Automatic Recognition of Facial Displays of Unfelt Emotions

Kaustubh Kulkarni*, Ciprian Adrian Corneanu*, Ikechukwu Ofodile*, *Student Member, IEEE,* Sergio Escalera, Xavier Baró, Sylwia Hyniewska, *Member, IEEE,* Jüri Allik, and Gholamreza Anbarjafari, *Senior Member, IEEE*

Abstract—Humans modify their facial expressions in order to communicate their internal states and sometimes to mislead observers regarding their true emotional states. Evidence in experimental psychology shows that discriminative facial responses are short and subtle. This suggests that such behavior would be easier to distinguish when captured in high resolution at an increased frame rate. We are proposing SASE-FE, the first dataset of facial expressions that are either congruent or incongruent with underlying emotion states. We show that overall the problem of recognizing whether facial movements are expressions of authentic emotions or not can be successfully addressed by learning spatio-temporal representations of the data. For this purpose, we propose a method that aggregates features along fiducial trajectories in a deeply learnt space. Performance of the proposed model shows that on average it is easier to distinguish among genuine facial expressions of emotion than among unfelt facial expressions of emotion and that certain emotion pairs such as contempt and disgust are more difficult to distinguish than the rest. Furthermore, the proposed methodology improves state of the art results on CK+ and OULU-CASIA datasets for video emotion recognition, and achieves competitive results when classifying facial action units on BP4D datase.

Index Terms—Affective Computing, Facial Expression Recognition, Unfelt Facial Expression of Emotion, Human Behaviour Analysis.

1 Introduction

In *"Lie to me"*, an American crime television drama, Dr. Cal Lightman, a genius scientist, is assisting investigators in the police departments to solve cases through his knowledge of applied psychology. This is mainly done through interpreting subtle facial expressions of emotion (FEE) and body language of alleged offenders in order to evaluate their authentic motivation or emotional experience.

However in real life, humans are very skilled in concealing their true affective states from others and displaying emotional expressions that are appropriate for a given social situation. Untrained observers tend to perform barely above chance level when asked to detect whether observed behaviours genuinely reflect underlying emotions [1], [2].

- *I. Ofodile and G. Anbarjafari are with the the iCV Research Group, Institute of Technology, University of Tartu, Tartu, Estonia. E-mail: {ike,shb}@icv.tuit.ut.ee*
- *K. Kulkarni is with the Computer Vision Center, Barcelona, Spain. E-mail: kaustubh14jr@gmail.com*
- *Ciprian A. Corneanu and S. Escalera are with the Computer Vision Center , University of Barcelona, Barcelona and University of Autonoma, Barcelona, Spain. E-mail: {kaustubh14jr,cipriancorneanu}@gmail.com, sergio@maia.ub.es*
- *X. Baró is with the Computer Vision Center and Universitat Oberta de Catalunya, Barcelona, Spain. Email: xbaro@uoc.edu*
- *S. Hyniewska is with the Institute of Physiology and Pathology of Hearing, Poland. E-mail: s.hyniewska@bath.ac.uk*
- *J. Allik is with Department of Psychology, University of Tartu and The Estonian Center of Behavioral and Health Sciences, Tartu, Estonia. E-mail: juri.allik@ut.ee*
- *G. Anbarjafari is also with Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep, Turkey.*
- ** Authors contributed equally in this work.*

Manuscript received July 13, 2017; revised Xxxxx XX, 2017.

Fig. 1: People may have difficulties in expressing emotions that look genuine when these do not correspond to the emotional state they are experiencing. In the case of smiling, differences can be observed in the contraction of the *orbicularis oculi* muscle around the eyes. *Left*: The lack of orbicularis oculi contraction has often been considered a marker of unfelt or even deceitful expressions. *Right*: A strong orbicularis oculi contraction, with very visible "crows feet" around the corners of the eyes, has often been considered a marker of genuine expressions.

This is a particularly difficult judgement when relying on visual cues only [3]. Even for professional psychologists it is difficult to recognise deceit in emotional displays as there are numerous factors that need to be considered [4], [5]. Although human perception is naturally biased in its interpretation of perceived facial displays (e.g. see [6]), an incorrect appraisal of the sincerity of observed facial displays can have detrimental consequences [7]. In the clinical context, credibility of patients is of great importance given the risk of simulated affective reactions, psychiatric syndromes [8] or the need to evaluate pain levels [9]. In the legal context, body language (including facial displays) is often consid-

ered a source of valuable information by judges and jurors [10], [11]. Beyond facts and evidence, facial displays, e.g. of remorse or anger in the defendant, are one factor that influences jurors in their verdicts [11]. These and other potential applications would benefit not only from improved human detection but also from the possibility to automatically discriminate between subtle facial expressions such as displays of genuine and unfelt emotional states. On top of legal and medical settings, improved human-computer interaction for assistive robotics [12]–[14], treatment of chronic disorders [15] and assisting investigation conducted by police forces [16]–[18] would be just a few.

An emotional display is considered unfelt (or masked) when it does not match a corresponding emotional state. There are three major ways in which emotional facial expressions are intentionally manipulated [19]: an expression is *simulated* when it is not accompanied by any genuine emotion, *masked* when the expression corresponding to the felt emotion is replaced by a falsified expression that corresponds to a different emotion, or *neutralized* when the expression of a true emotion is inhibited while the face remains neutral. All along this work, the term genuine FEE is used to denote FEEs congruent with the affective state, while the term unfelt FEE is used for denoting FEEs incongruent with the emotional state (aka masked) .

It has been argued that liers, deceivers and displayers of unfelt emotions would be betrayed by the leakage of their genuine emotional states through their nonverbal behaviour [4], [20], [21]. This is supposed to happen through subtle facial expressions of short duration, as well as changes in pitch, posture and body movement.

Studies on the unfelt display of emotion mostly originated based on Duchenne de Boulogne's work, a nineteenth century French scientist. He is considered the first to have differentiated facial actions observed in displays of felt and unfelt emotions [22], [23]. Part of his legacy concerns what is considered the typical genuine smile – often called a Duchenne smile. Duchenne smiles involve the contraction of the orbicularis oculi muscle (causing lifting of the cheeks and crow's feet around the eyes) together with the zygomaticus major muscle (pulling of lip corners upwards) [24]–[32] (see Fig. 1). In contrast, a masking smile (aka a non-Duchenne smile) can be used to conceal the experience of negative emotions [28], [32]–[35].

Although it has been argued that the orbicularis oculi activation is absent from masked facial expressions of enjoyment, empirical evidence is not conclusive. For example, in a database presenting 105 posed smiles 67% of them were accompanied by the orbicularis oculi activation [36]. Another study showed that over 70% of untrained participants were able to activate the majority of eye region action units, although not one action at a time, as they managed to perform them through the reliance and co-activation of other action units. The poorest performance was for the deliberate activation of the *nasolabial furrow deepener*, which is often observed in sadness and which was performed successfully only by 20% while the orbiculari oculi by 60% of participants.

Although a variety of studies have focused on the evaluation of how genuine some FEEs might be while relying on the analysis of still, i.e. static, images, not much attention has been paid to dynamics as evaluated in a sequence of frames [37]–[43]. In a naturalistic setting, FEEs are always perceived as dynamic facial displays, and it is easier for humans to recognize facial behaviour in video sequences rather than in still images [44]–[46].

It has been asserted that while trying to simulate the expression of an unfelt emotion, cues of the actual felt emotion appeared along cues related to the masked expression, which made the overall pattern difficult to analyse [47]. Leakages of a genuine emotion have been observed more frequently in the upper part of the face, while cues the lower half of the face was more often manipulated in order to express an unfelt emotion [48]–[51].

In this work, we propose a new data corpus containing genuine and unfelt FEE. While numerous studies involving the analysis of genuine or truthful behaviours rely on video recordings of directed interviews, such as the work in [2], studies that analysed nonverbal behaviour while controlling for the emotional state of subjects are rare [49].

When designing experiments that require facial emotion displays as independent variables, posed facial expressions of subjects being instructed to act out a particular emotion are often used. This is thought to provide greater control over the stimuli than a spontaneous emotion display might, in the sense that other variables such as context and the physical appearance of subjects (even hair style or make-up) are much less variable and will not bias the observers in an uncontrolled way.

To record FEEs, participants are usually asked to practice the display of specific emotions. In order to achieve a display close to a genuine emotional expression, the process can be facilitated through the presentation of FEEs [52], [53], or other pictures [49] or videos inducing emotions in line with the ones to be expressed [54], or mental imagery and related theatre techniques [55]. Such paradigms have been frequently used for recording and creating emotional expression databases [53], [55]–[58].

In addition to the published dataset, we propose a complete methodology that has the capacity to recognise unfelt FEEs and generalises to standard public emotion recognition datasets. We first train a Convolutional Neural Network (CNN) to learn a static representation from still images and then pull features from this representation space along facial landmark trajectories. From these landmark trajectories we build final features from sequences of varying length using a Fisher Vector encoding which we use to train a SVM for final classification. State-of-the-art results are presented on CK+ and Oulu-Casia, two datasets containing posed FEEs. Moreover, close to state-of-the-art results are shown on a more difficult problem of recognising spontaneous facial Action Units on BP4D-Spontaneous. We finally provide benchmarking and outperform the methods from the recent ChaLearn Challenge [59] on the proposed SASE-FE dataset.

