Towards recognizing facial expressions at deeper level: Discriminating genuine and fake smiles from a sequence of images

Minh Tu Nguyen[1], Quoc Khanh Nguyen[1], Kotani Kazunori[2], Prarinya Siritanawan[2]
[1]Le Quy Don Technical University, Hanoi
Email: tunguyenhsgs@gmail.com, khanh29bk@mta.edu.vn
[2]Japan Advanced Institute of Science and Technology
Email: {ikko,prarinya}@jaist.ac.jp

Abstract – **Understanding human's emotions is an important task and has application in a variety of fields. Because of that, facial emotion recognition or facial expressions recognition (FER) has gained many attentions of researchers, with different methods proposed, by using multiple sensors, using applying vision approaches from conventional FER to a deep-learning-based system. Although those methods have succeeded in recognizing facial expressions by analyzing the image or combining sequences of frames then concatenate with the audio extracted from a video, however, recognizing real emotion at a deeper level is still a challenge. We can detect a person is smiling, yet to say whether that smile is spontaneous or frustrated is difficult even for us, human. This paper focuses on the study of existing FER methods in discriminating real from fake smiles to get closer to detect deep emotion of a person from a given video. By the end of the paper, we conduct experiments of several models, the best of which uses bidirectional LSTM with attention mechanism on a combination of representations of a face image, gives 98% accuracy on MAHNOB database. The model was tested on SPOS and MMI and gave 87% and 97% accuracy respectively.**

Keywords – face analysis, smile classification, deep learning, facial emotion recognition.

I. INTRODUCTION

Facial expression is a genuine way to express human emotions. However, people sometimes do not act out as how they truly feel, especially when it comes to their smiling, since it is the most frequently shown expression, and usually expressed to signal many states of emotions (enjoying, embarrassed,…), also it is the easiest emotional facial expression to pose voluntarily. In recent years, many studies have been conducted to predict human emotions through facial expressions. However, most of them focused on identifying general emotions, which can be easily recognized at first glance instead of understanding of human emotions at a deeper level. Oddly, despite the importance and the potential of the topic, only a small amount of researches has been carried out to discriminate genuine smiles versus fake ones.

Recognizing spontaneous versus fake expressions is one of the hardest jobs for a human's brain. The way to perceive the true feelings of others is mainly by empathy, but it can still cause countless misunderstandings. Machines, however, do not have much empathy as a normal person (or at least not yet), so in order for computers to handle such a difficult task based on a computer vision, they need to be given rules to learn. Those tasks to recognize human's emotions (including distinguish genuine and fake smiles) is a hot topic in the field of psychological and brain & cognition. Hence, several psychological and neuroscience researches have been conducted to make things more representational and less abstract, based on physical movements of facial muscles, computer vision scientists thence can implement those specific rules to train machines. For example, many proposed methods have taken advantages of *Ekman*'s work [1], to classify human's emotion by applying different techniques to guide machines such as learning from geometric features [2][3] or appearance features [4][5][6][7], whereas the features are handcrafted. In addition to conventional methods, with the rapid development of deep learning, many deep learning-based approaches have been published, most of which use CNN to extract features to feed the classifier, or another RNN/LSTM network as a member of a sequence to detect the facial expressions at the basic level. A further study about smiles by *Ekman* identified 18 different types of smiling by visualizing some specific differences on the face and accompanying action units [8]. Another research about smile [9] indicates a smile of joy called *Duchenne smile* would include several muscles that affect the cheek or the eyes aperture,… This activation is called the *Duchenne marker*. In the fields of computer vision and pattern recognition, there are several studies that use this indicator to detect fake smiles, as in [10] using a CNN to explore patterns between frustrated and delighted smiles. However, it has recently been found that these muscles can be active or inactive under both genuine and posed expressions with comparable frequencies [14]. Moreover, these works consider only one static image for classification or analysis, whereas [15] conducted by *Namba et al.* has shown that even though the composition of the Action Units (AUs) (in one frame) indicates a genuine smile, the expression order and timing when considering the whole sequence could be different between spontaneous smiles and fake ones. Also, several characteristics of real and fake smiles, such as symmetry, speed, and timing are examined in [16]. Not only the order and timing, but we can infer a lot from the facial regions. In [19], *Dibeklio˘glu et al.* have pointed out that the eyelid movements contain significantly useful information for telling spontaneous versus posed enjoyment smiles apart.

