IMAGE RECONSTRUCTION AND MATCH FINDING IN DATABASE

A Thesis Submitted in Partial Fulfillment of the Requirements for the award of the degree of

Bachelor of Science in Electrical and Computer Engineering
(Computer Engineering)
By

ANTENEH GETACHEW ETS 0138/09

BEDILU BALCHA ETS 0173/09

YISHAK TADELE ETS1021/09

Under the guidance of

SOLOMON ZEMENE (PhD)
Asst. Professor

ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY

SEPTEMBER, 2021

EXAMINING COMMITTEE APPROVAL SHEET

IMAGE RECONSTRUCTION AND MATCH FINDING IN DATABASE

By

ANTENEH GETACHEW ETS0138/09

BEDILU BALCHA ETS0173/09

YISHAK TADELE ETS1021/09

Approved by the examining committee members:

	Name	Academic Rank	Signature	Date
Advisor:	_____	_____	_____	_____
Co-Advisor:	_____	_____	_____	_____
Examiner:	_____	_____	_____	_____
Examiner:	_____	_____	_____	_____

	Name	Signature	Date
DC Chairperson:	_____	_____	_____
Associate Dean for Under Graduate Programs:	_____	_____	_____

ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY
COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING
DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

Certificate

This is to certify that the thesis entitled ***"Image Reconstruction and Match Finding in Database****" is submitted by* **ANTENEH GETACHEW, BEDILU BALCHA, AND YISHAK TADELE** *for the award of the degree of Bachelor of Science in Electrical and Computer Engineering (Computer Engineering), Addis Ababa Science and Technology University is a record of original work carried out under my supervision and they fulfil the requirements of the regulations laid down by the University and meets the accepted standards with respect to originality and quality.*

The results of the thesis have neither partially nor fully been submitted to any other University or Institute for the award of any Degree or Diploma.

Name of Advisor: SOLOMON ZEMENE (PhD.) **Signature:**

Head of Department: Mr. Fisha Abayneh **Signature:**

ACKNOWLEDGEMENT

We would like to thank Dereje Engida (PhD.), President, and Tarekegn Berhanu (PhD.), Academic Vice-President, for providing all of the essential infrastructures for us to do and finish the project work.

We would like to express our heartfelt gratitude to Sultan Feisso (Dr.), Dean of the College of Electrical and Mechanical Engineering, for inspiring us and assisting us in completing project work.

We would like to take this occasion to thank our adored, dynamic, and role model Mr. Fisha Abayneh, Head of Department, for his unwavering support in providing all facilities, motivation, and encouragement in all parts of project completion.

We owe a debt of gratitude to project supervisor Assistant Professor Solomon Zemene (PhD), Deputy Head for High-Performance Computing and Big Data Analytics CoE Addis Ababa Science and Technology University, for his insightful and constructive comments during the planning and implementation of this project. His willingness to offer so freely of his time, as well as his passion for work and diligent character, has been a source of inspiration for us.

We would like to convey our heartfelt gratitude to Mr. Yonas Tesfaye of the Computer Department Head-Advisor for his excellent comments during our project presentation and report preparation.

On the path of life, no one walks alone. Where do we begin to express our gratitude to those who joined us, walked beside us, and assisted us in a variety of ways? We would like to extend our heartfelt gratitude to everyone who contributed to the successful completion of this project work, both directly and indirectly.

Project Associates

(Anteneh Getachew, Bedilu Balcha and Yishak Tadele)

ABSTRACT

Finding a way to recognize people from their sketches is of great importance for law enforcement. Automatic retrieval of photos of suspects from police mug-shot databases using sketches drawn by artists can quickly help police to narrow down potential suspects, and facilitate the investigation process. In this thesis, we have developed a deep learning model based on a famous Siamese Network architecture, which takes a sketch and a photo as an input and gives out a similarity score as an output using deep features. In addition to that, we have collected about 1800 unique local images of 200 people with different posts from different angles. Since our model architecture takes photo-sketch image pairs we have converted half of these images into sketches using the CycleGAN model.

The model architecture that we have built works fairly well since it captures promising features with this small amount of dataset that helps to give a similarity score of the given input images. In order to evaluate how the model works, we have selected the cost function loss as a way to make sure that our model works the job of finding similarities. Findings suggest that the use of the Siamese network in image similarity, in general, helps to capture promising feature representations without the need for a huge amount of training data that is required in the classical deep learning models, and the use of spatial pyramid pooling layers is also helpful to improve performance, which allows feeding different size and domain images without the need of having the fixed-size and same domain images as in classical deep learning.

Photo-sketch image matching can be applied in a lot of applications especially in the law enforcement area to identify criminals using the sketch of suspects' faces, which is based on the eye witness's description, and for finding lost people by automatically retrieve photos and other details from the country's public database based on the sketch drawn, and for entertainment.

TABLE OF CONTENTS

EXAMINING COMMITTEE APPROVAL SHEET ii

Certificate iii

ACKNOWLEDGEMENT iv

ABSTRACT v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES ix

LIST OF ABBREVIATIONS x

CHAPTER ONE 1

1. INTRODUCTION 1

1.1. Background 1

1.2. Statement of the Problem 2

1.3. Objectives 2

1.3.1. General Objective 2

1.3.2. Specific Objectives 2

1.4. Significance of the Project 3

1.5. Scope and Limitation of the Project 3

CHAPTER TWO 5

2. LITERATURE REVIEW 5

2.1. Deep Learning 5

2.2. Spatial Pyramid Pooling(SPP) 5

2.3. Generative Adversarial Networks 6

2.4. Siamese Neural Networks 6

2.5. Triplet Network 7

2.6. Sketch-Image Recognition 7

2.6.1. Viewed Sketch-Based Image Recognition Methods 8

2.6.2. Forensic Sketch-Based Methods 9

2.7. Deep Face Recognition 9

2.8. Deep Layer Aggregation 10

CHAPTER THREE 11
3. HARDWARE AND SOFTWARE USED 11
3.1. Hardware Components/Tools Used 11
3.2. Software Tools and Programming Languages 11
CHAPTER FOUR 12
4. METHODOLOGY AND SYSTEM DESIGN 12
4.1. Data Collection 12
4.2. Data Preprocessing 13
4.3. Generating Sketch 14
4.4. Model Selection 16
4.5. Model Design and Training 18
CHAPTER FIVE 22
5. RESULTS AND DISCUSSION 22
5.1. Result 22
5.2. Discussion 24
CHAPTER SIX 25
6. CONCLUSIONS AND SCOPE FOR FUTURE WORK 25
REFERENCES 26

LIST OF FIGURES

- Figure 2. An Example flowchart for traditional sketch methods
- Figure 4.1 Collected Photo Images
- Figure 4.2 Collected Sketch Images
- Figure 4.3 Cycle GAN Architecture
- Figure 4.4 Cycle GAN output results during the training phase.
- Figure 4.5 Final Generated Image Set for Each Entity
- Figure 4.6 Siamese Network with Shared Parameters
- Figure 4.7 Siamese Network with unshared parameters(Pseudo-Siamese) as a Binary Classifier
- Figure 4.8 Spatial Pyramid Pooling layer for fixed-length output representation
- Figure 4.9 Siamese Network with SPP model to give image similarity score
- Figure 5.1 Number of Epochs vs Loss
- Figure 5.2 Similarity Score for Test Samples

LIST OF TABLES

- *Table 4.1 Summary of Our Model Architecture*

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
CycleGAN	Cycle Generative Adversarial Network
OpenCV	Open Source Computer Vision Library
HAOG	Histogram of Averaged Oriented Gradients
SINT	Siamese INstance search Tracker
DNNs	Deep Neural Networks
CUFS	CUHK Face Sketch Database
CUFSF	CUHK Face Sketch FERET Database
CUHK	China University of Hong Kong
ECG-SPP-net	Electrocardiogram-Spatial Pyramid Pooling-Network
GAN	Generative Adversarial Network
CPU	Central Processing Unit
RAM	Random Access Memory
SPP	Spatial Pyramid Pooling
PC	Personal Computer
GPU	Graphics Processing Unit
GHz	GigaHertz
PhD	Doctor of Philosophy
HP	Hewlett-Packard
RGB	Red Green Blue

3D	3-Dimension
2D	2-Dimension
HD	High Definition
ANN	Artificial Neural Network
VGGFace	Visual Geometry Group Face Model
GB	GigaByte
HOG	Histogram of Oriented Gradients
LBP	Local Binary Pattern

CHAPTER ONE

1. INTRODUCTION

1.1. Background

In the early days, people could identify each other due to the limited human population and low mobility which resulted from the fact that there were no extensible ways for transportation. Currently, with a rapid increase in the world population and the development of various transportation methods, leading to an extensible mobility rate, it is important to have a way for identity management. Deep learning has gotten a lot of attention in various applications for identity management using iris, face, and fingerprint patterns as unique identifying features.

