Holistic Siamese Model Optimized for Aged Face-Sketch Similarity Detection

Ganesh Shukla
dept. Electronics and Telecommunication
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
ganeshukla79@gmail.com

Bhargav Desai
dept. Electronics and Telecommunication
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
desaibhargav98@gmail.com

Parth Mehta
dept. Electronics and Telecommunication
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
parthmehta15@gmail.com

Sunil Karamchandani
dept. Electronics and Telecommunication
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
skaramchandani@rediffmail.com

***Abstract*—The proposed work presents an Holistic approach for Face-sketch matching using a Siamese scenario. Proposed scheme is simple yet effective in overcoming the computational complexity of the feature based scheme and also provides a F1 score of 0.73 and accuracy of nearly 74%. The results are in demanding circumstances of using ageing effects in sketches. Optimum model from Siamese linkages is acquired with augmented sketches on CUHK dataset. The sketches are trained under augmented versions of zooming, distortion and cropped versions of them. The validation curve with merely L1 regularization peaks at 73% while it finds one at 77% with a combined L1 and L2 regularization for an average batch size of eight. The efficacy of the regularization in conjunction with augmentation prevents overfitting and provides effective results in ageing sketches.**

Keywords—Siamese network, holistic face-sketch recognition, convolution layer, augmented aged sketches

I. INTRODUCTION

It comes as no surprise that face recognition has been amongst the most thoroughly investigated problems within the computer vision fraternity. Face recognition, as an application, has achieved widespread acceptance worldwide in areas like security and attendance. The approaches used for face recognition today can be broadly categorized into two [1] Feature Based and Holistic Based.

Feature based approaches make use of spatial and statistical features like SIFT to match faces [2]. The usage of SIFT (Scale-Invariant Feature Transform stems from its relative invariance to rotation. So for example, if a candidate image (image to be recognized) is tilted by some degree, the image is still correctly matched. This would make a face recognition system robust, however, SIFT has been shown to fail for rotations exceeding 10-12 degrees [3]. Other feature based approaches involve LBP (Local Binary Pattern), LRBP (Local Radon Binary Pattern), ILBP (Improved LBP), MLBP (Modi- fied LBP), MDS (Multidimensional Scaling), facial landmarks or a combination of the above [4] [5] [6] [7] [8] [9] [10] [11]. On the other hand, holistic face recognition, has achieved far less scrutiny even though a large body of research supports the hypothesis that the human visual system does not process a face as a collection of separable facial features but as an integrated perceptual whole [12]. It naturally follows, that the goal of a holistic face recognition system is to utilize global information from faces to perform face recognition. The global information from faces is fundamentally represented by a small number of features, which are directly derived from the pixel information of face images. These small numbers of features distinctly capture the variance among different individual faces and therefore are used to uniquely identify individuals [13].

For example, let us consider the problem of matching a forensic sketch to a database of mugshot images addressed in this work. This is an important first step for apprehending criminals, where we propose that obtaining a match between a computerized or hand-drawn sketches with the digital face images already present in the criminal database can be looked at as a holistic face recognition problem since we do not have access to a complete face but only an outline sketch resembling the face as shown in Fig. 1.

Fig. 1. Criminal Faces against their Forensic Sketches [6] [7]

What makes forensic sketch matching a holistic task even more so, is the fact that oftentimes the ground truth image in the database of a criminal becomes outdated due to ageing. For example such a scenario can arise if the said criminal is caught, released and after a few quiet years or decades, he again engages in culpable activities. In such a scenario, local features taken from the sketch face would not match the local features taken from the database image and consequently, a feature based approach would inevitably perform poorly. This calls for a method to process global face information such as structure and other facial characteristics to find a befitting match. In the view of this, we use a Siamese Network.

