Changes in Facial Expression as Biometric

A Database and Benchmarks of Identification

Haamer, Rain Eric; Kulkarni, Kaustubh; Imanpour, Nasrin; Haque, Mohammad Ahsanul; Avots, Egils; Breisch, Michelle; Nasrollahi, Kamal; Guerrero, Sergio Escalera; Ozcinar, Cagri; Baro, Xavier; Naghsh-Nilchi, Ahmad Reza; Moeslund, Thomas B.; Anbarjafari, Gholamreza

Published in:
13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)

DOI (link to publication from Publisher):
10.1109/FG.2018.00098

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Changes in Facial Expression as Biometric: A Database and Benchmarks of Identification

Rain Eric Haamer*, Kaustubh Kulkarni†‡‡, Nasrin Imanpour‡§, Mohammad A. Haque‡, Egils Avots***, Michelle Breisch*, Kamal Nasrollahi‡, Sergio Escalera†††‡, Cagri Ozcinar¶, Xavier Baro†‖, Ahmad R. Naghsh-Nilchi§, Thomas B. Moeslund‡ and Golamreza Anbarjafari***††

*iCV Lab, University of Tartu, Estonia

†Computer Vision Center (CVC), Barcelona, Spain

††University of Barcelona, Spain

‡‡Universitat Autonoma de Barcelona, Spain

‡Visual Analysis of People Lab, Aalborg University, Denmark

§University of Isfahan, Iran

¶Trinity College Dublin, Ireland

‖Open University of Catalonia, Spain

**GoSwift Inc., Estonia

††Corresponding author's email: shb@icv.tuit.ut.ee

***Abstract*—Facial dynamics can be considered as unique signatures for discrimination between people. These have started to become important topic since many devices have the possibility of unlocking using face recognition or verification. In this work, we evaluate the efficacy of the transition frames of video in emotion as compared to the peak emotion frames for identification. For experiments with transition frames we extract features from each frame of the video from a fine-tuned VGG-Face Convolutional Neural Network (CNN) and geometric features from facial landmark points. To model the temporal context of the transition frames we train a Long-Short Term Memory (LSTM) on the geometric and the CNN features. Furthermore, we employ two fusion strategies: first, an early fusion, in which the geometric and the CNN features are stacked and fed to the LSTM. Second, a late fusion, in which the prediction of the LSTMs, trained independently on the two features, are stacked and used with a Support Vector Machine (SVM). Experimental results show that the late fusion strategy gives the best results and the transition frames give better identification results as compared to the peak emotion frames.**

I. INTRODUCTION

The human need for highly privatized systems for accessing personal information is evident. With advances in technology, biometric systems have emerged to identify individuals using biological features. Examples of biological features as biometric traits include face, fingerprint, iris, retina, hand geometry, voice, signature, gait, heart signal and so on [5], [13], [21], [30]. While considering facial biometric, face recognition aims for automatically identifying individuals from their digital images or video frames using physiological features. On the contrary, face verification certifies the user identity in terms of a predefined claimed identity. Over the past decades, significant efforts have been made to develop face recognition algorithms with 2D frontal face images/videos. Face recognition from images uses to include three main steps: face detection, feature extraction, and classification [14].

Former studies have shown that, when a large number of representative training images are available, computer algorithms are able to recognize even better than humans [6], [7], [25], [27], [33]. These algorithms, before recognizing, represent the face in the feature space, and perhaps because they are able to extract information from training images about the changes caused by different conditions they outperform humans in recognition [25].

In this work, we aim at utilizing facial dynamics that are visible during changes of facial expressions from one motion to the other, to assist regular face recognition systems. There are some works on using facial expression for face recognition [3]. Studies in this field contain facial action unit coding during a course of an expression. Facial action units are unique for each person [4]. They also do not change due to aging [4].

