International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Video based Face Recognition with limited resources

Yogini Patil, V. M. Barkade

Abstract*: Face recognition(FR) has multi-domain applications and video based FR is good area of research in terms of accuracy and performance. Recent research has proved that Convolutional Neural Network (CNN) is a one of best solution for object detection and recognition as it extracts features from on its own and performs classification as well. As we go for higher accuracy models, size of network increases and it requires more time to process a frame or video as it involves more computations. This paper aims at building a FR model which is smaller in network size, requires limited resources while building and still achieves good accuracy. The system uses combination of deep learning and machine learning based solution. FR system is built with CNN-Suport Vector Machine(SVM) model where CNN performs feature extraction and SVM performs classification task. Results shows that CNN-SVM model gives higher accuracy (94.05% validation and 90.17% testing) compared to conventional CNN-softmax model (93.37% validation and 88.77% testing) with a small network size and also requires less training time. Results can be improved by using cross validation techniques.

***Index Terms*: Face Recognition from Video, Convolutional Neural Network, Support Vector Machine, Feature Extraction, Classification**

I. INTRODUCTION

Face recognition is a task of identifying a face from digital media like still image or video sequence. Face recognition from images has been widely studied in past few years to achieve impressive performance. Regular use of cameras for video recording is changing the direction of face recognition research from still images to videos sequences. (i) Easy availability and enormous use of mobile devices for capturing videos and (ii) necessity of using surveillance cameras at public places has led to very big quantity of video being captured regularly. Video encompasses additional information in comparison with face images such as temporal and multi-view information and are more tough to recognize faces from as there are variations in pose, expression, illumination etc. [19]. All these facts makes face recognition from video an interesting area of research. Face recognition technique is used as base and extended in variety of applications across multiple disciplines. Like in security domain a surveillance system coupled with face recognition, access control for electronic transactions as biometric technique. Smart homes, human-machine interaction, interactive movies, computer games, image/ video search are few more systems using face recognition [19].

Face recognition is part of Computer vision technology and consists of two major processes - face (object) detection and face (object) recognition. Different approaches have been proposed for face detection in [8] [10] [13] [14] which can be classified as knowledge based, feature invariant, template based and appearance based methods. Most of the recent methods which are proved very effective are appearance based methods and uses machine learning and deep learning techniques like multi-layer perceptron, Principal Component Analysis(PCA), SVM, Naïve Bayes, Hidden Markov Model (HMM) etc. to overcome challenges like speed, pose variation, scale etc. There are different machine learning and deep learning approaches [3] [6] [7] [9] [11] [12] proposed for face recognition which have drawbacks that it requires more computation time and are not efficient across real-time systems. Building these face recognition models requires considerable amount of resources in terms of training dataset, graphical processing unit (GPU) capacity, training time and network size. Also, when size of network is large, it takes more time to process a frame or video which affects performance of system [18]. In order to solve this problem, a combinational approach is defined for face recognition problems using machine learning and deep learning based solution with limited resources like training dataset and GPU capacity. Feature extraction and classification are the two major processes of the solution. Traditional CNN-softmax model and combinational CNN-SVM model are trained on same dataset and results are compared. Rest of the sections of this research paper are organized as – section II contains literature survey of few recent video based face recognition systems. Section III gives details of methodology in terms of system architecture and algorithm. Section IV specifies experimental results and section V derives conclusion.

II. LITERATURE SURVEY

In [1], author proposed a deep learning methodology to deal with computer vision problem where objects are need to be detected and recognized from real time videos. Solution defines a learning model using combination of CNN and SVM. CNN is used to perform feature extraction from video.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Yogini Patil*, Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune, (M.H.), India.
V. M. Barkade, Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Pune, (M.H.), India.

SVM classifier uses these extracted features to perform classification process. In [2], proposed an architecture for Object detection and is named as YOLO (You Only Look Once). Object detection is modelled as regression and solved by using CNN model which performs object detection in single evaluation from full images. Model outputs bounding boxes around objects and each bounding box has corresponding class probability. It is very fast solution with reduced false positives on background.

In [3], procedure to assemble large scale dataset is presented and new assembled dataset is used for training CNN for face recognition and results are compared to benchmark datasets like Labelled Faces in Wild (LFW) and You Tube Faces (YTF). It is observed that results obtained are comparable to state of the art results.

In [4], author proposed a CNN-SVM model for image classification and compared its results with conventional CNN-Softmax model.

