Deep Learning Approach to convert Facial Composites to Videos to Aid Forensic Police Work

Authors:

Mr Parth Rangarajan

(LY B.Tech Student at Vishwakarma University, Pune, Maharashtra, India - 411048)

Prof.

(Professor at Vishwakarma University, Pune, Maharashtra, India - 411048)

**

Abstract:

Facial composites are sketches of criminals made by professional artists to help catch criminals. The artist takes inputs from multiple eyewitnesses and then based on their perception of how the suspected criminal may look they go ahead and create a Facial Composite. The police then put out these posters around town and circulate them among informants to catch the wanted criminal. Most of the facial composites drawn are of the subject in a neutral mood without any expressions. Also, identification marks can be underplayed. This also requires a lot of manual work of printing the sketches in newspapers or posting them on social media. These posters can be taken down by notorious groups of people thus making this process of catching the criminal very futile.

Having created a facial composite, it can be very tricky to use the same sketch for identification because most criminals, after committing the crime, tend to change their appearance with a haircut or removal of a visible tattoo [1]. Thus having a video of a person from their sketch could be a very valuable resource in searching for fugitives. The approach that is to be explained tackles such a use case wherein a user can input the sketch of a perpetrator. Once the sketch is received, some computations later, the outputs should contain a video of the same miscreant. The other idea is to capture how the culprit may look while exhibiting contempt,anger,happiness,blowing-out-cheeks, eye-close, and eye-up moods.

Keywords:

CNN, Face Detection, Face Recognition, ML/ Machine Learning, Deep Learning, Generative Adversarial Networks.

Introduction:

Witnesses to a crime are asked to create an image of a person committing a crime. These pictures are called Facial Composites and police use them to catch the criminal in question. The first know Facial Composite used to catch a criminal was in the year 1971 trying to convict a criminal who hijacked an aeroplane[4]. Capturing criminals when starting out with very little information about them has been a difficult process and 'composite identification rates are often low'. Once the composites are made and the pool of criminals is brought up by the police, the facial composites often lead to misclassification of the criminal of up to '27% of eyewitness misidentifications involved facial composite sketches'[5].

Facial Composites do not leave much to the imagination. They are created by sketch artists who take inputs from several eyewitnesses. I contend that these composites are just snapshots of suspected perpetrators and what they looked like at that moment in time. Most criminals change their appearance immediately after committing the crime to reduce their chance of being caught [1] thereby rendering the entire process of creating and compiling the sketch quite futile. There are many software that have been created to solve this problem like theEvoFIT system [1] which will be discussed in the subsequent sections of the paper.

Once created the GIF/video of the Facial Composite can be shared digitally and so it further reduces the legwork involved. In this digital age of fast connectivity, a GIF/video will travel and reach people faster and wider.

People can learn important information by recognising others' facial expressions [11]. For instance, accurate detection of expression-based visual information makes it possible for humans to anticipate events, respond to them, and infer the emotions of others.

Thus having a video of a suspected criminal exhibiting a number of facial expressions [2] will help the eyewitnesses recognize the suspected criminal from a police lineup better.

Related Work:

The CUHK Face Sketch Database (CUFS) [3], used in my 'novel' approach to convert Facial Composites(sketches) to video, is a database that comprises 188 faces from the Chinese University of Hong Kong (CUHK) students. It has been created specifically for research on 'face sketch synthesis and face sketch recognition'.Most of the work based on this dataset is strictly pertaining to those aspects only. Some papers focus on the same domain as this paper- criminal investigations and forensics.

To name a few, Bulbule, Sampada and Sutaone[19] develop a method to identify an individual not based on their biometrics, but on Nodal Points received from their faces. They used Facial Recognition techniques that focus on identifying landmarks that make one's face unique. These nodal points or landmarks include the length of the nose, the width of lips, the distance between eyes, the shape of the jawline, the depth of the eyeball socket, etc.