Since most of the existing approaches to recognize facial expressions or emotions based on the composition of AUs, the difference lies mostly on the context taken into consideration and the techniques used to extract features. This paper aims to review these two things: those methods used to extract features and the context in which these extracted features play. The structure of the paper is as follow:

- This paper first divides existed features extracting methods to two main categories: The conventional

978-1-7281-5163-2/19/$31.00 ©2019 IEEE

hand-crafted and deep learning methods to extract features *(Section II)*.

- Then we show a quick overview of approaches to consider temporal features to recognize facial expressions (with features extracted automatically using CNN or hand-crafted) *(Section III)*.
- Afterwards, we discuss the discriminating genuine versus posed smiles task, its problems and our method of conducting experiments *(Section IV)*.
- Finally, we present some experimental results of some models *(Section V)*.

II. EXTRACTING FEATURES METHODS

In the literature [20], *Byoung Chul Ko* has wrapped up approaches from conventional to recent advanced FER to compare and make a clear, detailed review, yet with the main purpose is to compare the accuracy and resource requirements. In this paper, we only want to have a quick look at the difference in calculating features between two approaches, from which to analyze which methods of calculating features might perform better.

A. Conventional FER approaches

Byoung has pointed in his work the common among these conventional methods is to detect faces and extract various types of features on the target face.

1) Geometric features

The geometric features are constructed based on the connection between facial parts. In this method, only geometrical information is taken into account while facial texture information is not considered. In [2], 77 facial landmarks are used to generate 13 high-level facial shape features, which then are normalized and feed into the classifier. *Ghimire and Lee* in [3] used two types of geometric features based on 52 facial landmarks to build the feature pool, one takes into account the tracking result of single facial landmark $(L = 52)$, while the other considers the tracking result of pairs of facial landmarks $(M = L * (L-1)/2 = 1326)$, which end up with the total of $L + M = 1378$ feature vectors, which are then fed to the AdaBoost feature selection to filter out a set of feature vectors that is adequate for recognizing facial expressions. This method of extracting features have the advantages of being fast, and follow the needs of the rules from *Facial Acting Code System*, which considers the movements of the face muscles, but it might pass such important texture features as the depth of the eye, the change of the pupils over time, etc., which will be clarified in *section IV*.

2) Appearance features

This kind of features is usually extracted from a face region, either it is global features [4] or region-specific features [5][6]. In [7], the entire face is divided into multiple regions, and an incremental search method is implemented to detect crucial regions.

3) Hybrid features

To overcome the weaknesses of the two previous methods, this approach is proposed by combining both geometric and appearance features [21].

B. Features extractor using deep learning

Convolutional Neural Network (CNN) has been a break-through in the field of computer vision recent decades. CNN is beneficial in terms of removing or reducing the dependence of pre-processing techniques and enable learning directly from raw images. The modern ConvNet was introduced in 1998 in a paper by *LeCun et al.* [22]. A CNN usually consists of two components: the extraction part followed by the classification part, which is made up from three types of heterogeneous layers: convolution and pooling layers for extraction part and a fully connected layer serves as a classifier on top of extracted features.

Most of the existing advanced deep-learning based FER approaches use CNN to extract features. The main difference is design of network or the number of layers in each network. *Jung et al.* [23] in their study deployed two types of convolutional neural networks: one aims for appearance features, one aims for geometry features by taking into consideration facial landmark points. One integration approach is then implemented to combine these two models to boost the performance of recognizing facial expressions. *Jingwei Yan et al.* used VGG-Face model to extract a 4096-dimensional facial feature map from a given image [24].