In [5], Trunk Branch Ensemble Convolution Neural Network (TBE-CNN) with improved triplet loss function is used. This approach is effective in dealing with problem of occlusion, pose variation and blur images for video based face recognition. This model consists of one main network (trunk) and multiple branch networks. Features of general face area are extracted by main network while features from cropped face area are extracted by branch networks.

Study presented in this paper is an extension of combinational approach used in [1] for object recognition from video. For face detection, faced face detection model which is heavily based on YOLO object detection model [2] is used which will output a single class probability and associated bounding box [18]. Feature extraction using CNN makes use of squared hinge loss to implement SVM for classification at last layer of CNN [4].

III. METHODOLOGY

A. *Architecture*

Face Recognition System (FRS) is explained in Fig. 1. Input to FRS is a video file from which frames are captured. Each frame is processed to detect and crop face area. Cropped face area is then given as input to learning model which extracts facial features and performs classification.

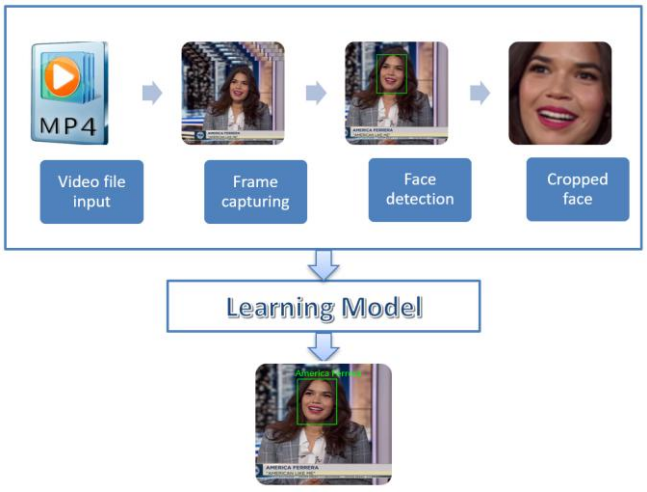


Figure 1. Face Recognition System

Workflow of FRS:

1. Input video file.
2. Capture frames from video file.
3. Detect face and crop face area.
4. Pre-process face area to resize to required shape and perform normalization.
5. Use CNN-SVM learning model to extract features on its own and perform classification.
6. Annotate frame with bounding box around face and recognized person name over it.

High level architecture of system is given in Fig. 2.

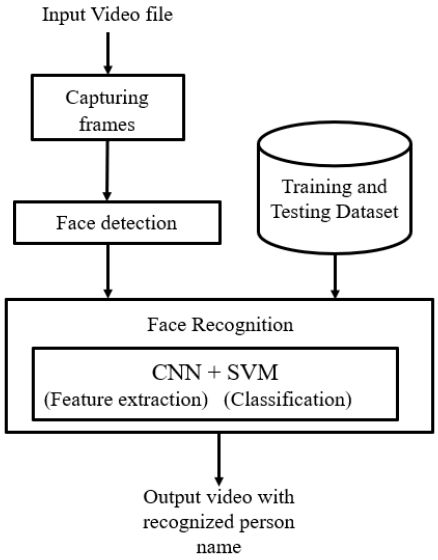


Figure 2. FRS Architecture

FRS accepts a video file as input. Output of FRS is a video file with bounding box drawn around face and name of recognized person written over it.

Dataset is set of images used to train learning model. It is partitioned into two disjoint sets for training and testing. Model is trained using FaceScrub dataset. FaceScrub dataset contains 106,863 images of 530 people [15].

URL: http://vintage.winklerbros.net/facescrub.html

Face detection architecture is heavily based on YOLO model. It is CNN model which accepts 288x288 input image and consists of series of convolutional and pooling layers. Model's output is a 9x9 grid. Each grid cell is in charge of predicting whether a face is inside that cell. Each grid cell has 5 associated values. The first one is the probability p of that cell containing the center of a face. The other 4 values are the (x center, y center, width, height) of the detected face (relative to the cell)

Face recognition is done using combinational approach of CNN-SVM. CNN performs feature extraction and SVM performs classification.

For human face, features could be position and shape of eyes, nose, lips, chin, eyebrow etc. Every person has unique facial features. As human face is flexible and deformable object, these features (measurements) varies slightly with different pose and expression.

Retrieval Number: B3375078219/2019©BEIESP
DOI: 10.35940/ijrte.B3375.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Extracting right set of features is very crucial because classification task is entirely dependent on it. CNN has a great property of extracting features on its own. In CNN, every layer applies set of filters and activations on input image which produces feature maps.