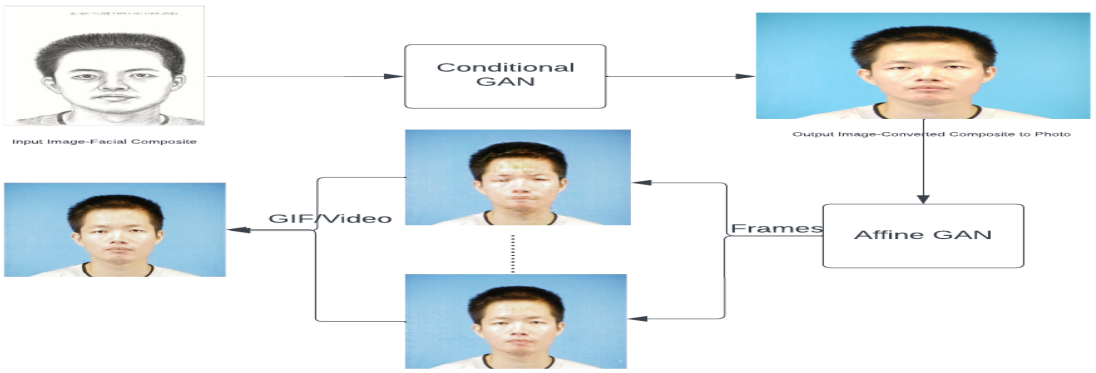

Figure 1. The pipeline of the proposed solution

Since there are numerous databases available for the whole face, but none specifically for the components of the face that can be utilised for a forensic application, they suggest developing a database for features like the nose, mouth, and eyes created using an algorithm called ViolaJones [19]. After the component database is created, both the databases of full face and component face are segregated into training and testing datasets. Each component from the component database is trained using A Convolutional Neural Network(CNN or ConvNet). The component of the face can now be used to identify a person using this trained network. The model is subsequently evaluated using several parameters.

Frowd, C. D., Pitchford, M., Bruce, V., McIntyre, A. H., & Hancock, P. J. B. (2010, November 9)[1] talk about Peter Hancock's 1990 invention, the EvoFIT system, which has been a software under development ever since. Principal Component Analysis(PCA) was used as the foundation for EvoFIT, a complete facial composite-creating system. By capturing variations in feature form and greyscale colouring, PCA makes it possible to synthesise additional faces, at first using random features. Users then provided a goodness of fit rating for each face and a Genetic Algorithm (GA) combines preferences to produce more items for selection. After repetitions the set progressively resembles each other and the target face. The best likeness produced was saved as the composite. Later, to address the issue that composites frequently resemble one another and that this lack of distinctiveness might make recognition challenging (for members of the public, etc.), an upgrade was introduced by enhancing face individuality.

Proposed Method:

This section presents the proposed method for the given use case better illustrated by Figure 1. The solution is divided into 2 broad frameworks, **Stage 1. Facial Composite to Photo** and **Stage 2. Photo to Video**. The video in Stage 2 is rather a 4-8 seconds gif of the suspected perpetrator in a multitude of moods. These include moods-like contempt, anger, happiness etc.

The moods will further help understand how the supposed culprit will look in real life. Before moving forward with the framework some key components that are used in the framework are Convolutional Neural Networks, Generative Adversarial Networks and the UNet Architecture.

CNNs are used as a basis for the building of this application. CNN employs the following procedure:

1. To select out specific dimensions, convolution applies filters.
2. Pooling aids in the extraction of important spatial patterns.
3. A fully connected layer flattens the last convolution or pooling layer and densely connects all of the flattened layer's nodes to all the nodes in the output layer.

A game-theoretic scenario in which a generator network must compete with an attacker(discriminator) serves as the foundation for Generative Adversarial Networks. Samples are generated directly by the generator network. The discriminator network, which is its rival, makes an effort to differentiate between samples taken from the training data and those taken from the generator.

UNet, which developed from the conventional convolutional neural network, was created and used for the first time in 2015 to process images used in biomedicine. In biomedical applications, it is necessary to identify both the presence of a disease and the location of the abnormality which UNet is devoted toward. It can localise and identify borders since every pixel is classified, ensuring that the input and output are of the same size.

Stage 1 Facial Composite to Photo: This section of the model is based on a Conditional Generative Adversarial Network based on [6] that inputs a 256x256 black-and-white sketch image and predicts the coloured version of the image without knowing the original black-and-white sketch. The generator model inputs a 256x256 black-and-white sketch image, and a conditional generative adversarial network predicts the coloured version of the image without knowing the original black-and-white sketch. For the generator model summary see Figure 2.

Layer (type)	Output Shape	Param #	Connected to
input_image (InputLayer)	[(None, 256, 256, 3)	0	
target_image (InputLayer)	[(None, 256, 256, 3)	0	
concatenate_15 (Concatenate)	(None, 256, 256, 6)	0	input_image[0][0] target_image[0][0]
sequential_32 (Sequential)	(None, 128, 128, 64)	6144	concatenate_15[0][0]
sequential_33 (Sequential)	(None, 64, 64, 128)	131584	sequential_32[0][0]
sequential_34 (Sequential)	(None, 32, 32, 256)	525312	sequential_33[0][0]
zero_padding2d (ZeroPadding2D)	(None, 34, 34, 256)	0	sequential_34[0][0]
conv2d_20 (Conv2D)	(None, 31, 31, 512)	2097152	zero_padding2d[0][0]
batch_normalization_32 (BatchNo	(None, 31, 31, 512)	2048	conv2d_20[0][0]
leaky_re_lu_20 (LeakyReLU)	(None, 31, 31, 512)	0	batch_normalization_32[0][0]
zero_padding2d_1 (ZeroPadding2D	(None, 33, 33, 512)	0	leaky_re_lu_20[0][0]
conv2d_21 (Conv2D)	(None, 30, 30, 1)	8193	zero_padding2d_1[0][0]