III. USING RECURRENT NEURAL NETWORKS TO RECOGNIZE FER FROM SEQUENCE OF IMAGES

As mentioned previously, temporal variations in the facial components throughout the video might indicate human's emotions differently compared to recognizing the emotions frame-by-frame. Methods that use only CNN will miss these features. Therefore, the use of RNN is introduced. The pipelines of methods using RNN is as follows: the features are extracted using one of the methods described in *section II*, then these features, instead of being fed directly to the classifier, will be passed into a recurrent neural network (RNN). The way to combine these two networks may differ, however. It might be merged as one network and be trained jointly or, the sequences of features extracted by a ConvNet would be passed to a separate RNN.

RNN is an efficient way to preserve information because at each time t, nodes with recurrent edges receive input from not only the current data point $x^{(t)}$ but also hidden node value in the network's previous state $h^{(t-1)}$. The output $\hat{y}^{(t)}$ at each time t is computed given the hidden node values $h^{(t)}$ at a corresponding time. The computation at each time step on the forward pass is specified in equations (1) and (2).

$$h^{(t)} = \sigma(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b_h) \quad (1)$$

$$\hat{y}^{(t)} = softmax(W^{yh}h^{(t)} + b_y) \quad (2)$$

where b_h and b_y are bias parameters; W^{hh} is the matrix of recurrent weights between the hidden layer (h) and itself (h) at next time step; W^{hx} is the matrix of conventional weights between the hidden layer (h) and the input (x).

The network is trained pretty much alike a feed forward network, which means also calculate the gradients to minimize the error loss. The difference is that RNN training procedure is executed through multiple time steps with backpropagation, called backpropagation through time (BBTT). One problem when training basic RNN is vanishing (and exploding) gradients, especially when it comes to long-term dependency, as stated in the study of *Bengio et al.* [27]. LSTM was proposed a solution to the vanishing gradient problem [28] and has actually proved its efficiency.

Many studies have used RNN or LSTM in the task of recognizing facial expressions, such as *Kahou et al.* proposed

using a recurrent network on features extracted by a convolutional network [25]. In this work the authors reasoned that a combination of CNN and RNN could outperform a previously CNN-only approach by analyzing temporal features. *Kim et al.* in their work designs a structure with similar scheme, whereas they feed the spatial features, which is also extracted using a CNN, into a LSTM model [29]. *Chu et al.* also followed the idea, stacked LSTMs on top of spatial image characteristics extracted by a CNN, the outputs of CNNs and LSTMs are then fused into a single network to create a prediction of 12 AUs for each frame [30].

Not only previous data can be used to infer the output at time t, but lateral data can be taken into account too, by using bidirectional RNN. A bidirectional RNN allows, at a point in time, to take information from both earlier and later in the sequence. The structure of bidirectional RNN is illustrated in *Figure 1*.

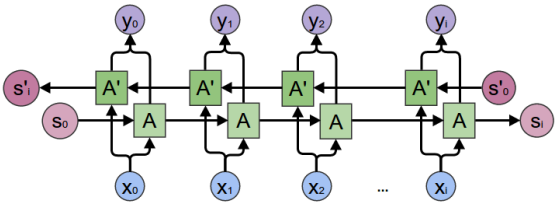

Fig 1. Structure of a bidirectional recurrent network

Graves et al. examined the usage of both bidirectional and original LSTM [31]. The authors came to a conclusion that bidirectional network gives notably better performance than the unidirectional. A research of *Jingwei et al.* also presented a joint convolutional bidirectional LSTM to recognize facial expressions, which includes concatenating appearance features and spatial dependency information in the last fully connected layer [24]. *Qianjin et al.* used an attention-based LSTM model along with a CNN, whereas the features generated by these two models are concatenated for classification [32]. Although this work studied on time-series data, but the idea of using attention mechanism can be inherited to apply to FER task as well. Attention-based LSTM allows the network to learn which temporal features are more important and play more critical roles in understanding the emotions throughout the whole sequence of frames. This approach might be especially useful when considering a long sequence of frames with the facial expressions vary from frames to frames (for example, one person express multiple expressions at different time throughout the video) to conclude the general emotions of the target. The experiments conducted in *section V* shows that using attention mechanism still gives a slightly better result in discriminating genuine and fake smiles when considering frames of smiling only.