The rest of the paper is organised as follows: in Section 2 we describe related work in FEEs recognition, in Section 3 we introduce the new SASE-FE dataset, in Section 4 we detail the proposed methodology, and Section 5 concludes the paper.

2 Related Work

This section first reviews main works on recognition of FEE, and then recognition of genuine and unleft FEE.

2.1 Recognizing Facial Expressions of Emotion

Automatic facial expression recognition (AFER) has been an active field of research for a long time. In general, a facial expression recognition system consists of four main steps. First the face is localised and extracted from the background. Then, facial geometry is estimated. Based on it, alignment methods can be used to reduce variance of local and global descriptors to rigid and non-rigid variations. Finally, representations of the face are computed either globally, where global features extract information from the whole facial region, or locally, and models are trained for classification or regression problems.

Features can be split into static and dynamic, with static features describing a single frame or image and dynamic ones including temporal information. Predesigned features can also be divided into appearance and geometrical. Appearance features use the intensity information of the image, while geometrical ones measure distances, deformations, curvatures and other geometric properties. This is not the case for learned features, for which the nature of the extracted information is usually unknown.

Geometric features describe faces through distances and shapes. These can be distances between fiducial points [60] or deformation parameters of a mesh model [61], [62]. In the dynamic case the goal is to describe how the face geometry changes over time. Facial motions are estimated from color or intensity information, usually through Optical flow [63]. Other descriptors such as Motion History Images (MHI) and Free-Form Deformations (FFDs) are also used [64]. Although geometrical features are effective for describing facial expressions, they fail to detect subtler characteristics like wrinkles, furrows or skin texture changes. Appearance features are more stable to noise, allowing for the detection of a more complete set of facial expressions, being particularly important for detecting micro-expressions.

Global appearance features are based on standard feature descriptors extracted on the whole facial region. Usually these descriptors are applied either over the whole facial patch or at each cell of a grid. Some examples include Gabor filters [65], Local Binary Pattern (LBP) [66], [67], Pyramids of Histograms of Gradients (PHOG) [68] and Multi-Scale Dense SIFT (MSDF) [69]. Learned features are usually trained through a joint feature learning and classification pipeline. The resulting features usually cannot be classified as local or global. For instance, in the case of Convolutional Neural Networks (CNN), multiple convolution and pooling layers may lead to higher-level features comprising the whole face, or to a pool of local features. This may happen implicitly, due to the complexity of the problem, or by design, due to the topology of the network. In other cases, this locality may be hand-crafted by restricting the input data.

Expression recognition methods can also be grouped into static and dynamic. Static models evaluate each frame independently, using classification techniques such as Bayesian Network Classifiers (BNC) [61], [70], Neural Networks (NN) [71], Support Vector Machines (SVM) [62] and Random Forests (RF) [72]. More recently, deep learning architectures have been used to jointly perform feature extraction and recognition. These approaches often use pre-training [73], an unsupervised layer-wise training step that allows for much larger, unlabelled datasets to be used. CNNs are by far the dominant approach [74]–[76]. It is a common approach to make use of domain knowledge for building specific CNN architectures for facial expression recognition. For example, in AU-aware Deep Networks [77], a common convolutional plus pooling step extracts an over-complete representation of expression features, from which receptive fields map the relevant features for each expression. Each receptive field is fed to a DBN to obtain a non-linear feature representation, using an SVM to detect each expression independently. In [78] a two-step iterative process is used to train Boosted DBN (BDBN) where each DBN learns a non-linear feature from a face patch, jointly performing feature learning, selection and classifier training.

Dynamic models take into account features extracted independently from each frame to model the evolution of the expression over time. Probabilistic Graphical Models, such as Hidden Markov Models (HMM) [79], are common. Other techniques use Recurrent Neural Network (RNN) architectures, such as Long Short Term Memory (LSTM) networks [63]. Some approaches classify each frame independently (e.g. with SVM classifiers [80]), using the prediction averages to determine the final facial expression. Intermediate approaches are also proposed where motion features between contiguous frames are extracted from interest regions, afterwards using static classification techniques [61]. For example, statistical information can be encoded at the frame-level into Riemannian manifolds [81].

2.2 Recognizing Genuine and Unfelt Facial Expressions of Emotion: Experimental psychology

Psychologists differentiate facial displays that are produced involuntarily, e.g. automatic "expressions" of felt emotional states, and displays produced voluntarily, e.g. for social purposes and not as a leakage of an experienced emotional state. Unfelt displays of emotions are often necessary for social acceptance. First, individuals judge whether any facial display is appropriate for any given situation, which leads to displaying unfelt but socially expected emotions for a healthy identity construction [82]. Second, societal norms tend to accentuate the need for positive emotions and positive emotional displays, while devaluing native displays or even expecting individuals to inhibit any negative emotional experience [83]. Third, some displays are facilitated in situations where they could provide social support, such as in the case of sadness [84]. Finally, they can be used prosocially, e.g. in order not to hurt other individuals' feelings [85]. Already in early childhood individuals start to learn which facial displays are appropriate in different daily life situations [86], [87] and become very skilled displayers of unfelt emotions [88]. Human observers are less skilled in detecting displays of unfelt emotions [4]. Psychologists accentuate the fact that there is no golden channel for consistent deception detection and no single cue at the non-verbal, verbal or physiological level is currently considered

sufficient (see [89] for additional information). Whereas for a general deception detection several prominent cues are expected to co-occur, e.g, illustrators, blink and pause rate, speech rate, vague descriptions, repeated details, contextual embedding, reproduction of conversations, and emotional 'leakage' in the face [89]; the subjective experience behind emotional displays might be judged by additional rules. So far unfelt emotion research is less advanced than that on pure deceit detection, and only a few emotions have been studied by psychologists in terms of limitations observed in voluntarily produced displays of unfelt emotions.

2.3 Recognizing Genuine and Unfelt Facial Expressions of Emotion: Affective computing

Emotion perception by humans or machines stands for the interpretation of particular representations of personal feelings and affects expressed by individuals, which may take different forms based on the circumstances governing their behaviour at the time-stamp at which they are evaluated [90], [91].

Amongst audiovisual sources of information bearing clues to the emotions being expressed, the ones extracted from single or multiple samples of facial configurations, i.e. facial expressions, provide the most reliable basis for devising the set of criteria to be incorporated into the foregoing analysis [47], [92] and are, therefore, the most popular alternatives utilised in numerous contexts, such as forensic investigation and security. These settings often rely on the assessment of the correspondence of the displayed expression to the actual one.

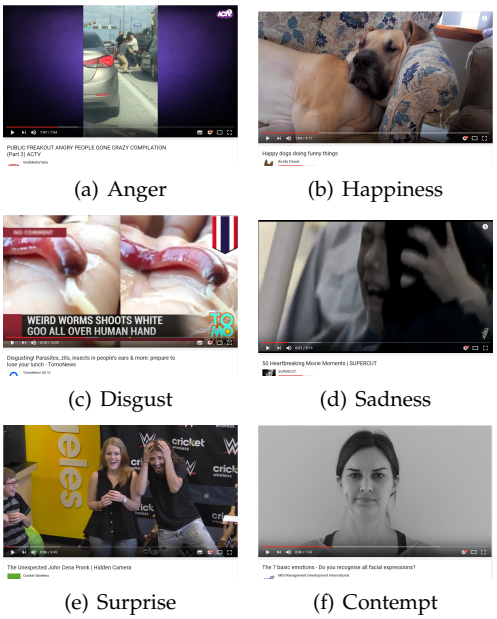
3 SASE-FE Dataset

A number of affective portrayal databases exist; however, none meets the required criteria for our analysis of controlled genuine and unfelt emotional displays presented in high resolution at an increased frame rate. To answer those needs, the SASE-FE database was created.

The SASE-FE database consists of 643 different videos which had been recorded with a high resolution GoPro-Hero camera. From the inital 648 recordings, 5 were eliminated post-hoc as the participants did not completely meet the defined protocol criteria. As indicated in Table 1, 54 participants of ages 19-36 were recorded. The reasoning behind the choice of such a young sample is that older adults have different, more positive responses than young adults about feelings and they are quicker to regulate negative emotional states than younger adults [93], [94].

Participants signed a written informed consent form after the experimental and recording procedures were explained. All participants agreed for their data to be released for research purposes and all data can be accessed by contacting the authors. The data collection and its use are based by the ethical rules stated by University of Tartu, Estonia.

For each recording, participants were asked to act two FEEs in a sequence, a genuine and an unfelt one. The participants displayed six universal expressions: Happiness, Sadness, Anger, Disgust, Contempt and Surprise. The subjects were asked if they felt the emotion and the large majority confirmed, but no recording of their answer was made. To increase the chances of distinguishing between the two FEEs presented in a sequence, two emotions were chosen based on their visual and conceptual differences as observed on the two dimensions of valence and arousal [95]–[97]. Thus a visual contrast was created by asking participants to act Happy after being Sad, Surprised after being Sad, Disgusted after being Happy, Sad after being Happy, Angry after being Happy, and Contemptuous after being Happy [98], [99]. For eliciting emotion, subjects were shown videos in line with the target emotion. Emotion elicitation through videos is a well established process in emotion science research [100]. Videos were short scenes from YouTube selected by psychologists . Fig. 2 shows captures from videos that have been used for inducing specific emotions in the participants.