Total params: 2,770,433
Trainable params: 2,768,641
Non-trainable params: 1,792

Figure 3. Model summary for discriminator model

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None, 256, 256, 3)	0	
sequential_17 (Sequential)	(None, 128, 128, 64)	3072	input_2[0][0]
sequential_18 (Sequential)	(None, 64, 64, 128)	131584	sequential_17[1][0]
sequential_19 (Sequential)	(None, 32, 32, 256)	525312	sequential_18[1][0]
sequential_20 (Sequential)	(None, 16, 16, 512)	2099200	sequential_19[1][0]
sequential_21 (Sequential)	(None, 8, 8, 512)	4196352	sequential_20[1][0]
sequential_22 (Sequential)	(None, 4, 4, 512)	4196352	sequential_21[1][0]
sequential_23 (Sequential)	(None, 2, 2, 512)	4196352	sequential_22[1][0]
sequential_24 (Sequential)	(None, 1, 1, 512)	4196352	sequential_23[1][0]
sequential_25 (Sequential)	(None, 2, 2, 512)	4196352	sequential_24[1][0]
concatenate_7 (Concatenate)	(None, 2, 2, 1024)	0	sequential_25[1][0] sequential_23[1][0]
sequential_26 (Sequential)	(None, 4, 4, 512)	8390656	concatenate_7[0][0]
concatenate_8 (Concatenate)	(None, 4, 4, 1024)	0	sequential_26[1][0] sequential_22[1][0]
sequential_27 (Sequential)	(None, 8, 8, 512)	8390656	concatenate_8[0][0]
concatenate_9 (Concatenate)	(None, 8, 8, 1024)	0	sequential_27[1][0] sequential_21[1][0]
sequential_28 (Sequential)	(None, 16, 16, 512)	8390656	concatenate_9[0][0]
concatenate_10 (Concatenate)	(None, 16, 16, 1024)	0	sequential_28[1][0] sequential_20[1][0]
sequential_29 (Sequential)	(None, 32, 32, 256)	4195328	concatenate_10[0][0]
concatenate_11 (Concatenate)	(None, 32, 32, 512)	0	sequential_29[1][0] sequential_19[1][0]
sequential_30 (Sequential)	(None, 64, 64, 128)	1049088	concatenate_11[0][0]
concatenate_12 (Concatenate)	(None, 64, 64, 256)	0	sequential_30[1][0] sequential_18[1][0]
sequential_31 (Sequential)	(None, 128, 128, 64)	262400	concatenate_12[0][0]
concatenate_13 (Concatenate)	(None, 128, 128, 128	0	sequential_31[1][0] sequential_17[1][0]
conv2d_transpose_16 (Conv2DTran	(None, 256, 256, 3)	6147	concatenate_13[0][0]

Total params: 54,425,859
Trainable params: 54,414,979
Non-trainable params: 10,880

Figure 2. Model summary for generator model

The main goal of the discriminator model is finding out which image comes from the actual training dataset and which is a generator model output. In essence, it is 2 models fighting to prove each other wrong. The generator model tries to fool the discriminator model into thinking its output is actually from the dataset while the discriminator classifies those that it thinks are not from the dataset and those which are. When compared to the Generator model, the architecture is less complex. There are sequential levels in order, each followed by a padding layer. Then a Conv2d layer is used to create a convolution kernel, which is combined with the other input layers to produce a tensor. The 2d tensor is then subjected to Batch Normalisation to standardise the inputs to a layer for each mini-batch. Some outputs can now go beyond zero so the model will not be able to catch these. Therefore Leaky ReLU will modify the model to allow small negative values when the input is less than zero. A Conv1d is applied to get the output tensor. See Figure 3 for the entire discriminator model summary.

