Automatic Recognition of Deceptive Facial Expressions of Emotion

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

Abstract—Humans modify facial expressions in order to mislead observers regarding their true emotional states. Being able to recognize the authenticity of emotional displays is notoriously difficult for human observers. 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 genuine and deceptive facial expressions of emotions for automatic recognition[1] . We show that overall the problem of recognizing deceptive facial expressions 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 feature space. Interesting additional results show that on average it is easier to distinguish among genuine expressions than deceptive ones and that certain emotion pairs are more difficult to distinguish than others.

Index Terms—Affective Computing, Facial Expression Recognition, Expressed Emotion, Fake Emotion Recognition, Human Behavior 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 and body language of alleged offenders.

However in real life, humans are very skilled in concealing their true affective states from others. Untrained observers tend to perform barely above chance level when asked to detect such behaviour [1], [2]. This is particularly the case 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].

Many potential applications would benefit from the ability of automatically discriminating between deceptive and genuine facial emotional displays. Improved human-computer interaction, improved human-robot interaction for assistive robotics [6]–[8], treatment of chronic disorders [9] and assisting investigation conducted by police forces [10]–[12] would be just a few.

An emotional display is considered deceptive when not accompanying a corresponding emotional state. There are three major ways in which emotional facial expressions are intentionally manipulated [13]: an expression is *simulated* when it is not accompanied by any genuine emotion, *masked* when the expression corresponding to the felt emotion is

Fig. 1: People may have difficulties in displaying expressions that look genuine when lying. In the case of smiling, differences can be observed in the contraction of the *orbicularis oculi* muscle around the eyes. *Left*: no orbicularis oculi contraction, a marker of fake expression. *Right*: strong orbicularis oculi contraction, with very visible "crows feet" around the corners of the eyes, a marker of genuine expression.

- *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, 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 Department of Psychology, University of Bath, Bath, UK. 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.

[1] *A sample video of the dataset is uploaded as supplementary material. Complete dataset will be publicly available after publication of the paper.*

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.

It has been argued that deceivers would be betrayed by the leakage of their genuine emotional states through their nonverbal behaviour [4], [14], [15]. 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 deceptive 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 genuine and deceptive displays of emotions [16]. 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) [17]–[24] (see Fig. 1). In contrast, a deceptive smile (aka a non-Duchenne smile) can be used to conceal the experience of negative emotions (a masking smile) [14], [21], [24]–[27].

Although it has been argued that the orbicularis oculi activation is absent from deceptive facial expressions of enjoyment, empirical evidence is not conclusive. For example, one study showed that when looking at 105 posed smiles, only 67% were accompanied by the orbicularis oculi activation [28]. 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 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.

Apart from short involuntary expressions, blinking is reported to be another important cue to deception. However, the exact conditions affecting blinking are still debated and the link to emotions is not completely clear [4], [29]. Therefore even the tracking of the most reliable signals according to the literature seems insufficient per se, and more complex behaviour analysis is required, without any reliance on specific facial action units or any other well-defined cues.

Although a variety of studies have focused on the evaluation of deceptive facial expressions of emotion based on still, i.e. static, images, not much attention has been paid to dynamics as evaluated in a sequence of frames [30]–[35]. In a naturalistic setting, facial expressions of emotions 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 [36]–[38].

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 deceptive expression, which made the overall pattern difficult to analyze [39]. Leakages of the genuine expression of emotion have been observed more frequently in the upper part of the face, while cues related to the deceptive expression of emotion in the lower half [40]–[43]. It has been also been suggested that a deceptive expression of emotion is accompanied by an increase in blink rate relative to genuine expressions [4]. However, blink rate seems to be a very inconsistent measure, easily influenced by any change in the context or task at hand [29], [41].

In this work, we propose a new data corpus containing genuine and deceptive universal facial expressions of emotion. While numerous studies involving videos of genuine and deceitful behaviours focused on cues of deceit in directed interviews, such as the work in [2], the studies that analysed cues of deceit while controlling for the emotional state of subjects are rare [41].

When designing experiments that require facial emotion displays as independent variables, psychologists often opt to use posed facial expressions of subjects being instructed to act out a particular emotion. 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 the facial expressions, participants are usually asked to practice the display of such emotions, and in order to achieve a display closer to the genuine emotional expression, external factors can be used to facilitate the process, such as pictures [41] or videos inducing emotions in line with the ones to be expressed [44], facial expressions of emotions [45], [46] or mental imagery and related theatre techniques [47]. Such paradigms have been frequently used for recording and creating emotional expression databases [46]–[52].

The rest of the paper is organised as follows: in Section 2 we describe related work in facial expression recognition, in Section 3 we introduce the new SASE-FE dataset, in Section 4 we detail the proposed methodology, in Section 5 we present and discuss experimental results and in the final section we conclude and propose future lines of research.

2 Related Work

In this section we first review related work on the simpler but related problem of facial expression recognition as well as deceptive facial expression recognition.

2.1 Recognizing Facial Expressions of Emotion

Automatic facial expression recognition (AFER) has been an active field of research for a long time. The first important method to be proposed dates back to the end of the 1970s [53]. Limitations in computing power and lack of labelled data have greatly slowed progress in the following decades. It is only with the pioneering work of first Silvan Tomkins [54] and later Paul Ekman [55] that research on facial expression recognition became prominent. TTechnological improvements and a set of new datasets led to the revival of field at the beginning of the 2000s [56]. The majority of the early work focused on geometrical representations of the face and hand-crafted features used to train classifiers able to discriminate between a limited set of greatly exaggerated expressions of emotions. In the early years, 2D face analysis from RGB images in a static context was dominant, but since then a great number of additional modalities has

been proposed like 3D or thermal. Dynamic analysis in sequences has become standard and an increasing number of datasets with richer sets of labelled expressions made publicly available. The interested reader can refer to many of the excellent surveys of the field published in recent years [57], [58].