Tubbs et al. calculated the Euclidean distance of a few facial feature points (like the corner of an eye or lip) during the course of an expression (in consecutive frames) and ranked them according to how far they move [28]. This method is sufficient to describe an emotion and is computationally cheap because of using only some 2D feature points. For experiments they recorded a database from a neutral expression to a prompted expression (including angry, disgusted, happy, sad, surprised, and scared) and then again back to a neutral expression. They achieved at best an error of only 3.4%, with an average error of 33.28% with 309,210 authentication attempts on 33 user profiles. Gavrilescu used neural network, which takes detected faces in multiple frames as input and predicts each individual's identity based on facial expression behavioural map extracted from segmented face [10]. This face expression based recognition alongside with Principal

978-1-5386-2335-0/18/$31.00

Component Analysis (PCA)-based face recognition algorithm are fed into a decision layer of a neural net to determine the identity of each individual. They achieved 85% accuracy with only the facial expression based method, but when combined with the PCA-based face recognition algorithm they achieved 94.5% accuracy on Honda/UCSD Video database and 92.9% on YouTube Faces database. Early studies used the first and last frames during an expression for human identification [24], [29], [32].

In this study we extend the method of Haque et al. [15] for human identification based on universal facial expressions (we call this emotion based face recognition). Coding facial action units in the course of an expression implies the temporal analysis of facial action units. This can be done more effectively via a Recurrent Neural Network (RNN) which exploit the temporal axis information from facial video. On the other hand these facial action units can be extracted automatically and in a more efficient way using a Convolutional Neural Network (CNN). We also hypothesize that transitions between different expressions can also be unique for each person, and therefore, it can be exploited as another discriminative information for secure human identification. Our experiments verifies the validity and effectiveness of this idea (we call this transition based face recognition). To the best of our knowledge, this is the first work that aims for learning chances of expression transition features for face recognition. We also used a late fusion strategy to examine if emotions and transition based recognition methods can be complementary. Given that there is no public dataset containing transitions between different facial expressions, we collected a new facial database to be publicly available, where the subjects are changing their facial expression in a sequence. The main contributions of this paper are:

- Developing the first sequential facial expression dataset for face recognition[1];
- Investigating on use of spatio-temporal emotion features to recognize users;
- Investigating on use of spatio-temporal transition features to recognize users;
- Employing late fusion of emotion and transition systems for boosting recognition rate.

The rest of the paper is organized as follows. Firstly, we introduce our new facial database in details in Section II. Our methodology for person identification is described in Section III. Section IV reports the experimental results. Finally Section V concludes the work.

II. DATABASE

Before any recording could take place, we set up a standard protocol, so all of the recordings would have very little undesired variances. This protocol was followed throughout each recording and provided the main emphasis for each decision.

[1]Please contact the corresponding author for the database download link.

Fig. 1. Overview of the recording setup from two different angles. The lighting conditions of the setup were altered for these pictures as the stationary light source did not fully illuminate the camera nor the recorder.

A. Equipment

For video capture, a Canon LEGRIA HF R66 Full HD Camcorder was used. All recordings were done in 1920×1080 with a frame rate of 25 fps [1]. All other video settings were left at the factory default, including auto focus and auto illumination enhancement. The camera was placed on a 1 m stand, which was only regulated when the height of the subject demanded it. A separate stand was used for holding the 2 spotlights and a small stool was provided to keep the location of the subjects fixed.

B. Setup

In order to keep the location, angle and perspective of the face uniform throughout the recordings, a standard distance of 75 ± 5 cm between the camera lens and the face of the subject was set for all sessions. An example of the recording setup with extra light sources can be seen on Fig. 1. The subjects were asked to limit the movement of their heads as only a frontal recording of each subject was made. The lighting conditions between different recording sessions were kept uniform by the use of 2 fixed diffused light sources. The fixed light sources were 50 cm behind the camera and angled $225°$ and $180°$ away clockwise from the subject. The pitches for both light sources were $45°$ and $0°$ respectively. All other light sources were either disabled or blocked as natural sunlight would have had an effect on the luminance of the videos. The venue for the recordings was a gray walled hallway with no windows and no distinct features which could affect the videos.

C. Sequences

This database is strictly comprised of natural posed faces where the subjects had no prior knowledge of how to perform certain expressions, nor were they allowed to practice in front of a mirror. Subjects were also prohibited from being under

Fig. 2. Samples of 6 different participants during recording showing expressions. The expressions from left to right, top to bottom are as follows: Angry, Surprised, Surprised, Neutral, Happy, Angry.

Fig. 3. Male sample for 2 recorded sequences. The 2 columns correspond to sequence 1 and 2, respectively.

the influence for the recording session as that could have had an affect on the final results [23].

