GENDER CLASSIFICATION IN LIVE VIDEOS

Jiale Chen, Sen Liu, Zhibo Chen

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China, Hefei 230027, China
Email : elsen@iat.ustc.edu.cn, chenzhibo@ustc.edu.cn

ABSTRACT

Human facial gender classification is an important task in live videos. However, it is still challenging in real applications due to motion blur, object occlusion and extreme illumination in real live videos. In this paper, we propose the Multi-Branch Voting CNN (MBV-CNN) framework which first detects and extracts the human face images in live videos, then apply adaptive brightness enhancement on each face image before feeding them into three CNN branches to settle the extreme illumination problem, and finally, we apply a majority voting scheme to reduce the influences from motion blur, object occlusion to further improve classification accuracy. Our method significantly outperforms the state-of-the-art solutions on the LFW dataset and our collected real-world live videos dataset called Gender Classification for Live Videos (GCLV), with respectively averaging 98.11% and 95.36% classification accuracy.

Index Terms—Gender classification, Deep learning, Brightness enhancement, Face analysis

1. INTRODUCTION

With the rapid development of the Internet, live video applications have attracted massive users, which enables us to share daily life videos with others. These sharing actions continuously generate large amount of real-world videos and a great proportion of these videos mainly record human faces, so face analysis in videos is becoming more and more important in real application for video content inspection and recommendation. Gender classification is one of the most important video analysis tasks. As shown in Figure 1, classification tasks in live videos are more difficult than traditional videos. The difficulties mainly come from two aspects: 1) Most of live videos are captured by mobile devices rather than professional camera equipment, some

Fig. 1: Face images in live videos dataset. These face images show the challenges of gender classification in real world. The face images in the first row show extreme illumination in live videos and the face images in second row suffer from motion blur or object occlusion.

are even captured with the devices in motion. These videos are affected by motion blur or object occlusion. 2) The locations of live video capturing are quite casual, such as bedroom, living room, market, outdoor and etc. This leads to great variations in lighting conditions with the problem of extreme illumination which is great challenge for gender classification.

Gender classification primarily need efficient feature extraction and robust classifiers. Most traditional methods use handicraft features as input for gender classification, such as using geometric relations as feature representation [1]. Researchers also attempt to exploit the raw pixels [2], Haar-like wavelets [3], Local Binary Patterns (LBPs) [4] and Gabor wavelets [5] as invariant face descriptors. However, these traditional methods require accurate face localization and efficient face representation. Besides, the facial texture is prone to be impacted by extreme illumination, which may result in degrading the robustness of gender classification methods. As for classifiers, the support vector machine (SVM) and the AdaBoost are two popular algorithms for gender classification which have been widely studied in [4,6].

Regarding to deep learning techniques, the neural networks [7,8] have shown great performance in computer vision tasks like gender classification and outperform SVM and AdaBoost algorithms remarkably. For example, Zhang [7] adopt VGG-Face model [8] for robust facial gender and smile classification. Levi [9] use a CNN to estimate the gender and age attributes by using real-world face images

This work was supported by the National Key Research and Development Plan under Grant 2016YFC0801001, by the National Program on Key Basic Research Projects (973 Program) under Grant 2015CB351803, by the Natural Science Foundation of China (NSFC) under Grants 61571413, 61390514, and 61632001, and Intel ICRI MNC.

and achieves the state-of-the-art performance. However, video-based classification is more challenging than still-image classification affected by transient variations in face appearance or extreme illumination. The model proposed by [7] is not compatible with real-world images attributed to extreme illumination or other factors, while the model in [9] is not suitable for video data because some video frames lack face features.

Considering these challenges in video-based gender classification, we propose a robust method called Multi-Branch Voting CNN (MBV-CNN) for gender classification, which utilizes multi frames and adaptive brightness enhancement that are robust to extreme illumination. In addition, we exploit multi frames and majority voting scheme to conquer the influence from motion blur or object occlusion. Our main contributions are threefold:

1. We build a more challenging live videos dataset called Gender Classification for Live Videos (GCLV) for gender classification, which includes the live videos under extreme illumination, partial occlusion, camera motion.
2. Our algorithm utilizes adaptive brightness enhancement to solve extreme illumination in live videos.
3. We propose a novel network architecture called Multi-Branch Voting CNN and achieves the state-of-the-art performance both on the public available dataset and the GCLV dataset.