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 can be estimated. Based on it, alignment methods can be used to reduce variance of local and global descriptors to rigid and non-rigid variations. This greatly improves robustness to in-plane rotations or head pose. Finally, representations of the face are computed either globally, where global features extract information from the whole facial region, locally, and models are trained for classification or regression problems. In this section we focus on presenting state-of-the-art methods for building representations and learning models for facial expression analysis.

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 [59] or deformation parameters of a mesh model [60], [61]. 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 [62]. Other descriptors such as Motion History Images (MHI) and Free-Form Deformations (FFDs) are also used [63]. 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 [64], Local Binary Pattern (LBP) [65], [66], Pyramids of Histograms of Gradients (PHOG) [67] and Multi-Scale Dense SIFT (MSDF) [68]. 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) [60], [69], Neural Networks (NN) [70], Support Vector Machines (SVM) [61] and Random Forests (RF) [71]. More recently, deep learning architectures have been used to jointly perform feature extraction and recognition. These approaches often use pre-training [72], an unsupervised layer-wise training step that allows for much larger, unlabelled datasets to be used. CNNs are by far the dominant approach [73]–[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) [63], [79]–[81], are common technique. Other techniques use Recurrent Neural Network (RNN) architectures, such as Long Short Term Memory (LSTM) networks [62]. Other approaches classify each frame independently (e.g. with SVM classifiers [82]), 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 [60]. For example, statistical information can be encoded at the frame-level into Riemannian manifolds [83].

2.2 Recognizing Deceptive Facial Expressions of Emotion

Emotion perception by humans or computers stands for the interpretation of particular representations of personal feelings expressed by individuals, which may take different forms based on the circumstances governing their behaviour at the time-stamp at which they are evaluated [84], [85].

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 [39], [86] 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 well-controlled genuine/deceptive emotional displays presented in high resolution at an increased frame rate. To answer those needs, the new database of genuine and deceptive universal facial expressions of emotion called SASE-FE database was created.

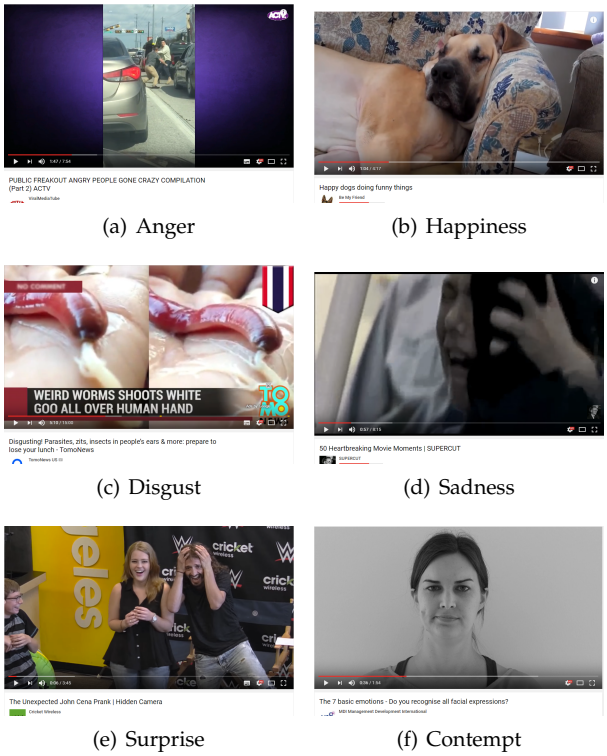
The SASE-FE database consists of 643 different videos which have been recorded with a high resolution GoPro-Hero camera. As indicated in Table 1, 54 participants of ages 19-36 were recorded. The reasoning behind the choice of such a young sample is that older adults have different, more positive responses than young adults about feelings and they are quicker to regulate negative emotional states than younger adults [87]–[89].

For each recording, participants were asked to act two facial expressions of emotions in a sequence, a genuine and a deceptive one (in our case a masked expression). The expressions are the six universal expressions, namely, Happiness, Sadness, Anger, Disgust, Contempt and Surprise. To increase the chances of distinguishing between the two facial expressions 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 [90]–[93]. Thus a 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 [94], [95]. For eliciting emotion, subjects were shown videos in line with the target emotion so as to increase the realism of their emotional portrayal. Emotion elicitation through videos is a well established process in emotion science research [96]. Videos were selected from YouTube. Fig. 2 shows a frame from one of the videos that have been used for inducing specific emotions in the participants.

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

None of the participants were aware of the fact that they would be asked to act 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, has been recorded by the same device with the same configuration. As a result, for each participant we have collected 12 different videos of which 6 are genuine facial expressions of emotion and other 6 are deceptive facial expressions of emotion. The length of captured facial expressions is not fixed. The process has been closely supervised by experimental psychologists so that the setup would result in realistic recordings of deceptive facial expressions of emotion. The summary of the SASE-FE dataset is provided in Table 1 and samples of frames are shown in Fig. 3.

It is important to note that while preparing the SASE-FE database, introduced and used in this work, external factors



(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.