In order to obtain the transitions between different emotions, a set of 5 emotional states - neutral [N], happy [H], surprised [S], angry [A] and unhappy [U], were chosen as seen on Fig. 2. These emotions were then combined into 2 different sequences, where the orders were N-H-U-A-S and N-A-H-S-U. The recordings of the 2 sets were done in succession and before each set, the subjects were given a brief summary of the set. This greatly reduced the amount of confusion when transitioning over to the second set.

Each of the emotions were recorded for 3 seconds, after which the subjects were notified to move on to the next expression. As each sequence of emotions comprised of 5 emotional states, the resulting videos were all 15.5 ± 0.5 sec long. For the sake of redundancy, 5-7 shots of both sequences were recorded for each person and only the best 5 were kept. If the subject was wearing glasses, a second set of recordings was made without the glasses, in order to remove the effect of facial obstructions in the final database.

Fig. 4. Female sample for 2 recorded sequences. The 2 columns correspond to sequence 1 and 2, respectively.

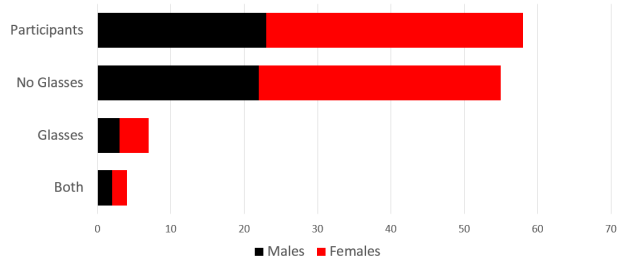

Fig. 5. The general distribution of genders and people who wore eye-wear during the recordings. The bottom bar is important as it shows the amount of participants who were willing participate in 2 recordings, one with glasses and one without.

D. Analysis

The final recordings were conducted for 61 different persons for a total of 630 videos. The 20 video discrepancy was caused by people who had recordings done with and without glasses (4.9%). The database has a slight majority of female participants (63.9%), which was caused by the nature of the recording setting.

The database also has a marginal majority of females who wore glasses (10.3%) compared to the minuscule 4.5% for males. The total number participants wearing glasses was 8.2%, with 40.0% of those having no without-glasses counterpart recordings as seen in Fig. 5.

The general age group of the recorded participants had a mean of 23.64 years with a standard deviation of 5.89. The minimum recorded age was 18 and the maximum was 54, though this was an unique case as the next oldest person recorded was 27.

III. METHODOLOGY

In this section, we describe our approach for face recognition based on expression transition sequences. It follows the standard CNN-LSTM architecture, as shown in Fig. 6.

TABLE I
RECORDING SPECIFICATIONS

Videos	630
Expressions	5
Mean Duration	15 s
Resolution	1920 x 1080
FPS	25
File type	MTS

TABLE II
PARTICIPANT DEMOGRAPHIC

	Participants	**Glasses**
Females	39	4
Males	22	1
Total	61	5

TABLE III
PARTICIPANT AGE GROUPS

	Age (y)
Mean	23.64
STD	5.89
Min	18
Max	54

Although face recognition from 2D face images/videos has reached its top performance, there are still challenges for avoiding performance reduction caused by intruders trying to mislead recognition systems by some fake evidences. Our approach exploits more diverse information of each individual considering their facial expressions. It is well known that the way each person makes an expression is unique for each person, and therefore, it can be used as a discriminative information to help common face recognition algorithms to be more secure. We also hypothesized that the transitions between different facial expressions can also be unique for each person. In next subsections, we describe face pre-processing step to extract the face and compute the facial landmark points, our method and network structure with some brief explanation of elements of our network structure, and how we can fuse emotion and transition based methods.

A. Method

A deep learning framework for facial video analysis entails two kinds of information processing: spatial information and temporal information. Spatial information comes from facial features in a single video frame. On the other hand, temporal information stands for the relationship between facial features revealed in consecutive video frames [2], [9], [22]. CNNs are well known for their great ability in learning abstract spatial features from a given image (single frame) [11]. On the other hand, a Recurrent Neural Network (RNN) can exploit the temporal axis information from facial video. So our network for human identification consists of a CNN followed by a RNN for spatial and temporal information analysis of facial expression videos.

