Individualized Fake Video Detection Using Facial and Head Dynamics

Zhexuan Li, Yaohan Ding, Yue Dai, Sanchayan Sarkar

March 5, 2021

Abstract

Recent work in deep learning has largely improved the reality of fake videos. However, only frame-level 2-D information is faked while temporal dynamics of the video remain untouched. Inspired by this, we will train individualized shallow and deep models to learn subject-specific dynamic features based on facial action unit intensities, mouth shapes and head rotations, and use these features to classify whether a new video segment is fake or real. Our results suggest that deep models outperform baseline SVM models on this task and the performance of models are related to the balance of fake and real data.

1 Introduction

Recent work in deep learning has largely improved the realism of fake content[1, 2, 3, 4]. Generally, state-of-the-art methods like deepfakes synthesize fake facial images and stack them into videos. In this way, only frame-level 2-D information in the video has manipulated while the dynamics information remain untouched, which means that the fake video of a subject is not expected to show similar dynamic patterns to the real videos. In addition, previous work [5] showed that facial and mouth movement dynamics is helpful in the task of subject identification, which implied that these features might be subject-specific. Based on these work, we propose to use facial, mouth and head movement dynamics as features to perform individualized deepfake video detection.

2 Literature Review

Researches on generative model brought significant improvement on deepfake techniques in past years. Inspired by the minimax two-player game, Generative Adversarial Nets architecture (GAN) [2] showed considerable potentials in image generating tasks. Based on GANs family, previous works also accomplished motion transferring [3] as well as face reenactment [1] between target and source persons. Additionally, recent GAN-based deepfake video generator [4] could transfer the faces between source and target actors as well. These advanced methods are able to generate considerably believable fake videos.

As the counter part, researches on detecting deepfake videos attracted growing attentions. One way of the works was tracing the manipulation tracks, [6] used a CNN based classifier to detect the wrapping artifacts features within the fake images and therefore recognizing authenticity. The approach was resource-saving and generalized yet might be potentially flaw to better forged fake examples with higher resolutions. On the other hand, another set of approaches were proposed to recognize the deepfake videos by using well engineered features. The POI-specific method in [7] distinguished fake videos by detecting unique facial action features from specific person. First they got frame based features of the videos abstracted by OpenFace2 [8] concatenated with position features. Then these frame features were compressed on temporal dimension by calculating correlations, hence converted to clip-level features. Then these clip-level features were fed into an SVM classifier for recognizing the authenticity of the video. The FaceCatcher [9] explored a fake video detecting system based on pair-wised biological signals extracted from several area of interests. The features were also fed into an SVM classifier working on the feature domains. These methods showed the potential of well-selected feature on recognizing authenticity. Additionally, there are some more integrated systems. [10] proposed a detecting method based on a convolutional LSTM structure, by combining a CNN frame feature abstractor and a following LSTM based classifier. The frame level features are embedded by the CNN and then fed to the LSTM classifier.

OpenFace2. Previous work [5] showed that facial and mouth movement dynamics was helpful in the task of subject identification, which implied that these features might be subject-specific. To represent facial features, we selected 16 Facial Action Units (AUs) used to describe the movements of specific facial muscles. To extract these features, we used OpenFace2 [8], a facial recognition application to provide an estimation of head and facial movements for each frame, including head poses, facial landmark locations and the intensities of AUs.

Statistical Recurrent Unit. The statistical recurrent unit (SRU) [11] is a recurrent neural unit. Instead of using gates or long-short term memory as GRU or LSTM, it only keeps running statistic average for sequential dependencies across sequence of inputs. We plan to use it as an alternative to LSTM as an extension.

In our proposed method in the following section, we want to leverage facial expression features to a temporal dynamics sensitive model, so that the model captures both effective frame-level facial expression features as well as temporal dynamic along the frame sequences. By adding temporal dynamics we expect a better performance than the previous works.