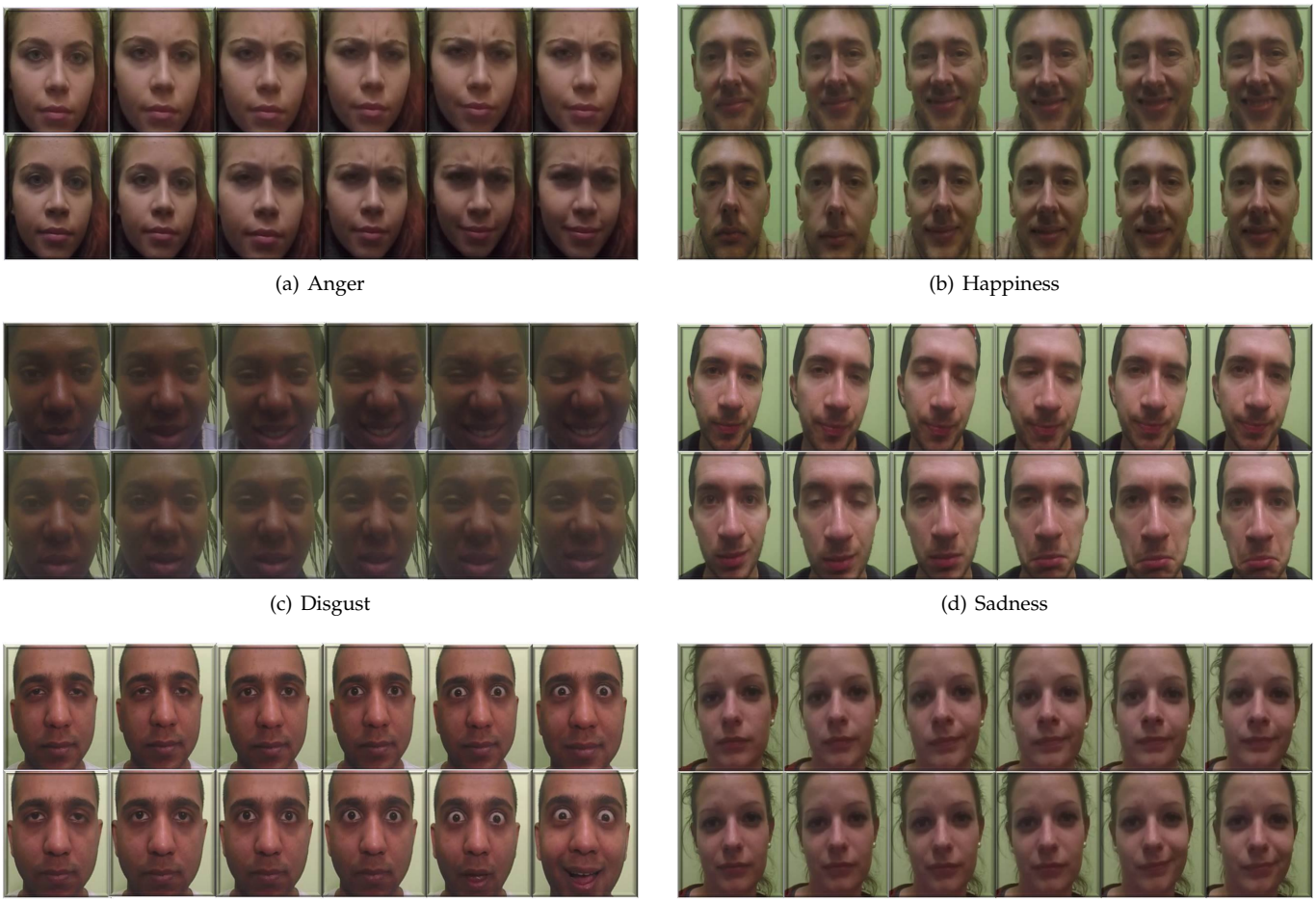
TABLE 1: Summary of SASE-FE database.

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/faked)	120,216/118,712

such as personality or mood of the participants have been 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.

4 The Proposed Method

In this section we are presenting and detailing the theoretical background of the methodology used for recognising deceptive facial expressions of emotion 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 facial expressions. Following this assumption, we consider learning an optimal spatio-temporal representation to be central for solving this problem. We first train a Convolutional Neural Network to learn a static representation from still images and then pull features from this representation space along facial landmark trajectories. Inspired by previous work in action recognition [97], a well studied sequence modelling problem, we build final features from



(a) Anger (b) Happiness

(c) Disgust (d) Sadness

(e) Surprise (f) Contempt

Fig. 3: Expressed emotion pair sequence showing acted (top) and fake (below) for each emotion in theSASE-FE dataset

sequences of varying length using a Fisher Vector encoding which we use to train an SVM for the final decision.
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 Convolutional Neural Networks. 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 [98] which proposes a regularisation function which helps use the face information to train the expression classification net.
We follow this section by first discussing the technique we have used to train a Convolutional Neural Network 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 Convolutional Neural Network for learning static spatial representation would follow the following steps: first, we fine tune the VGG-Face network for the facial expression recognition task [99]. We then use this fine tuned network to guide the learning of a so called emotion network (EMNet) [98]. Following [98] the EMNet can be 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. While θ_1 and θ_2 are the corresponding parameters to be estimated and I is the input image and O is the output before the softmax.
We follow the two step training proposed in [98]. 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 ouput of the VGG-Face to guide the training of the EMNet. In the first step, we will estimate the parameters of the convolution layers of the EMNet. In this step, the output of the VGG-Face will act as a regularisation for the emotion net. This step can be achieved by maximising the following loss function:

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

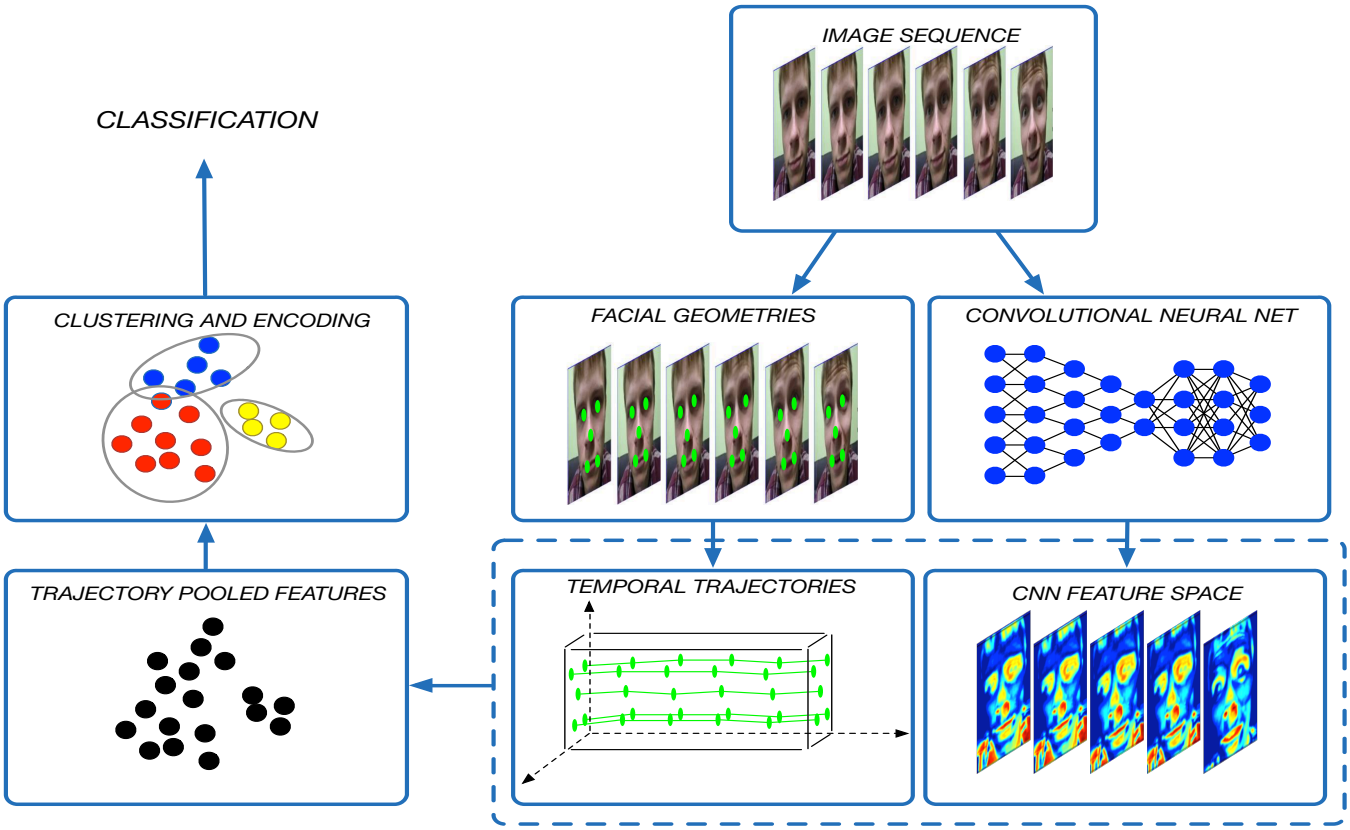


Fig. 4: Overview of the proposed method.

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

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

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

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. The detected fiducial points can be 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) [100] 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

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:

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

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:

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

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

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

where, $\gamma_k(m)$ is the posterior probability or the responsibility of assigning the observation $\vec{x}_m$ to component k. It is given as:

$$\gamma_k(m) = \frac{w_k \mathcal{N}(\vec{x}_m; \vec{\mu}_k, (\vec{\sigma}_k)^2)}{\sum_{i=1}^{K} w_i \mathcal{N}(\vec{x}_m; \vec{\mu}_i, (\vec{\sigma}_i)^2)} \tag{8}$$

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.

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+) [101] dataset and the Oulu-CASIA dataset [102].

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 [98], 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.

In the following sections we will 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 [99] for recognising facial expressions. Third, we use this network for guiding the training of a so called EMNet following work proposed in [98] (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 the 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. [103]. 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 esthetically pleasant way by using color information of the neighbouring visible regions and symmetry. Finally in all sythesized frontal faces we estimated the facial geometry, using a classical, robust facial alignment method [104] trained to find 68 points on the image (an example of the frontalization process is showed in Fig. 5).

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 [98]. We define a train, validation and a test set for the SASE-FE dataset. Since the training data is limited we augment the training set of the SASE-FE dataset with additional training data from the Oulu-CASIA [102] and CK+datasets. These experiments are denoted as *Data Augmentation*. For each fold 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 fake class and the 6 true 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 same as the one proposed in [98]. 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 fake and true emotions. This is because the training

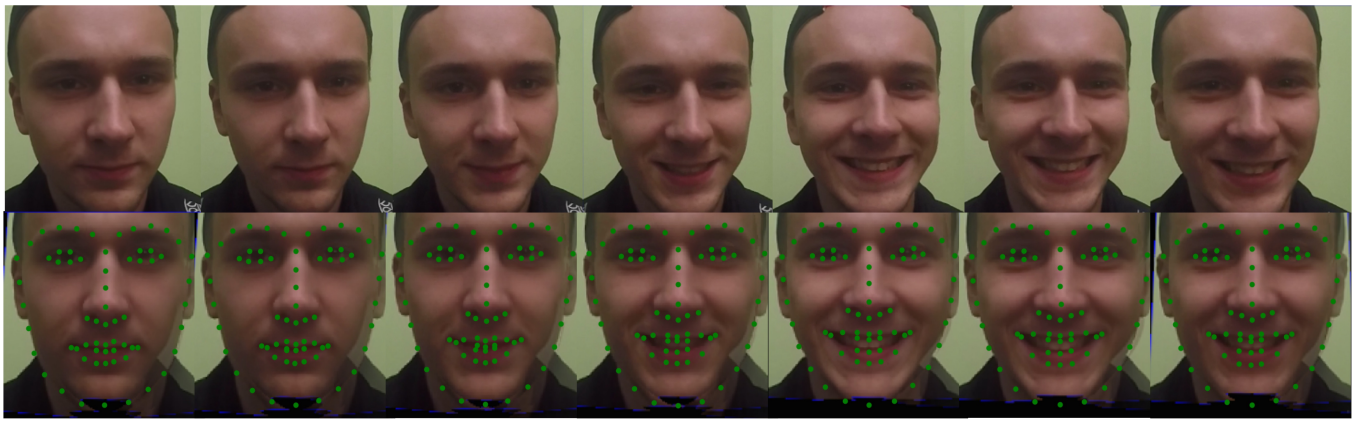


Fig. 5: Illustration of the pre-processing we perform on the data. Detected faces are first extracted and frontalized and facial landmarks localised for each image in the input video sequences.