(a) Anger (b) Happiness

(c) Disgust (d) Sadness

(e) Surprise (f) Contempt

Fig. 2: A screenshot of some of the videos that have been used to induce a specific basic emotion in participants.

Throughout the entire setup, participants were asked to start their portrayals from the neutral face. The length of facial expression was about 3-4 seconds. After each genuine FEE, participants were asked to display a neutral state again and then the expression of a second emotion, which was the opposite of the former.

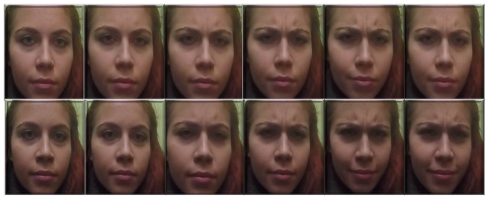
None of the participants were aware of the fact that they would be asked to display a second facial expression. The participant's first two seconds of behavior when performing a facial expression, and more exactly the opposite to the felt emotion, were recorded with the same device and the same configuration. As a result, for each participant we have collected 12 different videos of which 6 are genuine FEE and other 6 are unfelt FEE. The length of captured FEE is not fixed. The process has been closely supervised by experimental psychologists so that the setup would result in realistic recordings of genuine and unfelt FEE. The summary of the SASE-FE dataset is provided in Table 1.

It is important to note that while preparing the SASE-FE database, introduced and used in this work, external factors such as personality or mood of the participants have been

TABLE 1: Summary of SASE-FE database.

Subjects	# of persons	54
	gender distribution	female 41%, male 59%
	age distribution	19 - 36 years
	race distribution	Caucasian 77.8%, Asian 14.8%, African 7.4%
Videos	# of videos	643
	video length	3-4 sec
	resolution	1280 × 960
	#frames (acted/unfelt)	120,216/118,712

ignored, due to the fact that in order to eliminate such external factors several repetitions of the experiment would be necessary, but as a result the participant could start to learn to simulate the facial expressions better. Hence we have decided to ignore such external factors.



(a) Anger

(b) Happiness

(c) Surprise

Fig. 3: Selected examples of pairs of sequences showing genuine (top) and unfelt (below) FEEs of Anger, Happiness and Suprise from the SASE-FE dataset.

4 The Proposed Method

In this section, we present the methodology used for recognising unfelt FEEs from video sequences. As showed in the literature (see Sec. 1 and Sec. 2) most discriminative information is to be found in the dynamics of such FEEs. Following this assumption, we consider learning a discriminative spatio-temporal representation to be central for this problem. We first train a Convolutional Neural Network (CNN) to learn a static representation from still images and then pull features from this representation space along facial landmark trajectories. From these landmark trajectories and inspired by previous work in action recognition [101], a well studied sequence modelling problem, we build final features from sequences of varying length using a Fisher Vector encoding which we use to train a SVM for final classification.

Additionally, the amount of video data available is limited, which requires usage of advanced techniques when training high capacity models with millions of parameters such as CNNs. Fine-tuning existing deep architectures can alleviate this problem to a certain extent but these models might carry redundant information from the pre-trained application domain. In this paper, we use a recently proposed method [102] which proposes a regularisation function which helps using the face information to train the expression classification net.

We follow this section by first discussing the technique we have used to train a CNN on still images with a limited amount of data in Sec. 4.1. Then we show how we build a spatio-temporal representation from static features computed by the CNN in Sec. 4.2. The reader can refer to Fig. 4 for an overview of the proposed method. Specific implementation details will be presented in Sec. 5.1.

4.1 Using efficient knowledge transfer for training a CNN for facial expression recognition

Our proposed training procedure of the CNN for learning static spatial representation: first, we fine tune the VGG-Face network for the facial expression recognition task [103]. We then use this fine tuned network to guide the learning of a so called emotion network (EMNet) [102]. Following [102] the EMNet is denoted as:

$$O = h_{\theta_2}(g_{\theta_1}(I)) \,, \tag{1}$$

where h represents the fully connected layers and g represents the convolution layers, θ_2 and θ_1 are the corresponding parameters of the to be estimated of the fully connected layers and the convolution layers respectively, I is the input image and O is the output before the softmax.

We follow the two step training proposed in [102]. The basic motivation behind this training procedure is that the fine tuned VGG-Face network already gives a competitive performance on the emotion recognition task. We use the ouyput of the VGG-Face to guide the training of the EMNet. In the first step, we estimate the parameters of the only of the convolution layers of the EMNet. In this step, the output of the VGG-Face acts as a regularisation for the emotion net. This step is achieved by maximising the following loss function:

$$L_1 = \max_{\theta_1} \|g_{\theta_1(I)} - G(I)\|_2^2 \,, \tag{2}$$

where, $G(I)$ is the output of the *pool5* layer of the fine tuned VGG-Face network. In the second step we learn the parameters of the fully connected layer, θ_2 of the EMNet by training together the convolution layers, estimated in the previous step, and the fully connected layers. This step is achieved by minimizing the cross entropy loss:

$$L_2 = -\sum_{i=1}^{N}\sum_{j=1}^{M} l_{i,j} log\hat{l}_{i,j} \,, \tag{3}$$

where, $l_{i,j}$ is the ground truth label and $\hat{l}_{i,j}$ is the predicted label.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874996, IEEE Transactions on Affective Computing

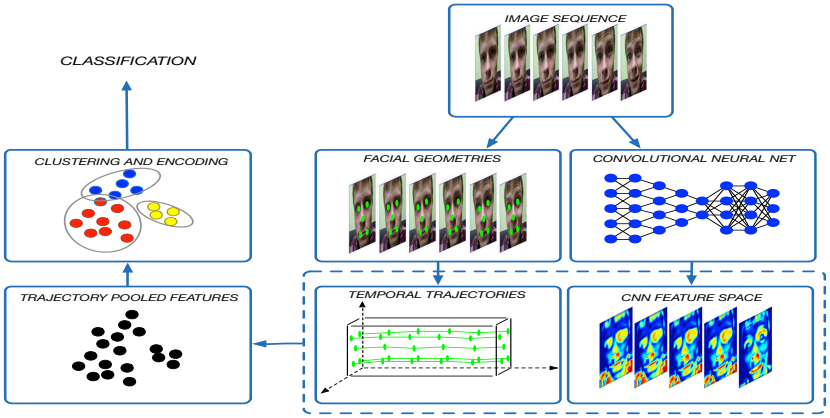


Fig. 4: Overview of the proposed method.

4.2 Learning a spatio-temporal representation

For learning a spatio-temporal representation of the facial video sequences we aggregate features computed by the EMNet along trajectories generated by facial geometries (we will name it TPF-FGT from Trajectory Pooled Features from Facial Geometry Trajectories). First we detect facial geometries in a form of a fixed set of fiducial points in the whole video sequence in a per-frame fashion. To compute the fiducial points we first frontalize all the cropped face with [104]. Then on this cropped frontalized faces we estimate the facial geometry with the with the facial alignment method [105]. This will output 68 fiducial landmark points on each image. The detected fiducial points are tracked across the sequence to form trajectories corresponding to specific locations on the face (e.g corners of the eyes, mouth, see Fig. 4 for an example). We pool features along these trajectories from the EMNet feature space. Such a pooling is advantageous because it captures the temporal relations between the frames. After reducing the dimensionality of the pooled features we learn a set of clusters over the distribution of the features using Gaussian Mixture Models (GMMs). Once the clusters are learned we use Fisher Vector (FV) [106] encoding to produce a compact feature vector for each sequence. The final vectors are used to train a linear classifier. In the rest of section we detail the main steps of the proposed method.

4.2.1 Trajectory pooled features

Given a sequence of images we can compute all corresponding facial geometries with the method previously presented. As each geometry is described by a fixed set of ordered points we can track these points along all the sequence to form trajectories. Along these trajectories we pool features from a feature space of choice. In our case, we use features computed at different layers of an EMNet.