II. PROPOSED SYSTEM DESIGN

The system utilizes a Siamese Network to correlate the face sketches with their corresponding images. The block diagram of a general Siamese Network is shown in Fig. 2

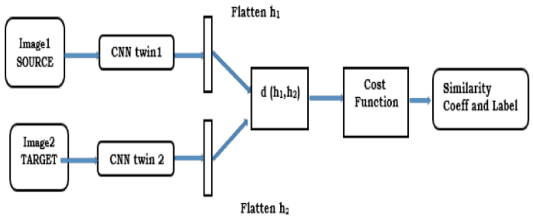

Fig. 2. General Siamese Linkage

Taking general framework of a Siamese Network, our proposed system architecture for forensic sketch matching is illustrated in Fig. 3.

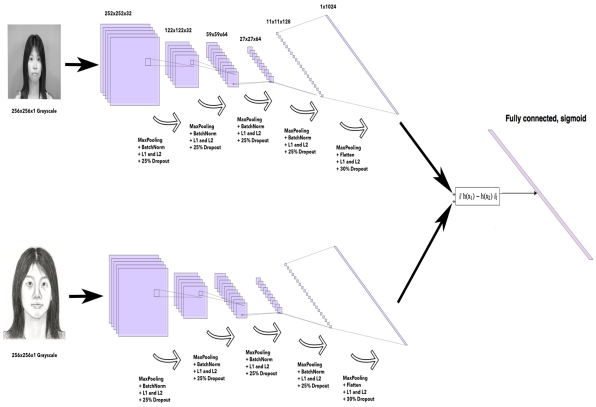

Fig. 3. Proposed Siamese Architecture for Face-Sketch Recognition

Both the CNNs (Convolutional Neural Networks) are completely symmetrical in every manner, the architecture of which with tensor output shapes at every level of hierarchy. The kernel size (or filter size), regularization and activation for each layer is tabulated in Table I.

TABLE I. CNN ARCHITECTURE OF SIAMESE NETWORKTABLE

Layer	Kernel (Filter) Size	Regularization	Activation
Conv2D_1	(5,5)	L1 = 0.008 or 0.01, L2 = 0.008 or 0.00	Rectified Linear Unit
Conv2D_2	(5,5)		
Conv2D_3	(3,3)		
Conv2D_4	(3,3)		
Conv2D_5	(3,3)		
Dense_1	Not applicable	L1 = 0.008 or 0.01, L2 = 0.008 or 0.00	Rectified Linear Unit
Output	Not applicable	L1 = 0.008 or 0.01, L2 = 0.005 or 0.00	Sigmoid

Fan et al [14] have used Siamese network but with increased computational complexity as their network is basically trained with Histogram of Averaged Oriented Gradients (HAOG) features. For the augmentation the researchers have used the CUHK database with just rotation at different angles, while the proposed work is a more realistic scenario of not only DZC (distorted, zoomed and cropped) sketches but our proposed algorithm is robust to sketches which have been aged and still provide a sufficiently high degree of accuracy. using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

A. Convolution Layer Design for proposed Siamese Linkage

The dataset used for the proposed model was the CUHK Face Sketch database [15] [16]. For each face, there is a sketch drawn by an artist based on a photo taken in a frontal pose, under normal lighting condition, and with a neutral expression. However, only using the images of the database, the proposed network has a tendency to overfit the data yielding 100% accuracy as shown by Fig. 4 below. In such a scenario if we were to take a sketch image, zoom or distort or skew it a little, the network will fail to map it to any image in the database. This is clearly undesirable as we cannot assume the sketches drawn will be exactly the same every time. In fact, even the same artist will fail to draw it with 100% accuracy and there will be variations in the strokes and scale. To counter this drawback, we use data augmentation on the CUHK dataset. So in a sense the proposed algorithm is trained to check the similarity between the sketch-image pair in conditions where the images are non-ideal such as zoomed, cropped ordistorted.

