International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952179

Video Based Person Re-Identification Through Selective Knowledge Distillation

Gudavalli Sai Abhilash, Kantheti Rajesh, Jangam Dileep Shaleem, Grandi Sai Sarath , Palli R Krishna Prasad
Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

ABSTRACT

The creation and deployment of face recognition models need to identify low-resolution faces with extremely low computational cost. To address this problem, a feasible solution is compressing a complex face model to achieve higher speed and lower memory at the cost of minimal performance drop. Inspired by that, this paper proposes a learning approach to recognize low-resolution faces via selective knowledge distillation in live video. In this approach, a two-stream convolution neural network (CNN) is first initialized to recognize high-resolution faces and resolution-degraded faces with a teacher stream and a student stream, respectively. The teacher stream is represented by a complex CNN for high-accuracy recognition, and the student stream is represented by a much simpler CNN for low-complexity recognition. To avoid significant performance drop at the student stream, we then selectively distil the most informative facial features from the teacher stream by solving a sparse graph optimization problem, which are then used to regularize the fine- tuning process of the student stream. In this way, the student stream is actually trained by simultaneously handling two tasks with limited computational resources approximating the most informative facial cues via feature regression, and recovering the missing facial cues via low-resolution face classification.

Keywords: Selective Knowledge Distillation, Convolutional Neural Networks, Teacher Stream, Student Stream, Sparse Representation, Feature Regression, Facial Cues

I. INTRODUCTION

One of the most common tasks in surveillance and forensic scenarios is to search for a face. Face is considered as the most important part of human body. Research shows that even face can speak and it has different words for different emotions. It plays a very crucial role for interacting with people in the society. It conveys people's identity, so it can be used as a key for security solutions in many organizations. Nowadays, face recognition system is getting increasing trend across the world for providing extremely safe and reliable security technology[10]. It is necessary to handle video data which usually lacks in quality compared to single images[2]. The quality refers to low-resolution face. W. W. Zou [7] addressed the problem of low-resolution image.

Low-resolution faces are handled in two ways in face recognition models. i) Converting low-resolution image into high-resolution using super-resolution. ii) Directly working on the low-resolution image by extracting the features from the image. Super-resolution technique requires high computational cost. Local patches are to be applied for missing features. The accuracy of face recognition depends on the algorithm or super-resolution technique. It requires more time for processing. By directly working on the low-resolution image, some of the features are missed and this may lead to incorrect results. G. Hinton [9] proposed distilling the knowledge in the low-resolution image. S. P. Mudunuri [8] proposed that variations in pose and illumination also results in incorrect face recognition.

In our approach, a high resolution face is given as input and its features are extracted by using a complex model. This model finally generates a dense layer with many multiple CNN layers in between for better extraction of features. The features are identified as objects of multiple classes which are represented in a sparse graph. The values in this graph are compared with the input video frames and the frames are labeled appropriately. Our model works on the LR version of the public YouTube Faces Database (YTF) and self-collected surveillance dataset. Our model works effectively even for the 224 X 224 resolution video frames.

The rest of this paper is organised as follows: section II reviews related works and section III presents proposed method. Section IV contains discussion and results and the paper is concluded in section V.

II. RELATED WORK

Trying to address LR video data mainly involves two specific challenges. First, the lack of spatial data, including resolution and image degradations such as out-of-focus, motion blur or compression artifacts. Second, one requires a strategy to compare two sets of still image face descriptors originating from two different face tracks.

Working with LR video data poses many challenges. The significant challenge is the lack of spatial data. The other major challenge is the strategy to compare two sets of images. Herrmann C, in his paper[2] used local patches to recover from the low resolution. In this patches are applied to in place of the missing features. The missing feature is identified with the help of CNN. These identified missing features are replaced with the nearest matching feature that is locally available. Thus, local patches help in reducing the number of comparisons further. Also it helps in improving the efficiency.

Chao Dong [3] proposed deep learning method for single image super-resolution. Initially, low-resolution image is given as input. The low-resolution image objects are represented in sparse graph. The points are stored in low-resolution dictionary. Missing features are replaced with patches. Model is trained to work on end-to-end mappings between the high/low-resolution images. After applying patches the objects are again represented in sparse-coding-based method. The points in the graph are stored in high-resolution dictionary. Finally, using all the points in the dictionary are represented as high-resolution image.