4.2.2 Fisher Vectors

The next step is to get a single vector representation of each emotion video. On this vector an SVM classifier is trained. We choose the Fisher Vector representation for this encoding [107]. Each TPF is an observation vector corresponding to each landmark trajectories. We denote all the observed TPFs in the training set as $\mathbf{X}$. We assume the trajectory pooled features (TPF) are drawn from a Gaussian Mixture Model (GMM). A K component GMM is computed over the training set of TPF . Assuming that the observations in $\mathbf{X}$ are statistically independent the log-likelihood of $\mathbf{X}$ given $\vec{\theta}$ is:

$$\log P(\mathbf{X}|\vec{\theta}) = \sum_{m=1}^{M} \log \sum_{k=1}^{K} w_k \mathcal{N}(\vec{x}_m; \vec{\mu}_k, (\vec{\sigma}_k)^2) \,, \qquad (4)$$

where $\sum_{k=1}^{K} w_k = 1$ and $\vec{\theta} = \{w_k, \vec{\mu}_k, (\vec{\sigma}_k)^2\}$. We assume diagonal covariance matrices. The parameters of the per-class GMMs are estimated with the Expectation maximization (EM) algorithm to optimize the maximum likelihood (ML) criterion. To keep the magnitude of the Fisher vector independent of the number of observations in $\mathbf{X}$ we normalize it by M. Now we can write the closed form formulas for the gradients of the log-likelihood $P(\mathbf{X}|\vec{\theta})$ w.r.t to the individual parameters of the GMM as:

$$\vec{\mathcal{J}}^{\mathbf{X}}_{w_k} = \frac{1}{M\sqrt{w_k}} \sum_{m=1}^{M} \gamma_k(m) - w_k \qquad (5)$$

$$\vec{\mathcal{J}}^{\mathbf{X}}_{\vec{\mu}_k} = \frac{1}{M\sqrt{w_k}} \sum_{m=1}^{M} \gamma_k(m) \left(\frac{\vec{x}_m - \vec{\mu}_k}{(\vec{\sigma}_k)^2} \right) \qquad (6)$$

$$\vec{\mathcal{J}}^{\mathbf{X}}_{(\vec{\sigma}_k)^2} = \frac{1}{M\sqrt{2w_k}} \sum_{m=1}^{M} \gamma_k(m) \left[\frac{(\vec{x}_m - \vec{\mu}_k)^2}{(\vec{\sigma}_k)^2} - 1 \right] , \qquad (7)$$

where $\gamma_k(m)$ is the posterior probability or the responsibility of assigning the observation $\vec{x}_m$ to component k.
Now the FV for each video is constructed by stacking together the derivatives computed w.r.t to the components of the GMM in a single vector. The details of all the close formed formulas can be found in the following paper [108].

5 Experimental Results and Discussions

The experimental results have been conducted on the introduced *SASE-FE* dataset. For comparison, we have replicated experiments on the *Extended Cohn Kanade* (CK+) [109] dataset and the Oulu-CASIA dataset [110] and for spontaneous expression recognition we provide results of the BP4D-Spontaneous dataset [111].
Due to its relatively small size and simplicity, the CK+

is one of the most popular benchmarking datasets in the field of facial expression analysis. It contains 327 sequences capturing frontal poses of 118 different subjects while performing facial expressions in a controlled environment. The facial expressions are acted. Subjects' ages range between 18 and 50 years old, consisting of 69% females and having relative ethnic diversity. Labels of presence of universal facial expressions and the Facial Action Units are provided. The Oulu-CASIA dataset provides facial expressions of primary emotions in three different illumination scenarios. It includes 80 subjects between 23 to 58 years old from whom 73.8% are males. Following other works [102], we only use the strong illumination partition of the data which consists of 480 video sequences (6 videos per subject). It has higher variation and constitutes a good complement to the CK+ for cross validating our method. We also test our method on the 12 action unit recognition problem of in the BP4D-Spontaneous dataset. In this dataset, there are 41 adults with 8 videos each giving a total of 328 videos. Each frame is annotated with 12 facial AUs. In contrast with all previous set-ups, recognizing AUs is a multi-label classification problem.

In the following sections we first discuss the implementation details of each step of the proposed methodology followed by discussion of the experimental results.

5.1 Implementation Details

The proposed methodology consists of the following steps: first, given a video sequence we extract faces from background, frontalize them and localize facial landmarks (see Fig. 5). Second, we fine-tune a pretrained VGG-Face deep network [103] for recognising facial expressions. Third, we use this network for guiding the training of a so called EMNet following work proposed in [102] (see also Sec. 4.1). This second network is used to compute static representations from still images. Fourth, we pool features from the previously computed static representation space along trajectories determined by the facial landmarks. Fifth, we compute fixed length descriptors for each video sequence using the Fisher Vector encoding. These final descriptors are then classified with a linear SVM. We use a leave-one-actor-out validation framework for all our experiments. For the theoretical framework of the spatio-temporal representation and the knowledge transfer training approach of the EMNet, please refer to Sec. 4. For a visual overview of the method see Fig. 4.

Preprocessing. We first extract faces from the video sequences. After faces are extracted we perform a frontalization which registers faces to a reference frontal face by using the method of Hassner et al. [104]. This removes variance in the data caused by rotations and scaling. This frontalization method estimates a projection matrix between a set of detected points on the input face and a reference face. This is then used to back-project input intensities to the reference coordinate system. Self-occluded regions are completed in an aesthetically pleasant way by using color information of the neighbouring visible regions and symmetry. Finally in all synthesised frontal faces we estimated the facial geometry, using a classical, robust facial alignment method [105] trained to find 68 points on the image (an example of the frontalization process is showed in Fig. 5).

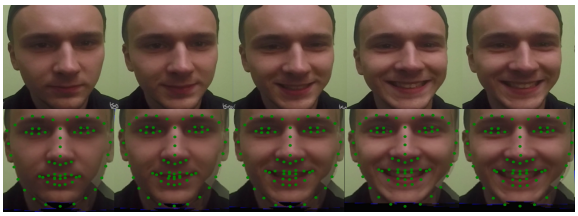


Fig. 5: Illustration of the pre-processing we perform on the data. Detected faces are first extracted and frontalized and facial landmarks localised for each image in the input sequences.

Fine-Tuning the VGG-Face. For all experiments, including fine tuning of the VGG-FACE are done in a 10-fold cross validation for the CK+ and Oulu-CASIA datasets to keep the experiments consistent with [102]. We define a train set of 40 actors, validation set of 5 actors and a test set of 5 actors for the SASE-FE dataset. This set is exactly similar to the partitions defined in [59]. Here we estimate the parameters of our proposed method on the validation set and final results are reported on the unseen test set. Here we also perform an additional experiment, since the training data is limited, we augment the training set of the SASE-FE dataset with additional training data from the Oulu-CASIA [110] and CK+ datasets. These experiments are denoted as *Data Augmentation*. The training is done for 200 epochs with a learning rate of 0.001. It is decreased every 50 epochs. The fully connected layers are randomly initialised with the Gaussian distribution. The min-batch size is 32 and the momentum is 0.9. The dropout is set to 0.5. From each frame the face is cropped and scaled to 224×224. The bottom two convolution layers are left unchanged. In the testing phase, if the CNN is able to recognise more than 50% of the frames in the video correctly then the video is deemed to be correctly classified. For the 6 genuine class and the 6 unfelt class experiment the network is trained for the 12 class problem, and the final fully connected layer is retrained with the appropriate number of classes.

Training the EMNet. The architecture of EMNet is the same as the one proposed in [102]. It consists of 5 convolutional layers each followed by a ReLU activation and a max pooling layer. The filter size of the convolutions layers is 3×3 and that of the pooling layer is 3×3 with a stride of 2. The output of each layer is $64, 128, 256, 512, 512$. Furthermore, we need to add another 1×1 convolutional layer to match the dimensionality of the output of the EMNet to the $pool5$ layer of the fine tuned VGG-Face net for the regularisation in the first step. We append a single fully connected layer of size 256. We just use one layer to prevent overfitting. We use this size of 256 for distinguishing between all multi-class experiments of classifying all emotions in the dataset. The size of the fully connected layer is further reduced to 128 for the binary classification experiment of distinguishing between genuine and unfelt FEEs. This is because the training data available for binary classification is much less than the training data for classifying all emotion.

Trajectory pooled features (TPF). The TPFs from the facial geometry trajectories (TPF-FGT) are aggregated in a rectangular region of pixel size 64×64 which we have

TABLE 2: Our method shows state-of-the-art results when compared with best performing setups on the CK+ dataset. This proves generalisation capacity of this approach.

Method	Accuracy(%)
AURF [77]	92.22
AUDN [113]	93.70
STM-Explet [114]	94.2
LOmo [115]	95.1
IDT+FV [116]	95.80
Deep Belief Network [78]	96.70
Zero-Bias-CNN [117]	98.4
Ours-Final	**98.7**

TABLE 3: Our method shows state-of-the-art results when compared with best performing setups on the Oulu-CASIA dataset. This proves the generalization capacity of such an approach.

Method	Accuracy (%)
DTAGN [118]	81.46
LOmo [115]	82.10
PPDN [119]	84.59
FN2EN [102]	87.71
Ours-Final	**89.60**

experimentally set. This size is scaled by a ratio of the size of the input image and the feature map from the corresponding layer of the neural network. For our experiments we use the TPF descriptors extracted from the conv5 of the EMNet. In order to train the Fisher vector for encoding we perform PCA to decorrelate the dimensions. We experimentally set the number of first principal components to 32.

Fisher Vectors encoding and classification. For encoding the TPFs into lower dimensional representations we used the Fisher Vector encoding. Its efficacy for video analysis has been proven for action recognition [112]. In order to train GMMs, we first decorrelate the dimensions of the TPFs with PCA and reduce its dimension to d. Then, we train a GMM with $k = 16$ mixtures. We can use a low value for k as compared to other papers in the literature because the trajectory computed on the landmarks is already discriminative as compared to the dense trajectory features. This enables us to construct a compact feature representation with FV which is also discriminative. Moreover, we square-root normalise followed by the $L2$ norm of each vector. The video is represented with a $2kd$ dimensional vector. We use the Fisher Vectors to train a linear SVM for classification. The value of the regularisation parameter is set to $C = 100$. The parameters K and C were set using the validation set and then tested on the unknown test set of the SASE-FE dataset.