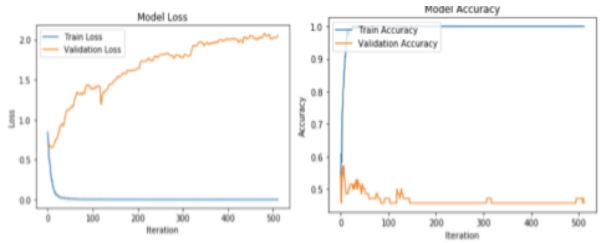

Fig. 4. Overfitting (training accuracy 100% and training loss 0.00004

B. Data Augmentation

The CUHK sketch images are made to undergo variations with augmentations to generate zoomed, skewed, cropped versions of the augmentations on the CUHK images. While the zoom is limited to 125% to 130% of the image size, the skewed and cropped images are randomly generated. The random distortions are performed within a 4x4 window of varying magnitude while the randomly cropped sketches contain between 80-90% of the original sketch. The proposed model also identifies sketches which have modified by giving them an ageing effect. This is the highlight of our proposed model. The proposed model hence creates six different sketches for every original sketch in the database. Thus taking the total number of images to 25 x 7 = 175. Some examples of augmented images are shown below in Fig. 5.

Fig. 5. Augmented Images cropped, distorted and zoomed and appended to CUHK database

C. Training on Augmented Data

Since the Siamese Network takes in two inputs (a sketch and database image pair) at a time as shown in Fig. 3, our input data needs to be in the form of a pair too. But if only true pairs are passed to the network, the network will only see the label '1' which is for true pairs and will come to the conclusion that it should predict all the values as '1' to minimize the loss. This is of course something we want to avoid. Consequently, our training data contains positive as well as negative pairs. For each of the 25 students, we have 6 augmented images of the sketches and 1 original sketch image. All of those images together with the database image will mean we have 7 positive pairs for a particular student. The negative pairs can be formed by pairing database image of student X with random sampling of all student's (except student X) images. To keep the data balanced, because we have 7 positive pairs, we will sample 7 negative pairs taking the total to 14 training pairs for a student. Following this for all the 25 students, we'll have 25 x 14 = 350 training pairs with corresponding labels (1 for positive pair, 0 for a negative pair). This is summarized for one student (student X) in Table II. The same data manipulation is applied for all the 25 students.

Once the data has been loaded in the said manner, we shuffle the data, take out 20% of the samples as validation set, and begin training.

D. Training the Siamese Network

Since our labels are 1 and 0, 1 indicating a positive match and 0 indicating a negative match, we use binary cross entropy loss as our loss function with ADAM optimizer (2) at a learning rate of 0.0006 trained for 1024 epochs since Siamese networks tend to exhibit slow convergence [8].

$$L(y,\overline{y}) = \frac{1}{N}\sum_{i=0}^{N}(y * \log(\overline{y}_i) + (1-y) * \log(1-\overline{y_i})) \quad (1)$$

where y represents the sketch and $\overline{y}$ the corresponding image each of size N.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (2)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (3)$$

with g_t as the gradient which varies batchwise.
Equation (2) and (3) represents the moving average of the gradient and the squared gradient for ADAM optimizer whose weight update equation is given by (4)

$$w_t = w_{t-1} - \eta \frac{\overline{m_t}}{\sqrt{\overline{v_t}} + \epsilon} \quad (3)$$

where η is the learning rate and $\overline{m_t}$ and $\overline{v_t}$ are the biased corrected estimators for the first and the second moment given by (5) and (6)

$$\overline{m_t} = \frac{m_t}{1-\beta_2^t} \quad (5)$$

$$\overline{v_t} = \frac{v_t}{1-\beta_2^t} \quad (6)$$

TABLE II. DATABASE CREATION FOR THE PROPOSED SIAMESE

For Student X			**Total**
No. of sketch images	6 (Augmented)	1 (Original)	7
No. of positive pairs	6 augmented sketch images paired with 6 copies of the student X's database image	1 original sketch image paired with student X's database image	7 sketch images paired with 7 copies of student X's database image
No. of negative pairs	Randomly sampled 7 sketch images of other students paired with student X's database image		7 sketch images of random students BUT paired with 7 copies of student X's database image
Total training pairs per student			14 training pairs for student X

The batch sizes and regularization were varied as hyperparameters (β_1) and (β_2) wherein only the best model was saved for every batch size by using Keras callbacks.