RNN is not the only way to take temporal features into account. In [12], a 3D-CNN architect is used to recognize human action whereas the 3rd dimension is time. Similar usage might still be applicable in this scenario. However, one problem when trying to approach this method is that the database for this specific task (discriminating genuine and fake smiles) might not be enough, since using this method requires training the model from scratch. So far, there have not yet been surveys on comparing 3D-CNN with Convolutional RNN/LSTM models, so we cannot conclude one structure would outperform the other. However, because of the problem related to the data when apply to this specific problem, we do not conduct experiments with this approach, but we do consider this as an intelligent way to handle spatio-temporal features, once a compatible dataset is available.

IV. Discriminating Genuine vs Non-Genuine Smiles

In the concept of distinguishing genuine versus fake smiles in the fields of computer vision, there are not many literature conducted. In this section we discuss the problems that a facial expressions recognition approach would meet, and propose our method of experimenting.

A. Former work

Aida Gutiérrez-García and Calvo did a study to investigate the threshold levels to recognize fake and genuine smiles that considers expressive changes in the eyes region [33]. This study pointed out that the smiling mouth is the most salient part of smiling and non-smiling faces, which means the importance of the eyes region is often overshadowed by the smiling mouth [17]. Hence, in expression recognition task, the smiling mouth attracts the most awareness among all regions of emotional faces, although the distinctive features lie around the eyes regions. And since each facial expression is categorized in one specific category only, when facial features intersect across categories of expressions, the smile is often linked to happiness uniquely. As a consequence, the model might be biased towards judging the face as happy even though it is not a genuine smile.

Now, those features extraction methods considering only the positions of landmarks and their displacements through time might still work to some extent, since the genuine smile would mostly use specific face muscles [8][9], yet there have been new findings that raised suspicions about the reliability of the *D-marker* [13]. Thus, processing based on geometric features might result to many false positives, since it might miss the processing of the eyes expression, and as indicated in [33], changes in the eyes region (including the movements of muscles around the eyes and the depth insides, the change of pupils,…) are more distinctive.

With CNN-based approach to extract spatial features, the facial representations might be deeper and more selective. The network might be able to learn by itself the weight of each receptive field and therefore know to which part it should pay attention, to categorize correctly. However, the major problem with this approach is that a great amount of training examples for deep learning algorithms is required, whereas currently-available datasets of genuine versus non-genuine smiles are quite insufficient. In [10], *Kumar et al.* used a CNN to recognize fake smiles, yet their model is built mainly to recognize facial expressions, and they used some testing images to test the model for the purpose of discriminating spontaneous versus posed smiles, due to the lack of dataset. This might work to some extent, however, the model trained on dataset whereas smiley vs. non-smiley faces are considerably distinctive might focus on visually highly salient factor (such as smiling mouth) to indicate a face as happy despite the fact that it is not a genuinely smiling face. Moreover, this research considered spatial characteristics only. But as stated in [15], the timing, speed, and order of expressions might indicate a genuinely happy face or not.

B. Method

Very few researches conduct experiments on temporal features to recognize spontaneous smiles [52]. Similarly, there are barely experiments on using RNN in this specific task, distinguish genuine smiles versus fake ones. Therefore,

in this work, we apply some RNN models to acquire temporal features. Furthermore, we tested on 2 types of features:

- Pass the sequence of spatial features of cropped normalized faces through an RNN to collect a set of features. *(1)*
- The representations of the difference between 2 cropped normalized faces of the 2 consecutive frames are merged with the features set *(1)*. *(2)*

These 2 sets of spatial features are then passed through an RNN. We then compare results on these 2 types of features, and with the results when applying directly some classifiers such as SVM and DecisionTree on stacked spatial representations.

First, let's follow the basic procedure to recognize the target's emotion from a video: Detect a face; Track that face until it disappear from the scene; Normalize faces cropped from the frames; Handle spatial and temporal features to recognize emotions.

