantages of our multimodal learning scheme for combining the feature level and the score level fusion in a hierarchical manner based on IG principles and GA optimization. Additionally, score level fusion has the advantage of re-weighting existing modalities to benefit from their individual expertise and performance on specific emotions.

Furthermore, our proposed system achieves better results than the baseline provided by the AFEW validation set [23] and also when compared with other approaches [44], [45], [46], as presented in Table V. The work described in [47] achieves a better score, by employing a massive amount of training data for fine-tuning the CNN features, while in our approach, we only used the training set available in the AFEW dataset, limited to 773 video samples.

VI. CONCLUSION

In this paper, we proposed a new framework for multimodal hierarchical emotion recognition, tested on a challenging dataset AFEW'16. We employed a Fisher vector representation for capturing the discriminative and temporal information across the frames in each video sample. This encoding was applied to different types of features (e.g. Dense SIFT, geometric, CNN, audio) enabling mapping them into a common space, where feature level fusion is performed achieving 45.6% accuracy when all the features vectors are used. Furthermore, we used information gain principles, for selecting the best combination of features to be fused, leading to an improved system performance of 47.5%. Next, we also applied a decision-level fusion approach on top of the best feature and modality combinations obtained through feature-level fusion. We optimized the modality weights for each emotional state using a genetic search algorithm, which lead to an overall 48.9% accuracy, while the best weights attained a score of 53.06% on the validation set, surpassing by 14% the dataset baseline of 38.81%.

As future work, we plan to further optimize our framework by fine-tuning our CNN features using several emotion datasets, in the training process.

ACKNOWLEDGMENT

This work was supported by the Horizon 2020 funded project MaTHiSiS (Managing Affective-learning THrough Intelligent atoms and Smart InteractionS) nr. 687772 (http://www.mathisis-project.eu/).

REFERENCES

[1] P. Ekman, "Facial expression and emotion," *American psychologist*, vol. 48, no. 4, pp. 384–392, 1993. 1, 2

[2] G. M. Nagi, R. Wirza, F. Khalid, and M. Taufik, "Region-based facial expression recognition in still images," *Journal of Information Processing Systems*, vol. 9, no. 1, 2013. 1

[3] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. of the 15th ACM on Int. Conf. on Multimodal interaction*, 2013, pp. 543–550. 1, 2

[4] 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 *IEEE Conf. on Computer Vision and Pattern Recognition-Workshops*, 2010, pp. 94–101. 1

[5] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. of the IEEE Int. Conf. on Automatic Face Gesture Recognition and Workshops*, 1998. 1

[6] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, no. 12, pp. 1615–1618, 2003. 1

[7] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Webbased database for facial expression analysis," in *Proc. of the IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2005, pp. 317–321. 1

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, vol. 2, 2011. 1

[9] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 1998, pp. 200–205. 1, 2

[10] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009. 1, 2

[11] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 4125–4128. 1, 2

[12] A. Asthana, J. Saragih, M. Wagner, and R. Goecke, "Evaluating AAM fitting methods for facial expression recognition," in *Proc. of the IEEE Int. Conf. on Affective Computing and Intelligent Interaction, ACII09*, 2009, pp. 598–605. 1

[13] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN)*, 2015. 1

[14] 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. of the 16th Int. Conf. on Multimodal Interaction*, 2014, pp. 494–501. 1, 2, 3

[15] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems*, 2012, pp. 2222–2230. 1, 2

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. of the British Machine Vision Conference (BMVC)*, 2015, pp. 1–12. 1, 2, 4

[17] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, N. Vidrascu, L. K. Amir, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Proceedings of IS-LTC*, 2006, pp. 240–245. 1

[18] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012. 1, 2

[19] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006. 1

[20] R. Kullback, S.and Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. 2, 6

[21] Z. Meng, S. Han, M. Chen, and Y. Tong, "Feature level fusion for bimodal facial action unit recognition," in *IEEE Int. Symposium on Multimedia (ISM)*, 2015. 2

[22] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," in *Proc. of the IEEE Int. Conf. on Pattern Recognition (ICPR)*, 2014. 2

[23] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *ACM Conf. on Multimodal Interaction (ICMI'13)*, 2013. 2, 9

[24] 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 (PAMI)*, vol. 31, no. 1, pp. 39–58, 2009. 2

[25] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011. 2

[26] R. T. Ionescu, M. Popescu, and C. Grozea, "Local learning to improve bag of visual words model for facial expression recognition," in *Workshop on challenges in representation learning at ICML*, 2013. 2

[27] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 6, pp. 915–928, 2007. 2

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. 2

[29] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proc. of the Int. Conf. on Multimodal Interaction*, 2016, pp. 445–450. 2

[30] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 3687–3691. 2

[31] D. Wu, L. Pigou, P.-J. Kindermans, L. Nam, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," 2016. 2

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. of the 28th Int. Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696. 2

[33] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 105, no. 3, pp. 222–245, 2013. 2, 5

[34] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2013, pp. 1–6. 3, 6

[35] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539. 4, 5

[36] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. of the British Machine Vision Conference (BMVC)*, 2013. 4

[37] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman, "A compact and discriminative face track descriptor," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1693–1700. 4, 5

[38] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, "Effective geometric features for human emotion recognition," in *IEEE 11th Int. Conf. on Signal Processing (ICSP)*, vol. 1, 2012, pp. 623–627. 5

[39] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in *Proc. of the Int. Conf. on Multimodal Interaction*, 2015, pp. 459–466. 5

[40] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the ACM Multimedia (MM)*, 2013, pp. 835–838. 5

[41] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Int. Conf. on Affective Computing and Intelligent Interaction*, 2011, pp. 415–424. 5

[42] D. Johnson and S. Sinanovic, "Symmetrizing the kullback-leibler distance," *IEEE Trans. on Information Theory*, 2000. 6

[43] T. Gehrig and H. K. Ekenel, "Why is facial expression analysis in the wild challenging?" in *Proc. of the 2013 on Emotion recognition in the wild Challenge and Workshop*, 2013, pp. 9–16. 8

[44] J. Gideon, B. Zhang, Z. Aldeneh, Y. Kim, S. Khorram, D. Le, and E. M. Provost, "Wild wild emotion: a multimodal ensemble approach," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 501–505. 9

[45] S. Chen, X. Li, Q. Jin, S. Zhang, and Y. Qin, "Video emotion recognition in the wild based on fusion of multimodal features," in *Proc. of the 18th ACM Int. Conf. on Multimodal Interaction*, 2016, pp. 494–500. 9

[46] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proc. of the 18th ACM Int. Conf. on Multimodal Interaction*, 2016, pp. 506–513. 9

[47] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. of the 18th ACM Int. Conf. on Multimodal Interaction*, 2016, pp. 433–436. 9

[48] 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*. ACM, 2016, pp. 427–432. 9