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 empirically found optimal. 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. Typically for action recognition experiments this size is 32×32. 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 need to perform PCA to decorrelate the dimensions. We found that picking the 32 first principal components is optimal.

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 [97]. 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 $C = 100$.

5.2 Discussion

In this section, we will discuss the experimental results obtained by our proposed method. For brevity, we have denoted both in the text and figures the genuine facial expressions labels by adding a *T* (for true) in front of the labels (e.g TSad) and the corresponding deceptive facial expressions by adding an *F* (for fake) in the same fashion (e.g FAnger). We start by discussing results on the *Cohn-Kanade* and the Oulu-CASIA datasets and then we discuss the results on the proposed SASE-FE dataset.

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

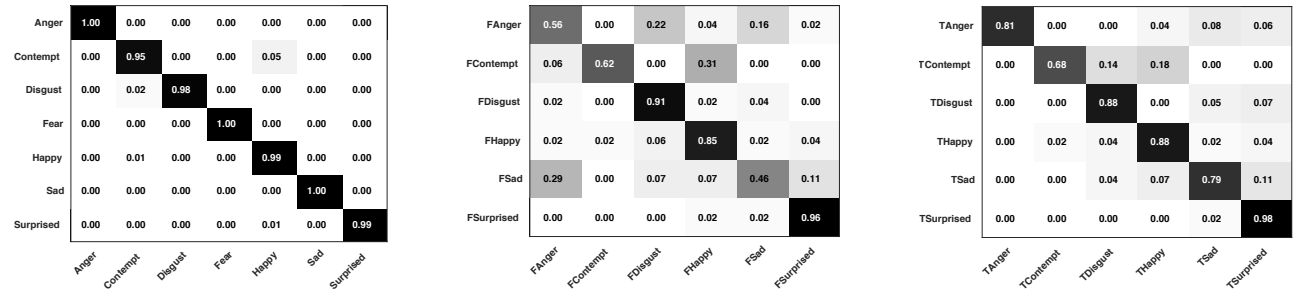
Method	Accuracy(%)
AURF [105]	92.22
AUDN [106]	93.70
STM-Explet [107]	94.2
LOmo [108]	95.1
IDT+FV [109]	95.80
Deep Belief Network [78]	96.70
Zero-Bias-CNN [110]	98.4
Ours-Final	**98.7**

TABLE 3: 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, [1] Fine-tune, no data augment, [2] Fine-tune, data augment.

	Method	Accuracy(%)
12 classes	SIFT+FV	12.1
	TPF-FGT(SIFT)+TPF-IDT(MBH)+FV	21.3
	VGG-Face [1]	43.5
	VGG-Face [2]	51.8
	TPF-FGT(VGG-Face)+FV	53.2
	TPF-FGT(VGG-Face)+FV+Aligned Faces	58.4
	TPF-FGT(VGG-Face)+FV+Aligned Faces+Data augmentation	62.7
	TPF-FGT(EMNet)+FV	69.1
	TPF-FGT(EMNet)+FV Aligned Faces	73.2
	TPF-FGT(EMNet)+FV Aligned Faces+Data augmentation	**75.6**
6 classes (true)	VGG[1]	71.4
	VGG [2]	78.4
	TPF-FGT(VGG-Face)+FV	79.6
	TPF-FGT(VGG-Face)+FV Aligned Faces	80.7
	TPF-FGT(VGG-Face) +FV Aligned Faces+ Data augmentation	82.4
	TPF-FGT(EMNet) +FV	83.2
	TPF-FGT(EMNet) +FV Aligned Faces	84.3
	TPF-FGT(EMNet) +FV Aligned Faces+ Data augmentation	**86.4**
6 classes (fake)	VGG [1]	49.9
	VGG [2]	67.4
	TPF-FGT(VGG-Face) + FV	68.3
	TPF-FGT(VGG-Face) + FV + Aligned Faces	70.4
	TPF-FGT(VGG-Face) + FV + Aligned Faces + Data augmentation	74.7
	TPF-FGT(EMNet) + FV	75.1
	TPF-FGT(EMNet) + FV + Aligned Faces	76.2
	TPF-FGT(EMNet)+FV Aligned Faces+Data augmentation (Ours-Final)	**77.1**

TABLE 4: Fake Emotions vs True Emotions classification performance on the SASE-FE dataset.

Emotion Pair	Accuracy Non Fake (%)	Accuracy Fake
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



(a) (b) (c)

Fig. 6: Confusion matrices for 6 classes classification. (a) 6 class classification on the CK+. (b) 6 class classification on the false subset of SASE-FE. (c) 6 class classification on the true subset of SASE-FE. Genuine facial expressions are labelled with an initial 'T' (for True) and deceptive facial expressions with an 'F' (for Fake).

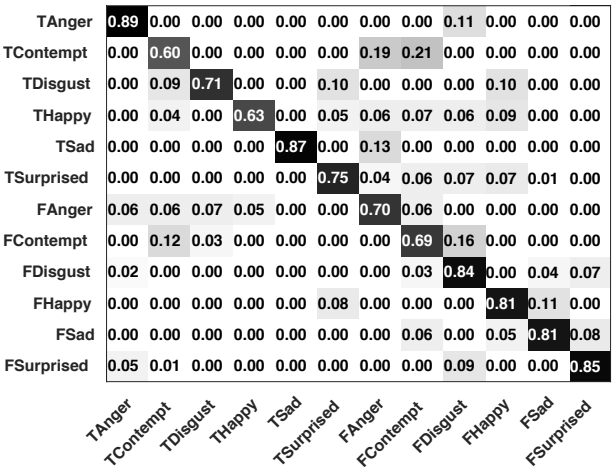


Fig. 7: Confusion matrix for 12 class classification on the SASE-FE dataset. Genuine facial expressions are labelled with an initial 'T' (for True) and deceptive facial expressions with an 'F' (for Fake).