In comparison with other physical identifying characteristics of a human, the face contains other useful information such as gender, emotion, ethnicity, and age in addition to identity [23]. For these reasons, face photo or face sketch images are used for investigation purposes in police departments, and for finding lost people in social media.

In this thesis, we have built a deep learning model based on Siamese network architecture to find a matching photo image from a given sketch image using deep features. The siamese network allows us to train a neural network without the need for a huge amount of data to achieve a reasonable amount of performance[69, 70]. In [24] the results showed that the expressiveness of the similarity metric learned by the fully convolutional Siamese network on ImageNet Video alone is enough to achieve very strong results, comparable or superior to recent state-of-the-art methods, which often are several orders of magnitude slower.

The use of the spatial pyramid pooling layer in the deep learning models helps to get rid of the requirements of having the same dimension of training images, which is the case in the classical deep learning image recognition as in [55, 57]. In addition to that, we have collected a dataset of more than 200 people with different poses and backgrounds and used the CycleGAN model to generate their corresponding sketches, which is used to train and evaluate our model. We choose the CycleGAN because quantitative comparisons against several prior methods demonstrate the superiority of the CycleGAN approach in translating images from one domain into a target domain

in the absence of paired examples[38]. And also because it is difficult to find an image-sketch pair dataset to train our model.

In addition to that, we have performed a comparative performance analysis among the Siamese Network with a contrastive loss, Siamese Network used for binary classification[20].

1.2. Statement of the Problem

The current criminal investigation method in our country's law enforcement is a very time-consuming process. It involves tiresome paperwork which yields low results that still consumes a huge amount of time and money. These problems on the law enforcement organizations made it hard to control and reduce the day-to-day major criminal activities.

When a crime is committed against a person they have a high probability of identifying the face of the criminal, which is used as major information for the investigation process. Currently, that information is recorded as a verbal description or translated to sketches with an artist's involvement. That sketch is used to search for the criminal manually/traditionally with no automation or computer support.

We aim to implement a system that takes the sketch as input and searches for similar images from the citizen's database and returns a similarity score minimizing the potential suspect list. This helps to speed up the process, by generating a result with very little time compared to the traditional searching. It also has the added advantage of searching for and comparing the pictures found in big datasets, which would be impossible if done manually.

1.3. Objectives

1.3.1. General Objective

The main objective of this project is as follows:

To build a Convolutional Neural Network model that helps to identify humans face images from the sketches.

1.3.2. Specific Objectives

We have three specific goals in mind to achieve and these are:

Localizing the semi-siamese neural network

Increasing similarity accuracy by combining a semi-siamese network and 2-channel Convolutional Neural Network.
Generating sketches using CycleGAN, from half of the photos collected

1.4. Significance of the Project

Nowadays, almost every nation is moving toward a digital era that involves digitizing law enforcement Agencies in order to sustain the development goals and keep the nation's sovereignty. starting from local law enforcement bureau to the Federal Police Departments, National Military Defense Force, Regional/State Special Military Force and Courts throughout the country included. Despite the development plans and some movements by tech startup companies, as a country, we still are performing everything in a traditional way and not using the power of technology to yield productive results.

This research project helps to fill this gap by providing a new way of investigating crime and an advanced file searching mechanism. The findings of this study directly benefit the law enforcement sector to be able to find suspects or lost people in a timely and cost-efficient manner. The application can be extended for medical document matching, to identify original documents from the fake ones and image-based file searching in large governmental and non-governmental Organizations.

1.5. Scope and Limitation of the Project

Scope: The scope of the study is limited to collecting the photos of 200 random people (10-16 unique photos/person) including artists, investors, influencers, doctors, individuals, and others between the ages of 14 and 80 from various social media platforms and google search engine. This data collection and data filtering(which is done by isolating those images with high contrast from dark ones and low contrast images that negatively affect our result) period will last for a maximum of 2 weeks and will end when either photo of 200 people are collected or the 2 week period has passed.

After the data collection for 2-3 weeks, further filtering will be done using a python script that can crop out the faces of the people from their photos, and organize them in folders for the next task. The next part of the work will be generating sketches from the collected dataset using the

CycleGAN model. Lastly, the sketch generated and the photos organized will be given as an input for our model to train it so that it will be able to find a similarity score for the input images.

Limitations: Even though this thesis project can be applied in a lot of application domains to facilitate and simplify the process we deal with every day to search for lost people, for law enforcement, and can be extended for medical document matching, to identify original documents from the fake ones.

As a limitation to our work, we have got two main problems. The first one is hardware resource and time limitation to train and collect a large amount of training data to achieve excellent performance in terms of accuracy. Hardware resource limitation comes from the fact that we could not find a GPU with enough RAM to load the model and the training data, so, this limits us to train the model with fewer epochs, when we come to the time limitation, we have not got enough time to do the project, this limits us from collecting an adequate amount of data for our model to learn more representative model parameters.

CHAPTER TWO

2. LITERATURE REVIEW

After a deep review of many papers and surfing the web (websites and journals) that are related to Image Reconstruction and Match Finding, we come up with the following review.

2.1. Deep Learning

Deep learning is a machine learning method that uses multiple layers to extract deep features from the input data. The feature extracting is performed progressively as the input passes through successive neural network layers. Deep learning models complex relationships among the data by learning and identifying multiple levels of representations. For example, low-level feature representations in the model such as edges and lines are used to define high-level features such as shapes and patterns. This hierarchical design of the feature is called deep architecture [63].

During the past years, deep learning techniques have been applied to a range of signal and information processing works, such as Visual and Audio Recognition, Natural language processing, Fraud Detection, Autonomous driving, etc.

Convolutional Neural Networks (CNNs) is one of the main categories of Deep learning methods. A convolutional neural network is a kind of feedforward neural network that is able to extract features from data by doing convolution operations.

[64] stated that CNN has the following advantages compared with the general artificial neural networks 1, Local connections. There is no full connection between a neuron and previous layer neurons, but the neuron will be connected to a small number of neurons. This helps to speed up the convergence by reducing the number of parameters. 2, Weight sharing. According to [65] a group of connections can share the same weights, which reduces parameters further. it also makes the feature search insensitive to feature location in the image. 3, Downsampling dimensionality reduction. A pooling layer in CNN uses the local correlation in image pixels to downsample the data size while retaining useful information.

2.2. Spatial Pyramid Pooling(SPP)

Kaiming He and his colleagues [3], explained and proved that SPP is a flexible solution for

handling different scales, sizes, and aspect ratios images and suggested a solution to train a deep network with a spatial pyramid pooling layer, which resulting SPP-net shows outstanding accuracy in classification/detection tasks and greatly accelerates DNN-based detection.