III. Results and Discussion

The proposed network was trained for different batch sizes using the CNN model proposed in Table I. The results are illustrated as a representative of the accuracy in Figure 6 where the batch size is varied from 1 to 32.

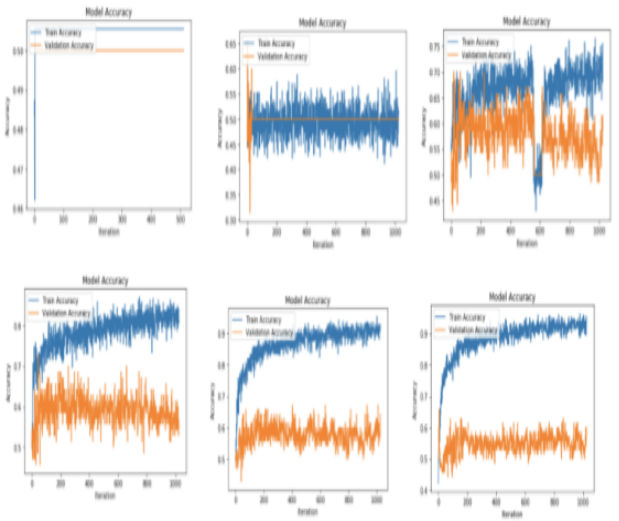

Fig. 6. Accuracy of the proposed model as a function of the batch sizes in powers of 2 from top left to bottom right

The proposed model exhibits the least loss for a batch size of 8 as suggested in Table III.

TABLE III. Validation Set Metrics for Different Batch Sizes with L1 and L2 Regularization

Batch Size	Best Accuracy (Train)	Best Accuracy (Validation)	Best Loss (Train)	Best Loss (Validation)
1	50.54%	50.00%	0.8199	0.8200
2	44.77%	65.714%	147.3092	130.8368
4	59.993%	70.00%	26.6541	25.6464
8	81.5%	72.857%	2.8944	3.0708
16	80.87%	65.714%	6.8173	6.9206
32	57.76%	65.714%	176.6134	172.1169

The proposed network was also trained for different regularizations, the results are illustrated in Fig. 7 and the best accuracies for each model are suggested in Table IV.

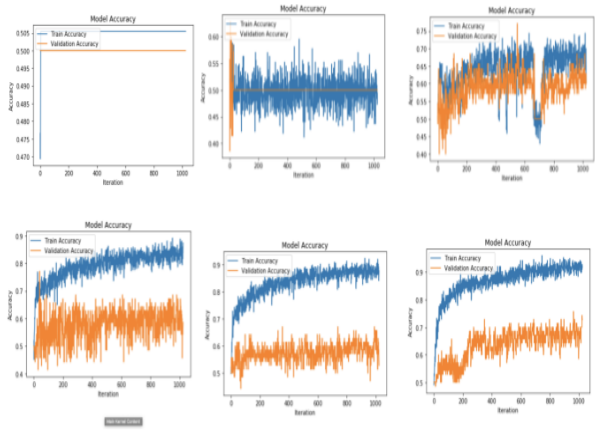

Fig. 7. Accuracy of the proposed model as a function of the batch sizes with regularization

TABLE IV. Validation Set Metrics for Different Batch Sizes with Merely L1

Batch Size	Best Accuracy (Train)	Best Accuracy (Validation)	Best Loss (Train)	Best Loss (Validation)
1	47.65%	50.00%	101.7246	14.7935
2	46.47%	62.857	87.9442	79.8880
4	66.43%	77.143%	3.1772	3.1879
8	67.15%	77.143%	11.8217	11.6286
16	83.39%	67.143%	2.5214	2.8090
32	90.25%	75.714%	1.8330	2.2369

The system proposes to identify sketches which have undergone an ageing process as shown in Fig.8.

Fig. 8. Test set made by 'ageing' the original sketches

The results on testing this data for the best models generated during the regularization and for optimum batch sizes are presented in Table V and the corresponding accuracies are plotted in Fig. 8.