Features learned from previous layer are again applied with set of filters and activations to produce feature maps of next layers. This way deeper layers produces high level features. These features act as better input to classification process as it combines low level and high level features.

Output feature vector from CNN is used as input for classification. SVM performs classification by separating hyperplane. When given labelled training samples in n dimensional space, SVM finds optimal hyperplane separating training samples such that given any new test sample of unknown class label, SVM predicts its class label accurately depending on its position in n dimensional space [16]. Support Vector Machines are very effective where number of dimensions is high. It is even more effective on data sets where number of dimensions is greater than the number of training samples. Also, Support Vector Machines are characterized with memory efficiency, speed and accuracy in comparison to other classification methods like k-nearest neighbour or deep neural networks [17].

B. *Algorithm*

CNN is a neural network with series of convolutional and pooling layers. There is no standard order of these layers. However, it varies for different model. Adam optimizer is used for parameter optimization and pseudocode for same is given below-

-Initialize weights w randomly
-while <termination condition> do
  -for <each training sample in a batch> do
    - Calculate feature vector by forward propagation
    - Calculate output at last layer of CNN and error
  -end for
  -Calculate total loss
  -Calculate gradient of loss which is partial derivative of loss with respect to weights
  -Update parameters $w_{new} = w_{old} + \eta * \text{gradient of loss}$ (where, η is learning rate)
-End while

Loss is calculated using squared hinge loss and below is equation for same-

$$\min \frac{1}{m} \|w\|^2 + C \sum_{i=1}^{m} \max(0, 1 - y_i . \hat{y}_i)^2 \qquad (1)$$

Where,
m - number of training examples,
w – parameters
C – constant penalty parameter
y_i – actual label
$\hat{y}_i$ – predicted label

IV. RESULTS

A. *Experimental Setup*

Dataset: Learning model is trained using images from FaceScrub dataset. FaceScrub dataset contains over 1 million images of 530 people [15]. However, most of the image URLs are broken. In order to have balanced data, selected 19 classes with 170 images per class. These classes includes 8 actresses and 10 actors. Dataset is partitioned into two disjoint sets for training and testing with 155 and 15 images per class respectively. FaceScrub dataset provides coordinates of face bounding boxes along with image. Hence, direct face image cropped using provided bounding boxes is considered for training instead of whole image.

This dataset is challenging for face recognition task as a single class contains images with variations in pose, illumination, expression and occlusion(makeup, specs, moustache and beard) which are general challenges faced in video based face recognition. Fig. 3 shows variations in face image for a single class.

(a) America Ferrera (b) Bill Hader

Figure 3. Images from FaceScrub dataset

Model: CNN architecture used in this study is detailed in Fig. 4.

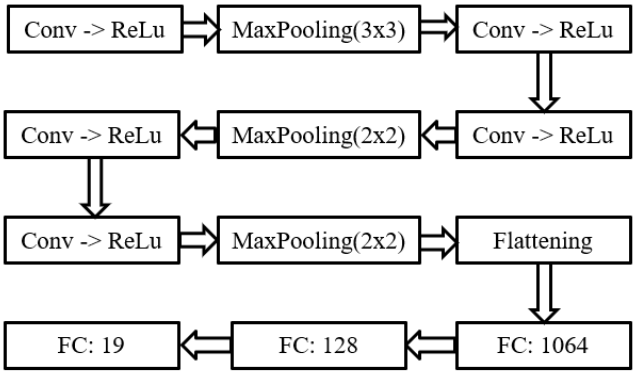


Figure 4. CNN Architecture

For CNN-softmax model, activation function at last fully connected layer is 'softmax', loss function is 'categorical crossentropy' and label encoding is in form of {0,1}. With same CNN architecture, CNN-SVM model uses no activation function at last fully connected layer (to make it linear SVM), squared hinge loss function and label encoding is in form of {1, -1}. Hyperparameters used in both models are given in table I. Size of network in terms of parameters for Face Detection model is 6,993,517 and that for Face Recognition model is 3,824,059. Implementation is done using python 3.6 with Tensorflow as backend and using Keras libraries. Hyper-parameters listed in table 1 were set manually. All experiments in this study were conducted on a laptop computer with Intel® Core™ i5-8250U CPU @ 1.60GHz 1.80GHz, 8GB of RAM, and 2GB NVIDIA GeForce MX150 card with 384 CUDA Cores.