5.2 Discussion

In this section, we discuss the experimental results obtained by our proposed method. For brevity, we have denoted both in the text and figures the genuine FEE labels by adding a *G* in front of the labels (e.g GSad) and the corresponding unfelt FEE by adding an *U* in the same fashion (e.g UAnger). We start by discussing results on the *Cohn-Kanade*, the Oulu-CASIA and BP4D-Spontaneous datasets and then we discuss the results on the proposed SASE-FE dataset.

TABLE 4: Emotion-wise comparison between our proposed method and [102] on the Oulu-CASIA dataset.

Emotion	Accuracy [102] (%)	Accuracy [Ours-Final] (%)
Anger	75.2	80.1
Disgust	87.3	88.0
Fear	94.9	95.1
Happiness	90.8	89.7
Sadness	88.4	91.3
Surprise	92.0	92.7
Average	87.7	89.6

5.2.1 CK+

The performance of several state-of-the-art methods and the performance of our final method is given in Table 2. We are able to come very close to the state of the art performance on this dataset.

In terms of methodology, [116] is the closest method to our proposed method. The authors of this paper implement the improved dense trajectories framework proposed for action recognition [120] for emotion recognition. We are able to improve their results by aggregating the feature maps along the fiducial points and computing the TPF-FGT features.

We observe that our method is better than methods which use a per frame feature representation rather than per-video as in our case [114], [115]. In [115], this per-frame feature is the concatenation of SIFT features computed around landmark points, head pose and local binary patterns (LBP). They propose a weakly supervised classifier which learns the events which define the emotion as hidden variables. The classifier is a support vector machine which was estimated using the multiple-kernel learning method. From the table we can observe that when landmarks are used along with the CNN feature maps we are able to top their performance. The rest of the methods listed in the table use deep learning techniques to classify emotions [77], [78], [117]. They design networks able to specifically learn facial AUs. We can observe that we out perform the best performing method [117] on the CK+ dataset.

5.2.2 Oulu-CASIA

We also , show the efficacy of our method on a more difficult dataset like the Oulu-CASIA dataset. In Table 3 we can observe that our method outperforms the previous best performance of [102] by 1.9%. In Table 4 we show the emotion-wise comparison between our proposed method and [102]. The two main differences between [102] and our method are that we align the faces and then add the TPFs for classification. In our experiments we observed that aligning the faces on the Oulu-CASIA dataset gave only very marginal improvement while once we add the TPFs for classification then we can get significant improvements. The improvements are especially observed in three emotions Anger, Disgust and Sadness. These emotions are typically confused between each other. This experiment shows that the temporal information is important for emotion recognition.

5.2.3 BP4D-Spontaneous

Considerably more challenging is the recognition of spontaneous expression of emotion. For this purpose we show

TABLE 5: This table presents the comparison of our method with the state-of-the-art on the BP4D dataset.

Method	Average F1-score
LSVM-HOG [121]	32.5
JPML [123]	45.9
AlexNet [121]	38.4
Ours-Final	43.6
Ours-Final + SF	46.8
Ours-Final + SF + CO	48.1
DRML [121]	48.3
CNN + LSTM [122]	**53.9**

results on the BP4D dataset. The evaluation is done in the 3-fold cross validation framework. The evaluation metrics is F1-segment score which is the harmonic mean of the precision and recall. We do the following steps to achieve the final results. First, we finetune the VGG-FACE network on the 12 action units. We sample 100 frames as positive and 200 frames as negative examples per sequence as done in [121]. Then we train the EMnet from VGG-FACE network to do AU recognition. From the EMnet we compute the TPF and then finally the SVM for classification of AUs. We compute a F1-segment score as opposed to F1-frame score as done in [121] because the trajectories on the landmark-points are computed over a 16 frame symmetric window around each frame. For each video in the dataset the first and the last 8 frames were discarded. We found that this window size was a good choice. If a large window was used then the Fisher vectors which are constructed for the segments are not discriminative.

The results of comparison of our framework with the state of the art are presented in Table 5. As we can see the method trained to recognise a single emotion label does not perform competitively as compared to the state-of-the-art. This is because the methods which are designed to do AU recognition are trained via local patches as opposed to the trajectories from all the face landmarks. Since we know the location of the action units we automatically selected the trajectories to train the final SVM. For example if the AU is a lip corner depressor we choose the trajectories from the patch where the action unit is most likely to occur. We know this location because of the landmark points. This result is represented as $Ours - Final + SF$ in table 5. Additionally AUs can co-occur. Therefore, we weight the final recognition scores of the SVM with the co-occurrence probability of the AU. We estimate this probability matrix from the training data. This result is shown as *Proposed+ SF+ CO* in table 5. This way we can show that our method is competitive for dynamic spontaneous AU recognition. If one explicitly estimates the spatial representation temporal modelling and AU correlation then this method can achieve a higher accuracy. This is done with a CNN and LSTM in [122].

5.2.4 SASE-FE

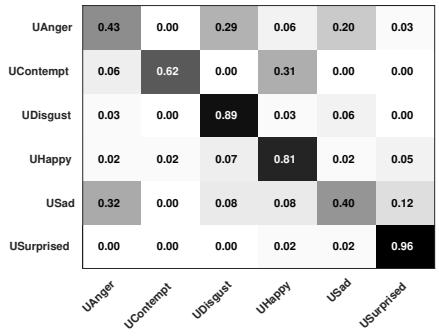
The set of experiments we present in this section has been designed with the purpose of exploring spatial and temporal representation for the proposed problem. We will show how results improve by increased use of domain knowledge for encoding temporal information and by using specially learned representations. Furthermore, we can see more improvement in the recognition results from learning a EMNet from a finetuned VGG-Facenet. For example, in the first conducted experiment we globally extract a hand-crafted descriptor (SIFT) and we disregard any temporal information. On the proposed dataset, this produces results slightly above chance. By computing local descriptors around Improved Dense Trajectories (IDT), a proven technique in the action recognition literature, we obtain a small improvement. While the tracked trajectories follow salient points, there is no guarantee that these points are fiducial points on the face. Because fiducial points are semantically representative on the facial geometry, they are usually best for capturing local variations due to changes of expression. This assumption is confirmed by extracting local descriptors around landmark trajectories produced by the facial geometry detector. In the final setup, the best performance is obtained by extracting the representation from a feature space produced by the EMNet CNN. In Table 8 we compare the performance between the TPF-FGT obtained from the last convolution layer of both the VGG-Face and EMNet. Since the EMNet is trained only for the emotion recognition domain the performance of the EMNet is higher than that of the VGG-Face.

In terms of the use of temporal information several comments can be made. In line with the literature, temporal information is essential in improving recognition of subtle facial expressions. What we are presenting is by no chance an exhaustive study. While a state-of-the-art method in producing compact representations of videos, Fisher Vectors encoding disregards some of the temporal information for compactness. Other, more powerful sequential learning methods, like Recurrent Neural Networks, might be employed with better results.

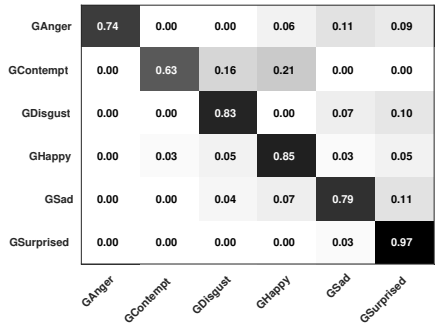
In Fig. 6 we present confusion matrices for a six class classification problem on the proposed dataset. We split the classification problem in two, training on the 6 genuine and the 6 unfelt emotions respectively. On the SASE-FE, several observations can be made. Both in the case of genuine and unfelt FEE classifications, the expressions that are easier to discriminate are Happiness and Surprise. This due to their particularly distinctive morphological patterns. The most difficult expression to distinguish is contempt, which is in alignment with the literature and with the result on the CK+, the benchmark dataset as previously explained. On average, the proposed method gets better results when trying to discriminate between the genuine emotions than when discriminating between the unfelt ones. This is to be expected, taking into account that when faking the expressions, the subjects are trying to hide a different emotional state. This will introduce particular morphological and dynamical changes that makes the problem more difficult. Particularly interesting is the difficulty the classifier has in recognizing unfelt sadness. The high level of confusion with unfelt anger should be noticed along with the fact that this is not the case for genuine emotions.

In Fig. 7 we present the confusion matrix for the problem of classifying between all 12 classes (genuine and unfelt jointly). This can be interpreted together with results in Table 8 where we present classification accuracies for each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2874996, IEEE Transactions on Affective Computing



(a)



(b)

Fig. 6: Confusion matrices for 6 classes classification. (a) 6 class classification on the unfelt subset of SASE-FE. (b) 6 class classification on the genuine subset of SASE-FE. Genuine FEEs are labelled with an initial 'G' and unfelt FEEs with an 'U'.