TABLE 5: 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 [111]	81.46
LOmo [108]	82.10
PPDN [112]	84.59
FN2EN [98]	87.71
Ours-Final	**89.60**

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, [109] is the closest method to our proposed method. The authors of this paper implement the improved dense trajectories framework proposed for action recognition [113] 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 [107], [108]. In [108], 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. To conclude, we find that landmark trajectories when combined with finely tuned CNN features perform better than landmarks trajectories combined with hand-crafted features. Secondly, it will be difficult to apply such weakly supervised methods on the SASE-FE dataset, as these facial expressions of fake emotions are not very well understood by psychologists.

The rest of the methods listed in the table use deep learning techniques to classify emotions [78], [105], [106], [110]. They design networks able to specifically learn facial action units. [110] is slightly better than our performance. This network is specifically designed to do emotion recognition while we try to adapt a state of the art performing face recognition network to emotion recognition. It will be an interesting future experiment to compute the TPFs from these specifically designed emotion recognition networks and see if these trajectories can capture spatio-temporal motion information.

In the confusion matrix of the CK+ as shown in Fig. 6(a), the results are along the expected lines. The most difficult emotion to recognise is contempt. This is because it is hard to act as well as because there are fewer examples in the dataset to model these expressions.

5.2.2 Oulu-CASIA

In this section, we compare our method with the state-of-the-art methods on the Oulu-CASIA dataset. We evaluate our method on the Oulu-CASIA dataset because we can see that the recognition accuracy of most of the methods on the CK+ dataset are saturated and in the same ball park, though we achieve state-of-the-art performance on CK+. General

TABLE 6: This table shows the emotion-wise comparison between our proposed method and [98] on the Oulu-CASIA dataset

Emotion	Accuracy [98] (%))	Accuracy **Ours** (%)
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
Overall	87.7	89.6

categorization of all the methods has been discussed in the CK+ dataset section. Therefore, to show the efficacy of our method we also show results on the Oulu-CASIA dataset. In Table 5 we can observe that our method betters the previous best performance by [98] by 1.9%. In Table 6 we show the emotion-wise comparison between our proposed method and [98]. The two main differences between [98] 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 can be 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 SASE-FE

The set of presented experiments has been designed with the purpose of exploring spatial and temporal representation for the proposed problem. One can see 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 handcrafted 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 3 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 both on the proposed dataset and on the CK+. On the proposed dataset we split the classification problem in two, training on the 6 true and the 6 fake emotions respectively. On the SASE-FE, several observations can be made. Both in the case of fake and true expression 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 align 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 true emotions than when discriminating between the fake 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 fake sadness. The high level of confusion with fake anger should be noticed along with the fact that this is not the case for true emotions.

In Fig. 7 we present the confusion matrix for the problem of classifying between all 12 classes (true and fake jointly). This can be interpreted together with results in Table 3 where we present classification accuracies for each pair (true/fake). When trained with all classes, the best results are obtained for true sadness and the worst for true contempt and fake contempt. In Table 4, overall accuracies of especially the fake 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 true and fake 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 fake expressed contempt is by chance, i.e. contempt is easier to fake, 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.

Fig. 8 and 9 show the feature maps with the landmark points superimposed. The descriptor computed along the landmark point help capture the temporal changes in the feature maps.

6 Conclusion

Previous research from psychology shows that discriminative facial behaviour for the deceptive facial expression of