Two distinct loss functions will be utilised for two models to independently calculate each model's loss. Finding the sigmoid cross-entropy loss (1) of the generator output and an array of ones is used to compute the generator loss. This means that we are training it to deceive the discriminator into producing an actual image, or a value of 1.

$$CE = -\sum_{i=1}^{C'=2} t_i log(f(s_i)) = -t_1 log(f(s_1)) - (1-t_1) log(1-f(s_1)) \quad (1)$$

We also include L1 loss (2) to account for the output's structural similarity to the target image. The authors of the original research advise[6] keeping LAMBDA's value at 100.

$$L1LossFunction = \sum_{i=1}^{n} |y_{true} - y_{predicted}| \quad (2)$$

Discriminator loss is the same sigmoid cross-entropy loss (1) of the real images and an array of ones and adds it with the cross-entropy loss of the output images of the generator model and array of zeros. From equation (3), the following parameters are used:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\left[\frac{\delta L}{\delta w_t}\right] v_t = \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\delta L}{\delta w_t}\right]^2 \quad (3)$$

1.ϵ - To prevent the 'division by 0' mistake when (vt -> 0), = a small +ve constant. (10-8).

2. β1 & β2- The decay rates.

3.α- Learning rate and step size parameter (0.001).

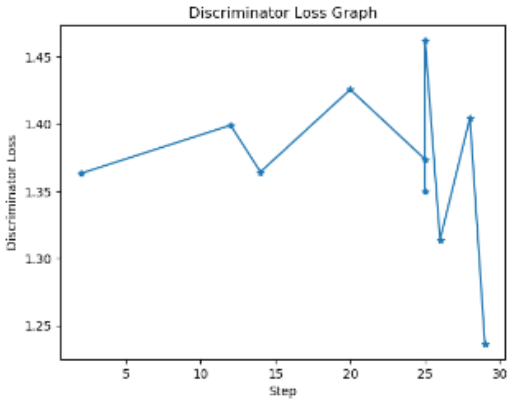

Figure 4. Discriminator Loss for 30 Epochs.

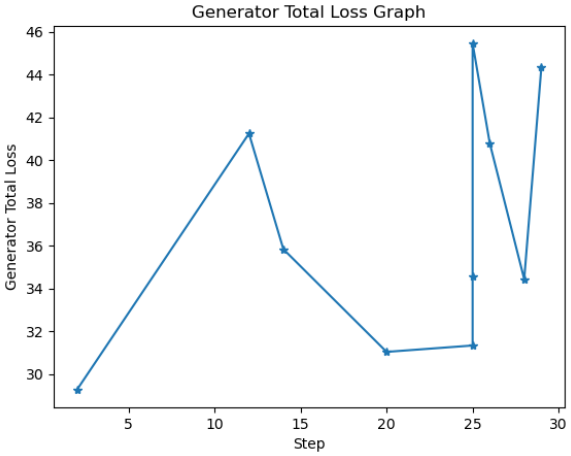

Figure 5. Generator Loss for 30 Epochs.

Stage 2. Photo to Video: This section deals with converting the colourized Facial Composite to a short video or a GIF. AffineGAN [2] is a completely novel approach full intensity-based procedure of an expression change from the neutral state to the peak that utilises an affine transformation in the latent space and assigns each facial image an expression intensity. In order to represent the relative intensity of the current expression in the complete neutral-to-peak operation, they assign each video frame a non-negative scalar. AffineGAN [2] generates a series of video frames with a growing succession of non-negative expression intensities based on an input image. For any training frame: there exists a latent space in which the codes of frames take the affine form,

$$ft = f0 + atf\Delta \quad (4)$$

,

Where ft and at are the latent code and expression intensity of the t-th frame, $f0$ is the latent code of a neutral frame, and $f\Delta$ encodes the direction to move from the neutral state to the current expression. All of them are learnt with the mere annotation of a neutral face frame per training video. The key is to relate the unknown expression intensity with the codes of the training frames, by which deriving it is possible based on the aforementioned affine transformation.

Figure 6. The results of Stage 1-Sketch to Image model trained on 188 images

The objective is to create a series of video frames.: $V := \{I_t\}_{t=0}^{T}$, which inputs a neutral face image $I_0 \in \mathbb{R}^{H \times W \times 3}$, where the image's height and width are H and W respectively. This sequence shows a change in expression starting from a neutral face. Each frame It is modelled at time t as a function of the input face $I0$ and an expression intensity $a_t \geq 0$, namely, $I_t = g(I_0, a_t)$. 'The larger at is, the further the generated expression It is from the neutral image $I0$ and the closer it is to the peak state of the expression (e.g., laugh loudly)'. When $at \approx 1$ the output frame g($I0, at$) achieves the peak expression state.

The entire mathematics behind the AffineGAN model and the model's architecture can be read in this paper [2].