S. Biswas [4] proposed Multi-dimensional scaling(MDS) method for matching low-resolution images. Low-resolution images are treated as probe images and high-resolution images as gallery images. Gallery images resolution are downsampled using Iterative Majorization algorithm. The gallery image resolution is downsampled until the resolution matches with the probe images and then mapping is performed on images. By matching the features the images are identified.

T. Uiboupin [5] proposed Image super-resolution (SR) technique using a dictionary pair. The features of the high-resolution image is stored in high-resolution dictionary and the features of low-resolution image is stored in a low-resolution dictionary. The missing features of low-resolution image is identified by applying patch-pairs. The low-resolution image is then converted into high-resolution image using a Hidden Markov Model. The low-resolution image is converted into a high-resolution image for improving face recognition.

For knowledge distillation the detection complexity poses new challenges in the form of region proposals, less voluminous labels and regression. To overcome these G. Chen [6] proposed a framework for learning compact and fast object detection networks with improved accuracy using knowledge distillation. For addressing class imbalance, weighted cross-entropy loss innovation and to handle regression component teacher bound loss is addressed.

III. DESIGN AND IMPLEMENTATION

A. DATASET COLLECTION

In this phase, two datasets are designed High-resolution image dataset and Low-resolution image dataset. All the high-resolution images of a person are stored in a folder and is converted into High-resolution image dataset. For creating low-resolution image dataset, person each high-resolution image is converted into multiple low- resolution images. We capture 20 different resolutions of the each high-resolution image and stored in a folder. For converting high-resolution image into low-resolution image a program is used. The program generates different image resolutions of the high-resolution image. All of these images are stored in a folder and the folder is converted into a Low-resolution image dataset. Each image in the dataset is resized to a dimension of 224 x 224 pixels.

B. TRAINING THE MODEL

All the images of the High-resolution image dataset is trained using a pre-trained complex face model. The knowledge from the high-resolution face is mined using a teacher stream. A Convolutional neural network (CNN) 3 x 3 filter is applied to mine knowledge. Most of the informative features are recognized in high-resolution face. For face recognition, the most relevant neurons are identified at higher hidden layers. The features are distilled based on classes. Inter-class similarity and Intra-class similarity are used for similarity between classes. Selective distillation is applied to compress similar features. All the identified features are represented in a sparse graph.

Low-resolution face does not contain much informative features compared to high-resolution face. A simple face model is used instead of a complex face model. All the images of the Low-resolution image dataset is trained using a simple face model. A CNN 3 x 3 filter is applied to mine knowledge. First all the features are distilled from the images. Then similar features are compressed to fine tune the process and also to reduce the computational cost. The loss is reduced using softmax loss function. Student stream works well for identifying face in low-resolution face.

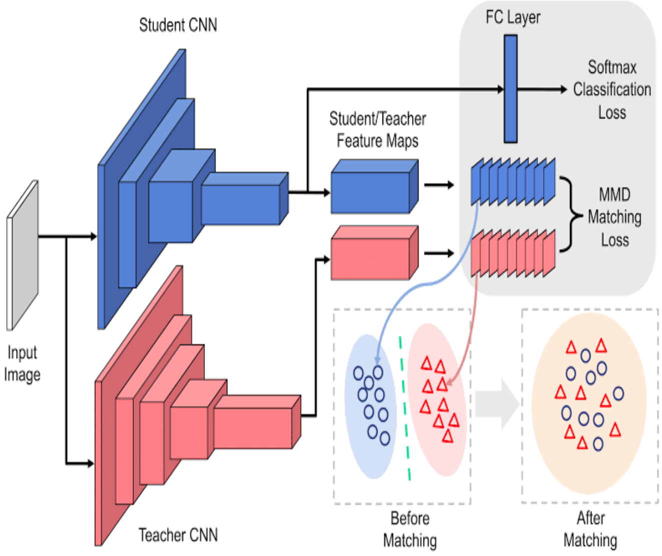

Figure 1. Working of teacher stream and student stream.