For CNN instead of designing and training the network from scratch, we fine tune from existing networks. In general fine-tuning of networks that are trained on a larger dataset can enrich the features. In this work we fine tuned the network suggested by [26]. It is a 16 layer network trained on 2.6M images, over 2.6K people. We obtained the features of the `fc7` layer and used them as input to Recurrent neural network (RNN) architecture.

RNNs are networks with feedback connections which makes them able to store representations of recent input events. Therefore, RNNs are very useful for learning temporal dependencies among consecutive frames of video. A critical issue with RNNs is exploding/vanishing gradients. Exploding gradients cause to oscillating weights. When there is long time lags between the relevant information and the point where it is needed, vanishing gradients make the training prohibitive or with low generalization capabilities.

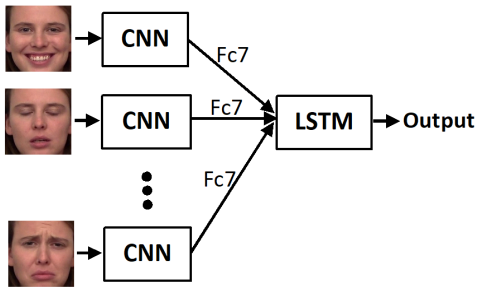

Fig. 6. The block diagram of CNN-LSTM based human identification system.

Hochreiter et al. introduced another recurrent network architecture called Long-Short Term Memory (LSTM) with an appropriate gradient-based learning algorithm [16]. LSTM can store and access information over long periods of time, thereby mitigating the vanishing gradient problem. It can learn long time lags of 1000 steps even in case of noisy, incompressible input sequences, while it can preserve its short time lag capabilities [16]. The structure of what is used as LSTM now is somehow different from the original LSTM. The different extensions of LSTM are explained in [12]. In this paper we used normal LSTM [17]. The repeating module in normal LSTM has four neural networks instead of one in standard RNN. The key of LSTMs is the cell state (long term memory) where information can flow along it unchanged. LSTM has the ability to remove or add information to the cell state, controlled by three structures called gates generating a number between zero and one to determine how much of information let through (1: Completely keep this; 0: completely get rid of this). As depicted in Fig. 7, the first step in LSTM is to decide what information is going to be erased from the cell state. This step is controlled by forget gate.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{1}$$

The next step is to decide what new information should be saved in the cell state which the amount is controlled by input gate.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C) \tag{3}$$

In the third step the old cell state updates into the new cell state according to the output of the first and second steps.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

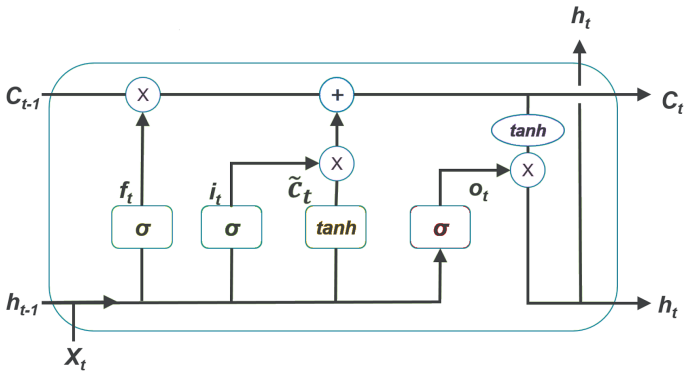

Fig. 7. LSTM repeating module.

Finally according to the new cell state the output is generated in forth step. The controlling gate of this step is called output gate.

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

B. Fusion

In our work, we perform both early fusion, Fig. 8, and late fusion, Fig. 9. In early fusion feature vectors obtained from different sources are concatenated into a single vector and then classifiers are trained on this concatenated vector. In late fusion the classifiers are trained on the the feature vectors from different sources and the scores w.r.t each classes are then combined to make a decision on the optimal class. There are different ways to combine the classifiers scores such as the sum rule and the product rule [20]. The problem with these rules is that is if difficult to estimate different weights for the scores belonging to each class. Therefore, in our work we stack the output of the different classifiers into a single vector and the train a Support Vector Machine (SVM) to make a combined decision [8].