Jia Li, Yujuan Si [10], built an ECG-SPP-net for the classification of heartbeats, and the simulation results showed that ECG-SPP-net can extract more representative features than traditional CNNs and has a higher classification accuracy.

2.3. Generative Adversarial Networks

Image generation [42], image editing [43], and representation learning [42, 44, 45] have all seen outstanding outcomes using Generative Adversarial Networks (GANs) [40, 41]. Text2image [46], image inpainting [47], and future prediction [48], as well as additional domains including movies [49] and 3D data [50], have recently adopted the same principle for conditional image generating applications.

The idea of an adversarial loss, which compels the generated images to be in principle indistinguishable from real photos, is the key to GANs' success. This loss is especially effective for image generation tasks, as this is precisely the goal that much of computer graphics seeks to achieve. To learn the mapping, we use an adversarial loss such that the translated images cannot be discriminated from those in the target domain.

CycleGAN [51] is a model designed to handle the challenge of image-to-image translation. The purpose of the image-to-image translation issue is to use a training set of matched picture pairs to learn the mapping between an input image and an output image. Obtaining paired examples, on the other hand, is not always possible. CycleGAN [51] uses cycle-consistent adversarial networks to learn this mapping without requiring paired input-output images.

2.4. Siamese Neural Networks

Gregory Koch and his colleagues [9] have presented a strategy for performing one-shot classification by first learning deep convolutional Siamese neural networks for verification and outlined new results comparing the performance of their networks to an existing state-of-the-art classifier developed for the Omniglot data set, to outperform all available baselines by a significant

margin and come close to the best numbers achieved by the previous authors.
In [14] proposed to train a Siamese network to identify candidate image locations that match the initial object appearance, dubbing their method SINT (Siamese INstance search Tracker).
Deep Siamese Conv-nets have previously been applied to a lot of deep learning tasks such as face verification [15,16,17], keypoint descriptor learning [18,19], and one-shot character recognition [20].

2.5. Triplet Network

Elad [12], introduced the Triplet network model in their work, which is a tool that employs a deep network to explicitly learn meaningful representation as the Siamese network does and the results on several datasets show that the learned representations are effective for classification in a way that is equivalent to a network that was specifically trained to categorize samples in which they believe this method should be investigated further, as it just requires knowing that two out of three images are sampled from the same class, rather than knowing what that class is.

2.6. Sketch-Image Recognition

The problem of automatically matching between face sketches and photos is how to fill the gaps, and it is not possible to directly apply face-photo recognition algorithms to face-sketch recognition problems. An example of the flowchart of the traditional sketch classification methods is shown in Figure 2.1 [1].
Yuchao [1] covered two major problems in the vision community, sketch image categorization and face sketch identification using face photographs since most past research has focused on these two topics separately, A novel approach for recognizing face sketches with images, which can help with law enforcement difficulties, has also been proposed, with state-of-the-art results.
Karen [21] investigated very deep convolutional networks (up to 19 weight layers) for large-scale image classification, demonstrating that representation depth improves classification accuracy and that state-of-the-art performance on the ImageNet challenge dataset can be achieved using a traditional ConvNet architecture with significantly increased depth.

According to what can be found, Uhl and da Vitoria Lobo [27] were the first to propose an automatic matching method for face sketches and photos [28]. According to the type of sketches used, studies on matching facial photos and sketch images can be classified as viewed or

forensic.

Viewed sketches are created by artists while they are looking at a photograph [29], as opposed to forensic sketches, which are created based on the description provided by the eyewitness(es), possibly days after the event. Forensic sketches can be deceptive because of errors in witness memory recall that result in inaccuracies in the sketch drawn by a forensic artist [30].

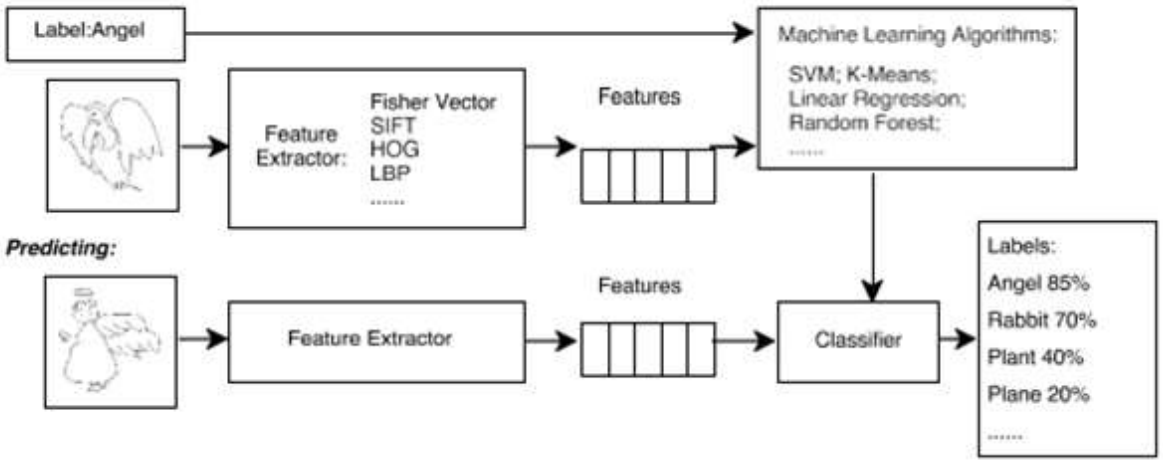

Figure 2.1 An Example flowchart for traditional sketch methods

However, due to the different nature between sketches and photos, it is not appropriate to apply the general image classification methods directly to sketches. Furthermore, the typical characteristics also add more difficulties to the task. To achieve robust visual representation for the sketch images, discriminative visual patterns in each sketch image must be identified. Deep learning methods proved to be able to incorporate the actual content of the image compared to its detailed pixel values [71].

2.6.1. Viewed Sketch-Based Image Recognition Methods

The majority of previous work has relied on viewed sketches and widely used and publicly available datasets (e.g., CUFS [25] and CUFSF [26]). The first approach focused on combining sketches and photos into a single domain, which was accomplished by either converting a face photo to a pseudo-sketch [31,32,33,25] or converting a sketch to a realistic photo [25,34]. This can significantly reduce the difference between photo and sketch, allowing for effective matching between the two, with most proposed face-photo recognition approaches being straightforward. However, synthesizing a sketch or photo from scratch remains an unsolved problem. Unsuccessful pseudo-images will complicate the matching process.

The second family of approaches, such as the Histogram of Averaged Oriented Gradients (HAOG) [35], reduce the modality gap during the feature extraction stage by designing new face descriptors that are more modality invariant. If the inter-modality difference between the extracted features is large, the classifiers' discriminative power will be reduced [26].

The third family uses methods such as Partial Least Squares to map images in different modalities to a common linear subspace in which they are highly correlated [36]. (PLS).

The viewed-sketch datasets are created under ideal conditions, which means that the faces to be studied are in a frontal pose with normal lighting, have neutral expressions, no occlusions, and the sketches closely resemble the photo with which they are paired. This has the advantage of allowing the cross-domain gap to be studied as a control variable. It is, however, unsuitable for use in real-world scenarios.

2.6.2. Forensic Sketch-Based Methods

Because real-world scenarios only involve forensic sketches, matching forensic sketches has become increasingly important, and achieving very good performance on viewed sketches does not necessarily imply that this methodology performs well on forensic sketch datasets [37].

Some forensic datasets have recently been published, and corresponding methods have been proposed. The majority of these works are built around descriptors that are unaffected by changes in image modalities.

In[39], a CNN-based framework for learning to automatically identify similar patches in SAR and optical pictures is given. A first evaluation yielded promising results, paving the door for the future development of SAR-optical tie point matching processes based on a learned generalized similarity measure.

