Siamese Network-Based System for Criminal Identification

Soham Dave, Parth Kansara, Vinaya Sawant and Shivam Mehta

September 5, 2022

Siamese Network-based system for criminal identification

Soham Dave[1], Parth Kansara[2], Dr. Vinaya Sawant[3], Shivam Mehta[4]

[1,2,3] *Department of Information Technology*

Dwarkadas J. Sanghvi College of Engineering

Mumbai, India

[4] *Technology Research and Innovation Office*

Mastek Ltd., Navi Mumbai, India

dave.soham2000@gmail.com

parthskansara@gmail.com

vinaya.sawant@djsce.ac.in

iamshivam27@gmail.com

Abstract

To discourage criminal activities, the large number of CCTV installations throughout the country play a crucial role. Through this paper, we propose an AI-based solution that can leverage these devices to remotely identify and report absconding criminals. Using the one-shot learning approach, we present a face recognition algorithm that yields accurate results even with low training data. The Siamese Network architecture is used to verify if the face embeddings of the image detected is the same as that of the criminal. Two parallel neural networks are designed to take one input each- one being the detected face and the other being an embedding from the dataset. The outputs of the two networks are compared to predict whether the detected face is the same as the input face or not. This algorithm is further integrated with an automated model for updating the information of the recognized criminal into the database along with updating the appropriate law enforcement authorities about the last known whereabouts.

Keywords:
Computer Vision, Machine Learning, Face Recognition

1. Introduction

Face detection and recognition is one of the most important biometric authentication technologies that may be used for both verification and surveillance. Face recognition is becoming a critical system for us as the number of scams rises every day. Face recognition has always been a difficult and time-consuming process. Its true test is to create an automated system that can recognize faces in the same way that humans can. Regardless, there is a limit to human capacity when it manages a large number of hidden to appearances tasks.

As a result, a pre-programmed automated electronic framework with greater identification accuracy and processing speed is required. The conventionally used face recognition algorithms have numerous drawbacks. Majority of these [2, 11, 12, 13, 14] require a balanced dataset with adequate representation of each class, and need hours of training on computationally extensive hardware, which is also costly. Thus, there is a restriction on the application of such algorithms in real life use-cases. In practical applications, it is not feasible to have a rich, balanced dataset that can produce good results with the conventional algorithms. In most cases, the authorities might have just one photograph that is used for the identification of the criminal. Thus, it becomes imperative to include an algorithm that can work with such scarce data.

One shot learning [3] is an object classification paradigm, which aims to resolve the drawback of the classical machine learning-based object classification algorithms that requires hundreds or thousands of training data samples. Using one-shot learning, specifically the Siamese network, it is possible to ascertain the similarity or dissimilarity between two images. It consists of two networks that each produce an output based on an input image. If the two outputs are at a Euclidean distance less than a predefined threshold, we can term the input images are similar. This allows us to match a test image with just one image in the training dataset, thus reducing the necessity for multiple images required by the conventional face recognition algorithms.

2. Related Work

This paper focuses on face recognition from images to identify criminals. Overall, this domain has received a massive attention in the past. The existing methods can be bifurcated into two types: using deep learning and without using deep learning.

A. *Image Descriptors for Face Recognition*

The methods that do not use deep learning extract a representation of the image of the face using a local image descriptor. These image descriptors are algorithms that take an input image and encode its information into a series of numbers. Outputs are usually in the vector format. Some commonly used image descriptors include SIFT [6], LBP [4] and HOG [7]. Such local descriptors are then summed to obtain an overall face descriptor via a pooling mechanism like the Fisher Vector [8,9].

B. *Deep Learning based Face Recognition*

A more widely used category of face recognition methods is the one that utilizes deep learning. The defining characteristic of such methods is the use of a CNN feature extractor, a learnable function obtained by composing several linear and non-linear operators. Some popular algorithms in this category include the DeepFace [10], which uses an ensemble of CNNs during the pre-processing phase, followed by a deep CNN trained on a dataset having 4 million images of 4000 unique individuals. This was extended by the DeepID, which was described in a series of papers [11, 12, 13, 14]. Several concepts were also included in this series, including using various CNNs [12], a Bayesian learning framework [15] to train a metric, multi-task learning over classification and verification [11], varied CNN architectures which branch a fully connected layer after each convolution layer [13], and very deep networks in [14]. Unlike DeepFace, DeepID uses a simple 2D affine alignment and is trained on a combination of two datasets, CelebFaces [12] and WDRef [15].