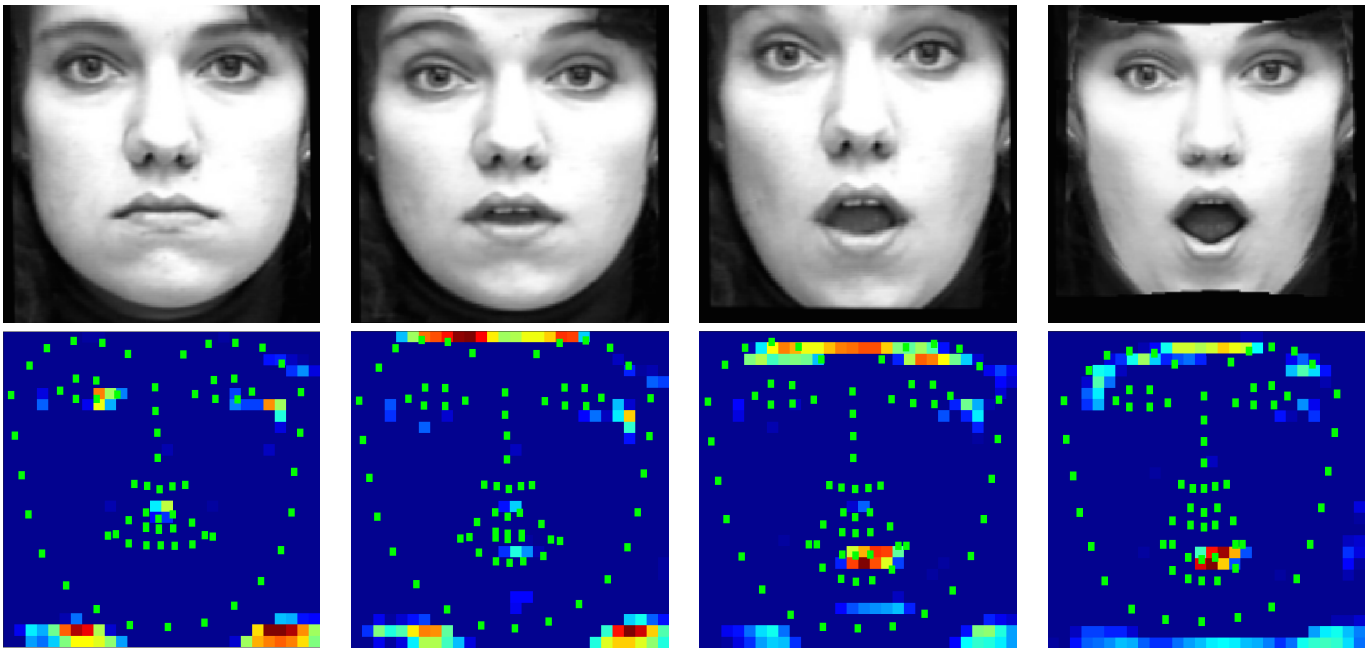


Fig. 8: Example of activation maps obtained from the *conv5* layer of the EMNet CNN for surprise facial expression of the CK+ dataset. As we can observe these feature maps activate around the lips as the emotion goes from neutral to the peak frame. The landmark points are superimposed on the feature maps. These temporal variations are captured with FGT-TPF descriptors. Best seen in colour.

emotion is subtle. For this reason, we provide for the first time a dataset capturing humans while expressing genuine and deceptive facial expressions of emotions 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 emotion recognition. We combine the features 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 show close to state of the art performance on the the publicly available CK+ and the Oulu-CASIA datasets. Furthermore, we provide several baselines on our SASE-FE dataset. We show that even though we obtain good results on the 6 class true and fake problem, the 12 class and the binary emotion pair classification problem still remains a challenge. This is because the distinguishing factors between the deceptive and genuine emotions 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 have the spatial resolution to detect minute changes in facial muscle movements, which required to differentiate and distinguish between deceptive facial expressions of emotions. Therefore, a CNN specific for the classification of genuine and deceptive facial expressions of emotions can be trained on the dataset collected at very high resolution.

Some future research directions would be to learn better temporal representations. Therefore, investigating the use of Recurrent Neural Nets or 3D-CNNs, specially trained on data collected at high fps as is done in the SASE-FE dataset, would be of interest.

Acknowledgments

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 Projects TIN2013-43478-P and 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° 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 *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 707–714.

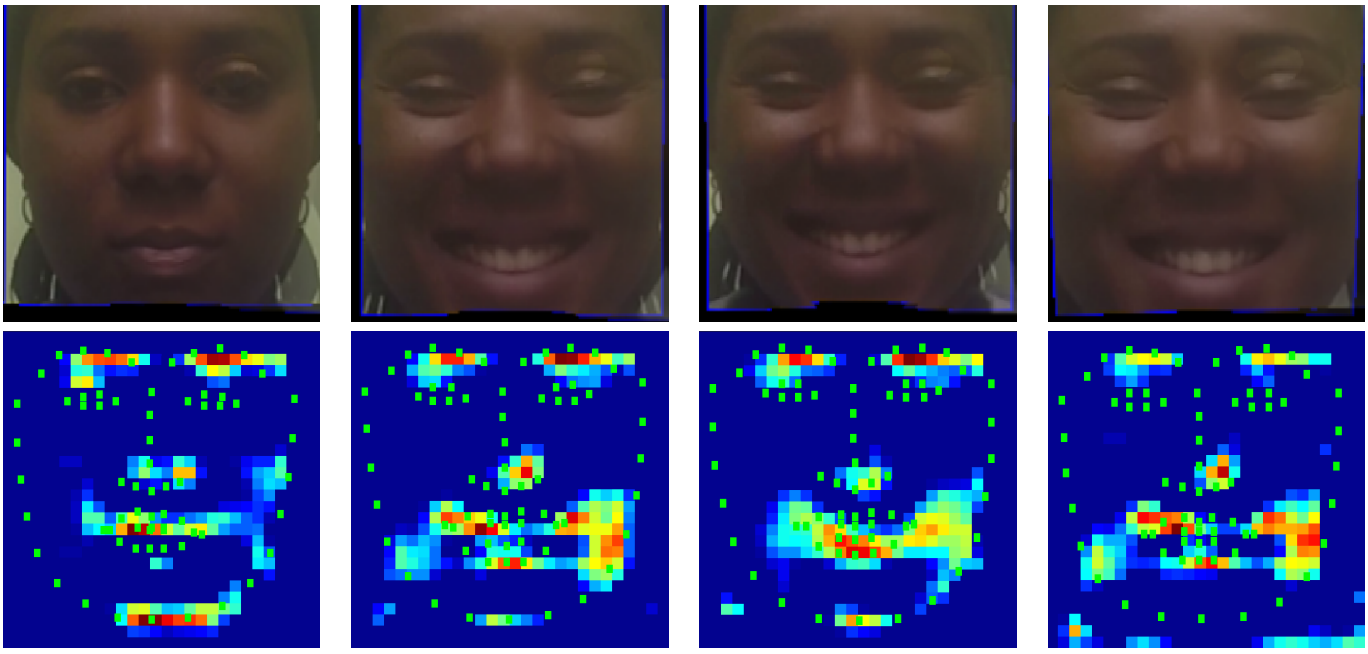


Fig. 9: Activation maps obtained from the *conv5* layer of the EMNet CNN for Happy facial expression of the SASE-FE dataset. The landmark points are superimposed on the feature maps in green. These temporal variations are captured with FGT-TPF descriptors as the feature map changes through time. Best seen in color.

[6] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, vol. 4. IEEE, 2002, pp. 4138–4142.

[7] T. Shibata and R. Irie, "Artificial emotional creature for human-robot interaction-a new direction for intelligent systems," in *Advanced Intelligent Mechatronics' 97. Final Program and Abstracts., IEEE/ASME International Conference on*. IEEE, 1997, p. 47.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] . F. W. Ekman, P., *Unmasking the face: A guide to recognizing emotions from facial clues*, 1975.

[14] C. Darwin, *The expression of the emotions in man and animals, New York: D*, 1872.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] P. Ekman, W. V. Friesen, and M. O'Sullivan, "Smiles when lying," *What the face reveals. New York: Oxford University Press. p*, pp. 201–216, 1997.