2.7. Deep Face Recognition

Omkar M. Parkhi and companions [8], have made two contributions: First, they devised a method for assembling a huge dataset with little label noise while reducing the amount of manual annotation required and one of the major concepts was to score the data supplied to the annotators using weaker classifiers, that was created for faces and the second contribution was to demonstrate that a deep CNN may obtain results comparable to the state of the art when trained properly and without any frills.

Another thing is Face Image Modality and Bingjie [13] presented approaches to face image modality recognition in his paper in order to expand the possibility of cross-modality research

and to handle new modality-mixed face images with a new database comprised of eight datasets with five face image modalities and over 50 thousand face images, as well as some face image modality recognition methods.

2.8. Deep Layer Aggregation

Fisher Yu and his colleagues [22] by relating architectures for aggregating channels, scales, and resolutions they identified the need for deeper aggregation and addressed it by iterative deep aggregation and hierarchical deep aggregation and their models are more accurate and make more efficient use of parameters and computation than baseline networks.

CHAPTER THREE

3. HARDWARE AND SOFTWARE USED

In this chapter, we will state the tools both hardware and software that were used in our entire project. Since the work required devices with a large amount of computing speed and memory, first, we tried to use our PC's potential in combination with distributed computing platforms found online. Below we described the specifications and types of the tools used in detail as follows:

3.1. Hardware Components/Tools Used

- NVIDIA Tesla K80 GPU, 16GB GPU, 12GB RAM; https://colab.research.google.com/
- NVIDIA Tesla P100 GPU, 12GB GPU, 13GB RAM ; https://www.kaggle.com/
- ASUS TUF FX505 Gaming PC; Processor: Intel(R) Core™ i7-8750H CPU @ 2.20 GHz x12 with 8GB RAM: GeForce GTX 1050 Ti
- HP pavilion PC; Processor: Intel(R) Core™ i5-7200U CPU @ 2.50 GHz x4; RAM: 4 GB
- HP PC; Processor: Intel(R) Core™ i7-7500U CPU @ 2.70 GHz 2.90 GHz; RAM: 8 GB

3.2. Software Tools and Programming Languages

- Python Programming language including the NumPy library, Pandas, MatPlotLib, and OpenCV which are widely used for machine learning and deep learning applications.
- Pytorch python deep learning Framework
- Keras Framework which integrated with TensorFlow
- Deep learning Networks such as Semi-Siamese Network, Cycle Gan, 2-Channel Convolutional Neural Network.
- HTML, CSS and JavaScript

CHAPTER FOUR

4. METHODOLOGY AND SYSTEM DESIGN

This chapter will cover the methods or techniques that have been used starting from the data collection phase up to the final result phase. And also, we will discuss in detail how the entire model has been designed to meet the specific and general objectives we have in mind. Finally, this section will answer the question of why we choose the methods/techniques used in the entire working process in terms of different dependent and independent variables or constraints like accuracy, performance, speed, time, computing capacity, and storage capacity.

4.1. Data Collection

Sketch recognition problem is a Supervised problem because the model has to learn deep features using the labels provided to it. So, we had to find a dataset with a sketch-image pair for each label, which in our case is the individual.

A few face sketch datasets for matching with photos have previously been introduced such as [52] and [53]. but those datasets are prepared in a frontal sitting position and have limited practicality in our problem, because we may have sketches generated from the artist's impression with a laughing face, slightly bent head, and other angry face impressions. so training the dataset with a diversified image set was necessary.

The other problem is the dataset consists of a small number of training and testing images. Using the different datasets and merging them is also not practical because it will bias the model. since all the datasets have their own specific background and colouring effect, merging them brings cross-domain problems. Merging the dataset will also affect the learning system because the images have different domains since they were captured in different backgrounds and different sitting positions.

We, therefore, collected the images ourselves. After the data is cleaned we obtained 10 - 16 images per person for 200 people, a total of around 2400 images. The images were collected from social media platforms, such as Telegram, Instagram, Facebook, and the search engine Google. We collected images that contain the picture of a single individual with different poses, different colouring effects, and different backgrounds. the samples are shown in Fig 4.1.

Figure 4.1 Collected Photo Images

In addition to the collected photos, after the cleaning, we obtained 225 image sketches. The sketches are randomly collected to account for multiple artist's impressions. The samples are shown in Fig 4.2.

Figure 4.2 Collected Sketch Images

4.2. Data Preprocessing

After the image is collected we perform primary data cleaning, that is to detect a human face, crop it and save it in another new directory within that person's directory. The face detection is performed using the python dlib library using the pre-trained model shape_predictor_68_face landmarks. We used OpenCV[61] to load and process the image. In a similar manner, the collected sketches are cropped, resized, and saved on a new folder inside the same directory. and the rest image is removed from the working directory and the cropped and resized images are prepared for the next phase. In this phase, some of the images and sketches were dropped because the model failed to detect their face. The reason is in some of the sketches and images

the people's faces are not clearly visible.

4.3. Generating Sketch

As we mentioned earlier obtaining a large sketch-image pair dataset is difficult, as people do not have multiple sketches most of the time. So, we used CycleGAN [38] to generate the sketches from the given photo images. Because CycleGAN outperforms other methods such as Neural Style transfer[62] in translating images from one domain to another domain[38].

The network consists of two generative and two discriminative networks. The generative networks will generate output and the discriminative models will check whether the generated images are real or fake. we set Domain A to be the picture and Domain B, the sketch as shown in Fig 4.3.

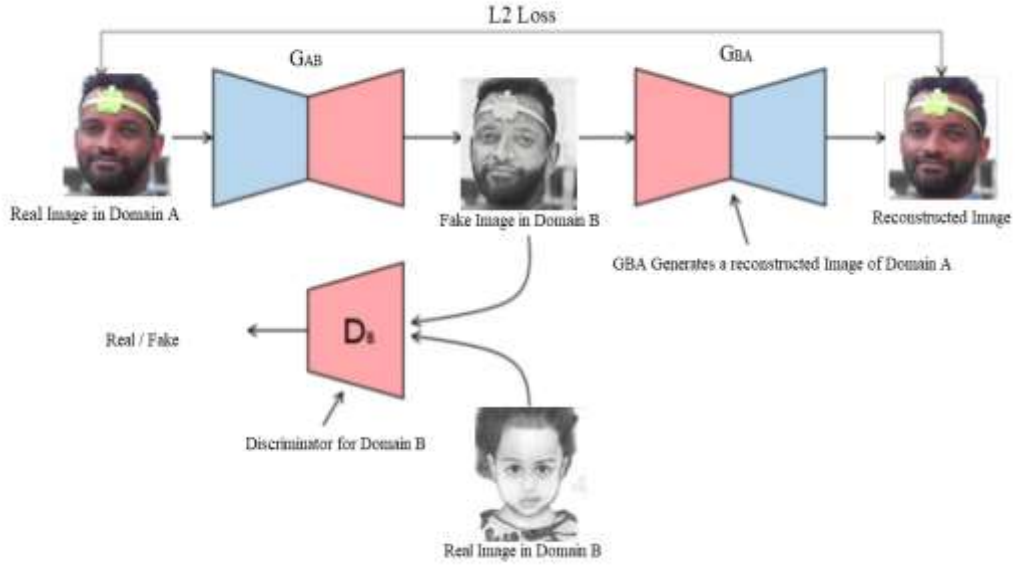

Figure 4.3 Cycle GAN Architecture

We trained the model using the collected 225 sketches in Domain B and 300 random photo images. Due to the memory and GPU processing limit, we trained the model iteratively by saving and loading the weight the next day. The results during the training stage are shown in Fig 4.4

Figure 4.4 Cycle GAN output results during the training phase.

After successive training, the model is used to generate sketches for the collected photo images. The generated sketches are shown in Fig 4.5.