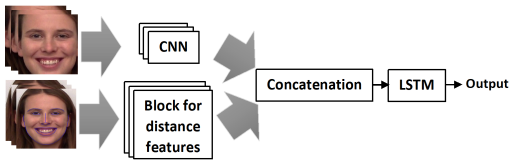

Fig. 8. Early Fusion.

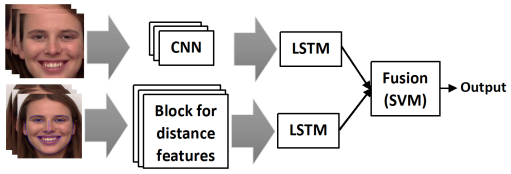

Fig. 9. Late Fusion.

Fig. 10. Examples of the face detection and the Facial landmark point detection for one actor at peak emotion frames.

IV. EXPERIMENTAL RESULTS

In this section, we perform the experimental evaluation of expression transition based face recognition on the proposed dataset. We compare the face recognition performance between the classifiers which are trained on the transition frames and peak emotion frames. The features used for the comparison are computed from two sources: the fine-tuned CNNs and the 68 facial landmark points extracted from each frame. To model temporal evolution of the transition frames we use a LSTM. The details of the experiments are as described in the following sections.

A. Implementation Details

From the collected dataset, the peak frames of all the emotions and the frames of the transitions between the emotions are annotated. Therefore, for each video in the dataset we get 5 frames corresponding to peak emotion frames and 4 pairs of start and end frames of transition between the 5 emotions. We compute the face recognition accuracy on the peak emotion frames and the emotions. We take 60% of videos of each subject as training 20% for validation and 20% testing which gives us a total of 376 videos for training, 126 videos for validation and 125 videos for testing respectively.

As pre-processing, we crop the faces from each frame of the dataset based on [31]. All frames are cropped to the size of 224×224. Then 68 facial landmark points are detected for all frames in the dataset as in [18]. In Fig. 10 we can see the cropped faces and landmark points. A feature vector is computed as a distance between all the pairs of landmark points for every frame in the dataset. Thus, giving a vector of dimension 2278 for each frame.

To extract the CNN based features we fine-tune a VGG-Face CNN. We fine-tune the two different VGG-Face CNN with the training data from the transition frames and the peak frames. All the frames are cropped to the size of 224×224. We keep the base learning rate 0.001 and the momentum as 0.9. The batch size is 20. We train the model for 10 epochs. All the convolution layers are trained at the base learning rate while the fully connected layers are trained at 10 times the

TABLE IV
TRANSITION RECOGNITION

Method	N-H	H-U	U-A	A-S	N-A	A-H	H-S	S-U	Mean
CNN-LSTM	97.8	96.4	97.4	99.1	98.2	97.3	97.6	97.3	97.6
Geom-LSTM	51.3	52.6	51.9	52.5	53.4	53.8	54.2	55.3	53.1
Early Fusion	93.4	92.4	93.4	94.1	92.2	91.1	92.1	93.2	92.7
Late Fusion	98.9	97.4	98.3	98.3	98.6	98.7	98.3	98.3	98.4

base learning rate. The training parameters were set on the validation set and the recognition accuracy is also computed on the test set.

Secondly, we train a LSTM to model the temporal context between the transition frames. This is because the transition frames are sequences of frames while the emotion frame is just a single peak frame of the emotion. We perform four sets of experiments with the LSTMs. In the first experiment, the output of the $fc7$ or the last layer of the fine-tuned VGG-Face CNN is used to extract a per frame feature vector for the transition frames. Then the LSTM is trained on the training set and the parameters are tuned on the validation set. The LSTM with trained with Adam optimizer [19]. The learning rate is started from 0.1 and decayed exponentially and the model is trained for 10 epochs. We choose the size of the hidden state for the LSTM to be 256. The dimensionality of the the per-frame vector is large i.e 4096 and given the amount of training data available we apply PCA to reduce the dimensionality of the vector. The size of the vector is set to 1024. In the second experiment, we input the per frame geometric features computed from the distances between the landmark points. Again, we reduce the dimensionality of the vector to the dimension of 512 with PCA. The rest of the parameters are same for training the LSTM. The third experiment we do is the early fusion of the the dimensionality reduced features from the CNN and the landmark points. These two vectors are concatenated to form a single vector and are then normalized with the $l2$ norm. The LSTM is then trained with the fused vector with the same parameters as before. In late fusion, the output of the CNN-LSTM and Geom-LSTM at each frame of the sequences are stacked and then normalized such that the sum of the squares is 1. Then on these stacked vectors a linear SVM is trained for prediction.