[22] 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.

[23] 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.

[24] 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.

[25] M. G. Frank, "Smiles, lies, and emotion." 2002.

[26] P. F. Ekman and V. Friesen, "Wv and o'sullivan, m.(1988)'smiles when lying'," *Journal of Personality and Social Psychology*, vol. 54, no. 3, pp. 414–420.

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

[28] K. L. Schmidt and J. F. Cohn, "Human facial expressions as adaptations: Evolutionary questions in facial expression research," *American journal of physical anthropology*, vol. 116, no. S33, pp. 3–24, 2001.

[29] E. H. Rauch, "Cues to deception: eye blinking," Ph.D. dissertation, San Francisco State University, 2015.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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 *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 809–813.

[43] 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 *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 833–838.

[44] 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 *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

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

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

[47] 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.

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

[49] 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.

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

[51] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[52] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[53] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *International joint conference on pattern recognition*, vol. 1978, 1978, pp. 408–410.

[54] S. S. Tomkins and R. McCarter, "What and where are the primary affects? some evidence for a theory," *Perceptual and motor skills*, vol. 18, no. 1, pp. 119–158, 1964.

[55] J. F. Cohn and P. & Ekman, "Methods for measuring facial action," *Handbook of Methods in Nonverbal Behavior Research*, pp. 45–90, 1982.

[56] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 46–53.

[57] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[58] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[59] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

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

[66] 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.

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

[68] 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.

[69] 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.

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

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

[72] 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.

[73] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *CVPR*, 2011, pp. 2857–2864.

[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] 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.

[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*, 2013, pp. 1–6.

[78] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.
[79] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3D dynamic faces using robust spatio-temporal shape features," in *FG*, 2011, pp. 414–421.
[80] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3D facial expressions and their temporal models," in *FG*, 2011, pp. 406–413.
[81] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden markov model for facial expression recognition," in *FG*, 2015.
[82] 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.
[83] 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.
[84] 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.
[85] 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 Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 664–674, 2011.
[86] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju, "Real-time automatic deceit detection from involuntary facial expressions," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–6.
[87] 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.
[88] J. J. Dahling and L. A. Perez, "Older worker, different actor? linking age and emotional labor strategies," *Personality and Individual Differences*, vol. 48, no. 5, pp. 574–578, 2010.
[89] 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.
[90] R. Plutchik, "Emotions, evolution, and adaptive processes," in *Feelings and emotions: the Loyola Symposium*. Academic Press, New York, 1970, pp. 3–24.
[91] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer vision and image understanding*, vol. 108, no. 1, pp. 116–134, 2007.
[92] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, 2017.
[93] 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.
[94] N. R. Whitesell and S. Harter, "Children's reports of conflict between simultaneous opposite-valence emotions," *Child Development*, pp. 673–682, 1989.
[95] Y. Y. Mathieu, "Annotation of emotions and feelings in texts," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 350–357.
[96] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.
[97] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1817–1824.
[98] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 118–126.
[99] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.
[100] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.
[101] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.
[102] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.
[103] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.
[104] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.
[105] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.
[106] ——, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.
[107] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749–1756.
[108] K. Sikka, G. Sharma, and M. Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 5580–5589.
[109] S. Afshar and A. A. Salah, "Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 1517–1525.
[110] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 19–27.
[111] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2983–2991.
[112] 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.
[113] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

Ikechukwu Ofodile obtained his BSc in Electrical and Electronics Engineering from Eastern Mediterranean University, North Cyprus. He is currently a MSc Student and a member of the intelligent computer vision (iCV) research group at the University of Tartu, Estonia. He is also a member of Philosopher, the Estonian Robocup team of the University of Tartu and a member of the Estonian student satellite project working on attitude determination and control of ESTCube-2. His research interests include machine learning, pattern recognition and HCI as well as control engineering and attitude control system design for nanosatellites and microsatellites.

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.

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 also a member of the Computer Vision Center at Campus UAB. 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.

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 (UOC).

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" (CISA). Afterward, she worked as an independent research Fellow of the Japan Society for the Promotion of Science (JSPS) 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 (CAAR).

Jüri Allik was Candidate of Science (PhD), University of Moscow (1976) and obtained PhD in psychology from the University of Tampere, Finland (1991). He has been Chairman of Estonian Science Foundation (2003-2009), Professor of Psychophysics (1992-2002) and Professor of Experimental Psychology (2002-) at the University of Tartu. He was also Dean of Faculty of Social Sciences (1996-2001), President (1988-1994) and Vice-President (1994-2001) of the Estonian Psychological Association. He served as a Foreign Member of the Finnish Academy of Science and Letters (1997). He has has received many awards including Estonian National Science Award in Social Sciences category (1998, 2005). 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 intelligent computer vision (iCV) research group in the Institute of Technology at the University of Tartu. He is also Deputy Scientific Coordinator of he European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307. He is Associate Editor and Guest Lead Editor of several journals, Special Issues and Book projects. He is an IEEE Senior member and the Vice Chair of Signal Processing / Circuits and Systems / Solid-State Circuits Joint Societies Chapter of IEEE Estonian section. He has got Estonian Research Council Grant (PUT638) in January 2015 and has been involved in many international industrial projects such as AIDesign, A.G.E., iRental and Virtual Fitting Room by Fits.Me Rakutan. He is expert in computer vision, human-robot interaction, graphical models and artificial intelligence. His work in image super resolution has been selected for the best paper award in 2012 by Electronics and Telecommunications Research Institute (ETRI) Journal, South Korea. He has supervised 9 MSc students and 5 PhD students. He has published over 100 scientific works. He has been in the organizing committee and technical committee of IEEE Signal Processing and Communications Applications Conference in 2013, 2014 and 2016 and TCP of conferences such as ICOSST, ICGIP, SampTA and SIU. He has been organizing challenges and workshops in FG17, CVPR17, and ICCV17.