Figure 4.5 Final Generated Image Set for Each Entity

4.4. Model Selection

Siamese Network is being applied diversely in a lot of applications in deep learning, where it is difficult to find a large amount of training data. It is a method for learning which employs a unique structure to naturally rank similarity between inputs[9]. We have selected a Siamese Network-Based model for this thesis to take advantage of the Siamese Network gives a reasonable performance with a reasonable amount of training data.

The Siamese network has two identical networks for two different samples of inputs, and the structures of the two CNN's are always the same, and the parameters can either be shared (Siamese) or unshared (pseudo-Siamese[39]) as shown in Fig 4.6, and Figure 4.7 respectively.

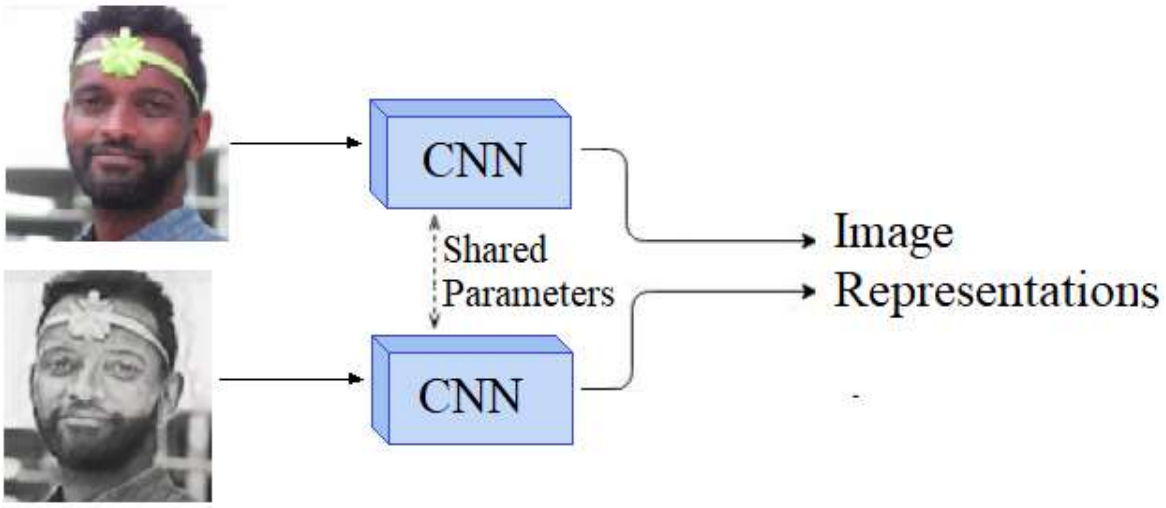

Figure 4.6 Siamese Network with Shared Parameters

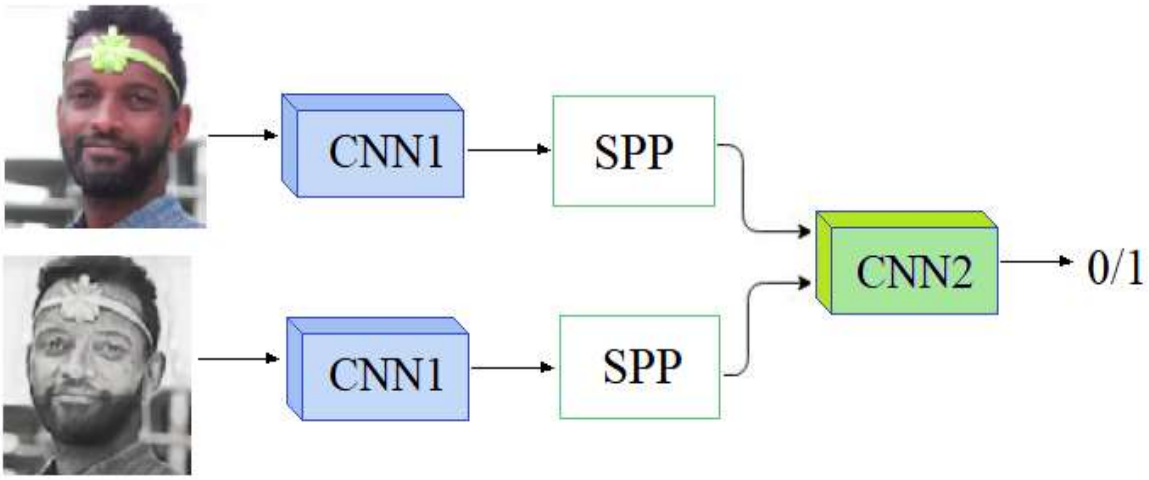

Figure 4.7 Siamese Network with unshared parameters(Pseudo-Siamese) as a Binary Classifier

On top of that, we equip the Siamese Network with a Spatial pyramid pooling layer since existing deep convolutional neural networks (CNNs) require a fixed size (e.g., 224×224) input image. This requirement is "artificial" and may reduce the recognition accuracy for the images of an arbitrary size/scale, as in[3] equipping the networks with spatial pyramid pooling helps to eliminate this requirement.

Spatial pyramid pooling [54], [55] as an extension of the Bag-of-Words (BoW) model [56], is one of the most successful methods in computer vision. It partitions the image into divisions from finder to coarser levels and aggregates local features in them as shown in Figure 4.6. In the leading and competing classification systems [57],[58],[59], and challenges in detection[60], spatial pyramid bundling has long been a significant element.

The pseudo siamese network is a siamese network that consists of two identical subnetworks with the same network architecture but with different model parameters, this network archives state-of-the-art performance for the generalized shot intent detection task[66]. The network architecture needs two different model parameters to be instantiated, so Due to the resource limitation (i.e RAM), we were unable to implement this network and perform a comparative analysis with the Siamese network to understand which architecture works well on our dataset.

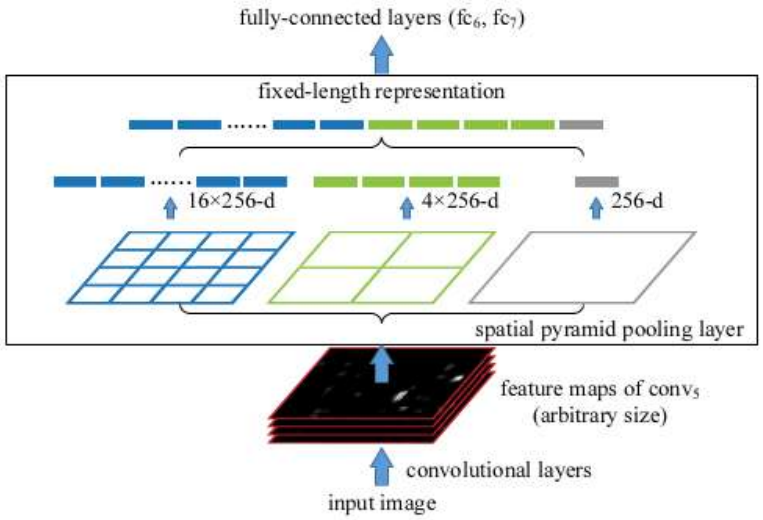

Figure 4.8 Spatial Pyramid Pooling layer for fixed-length output representation

4.5. Model Design and Training

After the model selection process, we have built the model which is a siamese network with a spatial pyramid pooling layer. The architecture of this model is shown in Fig 4.9. Because of the reasons that we have got hardware as well as time resource limitation, we took the first thirteen layers of the VGGFace model and applied transfer learning to initialize the pre-trained parameters as initial parameters to our model.

The reason why we use VGGFace is that it archives comparable results to the state-of-the-art using much fewer data and much simpler network architecture as compared to the state-of-the-art(i.e FaceNet with Alignment[67]) in face image recognition tasks that mean it is trained to capture deep feature representations[68].

On top of that, we added a seven-level spatial pyramid pooling(SPP) layer to get the same output feature size for any given input size images. The use of this kind of pooling layer helps to achieve cross-domain matching of input images. This layer allows us to feed a single-channel grayscale image, and a three-channel RGB image to compare them and find a similarity score. The output of SPP layers is gone through two Cov layers with Relu activation function and it flattens to give into the fully connected layers from where we got two feature vectors to be compared by the loss function.