B. Discussion

In this section, we discuss the results of the experiments the setup of which was described in the previous section. The results are tabulated in tables V and IV. All results are reported on the test partition of the dataset. The results are denoted in terms of bar graphs in Fig.11, where the first two bars are the recognition accuracies of the peak emotions in test set and the third is the recognition accuracy of the transitions between these peaks. We can observe that the transition frames consistently outperform the emotion frames.

In table IV, the first row gives the result of the CNN based features as input to the LSTM and the second row gives the geometric distance features as input to the LSTM. WFrom experimental results we observe that the CNN+LSTM obtains

TABLE V
EMOTION BASED FACE RECOGNITION

Method	N	H	U	A	S	Mean
CNN	96.1	97.1	96.4	96.3	96.7	96.5
Geom-SVM	20.1	19.9	18.9	21.2	21.1	20.2
Early Fusion	90.3	90.0	91.3	90.6	91.9	90.8
Late Fusion	96.2	96.4	95.9	96.1	96.4	96.2

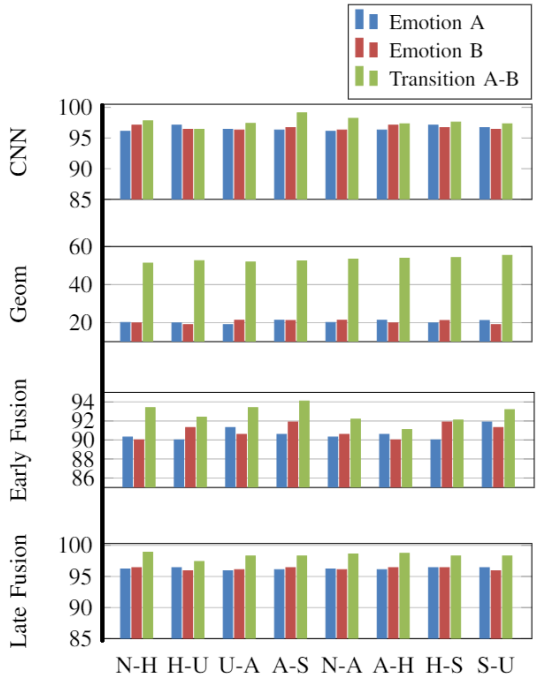

Fig. 11. Recognition accuracy.

well results as emotions does. This shows that the appearance transition features contain the discriminative information about face recognition. The next two rows give the recognition results of:

- Early fusion as in Fig. 8, where the geometric features are concatenated with the CNN features and a single LSTM is trained, and
- Late fusion as in Fig. 9, where a SVM is trained on the stacked output of the CNN+LSTM and Geom+LSTM.

It is important to note that when doing late fusion of the geometric and the CNN pipelines we observe that this pipeline gives the best accuracy of all experiments. This shows that there is some complimentary information to the appearance features in the geometric features that helps recognition. The early fusion under-performs in comparison to late fusion. In

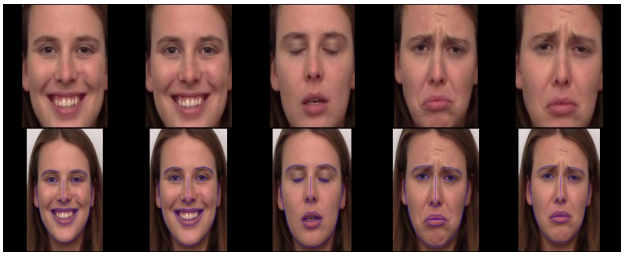

Fig. 12. Correctly recognised example for transition happy to sad.