C. *One-shot Learning*

Another variation of these deep learning methods is a method known as one-shot learning. Gregory Koch et al. [16] offer a unique supervised metric-based technique for character recognition using Siamese neural networks, then reuse the features of that network for one-shot learning without retraining. They use massive Siamese neural networks that can learn general visual characteristics to forecast unknown class distributions even when there is a scarcity of examples from these new distributions. The pairs chosen from the source data will then be trained using standard optimization methods and give a competitive strategy that does not depend on domain-specific knowledge, by utilizing deep learning techniques. The model identifies pairs of inputs based on whether they are of the same class or not. This model may then be used to pairwise evaluate fresh pictures against the test image. The pairing with the highest is then awarded the highest probability for the one-shot task. If the learnt characteristics are adequate to affirm or reject the identification of letters from one set of alphabets, they should be sufficient for other alphabets, assuming the model has been exposed to a range of alphabets to encourage variance among the learned features. They discovered that the new model outperformed a number of existing models, and they recommend that one-shot learning challenges be extended.

D. Face Recognition for Biometric Identification

There are several works that have attempted to apply face recognition algorithms for the task of biometric identification. The research in [21, 22] explains a classroom attendance management system created by applying Viola-Jones method for detection and Eigenface for identification. The work in [23] demonstrates an employee attendance system that leverages Haar-Cascade classifier for detection, along with PCA for identification. This method was able to achieve an accuracy of 68%, which is not very reliable. These papers have used the conventional algorithms for face recognition which either require large training datasets or compromise on the accuracy. The proposed method in this paper aims to resolve both these problems by providing good accuracy from a scarce training dataset.

3. Methodology

The proposed system can be better understood in two parts. First is the algorithm for criminal identification that is used. Second is the integration of this into an end-to-end system, that logs the information of the identified criminal into a central database and notifies the authorities via email, text messages and on the portal developed.

A. Criminal Identification

When a new criminal is entered into our database, we enter the name and the photo identification of the criminal, along with any other relevant details stored in a text format.

Face Extraction: Firstly, the image is converted to a grayscale image to reduce computation by eliminating the color information. As shown in Fig. 1, the image of the criminal is then given to a Haar Cascade classifier [1], to extract the face of the criminal from the entire image. We use a pre-trained model, which detects the eyes and faces in each image.

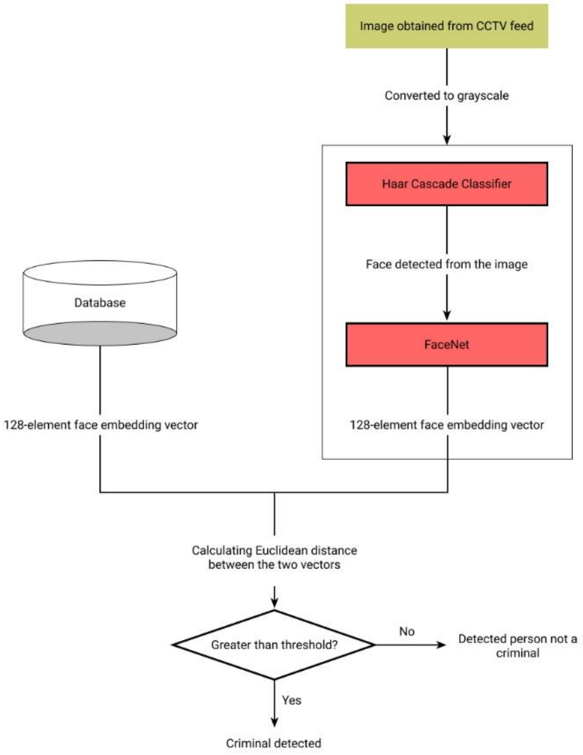


Image obtained from CCTV feed
Converted to grayscale
Haar Cascade Classifier
Face detected from the image
FaceNet
128-element face embedding vector
Database
128-element face embedding vector
Calculating Euclidean distance between the two vectors
Greater than threshold?
No
Detected person not a criminal
Yes
Criminal detected

Fig. 1 Training phase

The Haar cascade classifier has been trained on numerous ‘positive’ and ‘negative’ images. Positive images contain faces and help the classifier learn to detect them. Negative images do not have any inclusion of faces.