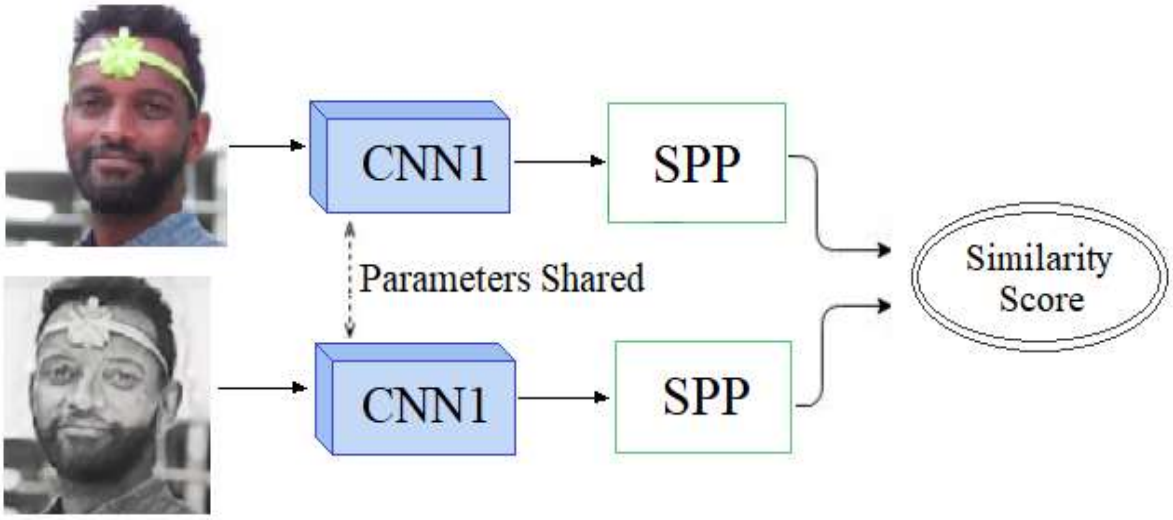

Figure 4.9 Siamese Network with SPP model to give image similarity score

The details of the architecture of our model are shown in Table 4.1. Based on this architecture first, we load the training and testing data, visualize sample images, and feed our model two 224x224 size images for the evaluation purpose, but the model can handle any input size. Then we load the pre-trained weights for the first thirteen convolution layers(Conv layers) from the DeepFace(i.e VGGFace), but the problem here was that we were working on Pytorch, whereas the original VGGFace was written in Keras. Hence, We map loaded weights from Keras into Pytorch Conv layers manually.

Next to this, we added a seven-level spatial pyramid pooling, and two Conv layers with a Relu activation function for 64 filters, where their parameters were randomly initialized. After this, we get a 64x1x64-d output and flattens it to 6x64-1d to go through three fully connected layers with Relu activation functions except for the last fully connected layer.

Table 4.1 Summary of Our Model Architecture

Operation Layer		Number of Filters (Neurons)	Filter Size	Stride Value	Padding Value
Input Image		-	-	-	1x1
Convolution Layer x2	Conv ReLU	64 -	3x3 -	1x1 1x1	1x1 -
Pooling Layer	Max pooling	-	2x2	2x2	0
Convolution Layer x2	Conv ReLU	128 -	3x3 -	1x1 1x1	1x1 0
Pooling Layer	Max pooling	-	2x2	2x2	0
Convolution Layer x3	Conv ReLU	256 -	3x3 -	1x1 1x1	1x1 0
Pooling Layer	Max pooling	-	2x2	2x2	0
Convolution Layer x3	Conv ReLU	512 -	3x3 -	1x1 -	1x1 0
Pooling Layer	Max pooling	-	2x2	2x2	0
Convolution Layer x3	Conv ReLU	512 -	3x3 -	1x1 -	1x1 0
Pooling Layer	Max pooling	-	2x2	2x2	0
Convolution Layer x2	Conv ReLU	512 -	3x3 -	1x1 -	1x1 0
Pooling Layer	SPP	512	5x5,4x4, 3x3, 2x2, 1x1, 3x2,1x3	1x1	0
Fully Connected Layer	Input Layer Hidden Layer 1 Hidden Layer 2 Output Layer 3	64x64 500 500 50	- - - -	- - - -	- - - -

Finally, we got two single-channel outputs size of 64x64, which are used by the loss function to learn their similarity or dissimilarity. In order to compare the performance of the model with these functions, we employed Contrastive loss functions which work. We trained this model

using 450 epochs, 0.005 learning rate, 32 batch size, and a total of 1600 photos and evaluated the model with other pictures.

The trained convolutional layer generates an optimized feature representation with reduced dimensions. We then pass the image and the sketch through the layer consecutively to generate their respective feature maps. Contrastive loss[7] is used to calculate the loss gained from each representation. The Equation for calculating contrastive loss is stated in Eq. 5.1, it will take the feature representation of the first and the second image, and the label which indicates the pair is in the same class or not. If they are in the same class the label will be 0, otherwise 1.

$$L(W,Y,X1,X2) = (1-Y)\frac{1}{2}(Dw)^2 + (Y)\frac{1}{2}\{max(0,m-Dw)\}^2 \quad (4.1)$$

To train the model we used Algorithm 4.1 which is obtained from [7] and modified to suit our case.

Algorithm

Step 1: For each input sample vector X_i, and prediction feature P_i do the following:

(a) Using prior knowledge find the set of samples $Sxi = \{Xi\}_{j=1}^{p}$, such that X_j is in the same class as X_i.

(b) Pair the sample, X_i with all the other training samples and label the pairs so that:

$Y_{ij} = 0$ if $X_i \in S_{X_i}$, and $Y_{ij} = 1$ otherwise.

Combine all the pairs to form the labeled training set.

Step 2: Repeat until convergence:

(a) For each pair (X_i, X_j) in the training set, do

i. if $Y_{ij} = 0$, then update W to decrease

$D = || Gw(X_i) - Gw(X_j)||_2$

ii. if $Y_{ij}=1$, then update W to increase

$D = || Gw(X_i)-Gw(X_j)||_2$

Step 3: Repeat for all (X_i,X_j) in Test set:

(a) Generate random (X_i,X_j,Y_{ij}) pair, with size L,

where $Y_{ij} = 1$, for $\frac{L}{2}$ and $Y_{ij} = 0$, for $\frac{L}{2}$

(b) For each pair (X_i,X_j), do:

i. get prediction score P_i, from the Model

CHAPTER FIVE

5. RESULTS AND DISCUSSION

5.1. Result

In order to evaluate the performance of the model, appropriate metrics should be selected to characterize the behaviour of the model. However, in this kind of problem, it is difficult to select which metrics fit the problem well. Since our problem is not a binary problem where the model trained to classify the input images are the same or not, instead the model gives us a similarity score of the given input images by computing the euclidean distance of the deep feature representations of them. After all, we selected the cost or loss of the model during the training as a metric to characterize how the model learns promising features that will be used in differentiating how much the given images(sketches and photos) are similar.

The main reason why we did not select accuracy as a performance metric is that since the model tried to give much the two given images are similar instead of checking whether they are the same or not. so we can not use accuracy as a performance metric. This reason limits us not to illustrate how much the model gets the correct answer, as per our knowledge we did not get any way to use accuracy as a metric in this kind of problem. The only way to cross-check how the model performs in our dataset is by plotting the cost function versus the number of iterations or echos.

Mapping of a higher dimension of the input image to a lower dimension is performed using the Siamese Network. The generated feature map is compared using the Contrastive loss function, which uses their Euclidean distance metric in the feature space.