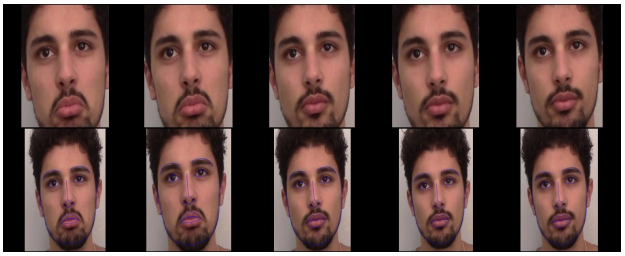

Fig. 13. Challenging cases for transition based face recognition.

early fusion the PCA reduced geometric features and the PCA reduced CNN features are concatenated into a single vector and a LSTM in trained. We observe that the dimensionality of the features after concatenation becomes large and given the limited number of sequences available to estimate the parameters of an LSTM causes the drop in accuracy.

In table V we show the experiments with the peak emotion frames. In the first row of the table fine-tuned VGG-Face is used for the peak emotion frames. In the second row, a SVM classifier is trained on the distance features computed from the single peak emotion frame. We can see from the second row of the table that the distance features computed from the single peak frame and with SVM for classification perform poorly as compared to the second row of the table IV where LSTM models the temporal context in the transition frames. We do the early fusion where a SVM is estimated on the concatenation of the CNN and the distance features and late fusion where a SVM is trained on the CNN and the Geom-SVM classifiers, respectively. We can see that the late fusion outperforms the early fusion. Moreover, we can also experimentally conclude from the mean values in the last columns of the table V and IV that the best result is obtained with transition frames. This shows the discriminative power of the expression transition frames for face recognition.

Finally, we do late fusion between the CNN classifier trained on the peak emotion frames and the CNN+LSTM and Geom+LSTM classifiers trained on the transition frames. The results are shown table VI. One can see from the first row of the table that the late fusion performs almost the same as the CNN+LSTM results while the late fusion with the peak emotion CNN helps the Geom+LSTM recognition accuracy. This shows the complementarity of both appearance and geometric features.

From the experimental results we observe that transition

Fig. 14. An example illustrating the confusion between the subjects for emotion base recognition

emotion based recognition performs with very high accuracy. One such case for which the recognition works is shown in Fig. 12. One must notice that the subjects acts the emotion with intensity. For a very few examples the method has misclassifications. One such case we observe is when the transitions between emotions are ambiguous we observe that the LSTM cannot make an accurate prediction. This is shown in Fig. 13. One can see in the figure that the subject does not act the emotion out with intensity especially anger, the last two frame to the right can look like neutral emotion and this can cause misclassification. Furthermore, we also show an example in which confusion occurs for when recognizing with only peak emotion frames in Fig. 14.

V. CONCLUSION

In this work the efficiency of the transition frames in emotion and peak emotion frames for face recognition was evaluated on proposed dataset. The extracted features from each transition frames using VGG-Face CNN and a geometric feature extractor were used as an input to LSTM algorithm for modeling the temporal context of the transition frames. Two fusion strategies were used. Early fusion containing the early concatenation of the geometric and the CNN feature, and a late fusion by SVM in the predictions of the LSTMs trained independently on mentioned two features. Moreover we also do a late fusion of the classifiers trained with the peak emotion frames and the classifiers trained with the transition frames. The experimental result showed that the transition frames outperform the peak emotion frames in face recognition.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work has also been partially supported by Estonian Research Council Grant (PUT638), the Scientific and Technological Research Council of Turkey (TBTAK) 1001 Project (116E097), and the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund. This project has also received funding from the European Unions Horizon 2020 research and innovation program under Marie Skodowska-Curie grant agreement No 6655919.

TABLE VI
LATE FUSION OF EMOTION BASED CLASSIFIERS AND TRANSITION BASED FEATURES

Late fusion Method	N-H	H-U	U-A	A-S	N-A	A-H	H-S	S-U	Mean
P(CNN)+T(CNN+LSTM)	97.6	96.1	97.5	98.7	97.6	97.9	97.2	97.6	97.5
P(CNN)+T(Geom+LSTM)	76.4	77.8	75.1	76.3	78.9	75.8	77.3	77.9	76.9

REFERENCES

[1] HD Camcorder Instruction Manual canon legria hf r66 full hd camcorder details. http://gdlp01.c-wss.com/gds/4/0300017444/02/hfr66-67-68-606-im2-p-en.pdf. Accessed: 2018-01-09.