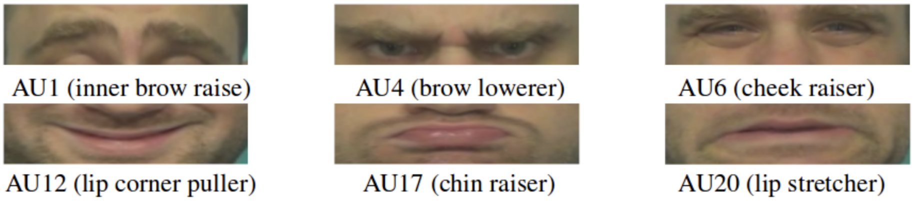

Figure 1: Example of 6 out of 16 Facial Action Units used in this project

3 Methods

3.1 Features

We extracted 20 frame level features using OpenFace2, namely facial action unit (AU) intensities, the head-rotation and mouth shape from 2D Videos. 16 AUs are employed: inner brow raiser (AU01), outer brow raiser (AU02), brow lowerer (AU04), upper lid raiser (AU05), cheek raiser (AU06), lid tightener (AU07), nose wrinkler (AU09), upper lip raiser (AU10), lip corner puller (AU12), dimpler (AU14), lip cor- ner depressor (AU15), chin raiser (AU17), lip stretcher (AU20), lip tightener (AU23), lip part (AU25), jaw drop (AU26). Example of AUs are shown in 2. For head rotation movements, we used head rotation along x-axis (pitch) and z-axis (roll). For mouth shape movements, we calculated the distance between lip corners as horizontal mouth shape and the distance between upper and lower lips as vertical mouth shape.

3.2 Models

Based on the previous discussion, there are two aspects of the problem : 1) Obtaining effective segment level feature representation for each videos and 2) Obtaining a good classification algorithm that works well for spatio-temporal classification. We implement the effective approach as baseline method from [7] and propose two improved approaches. The overall pipeline of the approaches is shown in the figure 2. All of the methods use frame level features extracted by OpenFace2. The baseline model first uses correlations to capture relationship between different frames within one clip, then classifies samples with a shallow classifier (SVM). And in our methods, to learn the pattern of temporal relationships directly from data, we use deep learning models instead: In method 1 we design an single SRU classifier and in method 2 we further explore a Siamese SRU encoder.

Baseline Model

For the baseline model, we implemented the approach proposed by [12]. To associate the frame-level features with each video clips, we used the Pearson Correlation Coefficients(PCC) to measure the collinearity between these features [7]. Each video will have C_{20}^{2} paired features, yielding 1 x 190 features for each video clip. These features are used in the person-specific two-class SVM to detect

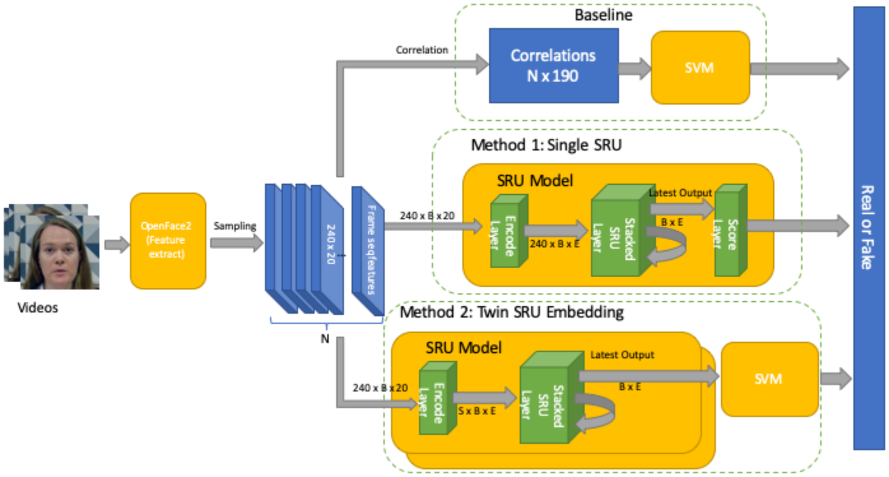

Figure 2: methods pipeline. Baseline part shows our correlation + SVM baseline model; method 1 shows our SRU classifier and method 2 shows our Siamese SRU + SVM approach.

fake videos.

SRU

We propose to use deep approaches to learn the temporal dynamics empirically. Inspired by the recurrent architecture utilized in [10], we want to employ a recurrent architecture. But convolution LSTMs are computationally expensive and require a huge amount of data. To avert that, we use the simpler recurrent architecture called the Statistical Recurrent Unit (SRU) [11]. As an un-gated architecture that keeps running moving average at multiple time-scales, it is a more efficient and computationally less expensive. Since we are interested in knowing the dynamics within the time serial features, maintaining a running statistics amongst the frame level representations might be more meaningful. The input to the architecture will be a sequence of frame level feature(i.e. $X = [x_1, x_2, x_3, ...x_T]$ vector of T time-steps and $\forall \mathrm{x}_i$ is a N length vector where $N = 20$). The inputs go through a linear layer then are fed into an SRU layer, and the final states of the SRU layer will be passed into the following dense and mirror linear layers for final classification. We developed individual SRU classifiers for each subject in our experiments.