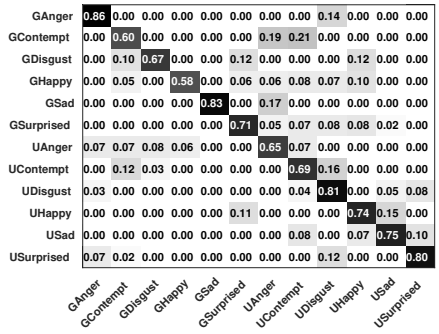


Fig. 7: Confusion matrix for 12 class classification on the SASE-FE dataset. Genuine FEE are labelled with an initial 'G' and unfelt FEE with an 'U' .

pair (genuine/unfelt). When trained with all classes, the best results are obtained for genuine sadness and the worst for genuine contempt and genuine contempt. In Table 6, overall accuracies of especially the unfelt ones remain low, which underlines again the difficulty of the problem and suggests more powerful sequential learning tools should be employed. Interestingly, it is easiest to discriminate between genuine and unfelt expressions of anger which is due to the fact that anger is recognised a lot by the activation of muscles in the eye region. Also the results show that the recognition rate of the unfelt expressed contempt is by chance, i.e. contempt is easier to unfelt, hence more difficult to detect, and this is due to the fact that the main facial features expressing this emotion are mainly around the mouth region which can be quickly and easily moved, whereas muscles around the eyes (which are important in expressing other emotions) are not instantly deformable by signals from brain.

Table 7 shows the comparison of the average recognition rate for a 12-class classification between recently proposed techniques reported in [59] and the proposed method. These results correspond to the winning methods from the ChaLearn international competition we organize at ICCV 2017. We outperform these winning methods. In this table, we can also observe that our proposed method outperforms the LSTM based approaches [124]. This is because in the temporal stage we used a hand tuned approach which requires fewer parameters to be tuned as compared to a LSTM. This advantage would be negated on a very large datasets but nevertheless it demonstrates the efficiency of our method.

TABLE 6: Genuine vs unfelt FEE classification performance on the SASE-FE dataset.

Emotion Pair	Accuracy Genuine (%)	Accuracy Unfelt (%)
Anger	72.5	66.3
Happiness	76.7	65.4
Sadness	71.5	61.3
Disgust	66.4	59.7
Contempt	63.4	58.3
Surprise	71.3	63.4

TABLE 7: The average recognition rate for 12 class classification between several state-of-the-art methods [59] and the proposed method; DA=Data augmentation.

Method	Accuracy
Rank-SVM [125]	66.67
LSTM-PB [124]	66.67
CBP-SVM [126]	65.00
HOG-LSTM [127]	61.70
CNN [128]	51.70
Ours-Final	68.7
Ours-Final + DA	**70.2**

6 Conclusion

Previous research from psychology suggests that discriminating the genuineness of feelings or intentions hidden behind facial expressions is not a well mastered skill. For this reason, we provide for the first time a dataset capturing humans while expressing genuine and unfelt facial expressions of emotion at high resolution and a high frame rate.

In this paper, we also propose a method inspired from action recognition and extend it to perform facial expression of emotion recognition. We combine the feature maps computed from the EMNet CNN with a facial landmark detector to compute spatio-temporal TPF descriptors. We encode these descriptors with Fisher vectors to get a single vector representation per video. The feature vector per video is used to train a linear SVM classifier. We outperform the state of the art performance on the the publicly available CK+

TABLE 8: Performance on the SASE-FE dataset. IDT = Improved dense Trajectories, FGT= Facial Geometry Trajectories, TPF-IDT = Trajectory Pooled Features along IDT, TPF-FGT = Trajectory Pooled Features along FGT, DA = Data Augmentation, [1] Fine-tune, no data augment, [2] Fine-tune, data augment.

	Method	Accuracy(%)
12 classes	SIFT+FV	12.2
	TPF-FGT(SIFT)+TPF-IDT(MBH)+FV	21.3
	VGG-Face[1]	39.5
	VGG-Face[2]	49.8
	TPF-FGT(VGG-Face)+FV	50.2
	TPF-FGT(VGG-Face)+FV+Aligned Faces	54.3
	TPF-FGT(VGG-Face)+FV+Aligned Faces+DA	60.3
	TPF-FGT(EMNet)+FV	65.7
	TPF-FGT(EMNet)+FV Aligned Faces	68.7
	TPF-FGT(EMNet)+FV Aligned Faces+DA	70.2
6 classes (genuine)	VGG[1]	65.2
	VGG [2]	71.7
	TPF-FGT(VGG-Face)+FV	73.7
	TPF-FGT(VGG-Face)+FV Aligned Faces	74.2
	TPF-FGT(VGG-Face) +FV Aligned Faces+ DA	76.5
	TPF-FGT(EMNet) +FV	77.2
	TPF-FGT(EMNet) +FV Aligned Faces	78.7
	TPF-FGT(EMNet) +FV Aligned Faces+ DA	80.3
6 classes (unfelt)	VGG[1]	42.7
	VGG[2]	59.2
	TPF-FGT(VGG-Face) + FV	62.3
	TPF-FGT(VGG-Face) + FV + Aligned Faces	64.2
	TPF-FGT(VGG-Face) + FV + Aligned Faces + DA	67.5
	TPF-FGT(EMNet) + FV	70.3
	TPF-FGT(EMNet) + FV + Aligned Faces	72.2
	TPF-FGT(EMNet)+FV Aligned Faces+DA (Ours-Final)	73.6

and Oulu-CASIA both containing posed FEEs, and show competitive results on the BP4D dataset for facial action unit recognition. Furthermore, we provide several baselines on our SASE-FE dataset. We also improve the results of the winning solutions of the recent ChaLearn competition about our dataset. We show that even though we obtain good results on the 6 class genuine and unfelt problem, the 12 class and the binary emotion pair classification problem still remains a challenge. This is because the distinguishing factors between the unfelt and genuine expressions occur in a very short part of the whole emotion and are a challenge to model.

This preliminary analysis opens several future lines of research. Our experiments showed two most important problems of current state of the art methods. Firstly, current state of the art CNNs, such as VGG-Face, do not work at the required spatial resolution to detect minute changes in facial muscle movements, which are required to differentiate and distinguish between unfelt FEEs. Secondly, alternative temporal analysis strategies could be considered to analyse SASE-FE at high fps, which may include variants of Recurrent Neural Nets or 3D-CNNs approaches.

ACKNOWLEDGEMENTS

This work is supported Estonian Research Council Grant (PUT638), the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund, the Spanish Project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement N^o 665919. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] M. Hartwig and C. F. Bond Jr, "Why do lie-catchers fail? a lens model meta-analysis of human lie judgments," 2011.

[2] L. Ten Brinke, P. Khambatta, and D. R. Carney, "Physically scarce (vs. enriched) environments decrease the ability to tell lies successfully." *Journal of experimental psychology: general*, vol. 144, no. 5, p. 982, 2015.

[3] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[4] S. Porter and L. Ten Brinke, "Reading between the lies identifying concealed and falsified emotions in universal facial expressions," *Psychological Science*, vol. 19, no. 5, pp. 508–514, 2008.

[5] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, "Intelligent expressions of emotions," in *Int. Conf. on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 707–714.

[6] S. Hyniewska and W. Sato, "Facial feedback affects valence judgments of dynamic and static emotional expressions," *Frontiers in psychology*, vol. 6, p. 291, 2015.

[7] A. Baker, P. J. Black, and S. Porter, "10. the truth is written all over your face! involuntary aspects of emotional facial expressions," *The Expression of Emotion: Philosophical, Psychological and Legal Perspectives*, p. 219, 2016.

[8] R. Rogers, "Models of feigned mental illness." *Professional Psychology: Research and Practice*, vol. 21, no. 3, p. 182, 1990.

[9] S. Lautenbacher and M. Kunz, "Facial pain expression in dementia: a review of the experimental and clinical evidence," *Current Alzheimer Research*, vol. 14, no. 5, pp. 501–505, 2017.

[10] M. Davis, K. A. Markus, and S. B. Walters, "Judging the credibility of criminal suspect statements: does mode of presentation matter?" *Journal of Nonverbal Behavior*, vol. 30, no. 4, pp. 181–198, 2006.

[11] M. K. MacLin, C. Downs, O. H. MacLin, and H. M. Caspers, "The effect of defendant facial expression on mock juror decision-making: The power of remorse." *North American Journal of Psychology*, vol. 11, no. 2, 2009.

[12] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in *ICRA*, vol. 4. IEEE, 2002, pp. 4138–4142.

[13] K. M. Lee, W. Peng, S.-A. Jin, and C. Yan, "Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction," *Journal of communication*, vol. 56, no. 4, pp. 754–772, 2006.

[14] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 1, pp. 96–105, 2006.

[15] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 15–21.

[16] A. O. Aremu and G. A. Lawal, "A path model investigating the influence of some personal–psychological factors on the career aspirations of police trainees: a perspective from oyo state, nigeria," *Police Practice and Research: An International Journal*, vol. 10, no. 3, pp. 239–254, 2009.

[17] A. Vrij and S. Mann, "Who killed my relative? police officers' ability to detect real-life high-stake lies," *Psychology, crime and law*, vol. 7, no. 1-4, pp. 119–132, 2001.

[18] M. O'Sullivan, M. G. Frank, C. M. Hurley, and J. Tiwana, "Police lie detection accuracy: The effect of lie scenario," *Law and Human Behavior*, vol. 33, no. 6, pp. 530–538, 2009.