Even though we can not generalize the performance of the model, as usual in a lot of classification and recognition tasks using accuracy as a metrics, we have plotted the cost or loss versus the number of iterations or epochs, which can be seen in Fig 5.1 In addition to that, we demonstrated the result of the model for sample images and their corresponding images along with their similarity matrics is shown in Fig 5.2.

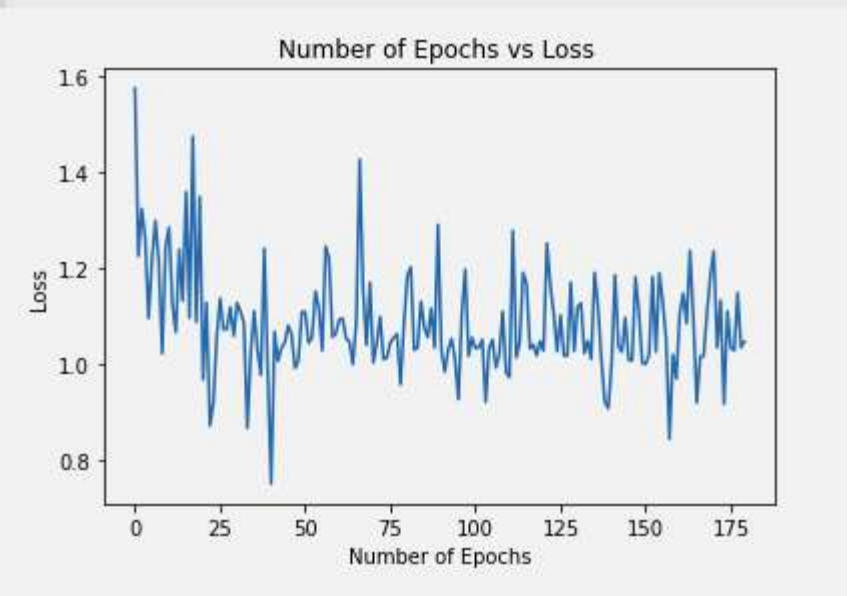

Figure 5.1 Number of Epochs vs Loss

As we can see in Figure 5.1 the cost function shows a decreasing trend, but the oscillation comes from the fact that we were working on a cross-domain image(photo and sketch) match, and we trained the model for only 176 epochs due to the computing resource and the time to finish the jobs. it takes about 9 hours just for these 175 epochs of training, this limits us from getting adequate performance on this problem.

Figure 5.2 Similarity Score for Test Samples

5.2. Discussion

Deep Learning methods require a large amount of computing power and storage size. We trained the model with the freely provided researching platforms we can find, which are Google's Colab and Kaggle. Due to the usage and resource limits, we have encountered memory problems several times. This forced us to train the model with a small batch size, 16 and a total of 440 epochs on the segmented training data.

The training took hours and even days to complete, and the usage allowed hours in Google Colab is less than 12 hours, so we had to save the model and reload again for the next train round. This limited us from evaluating the performance of different Siamese network architectures such as Pseudo-Siamese network and the state-of-the-art model, Triplet Siamese network.

The performance of the model is also affected by the transfer learning weights. We could not find a publicly available pre-trained model on the sketch dataset. So, we used DeepFace which is trained on the photos of individuals, using that as a filter to generate a feature map from the sketch image affects the model's performance.

CHAPTER SIX

6. CONCLUSIONS AND SCOPE FOR FUTURE WORK

In this thesis, we have implemented Siamese networks using Convolutional Neural Networks. We used a locally collected dataset to train and evaluate the model. Resource limitations have affected the result negatively, we trained and evaluated the model with the available resources and proved that the model is fully functional by plotting the cost with respect to the number of iterations to train the model. The plot showed the slightly decreasing trend of the cost with respect to the training iteration.

In the future, we want to build and evaluate multiple models with the collected and additional dataset. We want to explore the full potential of the Siamese Networks in the field of image similarity and match finding.

REFERENCES

[1] Yuchao Jiang, *Sketch Image Recognition Using Deep Features*. 2016.

[2] Koch, Gregory R, *Siamese Neural Networks for One-Shot Image Recognition*. 2015.

[3] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian, "Spatial
` Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." IEEE *Transactions on Pattern Analysis and Machine Intelligence*. 2014. 37.10.1109/TPAMI.2015.2389824.

[4] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2011 IEEE, 2011, pp, 513-520.

[5] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in Proceedings of the IEEE *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361.

[6] Y. Li, Y.-Z. Song, and S. Gong, "*Sketch recognition by ensemble matching of structured features.*" in BMVC. Citeseer, 2013.

[7] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Jun. 2006, vol. 2, pp. 1735–1742. doi: 10.1109/CVPR.2006.100

[8] Parkhi, Omkar & Vedaldi, Andrea & Zisserman, Andrew, *Deep Face Recognition*. 2015. 1. 41.1-41.12. 10.5244/C.29.41.

[9] Koch, G., Zemel, R. & Salakhutdinov, R, *Siamese Neural Networks for One-shot Image Recognition.* 2015.

[10] Li, Jia & Si, Yujuan & Lang, Liuqi & Liu, Lixun & Xu, Tao. (2018). *A Spatial Pyramid Pooling-Based Deep Convolutional Neural Network for the Classification of Electrocardiogram Beats*. Applied Sciences. 8. 1590. 10.3390/app8091590.

[11] M. Eitz, J. Hays, and M. Alexa, "*How do humans sketch objects?*"ACM Trans.

Graph., vol. 31, no. 4, pp. 44–1, 2012.

[12] Hoffer, Elad & Ailon, Nir, *Deep Metric Learning Using Triplet Network.* 2014. 10.1007/978-3-319-24261-3_7.

[13] Bingjie Liu, *Face Image Modality Recognition and Photo-Sketch Matching*. 2017.

[14] Tao, R., Gavves, E., Smeulders, A.W.M.: *Siamese instance search for tracking.* arXiv CoRR. 2016.

[15] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: *DeepFace: Closing the gap to human-level performance in face verification.* In: CVPR 2014. (2014) 1701–1708

[16] Schroff, F., Kalenichenko, D., Philbin, J.: *FaceNet: A unified embedding for face recognition and clustering*. In: CVPR 2015. (2015) 815–823

[17] Parkhi, O.M., Vedaldi, A., Zisserman, A.: *Deep face recognition*. BMVC 2015. 2015.

[18] Zagoruyko, S., Komodakis, N.: *Learning to compare image patches via convolutional neural networks.* In: CVPR 2015. 2015.

[19] Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: *Discriminative learning of deep convolutional feature point descriptors*. In: ICCV 2015. (2015) 118–126

[20] Koch, G., Zemel, R., Salakhutdinov, R.: *Siamese neural networks for one-shot image recognition*. In: ICML 2015 Deep Learning Workshop. 2015.

[21] Simonyan, K. and Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. The 3rd International Conference on Learning Representations (ICLR2015). 2015. https://arxiv.org/abs/1409.1556

[22] F. Yu, D. Wang, E. Shelhamer and T. Darrell, "Deep Layer Aggregation," 2018 IEEE/CVF *Conference on Computer Vision and Pattern Recognition*, 2018, pp.

2403-2412, doi: 10.1109/CVPR.2018.00255.

[23] A. K. Jain, A. A. Ross, and K. Nandakumar. “Introduction to Biometrics.” Springer, 2011.

[24] Bertinetto, Luca & Valmadre, Jack & Henriques, Joao & Vedaldi, Andrea & Torr, Philip. (2016). *Fully-Convolutional Siamese Networks for Object Tracking*. 9914. 850-865. 10.1007/978-3-319-48881-3_56.

[25] X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," in IEEE *Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955-1967, Nov. 2009, doi: 10.1109/TPAMI.2008.222.

[26] W. Zhang, X. Wang and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," CVPR 2011, 2011, pp. 513-520, doi: 10.1109/CVPR.2011.5995324.