1) Detect a face

Techniques for face detection task varies. [18] gives a survey on the most successful face detection methods. Although many state-of-the-art face detection methods with almost perfect accuracy are available, they normally are computationally intensive and comparably slow. Since detecting or recognizing faces is not our main task, we need to come through two more networks, to extract features and then infer emotions from those features using a LSTM model, hence we would expect a balance between effectiveness and efficiency. We chose the cascaded-CNN [26] which can achieve high accuracy within a fast speed. The cascade-CNN consists of 6 ConvNets working in cascaded in 3 stages. In each stage, one ConvNet is used for detecting faces vs. non-faces and the other ConvNet is used for bounding box calibration. The output of one stage takes input as the detection window position which is adjusted using the output of the previous stage. More details of the model can be found in [26].

2) Track a face

Since we consider the emotions of the target in a video, which means in a sequence of frames but not one single static images, we would need to follow the face from the beginning till the time it disappears from the frame. Tracking faces also have multiple methods proposed, under different constraints such as short-term or long-term tracking. In recognizing emotions, particularly distinguishing genuine versus fake smiles, we do not need to track one face throughout the whole video (long-term), but rather just until the face disappears from frames. With such databases as SPOS [38] or MMI [36], which contain only one subject throughout a video, the result and performance are not different from other methods, but with AFEW database [34], where each video might contain more than one subject and there might be a sudden change in the scene, this short-term tracking might be a better approach.

Tracked faces through those frames then make a sequence to predict genuine or non-genuine smiles. RPT (Reliable Patch Tracker) method [11] is a good way to go for tracking faces, under the assumption that the target's motion between consecutive frames is limited. In explicitly, the tracking procedure is working as follows: The face is detected (for example in this work, using cascaded-CNN), the bounding box which indicates the target's position, therefore, is extracted. The bounding box is fed into the RPT-based tracking algorithm, which will stop until there is a significant difference in the distance of the target's position in two two-consecutive frames. The tracking then stop and the extracted faces through frames are fed to next stage.

3) Normalize faces

Here we propose using a set of landmarks to aligned faces as in recognizing faces task. Several landmarks set and calculation algorithm is proposed such as [2][3]. We propose using the 68-landmark set, which is applied widely and has multiple support libraries. The face cropped and normalized has the size of $128 * 128$.

4) Handling features

Here we tested on 2 strategies of acquiring features:

- We first pass the normalized face from each frame through a CNN to collect representations of each image to collect features set *(1)*.
- Differences between 2 normalized faces from 2 consecutive frames are calculated, passed through the same CNN used to collect features set *(1)*, these representations are then merged with features from *(1)*, to collect features set *(2)*. In this case, the representation of the first frame is ignored.

These spatial features are then either fed to an RNN or directly to a classifier. Details of experimentations are described in *section V*.

V. EXPERIMENTS

A. Data preparation

1) Dataset

Although there are not small numbers of databases for facial expression recognition task. However, these databases rarely contain spontaneous smiles. MAHNOB database [35], which contains 22 subjects recorded in 4 sessions, consists of total 563 instances of laughter, 51 acted laughter. Some other databases such as MMI [36] or Cohn-Kanade [37] are not specifically collected for recognizing spontaneous smiles, but we still report results of trained model on MAHNOB dataset to report further results. MMI includes 74 posed smiles videos. Cohn-Kanade contains 69 sessions of posed smiles only. SPOS [38], which is quite a standard database for distinguish spontaneous expressions, consists of 66 spontaneous and 14 posed smiles. Finally, the AFEW [34] dataset is used to evaluate the model with real-life data. Details of databases are described in *Table I*.

TABLE I. OVERVIEW OF DATABASES

Databases	Spontaneous	Fake	Resolution	Avg frames/ session
MAHNOB [35]	563	51	720 × 576 pixels @25Hz	52
SPOS [38]	66	14	640 × 480 pixels @25Hz	53
MMI [36]	0	74	720 × 576 pixels @25Hz	22
Cohn-Kanade [37]	0	69	640 × 490 pixels @25Hz	11
AFEW [34]	52	0	720 × 576 pixels @25Hz	57

2) Training data

We train models on MAHNOB dataset only. All other databases are only used to test the trained model for further report. Due to the data imbalance, we include also remaining speaking (non-laughing) frames in this corpus as acted laughter, which gives us 563 sequences of laughter frames and 514 sequences of acted-laughter frames, each of which consists of about 10-700 frames. More concretely, the distribution of the data is illustrated in *Figure 2*.