[2] Marco Bellantonio, Mohammad A Haque, Pau Rodriguez, Kamal Nasrollahi, Taisi Telve, Sergio Escarela, Jordi Gonzalez, Thomas B Moeslund, Pejman Rasti, and Gholamreza Anbarjafari. Spatio-temporal pain recognition in cnn-based super-resolved facial images. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 151–162. Springer, 2016.

[3] Kyong I Chang, Kevin W Bowyer, and Patrick J Flynn. Multiple nose region matching for 3d face recognition under varying facial expression. *PAMI*, 28(10):1695–1700, 2006.

[4] Jeffrey F Cohn, Karen Schmidt, Ralph Gross, and Paul Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)*, 2002.

[5] Hasan Demirel and Gholamreza Anbarjafari. Data fusion boosted face recognition based on probability distribution functions in different colour channels. *EURASIP Journal on Advances in Signal Processing*, 2009:25, 2009.

[6] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):37, 2016.

[7] Changxing Ding and Dacheng Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144–152, 2017.

[8] Saso Džeroski and Bernard Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3):255–273, Mar 2004.

[9] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[10] Mihai Gavrilescu. Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks. *IET Biometrics*, 5(3):236–242, 2016.

[11] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[12] Alex Graves. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 37–45. Springer, 2012.

[13] Mohammad A Haque, Kamal Nasrollahi, and Thomas B Moeslund. Can contact-free measurement of heartbeat signal be used in forensics? In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 769–773. IEEE, 2015.

[14] Mohammad A Haque, Kamal Nasrollahi, and Thomas B Moeslund. Heartbeat signal from facial video for biometric recognition. In *Scandinavian Conference on Image Analysis*, pages 165–174. Springer, 2015.

[15] Mohammad A Haque, Kamal Nasrollahi, and Thomas B Moeslund. Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain? In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2017.

[16] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[17] Chaudhary Muhammad Aqdus Ilyas, Kamal Nasrollahi, Thomas B. Moeslund, Matthias Rehm, and Mohammad Ahsanul Haque. *Facial Expression Recognition for Traumatic Brain Injured Patients*, volume 4, page 1. SCITEPRESS Digital Library, 2018.

[18] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, June 2014.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[20] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.

[21] Juris Klonovs, Mohammad A. Haque, Volker Krueger, Kamal Nasrollahi, Karen Andersen-Ranberg, Thomas B. Moeslund, and Erika G. Spaich. *Distributed Computing and Monitoring Technologies for Older Patients*. SpringerBriefs in Computer Science. Springer International Publishing, Cham, 2016.

[22] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *TAC*, 2017.

[23] Melissa A Miller, Anya K Bershad, and Harriet de Wit. Drug effects on responses to emotional facial expressions: recent findings. *Behavioural pharmacology*, 26(6):571, 2015.

[24] Ye Ning and Terence Sim. Smile, youre on identity camera. In *ICPR*, 2008.

[25] Alice J. O'Toole, P. Jonathon Phillips, Fang Jiang, Janet Ayyad, Nils Penard, and Herve Abdi. Face recognition algorithms surpass humans matching faces over changes in illumination. *PAMI*, 29(9), 2007.

[26] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[27] Sima Soltanpour, Boubakeur Boufama, and QM Jonathan Wu. A survey of local feature methods for 3d face recognition. *Pattern Recognition*, 72:391–406, 2017.

[28] Dustyn James Tubbs and Khandaker Abir Rahman. Facial expression analysis as a means for additional biometric security in recognition systems. In *International Conference on Multimedia Communications, Services and Security (MCSS)*, 2015.

[29] Sergey Tulyakov, Thomas Slowe, Zhi Zhang, and Venu Govindaraju. Facial expression biometrics using tracker displacement features. In *CVPR*, 2007.

[30] J.A. Unar, Woo Chaw Seng, and Almas Abbasi. A review of biometric technology along with trends and prospects. *Pattern Recognition*, 47(8):2673–2688, 2014.

[31] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, June 2011.

[32] Stefanos Zafeiriou and Maja Pantic. Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011.

[33] Hailing Zhou, Ajmal Mian, Lei Wei, Doug Creighton, Mo Hossny, and Saeid Nahavandi. Recent advances on singlemodal and multimodal face recognition: a survey. *IEEE Transactions on Human-Machine Systems*, 44(6):701–716, 2014.