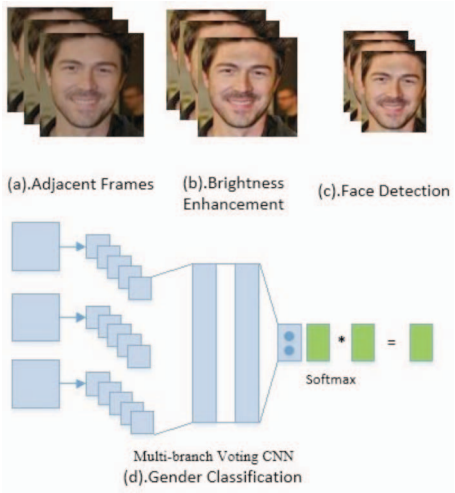


Fig. 2: The framework of our algorithm. (a) is a sample of our face images. (b) shows the result of brightness enhancement. (c) shows the face region detected from face image. (d) presents the architecture of Multi-Branch Voting CNN for gender classification.

The paper is organized as follows: Section 2 introduces our algorithm which is evaluated on public dataset and the GCLV dataset in Section 3. Finally, in Section 4, conclusions are presented.

2. APPROACH

In this section, we present the design details of the proposed CNN for gender classification and describe the network architecture and corresponding adaptive brightness enhancement algorithms. The framework of our algorithm is shown in Figure 2.

2.1. Multi-branch Voting CNN (MBV-CNN)

Our task is to classify gender in live videos. Inspired by multi frames in video stream, we consider multi frames as input for the network. Meanwhile, because of the transient variations in videos or extreme illumination, there are some frames lack of face features in videos. Therefore, we run the off-the-shelf face detector in dlib [10] to extract multi frames and apply adaptive brightness enhancement on each face image. Then we feed adjacent three face images into MBV-CNN.

As illustrated in Figure 3, we propose a new neural network architecture called Multi-Branch Voting CNN (MBV-CNN). It is based on VGG-16 architecture [11], which has remarkable performance on the ImageNet dataset. MBV-CNN contains three branches, which take three adjacent face images as input, one concatenate layer and three full-connected layers. Each branch in network has

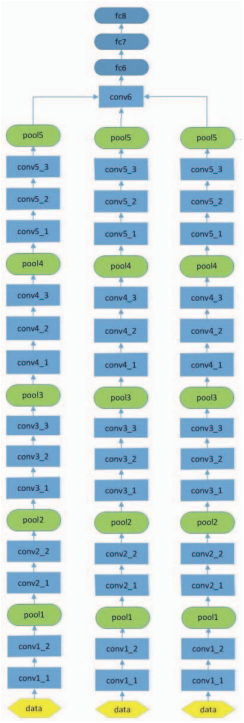


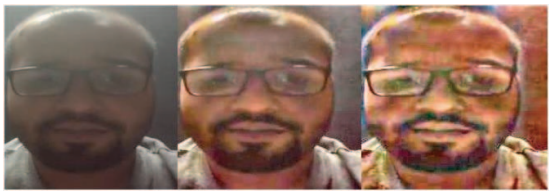
Fig. 3: The architecture of Multi-Branch Voting CNN

thirteen convolutional layers which are the same as VGG-16 and share the same weight. The outputs from these branches are concatenated and fed into a 1×1 convolutional layer for

dimensionality reduction. The first two full-connected layers have 4096 neurons and the third one has 2 neurons.