The first step in training the classifier is calculating the Haar features [1]. A Haar feature is a set of computations performed on consecutive rectangular areas of a detection window at a given position, as illustrated in Fig. 2. The calculation is adding up the pixel intensities in each region and then finding the difference of the sums. These features might be difficult to spot in a big picture. Integral pictures are useful in this situation since they decrease the number of processes required.

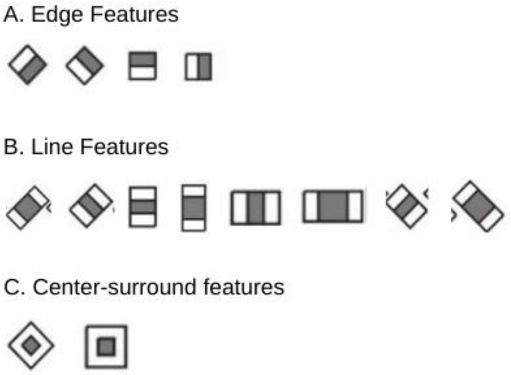
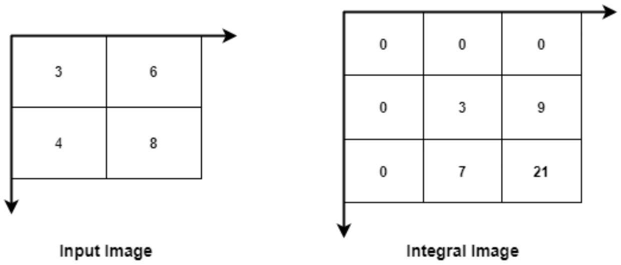


Fig. 2 Various Haar features

The total of the pixel values of the original image is defined as an integral image. The value of the integral image at any position (X,Y) is the total of the image's pixels above and to the left of the position (X,Y).



3	6
4	8

Input Image

0	0	0
0	3	9
0	7	21

Integral Image

Fig. 3 Calculation of integral image

A subset of the detected Haar features that can perform better are chosen by a feature selection technique. Simultaneously, the irrelevant features are eliminated. Adaboost is used to choose and train classifiers to use the best characteristics. Weak learners are first produced by sliding a window across the input picture and calculating Haar features for each part of the picture. This difference is contrasted with a threshold, which distinguishes objects from objects. These weak learners are subsequently merged into powerful, cascading learners.

A cascade classifier is made up of a series of stages, where each stage is a collection of weak learners. A window is moved over image which can help detect a face. At each stage, the classifier labels the specific region defined by the current location of the window as positive if a face is found and negative if a face is missing. If a negative label is generated, then the window is moved on. If a positive label is generated, then the same region is moved on to the next stage of classification. The requirements for getting a positive label get stricter with every passing stage. If a positive label is achieved in the final stage, the classifier declares a detected face in the image. Fig 4.1 illustrates the working of a cascade classifier. Using this, we can detect a face as shown in Fig 4.2.

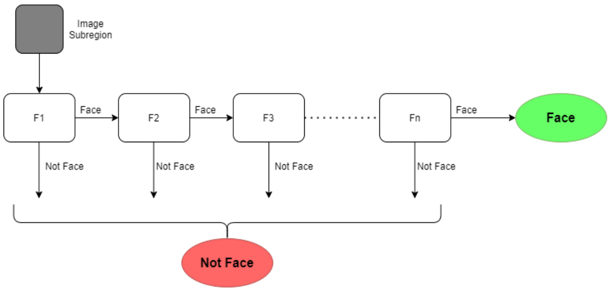


Fig. 4.1 Operation of cascading filters

Fig. 4.2 (a) Input Image

(b) Detected face

Next, this image is passed to a pre-trained Keras FaceNet model, which comprises the Siamese Network. It extracts high-quality features from the input image and predicts a 128-element vector representation of these features, called the face embedding. This is crucial information that we use to create our training dataset. Each criminal entered into the database has their image converted into a 128-element vector which is then mapped back with their names and added to the training dataset. This training dataset is updated every time a new criminal is entered the system.

FaceNet [2] is a deep neural network used for extracting features from an image of a person’s face. It utilizes convolutional layers to directly generate representations of the facial pixels. This network was trained to achieve invariance of lighting, poses and other variables on a huge data set. The Labelled Faces in the wild (LFW) dataset was used for training. This dataset comprises almost 13,000 pictures of different faces from the web, each with a name (label).

FaceNet generates a 128-dimensional embedding from images and inserts it into a feature space such that the squared distance between all faces of the same identity, irrespective of imaging conditions, is small, whereas the squared distance between a pair of faces from different people is large.