[19] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Englewood Cliffs, New Jersey: Prentice Hall, 1975.

[20] M. G. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies." *Journal of personality and social psychology*, vol. 72, no. 6, p. 1429, 1997.

[21] N. Abe, "The neurobiology of deception: evidence from neuroimaging and loss-of-function studies," *Current opinion in neurology*, vol. 22, no. 6, pp. 594–600, 2009.

[22] G. Duchenne de Bologne, "The mechanism of human facial expression (ra cuthbertson, trans.)," *Paris: Jules Renard*, 1862.

[23] S. A. Spence, T. F. Farrow, A. E. Herford, I. D. Wilkinson, Y. Zheng, and P. W. Woodruff, "Behavioural and functional anatomical correlates of deception in humans," *Neuroreport*, vol. 12, no. 13, pp. 2849–2853, 2001.

[24] M. J. Bernstein, S. G. Young, C. M. Brown, D. F. Sacco, and H. M. Claypool, "Adaptive responses to social exclusion social rejection improves detection of real and fake smiles," *Psychological Science*, vol. 19, no. 10, pp. 981–983, 2008.

[25] P. Ekman, R. J. Davidson, and W. V. Friesen, "The duchenne smile: Emotional expression and brain physiology: Ii." *Journal of personality and social psychology*, vol. 58, no. 2, p. 342, 1990.

[26] W. M. Brown and C. Moore, "Smile asymmetries and reputation as reliable indicators of likelihood to cooperate: An evolutionary analysis," in *11; 3*. Nova Science Publishers, 2002.

[27] M. G. Frank and P. Ekman, "Not all smiles are created equal: The differences between enjoyment and nonenjoyment smiles," *Humor-International Journal of Humor Research*, vol. 6, no. 1, pp. 9–26, 1993.

[28] P. Ekman, W. V. Friesen, and M. O'sullivan, "Smiles when lying." *Journal of personality and social psychology*, vol. 54, no. 3, p. 414, 1988.

[29] H. Demirel and G. Anbarjafari, "Data fusion boosted face recognition based on probability distribution functions in different colour channels," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 25, 2009.

[30] S. D. Gunnery, J. A. Hall, and M. A. Ruben, "The deliberate duchenne smile: Individual differences in expressive control," *Journal of Nonverbal Behavior*, vol. 37, no. 1, pp. 29–41, 2013.

[31] E. G. Krumhuber and A. S. Manstead, "Can duchenne smiles be feigned? new evidence on felt and false smiles." *Emotion*, vol. 9, no. 6, p. 807, 2009.

[32] M. Mehu, M. Mortillaro, T. Bänziger, and K. R. Scherer, "Reliable facial muscle activation enhances recognizability and credibility of emotional expression." *Emotion*, vol. 12, no. 4, p. 701, 2012.

[33] M. G. Frank, *An empirical reflection on the smile*. Lewiston, NY: E. Mellen Press, 2002, ch. Smiles, lies, and emotion., pp. 15–44.

[34] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[35] P. Gosselin, M. Perron, and M. Beaupré, "The voluntary control of facial action units in adults." *Emotion*, vol. 10, no. 2, p. 266, 2010.

[36] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *FG*. IEEE, 2000, pp. 46–53.

[37] Z. L. Boraston, B. Corden, L. K. Miles, D. H. Skuse, and S.-J. Blakemore, "Brief report: Perception of genuine and posed smiles by individuals with autism," *Journal of Autism and Developmental Disorders*, vol. 38, no. 3, pp. 574–580, 2008.

[38] V. Manera, M. Del Giudice, E. Grandi, and L. Colle, "Individual differences in the recognition of enjoyment smiles: No role for perceptual–attentional factors and autistic-like traits," *Frontiers in psychology*, vol. 2, p. 143, 2011.

[39] A. Uusberg, H. Uibo, K. Kreegipuu, M. Tamm, A. Raidvee, and J. Allik, "Unintentionality of affective attention across visual processing stages," *Frontiers in psychology*, vol. 4, 2013.

[40] M. Perron and A. Roy-Charland, "Analysis of eye movements in the judgment of enjoyment and non-enjoyment smiles," *Frontiers in psychology*, vol. 4, p. 659, 2013.

[41] J. Chartrand and P. Gosselin, "Judgement of authenticity of smiles and detection of facial indexes," *Canadian journal of experimental psychology= Revue canadienne de psychologie experimentale*, vol. 59, no. 3, pp. 179–189, 2005.

[42] A. Vrij, P. A. Granhag, and S. Porter, "Pitfalls and opportunities in nonverbal and verbal lie detection," *Psychological Science in the Public Interest*, vol. 11, no. 3, pp. 89–121, 2010.

[43] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas (me)ˆ 2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, p. 1, 2017.

[44] W. Sato, T. Kochiyama, S. Yoshikawa, E. Naito, and M. Matsumura, "Enhanced neural activity in response to dynamic facial expressions of emotion: an fmri study," *Cognitive Brain Research*, vol. 20, no. 1, pp. 81–91, 2004.

[45] E. G. Krumhuber, A. Kappas, and A. S. Manstead, "Effects of dynamic aspects of facial expressions: a review," *Emotion Review*, vol. 5, no. 1, pp. 41–46, 2013.

[46] R. E. Jack and P. G. Schyns, "The human face as a dynamic tool for social communication," *Current Biology*, vol. 25, no. 14, pp. R621–R634, 2015.

[47] M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements," *Scientific reports*, vol. 6, 2016.

[48] E. D. Ross, L. Shayya, A. Champlain, M. Monnot, and C. I. Prodan, "Decoding facial blends of emotion: Visual field, attentional and hemispheric biases," *Brain and cognition*, vol. 83, no. 3, pp. 252–261, 2013.

[49] S. Porter, L. Ten Brinke, and B. Wallace, "Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity," *Journal of Nonverbal Behavior*, vol. 36, no. 1, pp. 23–37, 2012.

[50] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *FG*. IEEE, 2017, pp. 809–813.

[51] C. Loob, P. Rasti, I. Lüsi, J. C. J. Junior, X. Baró, S. Escalera, T. Sapinski, D. Kaminska, and G. Anbarjafari, "Dominant and complementary multi-emotional facial expression recognition using c-support vector classification," in *FG*, 2017, pp. 833–838.

[52] P. Ekman, W. V. Friesen, and J. C. Hager, "Facs investigator's guide," *A human face*, p. 96, 2002.

[53] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[54] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *FG-Workshops*. IEEE, 2013, pp. 1–6.

[55] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[56] W. Gaebel and W. Wölwer, "Facial expression and emotional face recognition in schizophrenia and depression," *E. archives of psychiatry and clinical neuroscience*, vol. 242, no. 1, pp. 46–52, 1992.

[57] J. de Fockert and C. Wolfenstein, "Rapid extraction of mean identity from sets of faces," *The Quarterly Journal of Experimental Psychology*, vol. 62, no. 9, pp. 1716–1722, 2009.

[58] A. J. Calder, A. W. Young, J. Keane, and M. Dean, "Configural information in facial expression perception." *Journal of Experimental Psychology: Human perception & performance*, vol. 26, no. 2, p. 527, 2000.

[59] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, and G. Anbarjafari, "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *ICCVW*, vol. 4, no. 6, 2017.

[60] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. SMC-B*, vol. 36, no. 2, pp. 433–449, 2006.

[61] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *IVC*, no. 12, pp. 1856–1863, 2007.

[62] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *TIP*, vol. 16, pp. 172–187, 2007.

[63] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *IVC*, vol. 31, no. 2, pp. 153–163, 2013.

[64] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *TPAMI*, vol. 32, no. 11, pp. 1940–1954, 2010.

[65] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *FG*, 2011, pp. 298–305.

[66] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities," *CYB*, 2014.

[67] G. Anbarjafari, "Face recognition using color local binary pattern from mutually independent color channels," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 6, 2013.

[68] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *FG*, 2011, pp. 878–883.

[69] 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 *ICMI*, 2014, pp. 481–486.

[70] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, and T. S. Huang, "Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *CVPR*, 2003, pp. I–595–I–601.

[71] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *TPAMI*, vol. 23, pp. 97–115, 2001.

[72] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *FG*, 2015.

[73] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[74] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *ECCV*, 2012, vol. 7577, pp. 808–822.

[75] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baró, H. Demirel *et al.*, "Dominant and complementary emotion recognition from still images of faces," *IEEE Access*, vol. 6, pp. 26 391–26 403, 2018.

[76] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *ICCE*, 2014, pp. 564–567.

[77] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *FG*. IEEE, 2013, pp. 1–6.

[78] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *CVPR*, 2014, pp. 1805–1812.

[79] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden markov model for facial expression recognition," in *FG*, 2015.

[80] A. Geetha, V. Ramalingam, S. Palanivel, and B. Palaniappan, "Facial expression recognition–a real time approach," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 303–308, 2009.

[81] 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 *ICMI*, 2014, pp. 494–501.

[82] D. T. Robinson and L. Smith-Lovin, "Emotion display as a strategy for identity negotiation," *Motivation and Emotion*, vol. 23, no. 2, pp. 73–104, 1999.