Siamese SRU

As an extension to this proposed method, we want to propose a Twin-SRU architecture for the pre-training of our embeddings. The Siamese-SRU architecture is a shared weight architecture which will take <real,real> pairs and <real,fake> pairs as inputs and train a contrastive loss of euclidean distance which will try to maximize the similarity between the matching pairs and maximize the dissimilarity between the different pairs. The embedding learnt from this pre-training step will then be used as the input to train the SVM classifier. We believe this will give better results than just training the classifier with the features.

3.3 Evaluation

Considering the prediction of our method is either "real" or "fake", we define our task as a binary classification problem and the label fake as positive class. Therefore, we will evaluate our proposed methods in following aspects:

AUC-ROC. The Area Under Curve of Receiver Operating Characteristic curve is another commonly used evaluation matrix for binary classification tasks. The receiver operating characteristic curve is a graphical plot which reflect the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true positive rate, which is also called recall or sensitivity, indicates how much percentage of positive examples are correctly predicted (in our case, how much percentage of fake videos are correctly classified.) The false positive, which is also called false alarm as (1 - specificity), indicates how much percentage of negative examples are incorrectly predicted(in our case, how much percentage of real videos are incorrectly classified.) The ROC curve is the plot of FPR as X and TPR as Y under different threshold (usually the threshold is 0.5 considering the probabilities). And the Area Under Curve indicates the area under this ROC curve, the higher AUC, the better performance.

4 Data sets

All the data sets are downloaded from FaceForensics++ dataset [13]. The dataset contains 363 original sequences from 28 subjects in 16 different scenes as well as over 3000 manipulated videos on these subjects using DeepFakes. To generate fake videos, pairs of subjects were selected randomly and deep neural networks swapped the face of one subject onto the head of the other.

Similar to [7], original videos are partitioned into overlapping 10-second clips by sliding a window across the segment 5 frames at a time. Fake videos are sampled in a similar way with 2-second sliding windows. This yields 1000-2500 original clips and 270-2100 fake clips for each subject, with ratios of original-clip number to fake-clip number between 0.75 and 6.67.

5 Experiments and Results

5.1 Experiments

Feature Extraction
We processed all the videos using OpenFace2.0 Toolkit. This toolkit normalized the face across the video and then tracked head and facial movements. For head rolling and pitching pose and 16 AU intensities, we directly used the estimation output by the toolkit. For. mouth horizontally width and vertical openness, we calculated the distances between corresponding facial landmarks among the 68

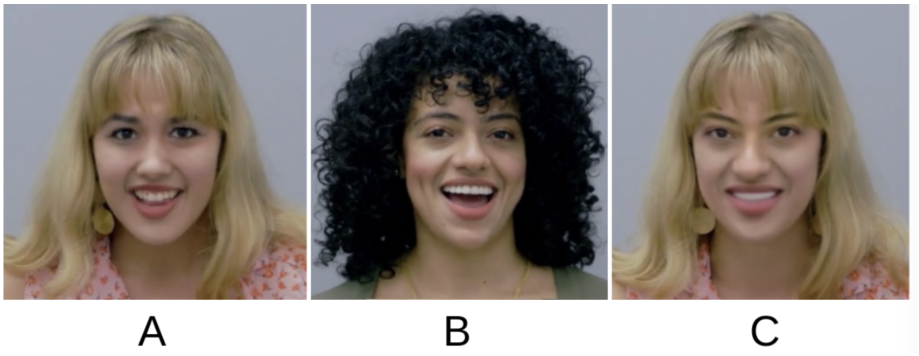

Figure 3: A: an original image of subject 16; B: an original image of subject 05; C: a fake image of subject 05 displaying features of the original actor (subject 16)

important facial landmarks tracked by the toolkit, which are the two lip-corner landmarks and two mid-lip landmarks.

Data Split

For each subject, we split all the video clips with a ratio of 80:20 for training and testing, with the guarantee that there is no overlapping in the training and testing video segments.

SVM (baseline)

For baseline model, we first calculated the correlation among 20 features for each video clip, resulting in a 1x190 feature matrix. Then, we implemented SVM with different kernels: Gaussian, linear and polynomial, among which linear kernels give the best performance after parameter optimization on C.

SRU classifier

We trained our SRU classifier based on shuffled $< Sequence, BinaryLabel >$ pairs. With identical training set as our baseline model, the SRU classifiers were trained using Adam Optimizer with weight decaying. The ROC-AUC curves were evaluated using logit outputs from final layer. The experiments were conducted for all subjects and models for different subjects were trained individually.