The L2 distance between pictures of the same identity is minimized, while the L2 distance between the facial images of distinct characters is maximized, using a loss function called triplet loss, as shown in Fig. 4. The designers came up with a clever triplet selection system that chooses three pictures at once. These pictures are divided into three categories:
1. anchor: an image of an arbitrary individual.
2. positive image: another image of the same individual.
3. negative image: an image of a different individual.

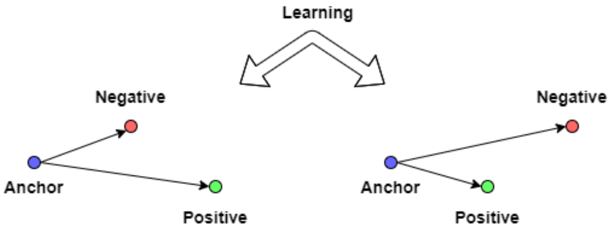


Fig. 5 Visualizing triplet loss function

Triplet loss function can be formally defined as:

$$\sum_i^N \left[\left\|f(x_i^a) - f\left(x_i^p\right)\right\|_2^2 - \left\|f(x_i^a) - f(x_i^n)\right\|_2^2 + \alpha\right]_+ \qquad (1)$$

Here, xi represents an image and f(xi) represents the embedding of an image. The superscript letters a, p, n over x correspond to anchor, positive and negative image respectively. α is used to represent the margin between the positive and negative pairs.

Two Euclidean distances are calculated: one between the anchor and the positive image and another between the anchor and the negative picture. The training procedure tries to minimize the former while increasing the latter, so that comparable pictures are near together, and different images are far apart in the embedding space.

Suppose the anchor image has a face embedding A = (x1, x2,,, x128) and the positive image has a face embedding B = (y1, y2,,, y128). Then the Euclidean distance between two points can be measured as:

$$\text{AB} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_{128} - y_{128})^2} \qquad (2)$$

Next, we run a Haar Cascade classifier on the CCTV feed to extract the face detected, as shown in Fig. 6. This face is then passed through the same FaceNet model to yield a corresponding 128-element vector, i.e., the face embedding. Now, we iteratively calculate the Euclidean distance between the new vector and the existing vectors in our dataset. If this distance is less than a specified threshold, the model will classify the two faces to be a match.

B. Integrated System

The criminal identification module is integrated into a portal. This portal makes it easy for the end-users to enter the criminal details into the system and receive updates. The UI has been designed to make it easier to use.

The system, as shown in Fig. 6, uses Firebase to maintain the database for the criminals. PyMySQL has been used to interact with this database and run queries on it. In case the criminal identification module gets a match, the system retrieves the details from the database and sends a notification using the Twilio Messaging API via email, text message to the mentioned point of contacts. A notification is sent out to all users on the portal aswell.

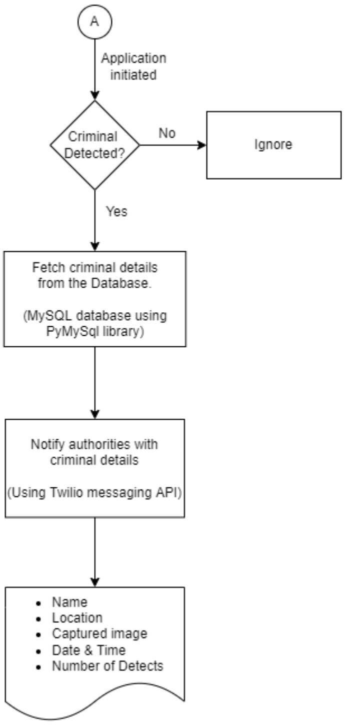

A
Application initiated
Criminal Detected?
No
Ignore
Yes
Fetch criminal details from the Database.
(MySQL database using PyMySql library)
Notify authorities with criminal details
(Using Twilio messaging API)
Name
Location
Captured image
Date & Time
Number of Detects

Fig. 6 Representation of the integrated system

4. Results

Our dataset consisted of 10 people; out of which 2 people were the authors and the remaining 8 were popular celebrities. Several iterations of the algorithm were performed to find out the optimum threshold for the Euclidean distance that can be used to label the pair of faces as matching.

During the training phase, we considered two separate cases:

1. Using 1 training image per class
2. Using 3 training images per class

To compare the accuracy with the conventional face recognition algorithms, we consider model using Local Binary Pattern Histograms (LBPH), similar to the model used by Ahonen et al. [20]. We train this model on the same dataset, with over 100 images for each person.