Input video is divided into frames using a program. Each second in the video is divided into 42 frames. All the frames are stored in a folder and is converted into a dataset. The dataset is then given as input to the model for testing. The features represented in the sparse graph are matched with the faces in the frames. CNN filter is applied on the frame for identifying features and are identified as objects. If the features are matched then the frame is labeled with person and if the features are not matched then the frame is labeled with unknown. All the frames labeled with person are the frames where person is matched.

IV. DISCUSSION AND RESULTS

Our approach gives a good overall recognition rate. We did our experiments using the video recorded in integrated webcam of the laptop. For a dataset of 10 high-resolution face images of a person, the recognition rate is around 75-80% for the captured frontal face image orientation. If there exists is a different orientation other than the image stored in the dataset then the recognition rate is low. The recognition of low-resolution face image generates better results.

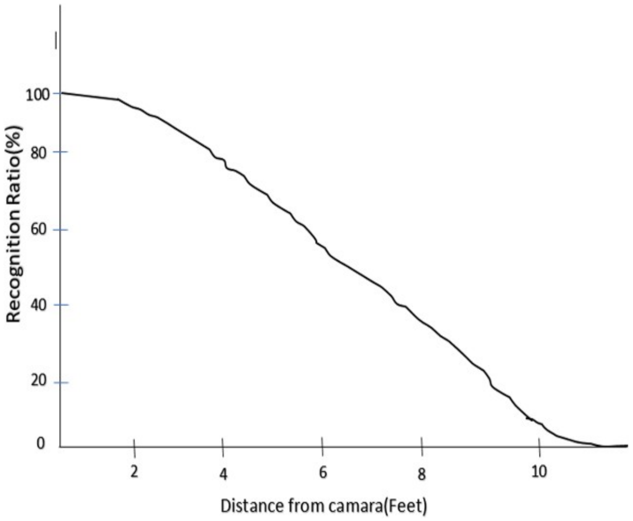

Figure 2

From fig2, we observe that as the distance between camera and human face increases, the recognition rate decreases. In order to overcome this, we use a better resolution camera and improved illumination.

V. CONCLUSION

This paper presents an approach to detect and recognize the faces which in turn can be effectively used to identify a person. This approach can be further enhanced for an Live video surveillance with a few code deviations and some hardware modifications. It can also be used to track and find the location of a person. Compared to existing methods, it gives us a much effective and simple solution for identifying a person. We can nearly double the recognition rate while halving the computational runtime by giving more training dataset images of each person.

VI. REFERENCES

[1]. Shiming Ge, Shengwei Zhao, Chenyu Li and Jia Li, "Low-resolution Face Recognition in the Wild via Selective Knowledge Distillation," in IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 2051-2062, 2018.

[2]. C. Herrmann, D. Willersinn, and J. Beyerer, "Low-resolution convolutional neural networks for video face recognition," in IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2016, pp. 221–227.

[3]. C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in European Conference on Computer Vision (ECCV). Springer, 2014, pp. 184–199.

[4]. S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 34, no. 10, pp. 2019– 2030, 2012.

[5]. T. Uiboupin, P. Rasti, G. Anbarjafari, and H. Demirel, "Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring," in IEEE Conference on Signal Processing and Communication Application, 2016, pp. 437–440. 6G. Chen, W. Choi, X. Yu, and et al., "Learning efficient object detection models

with knowledge distillation," in Neural Information Processing Systems (NIPS), 2017, pp. 742–751.

[6]. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," IEEE Transactions on Image Processing (TIP), vol. 21, no. 1, pp. 327–340, 2012.

[7]. S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 38, no. 5, pp. 1034–1040, 2016.

[8]. G. Hinton, J. Dean, and O. Vinyals, "Distilling the knowledge in a neural network," in Neural Information Processing Systems (NIPS) Workshop, 2014, pp. 1–9.

[9]. https://en.calameo.com/books/003014942376bd07d979c#

Cite this Article

Gudavalli Sai Abhilash, Kantheti Rajesh, Jangam Dileep Shaleem, Grandi Sai Sarath , Palli R Krishna Prasad, "Video Based Person Re-Identification Through Selective Knowledge Distillation", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 699-703, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952179
Journal URL : http://ijsrcseit.com/CSEIT1952179