Siamese SRU + SVM

In this experiment, we needed to $< real, real >$ and $< real, fake >$ pairs. So we sampled 40000 such similar and dissimilar pairs. Each pair is a data point for our classifier and each data point is (N x 20). We used the Adam optimizer for training with a learning rate of 0.0001 and a weight decay of 0.0001. The network was trained using a contrastive loss of euclidean pairwise distance. The output of a model is also a 20-dimensional vector for each data-point. This is the spatio-temporal embedding for each of the slips which is used as the input to our SVM classifier.The SVM we used is a linear SVM and was trained with multiple regularization parameters and the results for the best parameter configuration are reported in this report. This experiment was conducted for all of the subjects.

Table 1: AUC-ROC scores of 28 subjects using 3 models: SVM (baseline), SRU (statistical recurrent unit and SSS (Siamese SRU + SVM)

ID	SVM	SRU	SSS
1	0.65	0.68	0.78
2	0.90	0.47	0.73
3	0.85	0.97	0.67
4	0.96	0.90	0.84
5	0.05	0.33	0.82
6	0.59	0.89	0.84
7	0.49	0.94	0.77
8	0.68	0.31	0.40
9	0.68	0.96	0.72
10	0.37	0.58	0.67
11	0.91	0.84	0.87
12	0.88	0.73	0.91
13	0.85	0.81	0.77
14	0.71	0.89	0.92
15	0.51	0.35	0.78
16	0.34	0.74	0.72
17	0.67	0.08	0.84
18	0.68	0.99	0.84
19	0.34	0.65	0.92
20	0.35	0.98	0.72
21	0.83	0.78	0.77
22	0.59	0.91	0.89
23	0.54	0.56	0.71
24	0.72	0.97	0.89
25	0.62	0.47	0.89
26	0.56	0.73	0.84
27	0.99	0.39	0.78
28	0.64	0.71	0.79
avg	0.64	0.70	0.79

5.2 Results

As shown in Table 1, deep models on average outperformed baseline SVM models, which suggested that deep models learning frame-level features can better distinguish fake and real videos, compared with SVM models learning only correlation among features. Specifically, Siamese SRU performed the best and stayed very consistent among most of the subjects. One of the reason that the Siamese SRU performs better is because it is trained on over a lot of data samples, considering it takes pairs of clips and not individual clips for training. The model learnt therefore learns a better spatio-temporal representation for each person. With better representations as inputs, the SVM converges better as compared to the Baseline. Additionally, the Siamese SRU has half the weights as that of the original SRU thereby further reducing the model complexity. Finally, the contrastive loss used for separating the real and the fake pairs is useful in giving more discriminative feature representation than the original SRU.

Additionally, we noticed that there are some subjects where SVM and SRU performed poorly. Looking into the data, we found out that the performance of SRU is linearly correlated with the ratio of original and fake video clip numbers ($p=0.007$). Using only subjects with this ratio between 0.75 and 2, the AUC of SRU method will increase to 0.80, while the AUC of SVM slightly increases to 0.68. This suggested that SRU models are sensitive to the balance of fake and real video numbers.

6 Conclusions and Future Work

With the assumption that every subject has unique facial and head movements patterns that can be recognized by computer models, we conclude that : 1. Deep models work better on individualized deepfake video detection by capturing the temporal dynamics of facial and head movement features. 2. Based on the linear relationship between original/fake video clip ratio and AUC, we expect better performance with more data of each subject.

One constraint in our work is that we have very limited data per subject, so that we're using overlapping clips. In the future, we need to collect more original and fake data of each subject and sampling more carefully for a balanced data. In addition, we will also look into other features and feature selection strategy that can better represent the temporal facial dynamics.

A possible future work for this project is to analyze the proportion of the factors that contribute our evaluation. Occluding every feature and repeating the experiment for all the features save the occluded one, will give us an insight on which of the features are being learnt by the classifier.

References

[1] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933–5942.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[5] N. A. Fox, R. Gross, J. F. Cohn, and R. B. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Transactions on multimedia*, vol. 9, no. 4, pp. 701–714, 2007.

[6] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, vol. 2, 2018.

[7] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.

[8] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[9] U. A. Ciftci and I. Demir, “Fakecatcher: Detection of synthetic portrait videos using biological signals,” *arXiv preprint arXiv:1901.02212*, 2019.

[10] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[11] J. B. Oliva, B. Póczos, and J. Schneider, “The statistical recurrent unit,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2671–2680.

[12] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, “Dynamic multimodal measurement of depression severity using deep autoencoding,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, March 2018.

[13] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” *CoRR*, vol. abs/1901.08971, 2019. [Online]. Available: http://arxiv.org/abs/1901.08971