2.2. Adaptive Brightness Enhancement(ABE)

Influenced by different facial poses, camera angles and variable illumination conditions, the human faces in live videos display unfathomably complex and highly variable contrast and brightness. To handle this issue, we enhance the brightness of face images respectively for each branch. In numerous enhancement algorithms, contrast limited adaptive histogram equalization (CLAHE) [12] is commonly used due to its simplicity and relatively better performance. CLAHE divides face images into blocks, and histogram equalization is performed to each block. In order to overcome noise introduced by the histogram peaks, CLAHE cuts histogram at some threshold before computing cumulative distribution functions (CDF). The CLAHE has two key parameters to control image quality: block size (BS) and clip limit (CL). However, these parameters are manually adjusted without fixed optimal values. As shown in Figure 4, when we determine inappropriate parameters, the results of the CLAHE would be worse than that of original images.



(a) (b) (c)

Fig. 4: The performance of CLAHE on live video frames. (a) is original frames. (b) is appropriate enhancement on face images. (c) is the result of over-enhancement.

Therefore, we utilize different parameters of block size and clip limit for each branch at a time. The block size is M×M, so the local mapping function can be presented as:

$$m(i) = \frac{255 \times CDF(i)}{M \times M} \tag{1}$$

The histogram Hist(i) is the derivative of CDF(i) and it can be obtained as follow:

$$Hist(i) = \frac{d[m(i)]}{di} \cdot \frac{M \times M}{255} \tag{2}$$

In order to avoid noise, the histogram must satisfy the formula as follow:

$$Hist(i) = \begin{cases} Hist(i) + L, & Hist(i) < T \\ H_{max}, & Hist(i) \geq T \end{cases} \tag{3}$$

In this formula, T represents the clip limit and H_{max} represents the maximum value in histogram. Besides, to maintain the area of histogram, it adds L among all histogram bins.

In our experiment, we choose three different groups of T and M for the branches. Adaptive brightness enhancement could improve the robustness of face recognizing and adapt to various lighting conditions. Meanwhile, it is also an effective way to do data augmentation, which could improve the generalization ability of the neural network.

2.3. Majority Voting Scheme

Video-based classification is an ill-posed problem [13] due to video frames suffer from object occlusion or motion blur. Despite a robust classifier, misclassification of video frames still occurs. Therefore, we employ a majority voting scheme mentioned in [14] to vote for the best prediction result across multi video frames. When the face images we fed into network are suffered from motion blur or object occlusion, we can utilize other frames to determine our classification result. For each video, we choose three time positions and for each position, we extract three adjacent frames with half-second interval and feed into our network at a time. Then majority voting scheme is applied on the set of three prediction result for each input video and the final classification is the gender classes obtained most of votes. We integrate this voting technique in our network architecture to overcome some frames suffered from motion blur or object occlusion.

3. EXPERIMENTS

Our method is implemented using the TensorFlow open-source framework [15].

3.1. Datasets

In our experiments, we build a real-world live videos dataset called Gender Classification for Live Videos (GCLV) which is more challenging than public datasets. There are totally 5969 live videos in our dataset, including 3,025 male videos and 2,944 female videos. The gender labels for the videos have two classes (0 for female, 1 for male). The live videos are recorded in real-world environments by mobile devices and typically including various lighting conditions, partial occlusion, camera motion. Figure 1 shows some samples in our dataset.

To validate the effectiveness of our method, we also test our algorithm on the Labeled Faces in the Wild (LFW) dataset [16], which contains 13,233 face images (10,256 male and 2,977 female) from 5,749 celebrities collected from the web.

Since the number of gender classes in the LFW dataset is unbalanced and not big enough for training, we consider the IMDB and Wikipedia (IMDB-WIKI) dataset [17], the largest publicly available dataset for training. IMDB-WIKI dataset contains 524,230 face images and they were manually labeled with gender information. In our experiments, we remove the broken face images and select 320,000 images evenly.

3.2. Experimental Results

In this section, we show the experimental results on the GCLV and LFW dataset. For comparison, we choose the algorithm proposed by [9] with majority voting scheme as baseline, which is the application of the VGG-Face [8] for robust gender classification.

Table 1 shows the results of our method and baseline evaluated on the GCLV dataset. These results show that the MBV-CNN is remarkably better than VGG-Face, leading to large performance improvement to the GCLV dataset by 13.86%. This demonstrates that our method has very good compatibility to real-world gender classification.