[27] R. G. Uhl and N. da Vitoria Lobo, "A framework for recognizing a facial image from a police sketch," Proceedings CVPR IEEE *Computer Society Conference on Computer Vision and Pattern Recognition*, 1996, pp. 586-593, doi: 10.1109/CVPR.1996.517132.

[28] Lahlali, S.E. & Sadiq, Abdelalim & Mbarki, Samir. *A review of face sketch recognition systems.* 2015. 81. 255-265.

[29] S. Klum, H. Han, A. K. Jain and B. Klare, "Sketch-based face recognition: Forensic vs. composite sketches," *2013 International Conference on Biometrics (ICB)*, 2013, pp. 1-8, doi: 10.1109/ICB.2013.6612993.

[30] A. K. Jain, B. Klare and U. Park, "Face recognition: Some challenges in forensics," 2011 IEEE *International Conference on Automatic Face & Gesture Recognition (FG)*, 2011, pp. 726-733, doi: 10.1109/FG.2011.5771338.

[31] Xiaoou Tang and Xiaogang Wang, "Face sketch recognition," in IEEE *Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 50-57, Jan. 2004, doi: 10.1109/TCSVT.2003.818353.

[32] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu and Songde Ma, "A nonlinear approach for face sketch synthesis and recognition," 2005 IEEE *Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'05), 2005, pp. 1005-1010 vol. 1, doi: 10.1109/CVPR.2005.39.

[33] Xiaoou Tang and Xiaogang Wang, "Face sketch synthesis and recognition," Proceedings Ninth IEEE *International Conference on Computer Vision*, 2003, pp. 687-694 vol.1, doi: 10.1109/ICCV.2003.1238414.

[34] Yung-hui Li, M. Savvides and V. Bhagavatula, "Illumination Tolerant Face Recognition Using a Novel Face From Sketch Synthesis Approach and Advanced Correlation Filters," 2006 IEEE *International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, pp. II-II, doi: 10.1109/ICASSP.2006.1660353.

[35] H. K. Galoogahi and T. Sim, "Inter-modality Face Sketch Recognition," 2012 IEEE *International Conference on Multimedia and Expo*, 2012, pp. 224-229, doi: 10.1109/ICME.2012.128.

[36] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," CVPR 2011, 2011, pp. 593-600, doi: 10.1109/CVPR.2011.5995350.

[37] J. Choi, A. Sharma, D. W. Jacobs and L. S. Davis, "Data insufficiency in sketch versus photo face recognition," 2012 IEEE *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1-8, doi: 10.1109/CVPRW.2012.6239208.

[38] Zhu, Jun-Yan & Park, Taesung & Isola, Phillip & Efros, Alexei. *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.* 2017. 2242-2251. 10.1109/ICCV.2017.244.

[39] Lichao Mou, M. Schmitt, Yuanyuan Wang and Xiao Xiang Zhu, "*A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes,"* 2017 Joint Urban Remote Sensing Event (JURSE), 2017, pp. 1-4, doi: 10.1109/JURSE.2017.7924548.

[40] Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y. 2014. *Generative Adversarial Networks. Advances in Neural Information Processing Systems.* 3. 10.1145/3422622.

[41] Zhao, Junbo & Mathieu, Michael & Lecun, Yann. *Energy-based Generative Adversarial Network.* 2016.

[42] Radford, Alec & Metz, Luke & Chintala, Soumith. *Unsupervised Representation*

Learning with Deep Convolutional Generative Adversarial Networks. 2015.

[43] Zhu, Jun-Yan & Krähenbühl, Philipp & Shechtman, Eli & Efros, Alexei. 2016. *Generative Visual Manipulation on the Natural Image Manifold.* 9909. 10.1007/978-3-319-46454-1_36.

[44] Salimans, Tim & Goodfellow, Ian & Zaremba, Wojciech & Cheung, Vicki & Radford, Alec & Chen, Xi. *Improved Techniques for Training GANs*. 2016.

[45] Mathieu, Michael & Zhao, Junbo & Sprechmann, Pablo & Ramesh, Aditya & Lecun, Yann. *Disentangling factors of variation in deep representations using adversarial training.* 2016.

[46] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. *Generative adversarial text to image synthesis.* 2016.

[47] Pathak, Deepak & Krahenbuhl, Philipp & Donahue, Jeff & Darrell, Trevor & Efros, Alexei. *Context Encoders: Feature Learning by Inpainting.* 2016. 2536-2544. 10.1109/CVPR.2016.278.

[48] Mathieu, Michael & Couprie, Camille & Lecun, Yann. *Deep multi-scale video prediction beyond mean square error.* 2015.

[49] Vondrick, Carl & Pirsiavash, Hamed & Torralba, Antonio. *Generating Videos with Scene Dynamics.* 2016.

[50] Wu, Jiajun & Zhang, Chengkai & Xue, Tianfan & Freeman, William & Tenenbaum, Joshua. *Learning a Probabilistic Latent Space of Object Shapes via 3D*

Generative-Adversarial Modeling. 2016.

[51] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE *International Conference on Computer Vision (ICCV)*, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[52] L. Wang, V. A. Sindagi, and V. M. Patel, "High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks," *arXiv:1710.10182 [cs]*, Mar. 2018, Accessed: Sep. 15, 2021. [Online]. Available: http://arxiv.org/abs/1710.10182

[53] K. Panetta *et al.*, "A Comprehensive Database for Benchmarking Imaging Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509–520, Mar. 2020, doi: 10.1109/TPAMI.2018.2884458.

[54] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in ICCV, 2005.

[55] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in CVPR, 2006.

[56] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in ICCV, 2003.

[57] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in CVPR, 2009.

[58] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear

coding for image classification," in CVPR, 2010.

[59] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in ECCV, 2010.

[60] K. E. van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in ICCV, 2011.

[61] N. Mahamkali and V. Ayyasamy, "OpenCV for Computer Vision Applications," Mar. 2015.

[62] L. A. Gatys, A. S. Ecker, and M. Bethge. *Image style transfer using convolutional neural networks.* CVPR, 2016.

[63] S. Sremath Tirumala and A. Narayanan, "Hierarchical Data Classification Using Deep Neural Networks," Nov. 2015, vol. 9489, pp. 492–500. doi: 10.1007/978-3-319-26532-2_54.

[64] Z. Li, W. Yang, S. Peng, and F. Liu, *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects, arXiv:2004.02806 [cs, eess]*, Apr. 2020, Accessed: Sep. 15, 2021. [Online]. Available: http://arxiv.org/abs/2004.02806

[65] J.-C. Vialatte, V. Gripon, and G. Mercier, *Generalizing the Convolution Operator to extend CNNs to Irregular Domains,* arXiv:1606.01166 [cs], Oct. 2017, Accessed: Sep. 15, 2021. [Online]. Available: http://arxiv.org/abs/1606.01166

[66] Xia, C, Xiong, C, & Yu, P. S.. *Pseudo Siamese Network for Few-shot Intent Generation. ACM SIGIR*, (). Retrieved from https://par.nsf.gov/biblio/10228570.

[67] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: *A unified embedding for face recognition and clustering*. In Proc. CVPR, 2015.

[68] Parkhi, O., Vedaldi, A., & Zisserman, *A. Deep Face Recognition. BMVC*. 2015.

[69] A. Nandy, S. Haldar, S. Banerjee and S. Mitra, "A Survey on Applications of Siamese Neural Networks in Computer Vision," *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9153977.

[70] S. Roy, M. Harandi, R. Nock and R. Hartley, "Siamese Networks: The Tale of Two Manifolds," 2019 IEEE/*CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3046-3055, doi: 10.1109/ICCV.2019.00314.

[71] Gatys, L. A., Ecker, A. S. & Bethge, M. *A Neural Algorithm of Artistic Style*. 2015.