Detecting combined identities: Generated using Swapping

With the increasing reliability on the biometric systems, the interest in testing the vulnerabilities if these algorithms is also at a rise. Circumventing the biometric recognition system using presentation attacks or by creating fake identities to hide one's own identity or to present someone else's identity is no more a surprise to people. Using AI, we can now generate highly realistic fake videos. And thus we can create digital doppelgangers of anyone by just analysing their images. This can be very dangerous as well, since there are already multiple images of a person available online and if this would really work perfectly, our privacies are at high risk. In one of the papers,[1] the neural network is trained on mouth shapes for different sounds. And with that, the authors are able to regenerate a lip-synced video using an audio and target videos as input. There are multiple softwares and techniques freely available online which makes it even easier to create combined identities which might be a face overlapped onto another or a composite image resembling both original images. Morphing enables criminals to hide/change their identities easily. can be done in multiple ways. Vulnerabilities also involve facial reenactment which include transfering facial expression from one face to another. This can be used to modify the human expressions or lip movement on a human face. Putting words into someone's mouth can be very dangerous
Some of the major database sources can be:

1. Face Swap Apps (Snapchat)
2. Synthesizing Obama: Lip sync from audio [1]
3. Reenactment [2]

On one hand, where morphing can help create digital models of a person, it can also be used to violate a person's image and identity or be used to manipulate people. Therefore, we need to develop an algorithm to check if the image is modified or not.

Related work:

[3] used Deep CNNs i.e. VGG-19 and AlexNet to detect digital and print-scanned morphed face images. [4] gives demorphing techniques to reduce the risk of the attack. It's observed that they don't model emotions, so in some cases, the face can be too serious for a casual speech or vice-versa. Such modifications in the video generally blurs the lip.

Dataset:

I created my own dataset consisting of 100+ videos of 5-6 secs each collected from 16 individuals. They're used as our training dataset to train our model on faceswap images, and test them on Lip-sync images(extracted from the images provided by [1])

My Approach:

VGG16 + SVM => This approach uses VGG16 to extract features from individual frame of the video, the sequence of frame features are then taken into SVM for classification.

I chose to fine-tune the original pre-trained model of VGG16 which is already trained on a large and a diverse Imagenet dataset consisting of 1.2m images. VGG16 captures features like edges and curves in starting layers which is useful for our problem.

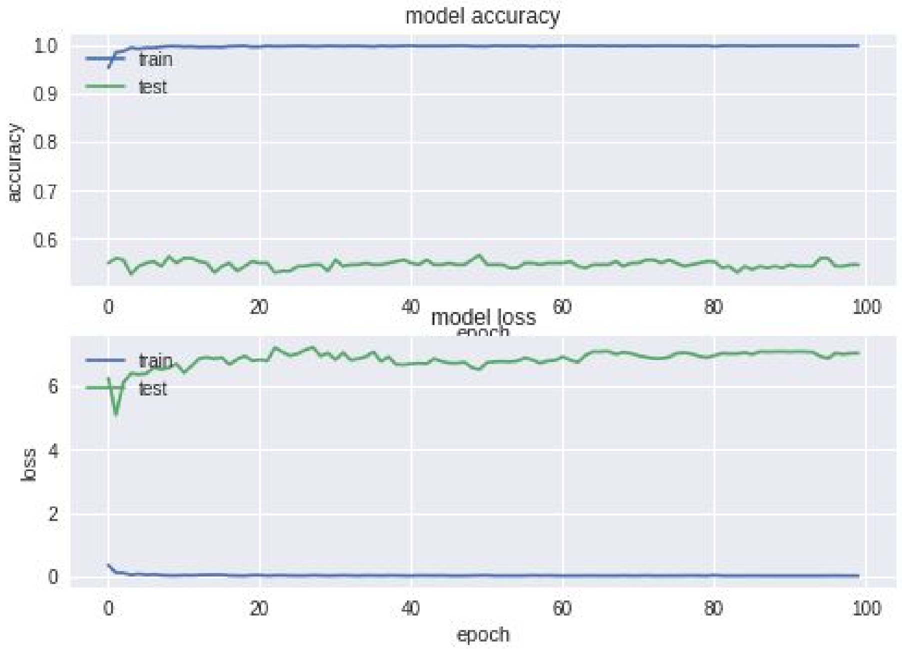
Now, since our dataset is small, it would have lead to overfitting since the last few layers are fully connected(FC) layers. One way to avoid overfitting is data augmentation like flipping, shifting, rotation, scaling, zoom, shear etc. and it yields us much better results. But since the dataset is small, we take the intermediate layer outputs and train it using Support Vector Machines since SVMs are very good for classification on small datasets.