[83] B. Bastian, P. Kuppens, K. De Roover, and E. Diener, "Is valuing positive emotion associated with life satisfaction?" *Emotion*, vol. 14, no. 4, p. 639, 2014.

[84] J. Zeman and J. Garber, "Display rules for anger, sadness, and pain: It depends on who is watching," *Child development*, vol. 67, no. 3, pp. 957–973, 1996.

[85] J. Gnepp and D. L. Hess, "Children's understanding of verbal and facial display rules." *Developmental psychology*, vol. 22, no. 1, p. 103, 1986.

[86] P. M. Cole and A. E. Jacobs, "From children's expressive control to emotion regulation: Looking back, looking ahead," *European Journal of Developmental Psychology*, pp. 1–20, 2018.

[87] P. W. Garner, "The relations of emotional role taking, affective/moral attributions, and emotional display rule knowledge to low-income school-age children's social competence," *Journal of Applied Developmental Psychology*, vol. 17, no. 1, pp. 19–36, 1996.

[88] L. ten Brinke and S. Porter, "Discovering deceit: Applying laboratory and field research in the search for truthful and deceptive behavior," in *Applied issues in investigative interviewing, eyewitness memory, and credibility assessment*. Springer, 2013, pp. 221–237.

[89] S. Porter and L. ten Brinke, "The truth about lies: What works in detecting high-stakes deception?" *Legal and criminological Psychology*, vol. 15, no. 1, pp. 57–75, 2010.

[90] E. Diener, S. Oishi, and R. E. Lucas, "Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life," *Annual review of psychology*, vol. 54, no. 1, pp. 403–425, 2003.

[91] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Trans. SMC-B*, vol. 41, no. 3, pp. 664–674, 2011.

[92] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju, "Real-time automatic deceit detection from involuntary facial expressions," in *CVPR*. IEEE, 2007, pp. 1–6.

[93] R. E. Ready, G. D. Santorelli, and M. A. Mather, "Judgment and classification of emotion terms by older and younger adults," *Aging & mental health*, pp. 1–9, 2016.

[94] D. M. Isaacowitz, "Mood regulation in real time: Age differences in the role of looking," *Current directions in psychological science*, vol. 21, no. 4, pp. 237–242, 2012.

[95] R. Plutchik, "Emotions, evolution, and adaptive processes," in *Feelings and emotions: the Loyola Symposium*. Academic Press, 1970, pp. 3–24.

[96] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE TAC*, 2017.

[97] J. T. Larsen and A. P. McGraw, "Further evidence for mixed emotions." *Journal of personality and social psychology*, vol. 100, no. 6, p. 1095, 2011.

[98] N. R. Whitesell and S. Harter, "Children's reports of conflict between simultaneous opposite-valence emotions," *Child Development*, pp. 673–682, 1989.

[99] Y. Y. Mathieu, "Annotation of emotions and feelings in texts," in *Int. Conf. on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 350–357.

[100] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[101] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015, pp. 4305–4314.

[102] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *FG*. IEEE, 2017, pp. 118–126.

[103] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[104] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *CVPR*, 2015, pp. 4295–4304.

[105] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.

[106] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[107] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *NIPS*. Cambridge, MA, USA: MIT Press, 1999, pp. 487–493. [Online]. Available: http://dl.acm.org/citation.cfm?id=340534.340715

[108] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, Dec. 2013.

[109] 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 *CVPRW*. IEEE, 2010, pp. 94–101.

[110] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *IVC*, vol. 29, no. 9, pp. 607–619, 2011.

[111] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *IVC*, vol. 32, no. 10, pp. 692–706, 2014.

[112] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *ICCV*, Dec 2013, pp. 1817–1824.

[113] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[114] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, 2014, pp. 1749–1756.

[115] K. Sikka, G. Sharma, and M. Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *CVPR*, June 2016, pp. 5580–5589.

[116] S. Afshar and A. A. Salah, "Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding," in *CVPRW*, June 2016, pp. 1517–1525.

[117] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *ICCVW*, Dec 2015, pp. 19–27.

[118] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, 2015, pp. 2983–2991.

[119] X. Zhao, X. Liang, L. Liu, T. Li, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," *arXiv preprint arXiv:1607.06997*, 2016.

[120] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[121] K. Zhao, W. S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *CVPR*, 2016, pp. 3391–3399.

[122] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *FG 2017*. IEEE, 2017, pp. 25–32.

[123] K. Zhao, W. S. Chu, F. D. la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit and holistic expression recognition," *IEEE TIP*, vol. 25, no. 8, pp. 3931–3946, 2016.

[124] J. Tani, M. Ito, and Y. Sugita, "Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using rnnpb," *Neural Networks*, vol. 17, no. 8, pp. 1273–1289, 2004.

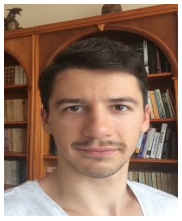
[125] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[126] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *CVPR*, 2016, pp. 317–326.

[127] W. Pei, T. Baltrusaitis, D. M. Tax, and L.-P. Morency, "Temporal attention-gated model for robust sequence classification," in *CVPR*. IEEE, 2017, pp. 820–829.

[128] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *ECCV*. Springer, 2016, pp. 414–428.

Kaustubh Kulkarni obtained in Bachelors in engineering from Mumbai university. He completed his MSc. from Auburn University, USA. Following which he worked at Siemens research labs in India and USA. He is in the process of getting his PhD from INRIA, Grenoble, France. Currently, he is working at the Computer Vision Center at Universitat Autònoma de Barcelona. He has experience working in medical image analysis, action recognition, speech recognition and emotion recognition.



Ciprian Adrian Corneanu got his MSc in Computer Vision from Universitat Autónoma de Barcelona in 2015. Currently he is a PhD student at the Universitat de Barcelona and a fellow of the Computer Vision Center from Universitat Autónoma de Barcelona. His main research interests include face and behavior analysis, affective computing, social signal processing and human computer interaction.

Ikechukwu Ofodile obtained his BSc and MSc from Eastern Mediterranean University and University of Tartu, respectively. He is currently a PhD Student and a member of the iCV Lab at the University of Tartu. His research interests include machine learning, pattern recognition and HCI as well as control engineering and attitude control system design for nanosatellites and microsatellites.

Xavier Baró received his B.S. degree in Computer Science at the UAB in 2003. In 2005 he obtained his M.S. degree in Computer Science at UAB, and in 2009 the PhD degree in Computer Engineering. At the present he is associate professor and researcher at the Computer Science, Multimedia and Telecommunications department at Universitat Oberta de Catalunya.

Sergio Escalera is an associate professor at the Department of Mathematics and Informatics, Universitat de Barcelona. He is an adjunct professor at Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University. He obtained the PhD degree on Multi-class visual categorization systems at the Computer Vision Center, UAB. He obtained the 2008 best Thesis award on Computer Science at Universitat Autònoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB and CVC. He is an expert in human behavior analysis in temporal series, statistical pattern recognition, visual object recognition, and HCI systems, with special interest in human pose recovery and behavior analysis from multi-modal data. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. He is Chair of IAPR TC-12: Multimedia and visual information systems.

Sylwia Hyniewska received a double PhD degree from the Telecom ParisTech Institute of Science and Technology and the University of Geneva. She finished her doctoral school at the "Swiss National Center for Affective Sciences". Afterward, she worked as an independent research Fellow of the Japan Society for the Promotion of Science at Kyoto University, where she collaborated with world-renowned specialists in social and emotion perception. Since 2014 she has worked at the University of Bath on topics related to emotion perception, virtual reality and pervasive devices and is a member of the Centre for Applied Autism Research. In 2017, she joined the Institute of Physiology and Pathology of Hearing, Poland. At the institute's Bioimaging Research Center, she studies neurofeedback applied to brain fingerprinting in attentional and affective tasks.

Jüri Allik was Candidate of Science (PhD), University of Moscow and obtained PhD in psychology from the University of Tampere, Finland. He has been Chairman of Estonian Science Foundation, Professor of Psychophysics and Professor of Experimental Psychology at the University of Tartu. He was also Dean of Faculty of Social Sciences, President and Vice-President of the Estonian Psychological Association. He served as a Foreign Member of the Finnish Academy of Science and Letters. He has has received many awards including Estonian National Science Award in Social Sciences category. He was a member of the Estonian Academy of Sciences. His research interests are psychology, perception, personality and neuroscience and his research works have received over 14,000 citations.

Gholamreza Anbarjafari is heading the iCV Lab in the Institute of Technology at the University of Tartu. He was also Deputy Scientific Coordinator of the European Network on Integrating Vision and Language COST Action (IC1307). He is an IEEE Senior member and the Chair of SP/CS/SSC Joint Societies Chapter of IEEE Estonian section. He has got Estonian Research Council Grant (PUT638) and the TÜBİTAK (116E097) in 2015 and 2016, respectively. He has been involved in many national and international industrial projects mainly related to affective computing. He is expert in computer vision, human-robot interaction, and human behaviour analysis. He has been in the TCP of SIU, ICOSST, ICGIP, SampTA and FG. He has been organizing challenges and workshops in FG17, CVPR17, and ICCV17. He is Associate Editor of SIVP and have organized several SI on human behaviour analysis in JIVP and MVAP.