The following table shows the comparison of the accuracy achieved in all cases.

Table I. Comparison of results

	Case 1	**Case 2**	**LBPH**
No. of classes	10	10	10
No. of training images	10	30	100
No. of testing images	90	70	20
Training accuracy	80.23%	89.35%	92.51%
Testing accuracy	82.17%	90.01%	93.63%

Evidently, as the number of training images was increased, the accuracy increased. At just 3 images per class, a reliable accuracy was achieved. Although this does not match the accuracy of a conventional model, the reasonably close numbers suggest that our model can be used in adverse cases where there is a lack of training data.

Further, the integrated system was tested by 8 users. Two of these were from technical background; three were from non-technical background but used a computer daily and the remaining three were people who rarely used a computer. Each user performed a task which consisted of uploading a criminal's image to the database and reading the notification of a detected criminal from the system.

After the exercise, each user was asked to fill out a survey which included questions about the UI, the ease of use and the usefulness of the system. Based on this survey, additional changes to the integrated system were incorporated to make it a better experience for the ultimate end-user.

5. Conclusion

In this paper, we employ one-shot learning using the FaceNet model. Faces are detected using the Haar Cascade classifier and then passed on to a pre-trained model which can compress the image into a 128-element vector representation. These 128-element vectors are compared using Euclidean distances and classified based on a threshold.

This algorithm is integrated into a system that makes it easier to log, detect and report absconding criminals. Targeting the specific problem of limited training images, a large-scale implementation of this system could prove highly effective.

References

[1] Viola, P. and Jones, M., 2001, December. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.

[2] Schroff, F., Kalenichenko, D. and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[3] Fei-Fei, L., Fergus, R. and Perona, P., 2006. One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence, 28(4), pp.594-611.

[4] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. AAAI, 2015.

[5] Parkhi, O.M., Vedaldi, A. and Zisserman, A., 2015. Deep face recognition.

[6] Gao, Y. and Lee, H.J., 2015. Cross-pose face recognition based on multiple virtual views and alignment error. Pattern Recognition Letters, 65, pp.170-176.

[7] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: Video shot retrieval for face sets. In Proc. CIVR, 2005

[8] O. M. Parkhi, K. Simonyan, A. Vedaldi, and A. Zisserman. A compact and discriminative face track descriptor. In Proc. CVPR, 2014.

[9] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In Proc. BMVC., 2013

[10] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deep-Face: Closing the gap to human-level performance in face verification. In Proc. CVPR, 2014

[11] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. NIPS, 2014.

[12] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In Proc. CVPR, 2014.

[13] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014.

[14] Y. Sun, L. Ding, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. CoRR, abs/1502.00873, 2015.

[15] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In Proc. ECCV, pages 566–579, 2012.

[16] Koch, G., Zemel, R. and Salakhutdinov, R., 2015, July. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop (Vol. 2).

[17] Guo, Y. and Zhang, L., 2017. One-shot face recognition by promoting underrepresented classes. arXiv preprint arXiv:1707.05574.

[18] Hariharan, B. and Girshick, R., 2017. Low-shot visual recognition by shrinking and hallucinating features. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3018-3027).

[19] Jadhav, A., Namboodiri, V.P. and Venkatesh, K.S., 2016, October. Deep attributes for one-shot face recognition. In European Conference on Computer Vision (pp. 516-523). Springer, Cham.

[20] Ahonen, T., Hadid, A. and Pietikäinen, M., 2004, May. Face recognition with local binary patterns. In European conference on computer vision (pp. 469-481). Springer, Berlin, Heidelberg.

[21] Balcoh, N.K., Yousaf, M.A., Ahmad, W. & Baig, M.I., 2012. Algorithm for Efficient Attendance Management: Face Recognition Based Approach. International Journal of Computer of Science Issues (IJCSI), IX(4), pp.146-50.

[22] Febrianto, A.J., 2012, Pengenalan Wajah Dengan Metode Principle Component Analysis

(PCA) Pada sistem Absensi Real Time, Tesis, Magister Teknik Elektro, Universitas Gadjah Mada, Yogyakarta.

[23] Tharanga, J.G.R., Samarakoon, S.M.C. & Karunarathne, T.A.P., 2013, Smart Attendance using Real Time Face Recognation (Smart-FR), SAITM Research Symposium on Engineering Advancement, Sri Lanka.