3) Data augmenting

We notice the length of each instance varies quite greatly. However, since in this specific task, changes between two consecutive frames might be significant. Therefore, we cannot augment data by clipping some middle frames out of the whole sequence. We, however, can determine which range of frames to extract to new data.

We also do some spatial augmentation throughout the video, but since we later will normalize these frames to calculate landmarks before feeding to a CNN to extract representations, we do not perform any rotation or flip.

We separate the data after augmenting to 80% used for training, and the rest 20% for testing. There is no overlap between two partitions.

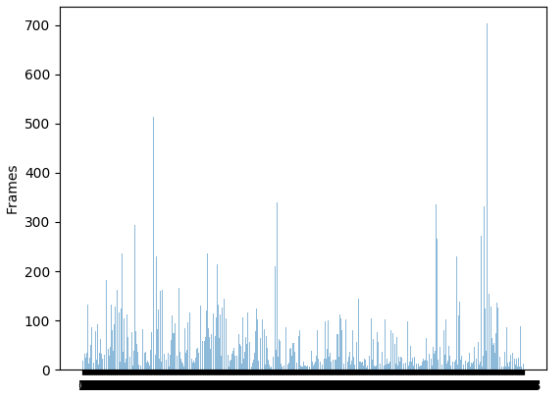

Fig 2. Length of each session in MAHNOB database

B. Results

We do not retrain the VGG-face (with last fully connected layers) to classify expressions, we apply the top-off model (i.e. the model with the last fc layers removed) directly to extract facial features (i.e. same features used for face recognizing task). *Table II* presents detailed results of different models on different dataset.

TABLE II. RESULTS ON DIFFERENT SPATIAL FEATURES SETS

VGG-face (1) means using VGG-face on cropped normalized faces to acquire features set (1). Similarly, VGG-face (2) means spatial features set used is featured set (2).

Model	Accuracy (%)					f1-score									
						MAHNOB		SPOS		MMI		Cohn-Kanade		AFEW	
	MAHNOB	SPOS	MMI	Cohn-Kanade	AFEW	0	1	0	1	0	1	0	1	0	1
VGG-face (1) + UniLSTM	94.62	65.00	34.21	30.43	53.26	0.94	0.95	0.32	0.76	0.51	-	0.47	-	0.48	0.57
VGG-face (2) + UniLSTM	97.18	69.43	35.07	34.29	52.07	0.96	0.95	0.34	0.80	0.51	-	0.47	-	0.47	0.55
VGG-face (1) + BiLSTM	95.00	53.12	96.05	95.65	53.12	0.96	0.97	0.23	0.66	0.98	-	0.98	-	0.71	0.24
VGG-face (2) + BiLSTM	96.05	54.08	96.83	95.07	55.20	0.96	0.97	0.23	0.67	0.98	-	0.98	-	0.73	0.28
VGG-face (1) + Attn-BiLSTM	96.77	82.50	97.12	97.10	58.70	0.96	0.97	0.22	0.90	0.99	-	0.99	-	0.68	0.61
VGG-face (2) + Attn-BiLSTM	**98.22**	**86.96**	97.37	98.06	62.27	**0.96**	**0.98**	0.22	**0.90**	0.99	-	0.99	-	0.70	**0.66**
VGG-face (1) + Attn-GRU	84.95	61.88	30.26	23.19	64.13	0.84	0.86	0.21	0.75	0.46	-	0.38	-	0.73	0.46
VGG-face (2) + Attn-GRU	85.12	61.35	32.73	24.52	**64.17**	0.84	0.87	0.21	0.75	0.47	-	0.38	-	**0.74**	0.46
VGG-face (1) + SVM	72.04	18.13	93.42	91.30	54.35	0.75	0.67	0.26	0.08	0.97	-	0.95	-	0.70	0.05
VGG-face (2) + SVM	72.11	18.72	93.52	92.63	53.62	0.75	0.67	0.26	0.08	0.97	-	0.96	-	0.70	0.05
VGG-face (1) + DecisionTree	95.70	55.00	**100.00**	98.55	58.67	0.95	0.96	0.29	0.67	1.00	-	0.99	-	0.64	0.51
VGG-face (2) + DecisionTree	95.94	55.90	**100.00**	**100.00**	54.67	0.95	0.96	0.29	0.67	1.00	-	0.99	-	0.64	0.51

The results show that the features set *(2)* gives better results in most cases. This features set usually gives worse results on AFEW dataset, which we think because AFEW dataset contains clips cut from movies, so there are a lot of sudden change scene in 2 consecutive frames.