TABLE V. TEST METRICS FOR THE TRAINED MODELS AS APPLIED TO AGED SKETCHES

Batch Size	Test Accuracy (L1 + L2)	Test Accuracy (Only L1)	F1 Score (L1 + L2)	F1 Score (Only L1)
1	51.34%	53.21%	0.6754	0.6928
2	50.18%	56.45%	0.5098	0.4050
4	60.89%	46.56%	0.6008	0.109
8	57.01%	63.47%	0.4775	0.6407
16	67.43%	51.13%	0.7226	0.4615
32	59.00%	73.89%	0.686	0.7326

The indicators evaluated for the optimum model for aged sketches are further briefed in Table VI and Table VII which provide detailed statistical measures in scenarios of both L1 and L2 regularization and with just L1 regularization respectively for further comparison.

It is observed from Table VI that initially for Batch size 1 the model overfits as previously discussed. However the readings show that the model shows erratic trends till batch size 16. The high validation accuracy observed for batch size 8 for augmented images is not consistent for aged sketch augmented images. This is further corroborated in Table VII.

The confusion matrix for the optimum model is shown in Fig. 9.

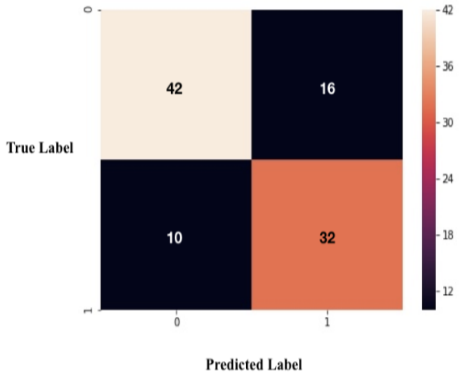

Fig. 9. Confusion Matrix for Optimum Model

TABLE VI. STATSTICS WITH BOTH L1 + L2 REGULARIZATION FOR OPTIMIZED MODEL

Batch Size	Accuracy (out of 1.00)	F1 Score	Precision	Recall	True Positives	True Negatives	False Positives	False Negatives
1	0.51	0.6754	0.51	1.0	51	0	49	0
2	0.50	0.4565	0.525	0.4038	21	29	19	31
4	0.54	0.6101	0.5454	0.6923	36	18	30	16
8	0.52	0.4666	0.5526	0.4038	21	31	17	31
16	0.58	0.6440	0.5757	0.7307	38	20	28	14
32	0.63	0.7131	0.5897	0.9019	46	17	32	5

TABLE VII. STATSTICS WITH MERELY L1 REGULARIZATION FOR OPTIMIZED MODEL

Batch Size	Accuracy (out of 1.00)	F1 Score	Precision	Recall	True Positives	True Negatives	False Positives	False Negatives
1	0.52	0.6842	0.5200	1.0	52	0	48	0
2	0.55	0.4155	0.6153	0.3137	16	39	10	35
4	0.47	0.1311	0.5714	0.074	4	43	3	50
8	0.62	0.6274	0.6153	0.6400	32	30	20	18
16	0.54	0.5208	0.6097	0.4545	25	29	16	30
32	0.74	0.7425	0.7619	0.7241	32	10	42	16

IV. Conclusion and Future Scope

A batch size of 8 seems to give the best results on the training set, validation set and the test set when using L1 and L2 regularization together. However, the same model is unable to handle 'aged' sketches with a reasonable accuracy. It gives 57.01% accuracy which is just slightly better than random guessing (50%). Other batch sizes either seem to stagnate or overfit/underfit as observed in Figure 6, 7 and 8. From Table 3, it is becomes clear that the model with batch size 8 has the least loss among the models for which the gradients did not go to zero from the beginning (i.e., the models that trained for a while before stagnating) along with having the highest training and validation accuracies amongst all models trained. When just L1 was used as regularization, batch size 8 tied in place with batch size 4 for the highest validation accuracy, however, batch size of 32 emerged as the top model due to highest training accuracy, second highest validation accuracy and lower loss than any other model. Overall, from Table 5, it is clear that the model of batch size 32 with only L1 regularization recorded a test accuracy of 74%and a F1 score of 0.73. In conclusion, this project makes a strong case towards using semantic or holistic abstractions of an image for a face recognition or a face matching task. As illustrated, model trained with holistic abstractions extracted by our Siamese Network are correctly able to match sketches to the image in the database with robustness and handling the nuances of sketch variations and age. As a future latitude we recommend to train the proposed architecture using triplet loss function.