Table I. Hyperparameters

Hyper-parameters	CNN-Softmax	CNN-SVM
Learning rate	1e-4	1e-4
Epoch	1850	1850
Batch size	32	32
SVM C	NA	Default (1)

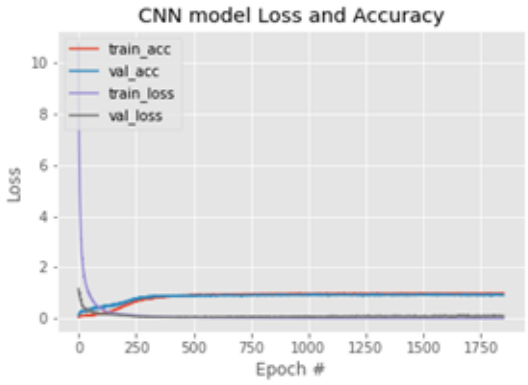
B. *Observations*

Table II compares accuracy and training time observed for both the models.

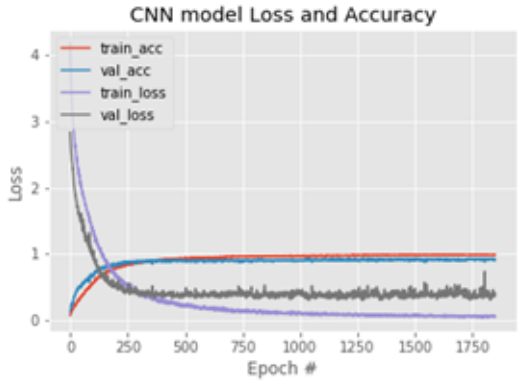
Table II. Comparison of Models

Model	Training Accuracy	Validation Accuracy	Testing Accuracy	Training Time
CNN-Softmax	100%	93.37%	88.77%	168 min
CNN-SVM	100%	94.05%	90.17%	152 min

Fig. 5 shows accuracy curves for (a) CNN-SVM and (b) CNN-softmax model.



(a) CNN-SVM Model



(b) CNN-softmax Model

Figure 5. Accuracy Curve

Figure 6 shows how feature map looks like at different convolutional layers. Figure 6 (a) is original image resized to 64x64. Figure 6 (b), (c), (d), (e), (f) are feature maps captured after 1st , 2nd , 3rd, 4th and 5th convolutional layers. It can be easily observed that after initial few convolutional layers image can be visually interpreted. However, as network goes deeper and more deep and complex features are extracted, it becomes less visually discriminative but acquires stronger discriminative features.

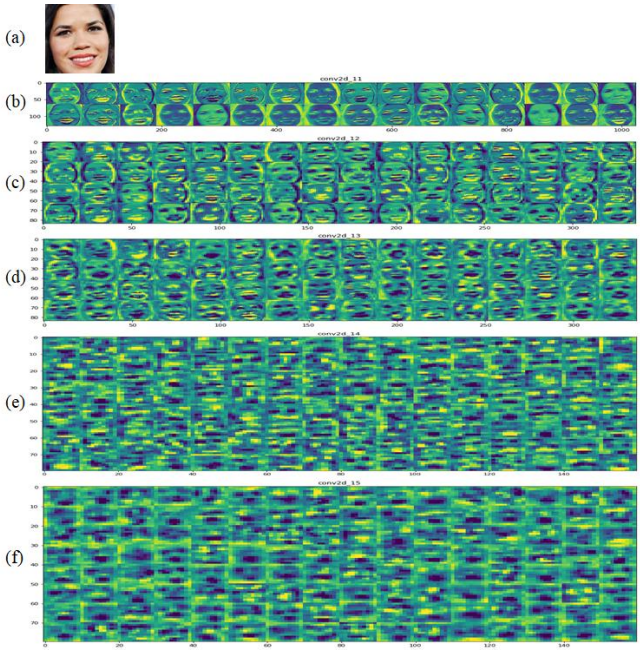


Figure 6.Features captured at different convolutional layers

V. CONCLUSION

As part of this study, implemented a Face Recognition System which accepts a video file, captures frame from it and processes each frame to detect face. Detected face is recognized using combinational approach of CNN-SVM. Output is given in form of video where bounding box is drawn around face and recognized person name is written over it. Observations from CNN-SVM model are compared with that of conventional CNN-softmax model. Results clearly shows that CNN-SVM model forms an effective solution for video based FR system. Accuracy of model can be further improved by using cross validation techniques.