After training, we used the test partition of the MAHNOB database with smiling mouth covered (*Figure 3*) to evaluate models. The results are shown in *Table III*. The accuracy might decrease significantly compare to original samples, but the accuracy is still higher than 80%, which means the model did not out leave out the importance of the more abstract features of the smiles.

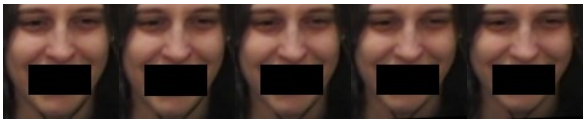

Fig 3. Some samples with smiling mouth covered

TABLE III. RESULTS ON SAMPLES WITH SMILING MOUTH COVERED

Model	Accuracy (%)	f1-score	
		0	1
VGG-face + BiLSTM	57.65	0.24	0.71
VGG-face (2) + Attn-BiLSTM	**79.85**	**0.31**	**0.81**
VGG-face + DecisionTree	67.43	0.25	0.79

VI. CONCLUSIONS

This paper presented a brief review of existing facial expressions recognition approaches and mention some problems when detecting human emotions at deeper level, more particularly, when distinguishing spontaneous versus posed smiles, and conducted an experiment on using RNN to extract temporal features in recognizing fake smiles.

VII. REFERENCES

[1] Ekman, P. and Friesen, W.V. (1978). "The Facial Action Coding System: A Technique for The Measurement of Facial Movement". Consulting Psychologists Press, San Francisco.

[2] Suk, M. and Prabhakaran, B. (2014). "Real-time mobile facial expression recognition system - A case study". Computer Vision and Pattern Recognition Workshops, Columbus, pp. 132–137.

[3] Ghimire, D. and Lee, J. (n.d). "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines".

[4] Happy, S.L., George, A. and Routray, A. (2012). "A real time facial expression classification system using local binary patterns". 4th International Conference on Intelligent Human Computer Interaction, Kharagpur, India, pp. 1–5.

[5] Siddiqi, M.H., Ali, R., Khan, A.M., Park, Y.T. and Lee, S. (2015). "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields". IEEE Trans. Image Proc.

[6] Khan, R.A., Meyer, A., Konik, H. and Bouakaz, S. (2013). "Framework for reliable, real-time facial expression recognition for low resolution images". Pattern Recognition. Lett.

[7] Ghimire, D., Jeong, S., Lee, J. and Park, S.H. (2017). "Facial expression recognition based on local region specific features and support vector machines". Multimed.

[8] Ekman, P. (1992). "Telling lies: Cues to deceit in the marketplace, politics, and marriage".

[9] Ekman, P. and Friesen, W.V. (1976). "Pictures of facial affect". Palo Alto, CA: Consulting Psychologists Press.

[10] Rajesh, K.G.A, Ravi, K.K and Gouta, S. (2017). "Discriminating Real from Fake Smile Using Convolution Neural Network". International Conference on Computational Intelligence in Data Science (ICCIDS).

[11] Li, Y., Zhu, J. and Hoi, S. C. (2015). "Reliable patch trackers robust visual tracking by exploiting reliable patches". IEEE Conference on Computer Vision and Pattern Recognition.

[12] Shuiwang, J., Wei, X., Ming, Y. and Kai, Y. (2010). "3D Convolutional Neural Networks for Human Action Recognition". 27th International Conference on Machine Learning.

[13] Schmidt, K.L., Bhattacharya, S. and Denlinger, R. (2009). "Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises". J. Nonverbal Behav.

[14] Krumhuber, E.G. and Manstead, A.S.R. (2009). "Can Duchenne smiles be feigned? New evidence on felt and false smiles".