References

[1] D. N. Parmar and B. B. Mehta, "Face Recognition Methods & Applications," in *International Journal Computer Technology & Applications*, vol. 4 (1), pp. 84-86, Feb 2013.

[2] S. Rakesh, K. Atal, A. Arora, P. Purkait, and B. Chanda, "Face Image Retrieval Based on Probe Sketch Using SIFT Feature Descriptors," *in Int. Conf. on Perception and Machine Intelligence, Springer, Berlin, Heidelberg,* vol. 7143*,* pp. 50-57, 2012.

[3] B. Klare, Z. Li and A. K. Jain, "Matching Forensic Sketches to Mug Shot Photos," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639-646, March 2011.

[4] H. Kiani Galoogahi and T. Sim, "Face sketch recognition by Local Radon Binary Pattern: LRBP," *2012 19th IEEE International Conference on Image Processing*, Orlando, FL, 2012, pp. 1837-1840.

[5] Yang, Bo and Songcan Chen, "A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image," *Neurocomputing* vol.120, pp. 365-379, November 2013

[6] X. Wei, H. Wang, G. Guo. H. Wan, "A General Weighted Multi-scale Method for Improving LBP for Face Recognition," *in Int. Conf. on Ubiquitous Computing and Ambient Intelligence, Springer Cham*, vol. 8867, pp. 532-539, 2014.

[7] Joshi. M, P. Gargi and S. Shamal, "Face Sketch Recognition Based On SIFT and MLBP," *in International Journal of Advanced Technology in Engineering and Science,* vol. 04, pp. 108-117, 2016.

[8] H. Momin, and J.R. Tapamo, "Automatic Detection of Face and Facial Landmarks for Face Recognition*," in International Conference on Signal Processing, Image Processing and Pattern Recognition,* vol. 260, pp. 244-253, Springer, Berlin, Heidelberg, 2011.

[9] A. Juhong and C. Pintavirooj, "Face recognition based on facial landmark detection*," in 10th Biomedical Engineering International Conference (BMEiCON),* Hokkaido, 2017, pp. 1-4.

[10] Q. Wang and K. L. Boyer, "Feature Learning by Multidimensional Scaling and Its Applications in Object Recognition*," in XXVI Conference on Graphics, Patterns and Images, Arequipa,* 2013, pp. 8-15.

[11] B. O'Connor and K. Roy, "Facial Recognition using Modified Local Binary Pattern and Random Forest," *in International Journal of Artificial Intelligence & Applications (IJAIA),* vol. 4, no. 6, November 2013.

[12] Z.A. Abbas and B. Duchaine, "The Role of Holistic Processing in Judgments of Facial Attractiveness," *in Perception*, vol. 37, no. 8, Aug. 2008, pp. 1187–1196.

[13] S. Karamizadeh, S.M. Abdullah and M. Zamani, "An Overview of Holistic Face Recognition," *in International Journal of Research in Computer and Communication Technology*, vol. 2, no. 9, pp. 738-741, September 2013.

[14] L. Fan, H. Liu and Y. Hou, "An Improved Siamese Network for Face Sketch Recognition," 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, 2019, pp. 1-7.

[15] S. Chopra, R. Hadsell and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," *in IEEE CVPR'05*, San Diego, CA, USA, 2005, vol. 1, pp. 539-546.

[16] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, and K. Koray, "*Matching networks* for *one shot learning*," in *Advances* in *Neural Information Processing Systems*, *pp. 3630– 3638*, December 2017.