REFERENCES

1. Nalinipriya, G., Balamurugan Baluswarny, Rizwan Patan, Suresh Kallam, Tamizharasi Gs, and M. Rajasekhara Babu. "To detect and Recognize Object from Videos for Computer Vision by Parallel Approach using Deep Learning." In 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 336-341. IEEE, 2018.
2. Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.
3. Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." In BMVC, vol. 1, no. 3, p. 6. 2015.
4. Agarap, Abien Fred. "An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification." arXiv preprint arXiv:1712.03541 (2017).
5. Ding, Changxing, and Dacheng Tao. "Trunk-branch ensemble convolutional neural networks for video-based face recognition." IEEE transactions on pattern analysis and machine intelligence 40, no. 4 (2018): 1002-1014.

6. Dhamija, Jalendu, Tanupriya Choudhury, Praveen Kumar, and Yogesh Singh Rathore. "An Advancement towards Efficient Face Recognition Using Live Video Feed:" For the Future"." In Computational Intelligence and Networks (CINE), 2017 3rd International Conference on, pp. 53-56. IEEE, 2017.
7. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815-823. 2015.
8. Pandit, Trupti M., P. M. Jadhav, and A. C. Phadke. "Suspicious object detection in surveillance videos for security applications." In Inventive Computation Technologies (ICICT), International Conference on, vol. 1, pp. 1-5. IEEE, 2016.
9. Le, Quoc V., Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby rochnow, and Andrew Y. Ng. "On optimization methods for deep learning." In Proceedings of the 28th International Conference on Machine Learning, pp. 265-272. Omnipress, 2011.
10. Lu, Ke, Zhengming Ding, Jidong Zhao, and Yue Wu. "Video-based face recognition." In Image and Signal Processing (CISP), 2010 3rd International Congress on, vol. 1, pp. 232-235. IEEE, 2010.
11. Hatimi, Hicham, Mohamed Fakir, and Mohamed Chabi. "Face recognition using a fuzzy approach and a multi-agent system from video sequences." In Computer Graphics, Imaging and Visualization (CGiV), 2016 13th International Conference on, pp. 442-447. IEEE, 2016.
12. Hu, Changbo, Josh Harguess, and J. K. Aggarwal. "Patch-based face recognition from video." In Image Processing (ICIP), 2009 16th IEEE International Conference on, pp. 3321-3324. IEEE, 2009.
13. Yang, Ming-Hsuan, David J. Kriegman, and Narendra Ahuja. "Detecting faces in images: A survey." IEEE Transactions on pattern analysis and machine intelligence 24, no. 1 (2002): 34-58.
14. Yang, Ming-Hsuan. "Recent advances in face detection." IEEE ICPR 2004 Tutorial (2004).
15. Ng, Hong-Wei, and Stefan Winkler. "A data-driven approach to cleaning large face datasets." In *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 343-347. IEEE, 2014.
16. Holehouse, A.S.,"12: Support Vector Machines (SVMs)," *Stanford Machine Learning*. [Online]. Available: http://www.holehouse.org/mlclass/index.html.
17. Ray, S. (2017, September 14). Understanding Support Vector Machine algorithm from examples (along with code). Retrieved from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
18. Itzcovich, I. (2019). faced: CPU Real Time face detection using Deep Learning. [online] Towards Data Science. Available at: https://towardsdatascience.com/faced-cpu-real-time-face-detectionusing- deep-learning-1488681c1602.
19. Patil, Y. and Barkade, V. (2019). Face Recognition from Video: An overview. International Journal of Management, Technology And Engineering, 9(03), pp.283-289.

AUTHORS PROFILE

Yogini Patil is a second year student of Master of Computer Engineering at JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, affiliated to Savitribai Phule Pune University. She is graduated with Bachelor of Computer Engineering from Walchand College of Engineering, Sangli. She has over 7 years of industrial experience in Software Testing from Banking and Financial Services (BFS) and Medical domain. She has published 2 research papers in reputed international conferences. Her research interests are Machine Learning, Deep Learning and Computer Vision.

Vaishali M Barkade is working as Assistant Professor in Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, affiliated to Savitribai Phule Pune University. She has completed M. Tech. in Information Technology from Bharati Vidyapeeth Deemed University College of Engineering and Bachelor of Computer Engineering from Pune University. She has 1.5 years of industrial experience and 15.5 years of teaching experience. She has published 30 research papers in reputed international journals and conferences and filed 2 patent. Her research interests are Distributed Computing, Internet of Things and Machine Learning.