Figure 7. The results of Stage 2-Image to GIF/Video

Putting together a Conditional Generative Adversarial Network based model and AffineGAN we are able to imagine how a criminal's facial composite will look in a video format thus bringing the sketch to life.

However, Stage 1 is unable to yield results that are promising as we can see in Figure 6.

This is due to the fact that there are not many images for the model to learn from which gives rise to two problems mainly.

A total of 198 Images were used from the CUHK Face Sketch Database (CUFS)[3] on which two models were built. Keeping a 70/30 split on the data, one model was trained on 130 images (58 for testing) and the other model was trained on 188 images and 10 were kept aside for testing. Thus little to no similarity between the ground-truth images and predicted images. With the addition of more images, there will definitely be an uptick in the performance of the model.

Another issue that needs tackling is the fact that all the images below to people from a single race i.e. those belonging to the Asian community. Thus the model and the architecture will perform well on perpetrators from the same community so there is a complete bias in the network. This will again be removed with the addition of more data and more diverse data for representing people from all over the world.

Figure 7 is the output of applying AffineGAN on the ground truth photo and not a predicted photo.

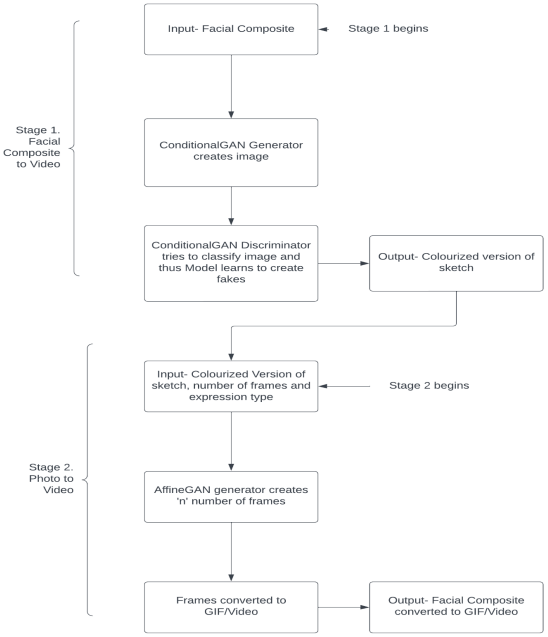

Figure 8. Complete flowchart of the use-case specified in the research paper

Conclusion:

Although my approach to creating a video from a sketch is completely focused on Police Composites and helping in aid of Police Work, it can be applied to a multitude of use cases.
The main idea was to make sure that the Facial Composites are being put to better use due to their low success rates in catching criminals. It can also be used as strong evidence against the perpetrator in trying to convict them.

References:

1. Frowd, C. D., Pitchford, M., Bruce, V., McIntyre, A. H., & Hancock, P. J. B. (2010, November 9). *Giving crime the 'evo': catching criminals using EvoFIT facial composites*. UCLan. Retrieved April 26, 2023, from https://evofit.co.uk/ .

2. Shen, G., Huang, W., Gan, C., Tan, M., Huang, J., Zhu, W., & Gong, B. (2019). Facial Image-to-Video Translation by a Hidden Affine Transformation. ACM Multimedia. https://doi.org/10.1145/3343031.3350981.

3. X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 31, 2009.

4. Facial composite. (2023, April 13). In *Wikipedia*. https://en.wikipedia.org/wiki/Facial_composite#:~:text=Facial%20composites%20are%20used%20mainly,ancient%20mummies%20or%20human%20remains.

5. Bulbule, Sampada & Sutaone, Mukul & Vyas, Vibha. (2019). Component-Based Face Recognition using CNN for Forensic Application. 1-7. 10.1109/ICCCNT45670.2019.8944841.

6. Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2016). Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv*. /abs/1611.07004.

7. Morkar, T. (2022, April 8). Learning to Build a Model for Sketch-to-Color Image Generation using Conditional GANs | Towards Data Science. Medium. https://towardsdatascience.com/generative-adversarial-networks-gans-89ef35a60b69

8. Leios Labs. (2021, March 2). What are affine transformations? [Video]. YouTube. https://www.youtube.com/watch?v=E3Phj6J287o

9. pix2pix: Image-to-image translation with a conditional GAN. (n.d.). TensorFlow. https://www.tensorflow.org/tutorials/generative/pix2pix.

10. Guarnera, M., Hichy, Z., Cascio, M. I., & Carrubba, S. (2015). Facial Expressions and Ability to Recognize Emotions From Eyes or Mouth in Children. *Europe's Journal of Psychology*, *11*(2), 183-196. https://doi.org/10.5964/ejop.v11i2.890