[15] Shushi, N., Shoko, M., Russell S.K., Makoto, M. and Takashi, N. (2016). "Spontaneous Facial Expressions Are Different from Posed Facial Expressions: Morphological Properties and Dynamic Sequences". Springer Science+Business Media New York.

[16] Ekman, P., Hager, J.C. and Friesen, W.V., (1981). "The symmetry of emotional and deliberate facial actions". Psychophysiology 18.

[17] Calvo, M.G., Fernández-Martín, A. and Nummenmaa, L. (2013). "A smile biases the recognition of eye expressions: Configural projection from a salient mouth". The Quarterly Journal of Experimental Psychology.

[18] Zafeiriou, S., Zhang, C. and Zhang, Z. (2015). "A survey on face detection in the wild: past, present and future". Computer Vision and Image Understanding.

[19] Dibeklio˘glu, H., Valenti, R., Salah, A.A. and Gevers, T. (2010). "Eyes do not lie: Spontaneous versus posed smiles". ACM Multimedia.

[20] Byoung, C.K.. (2018). "A Brief Review of Facial Emotion Recognition Based on Visual Information".

[21] Benitez-Quiroz C.F., Srinivasan, R. and Martinez A.M. (2016). "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild". IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas.

[22] Yann, L., Patrick, H., Léon, B. and Yoshua, B. (1998). "Object recognition with Gradient-based learning".

[23] Jung, H., Lee, S., Yim, J., Park, S. and Kim, J. (2015). "Joint fine-tuning in deep neural networks for facial expression recognition". IEEE International Conference on Computer Vision, Santiago.

[24] Jingwe, Y., Wenmin, Z., Zhen, C. and Peng, S., (2018). "A Joint Convolutional Bidirectional LSTM Framework for Facial Expression Recognition".

[25] Kahou, S.E., Michalski, V., Konda, K. (2015). "Recurrent neural networks for emotion recognition in video". International Conference on Multimodal Interaction, Seattle.

[26] Li, H., Lin, Z., Shen, X., Brandt, J. and Hua, G. (2015). "A convolutional neural network cascade for face detection". IEEE Conference on Computer Vision and Pattern Recognition.

[27] Yoshua, B., Patrice, S. and Paolo, F. (1994). "Learning Long-Term Dependencies with Gradient Descent is Difficult". IEEE Transactions on Neural Networks.

[28] Sepp, H. and Jurgen, S. (1997). "Long Short-Term Memory".

[29] Kim, D.H., Baddar, W., Jang, J. and Ro, Y.M. (2017). "Multi-objective based Spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition". IEEE Trans. Affect. Comput.

[30] Chu, W.S., Torre, F.D. and Cohn, J.F. (2017). "Learning spatial and temporal cues for multi-label facial action unit detection". 12th IEEE International Conference on Automatic Face and Gesture Recognition, Washington, D.C.

[31] Graves, A., Mayer C., Wimmer M., Schmidhuber J. and Radig B. (2008). "Facial expression recognition with recurrent neural networks". International Workshop on Cognition for Technical Systems.

[32] Qianjin, D., Weixi, G., Lin, Z. and Shao-Lun, H. (2018). "Attention-based LSTM-CNNs For Time-series Classification".

[33] Aida G., and Manuel G.C. (2015). "Discrimination thresholds for smiles in genuine versus blended facial expressions".

[34] Abhinav, D., Roland, G., Simon, L. and Tom, G. (2011). "Acted Facial Expressions In The Wild Database".

[35] Soleymani, M., Lichtenauer, J., Pun, T. and Pantic, M. (2011). "A Multi-Modal Affective Database for Affect Recognition and Implicit Tagging", IEEE Transactions on Affective Computing, Special Issue: Naturalistic Affect Resources.

[36] Valstar, M.F. and Pantic, M. (2010). "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database", International Language Resourcesand Evaluation Conference.

[37] Kanade, T., Cohn, J. F. and Tian, Y. (2000). "Comprehensive database for facial expression analysis", Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition.

[38] Pfister, T., Xiaobai, L., Guoying, Z. and Pietikainen, M. (2011). "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework". IEEE International Conference on Computer Vision Workshops.