I used keras, which is a simple neural net library built on the top of Tensorflow to help others prototype ideas quickly.

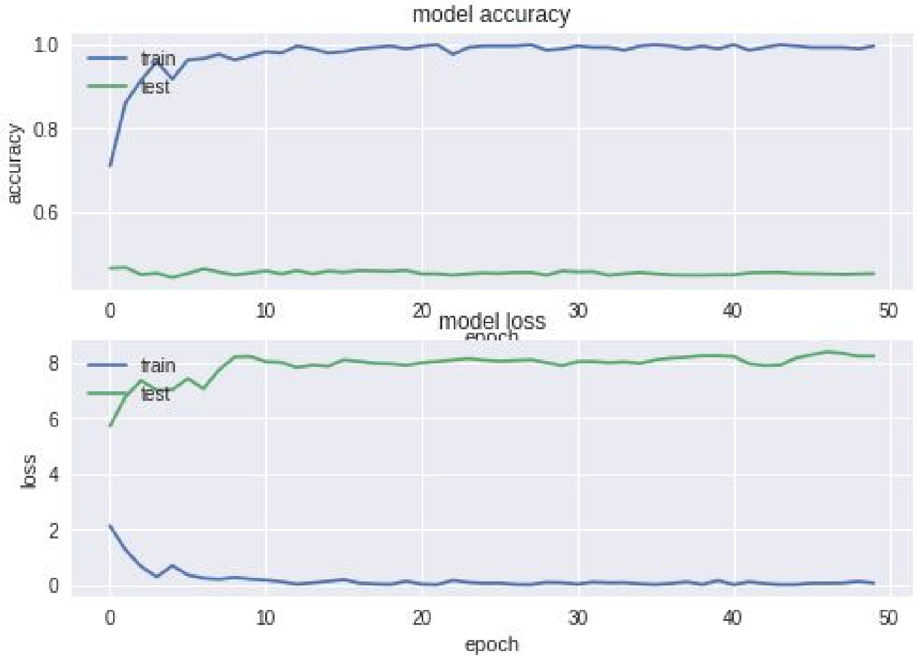
Results:

Dataset - 100 videos - 5000+ frames extracted for training purposes.
150 lipsync images clipped from youtube video for testing.
Architecture - VGG16 trained with SVM as the last layer.
I performed two experiments:

1. Trained on 2716 images of Face-swap dataset and Tested on 156 images extracted from lip-sync video.
   - Accuracy = 54.605%



2. Trained on 156 images extracted from lip-sync video and tested on 2716 images of Face-swap dataset
    - Accuracy = 45.30%



Conclusion

Just using faceswap data we can identify images which are lip-synced or reenacted. Accuracy is currently fairly low and is justified due to less amount of training and testing data which will be considered as the future scope of the project.

References:

[1] http://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf
[2] http://niessnerlab.org/papers/2016/1facetoface/thies2016face.pdf
[3]http://openaccess.thecvf.com/content_cvpr_2017_workshops/w28/papers/Busch_Transferable_Deep-CNN_Features_CVPR_2017_paper.pdf
[4]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8119561&tag=1

Link to repository:
https://github.com/aksh98/Advanced_Biometrics_Assignments/blob/master/BiometricsProject.ipynb