Table 1. Gender classification results on the GCLV dataset

Neural network	Male	Female	Total
VGG-Face	0.8180	0.8120	0.8150
MBV-CNN	0.9710	0.9362	**0.9536**

.

We also do the evaluation on the LFW dataset. While there are only single images but not videos in the LFW dataset, we could not directly apply the MBV-CNN method. Instead, we use the MBV-CNN without the majority voting scheme, called MB-CNN, which feed the same image into the multi branches with adaptive brightness enhancement. As table 2 presents, MB-CNN still outperforms VGG-Face. It again verifies the effectiveness of Multi-Branch architecture and adaptive brightness enhancement.

Table 2. Gender classification results on the LFW dataset

Neural network	Male	Female	Total
VGG-Face	0.9543	0.9367	0.9445
MB-CNN	0.9785	0.9835	**0.9811**

3.3. Analysis of Adaptive Brightness Enhancement and Majority Voting Scheme

To further examine how majority voting scheme and adaptive brightness enhancement impact on the final prediction performance, we quantitatively analyze this two parts.

Table 3 presents the results in which only the IMDB-WIKI dataset is used for training and our live dataset for testing. As the results show, we first train VGG-Face CNN and the MBV-CNN without adaptive brightness enhancement and majority voting scheme; here we name it MB-CNN without ABE. This method achieves the accuracy by 81% and 87.75% respectively. Then our networks perform better when majority voting scheme was applied, reaching up to 85.20% and 91.70% accuracy, which shows that majority voting scheme are robust to motion blur or object occlusion in live videos. In addition, it demonstrates that multi-branch network outperforms single-branch network.

As explained in Section 2, we adopted adaptive brightness enhancement on MBV-CNN and VGG-Face CNN, which obtained additional boost to 87.48% accuracy on VGG-Face, while the accuracy of MBV-CNN increases to 92.42%. This confirms that our adaptive brightness enhancement is benefit to solve extreme illumination in live videos and promote the robustness of our networks. This can be explained by the fact that face images with adaptive brightness enhancement are more efficient to extract face feature. Finally, as illustrated in Table 1, the method MBV-CNN fine-tuned on the GCLV dataset achieves the best result of 95.4%. In conclusion, different adaptive brightness enhancement based on multi branches can effectively improve the gender classification accuracy.

Table 3. The gender classification performance of MBV-CNN on the GCLV dataset

Neural network	Male	Female	Total
VGG-Face	0.8180	0.8120	0.8150
+ Majority Voting Scheme	0.8527	0.8513	**0.8520**
+ Brightness Enhancement	0.8792	0.8752	**0.8748**
MB-CNN without ABE	0.8450	0.9098	0.8775
+ Majority Voting Scheme	0.8795	0.9441	**0.9170**
+ Brightness Enhancement	0.8895	0.9592	**0.9242**

4. CONCLUSIONS

In this paper, we propose a new neural network architecture called Multi-Branch Voting CNN (MBV-CNN) for gender classification. To be robust for live videos with extreme illumination, adaptive brightness enhancement for input is utilized. Considering the misclassification of video frames with motion blur or object occlusion often occurs, majority voting scheme for evaluation is also added in our framework. Experimental results conducted on the LFW dataset and the GCLV dataset show that our algorithm achieves significant improvements over previous methods. These results verify that multi-branch method is more robust to image quality than single-branch method and adaptive brightness enhancement enables to adapt to extreme illumination in live videos.

Multi-Branch Voting CNN is also a promising approach for other human facial recognition tasks. The further exploration is also necessary to overcome motion blur and object occlusion in live videos. We will also try to optimize the architecture for real-time classification.

5. REFERENCES

[1] Ng, C., Tay, Y., Goi, B.M., Recognizing human gender in computer vision: A survey, in: PRICAI 2012: Trends in Artificial Intelligence. volume7458 of Lecture Notes in Computer Science, pp. 335–346, 2012.

[2] Moghaddam, B., Yang, M.H., Learning gender with support faces. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 707–711, 2002.

[3] Shakhnarovich, G., Viola, P.A., Moghaddam, B., A unified learning framework for real time face detection and classification, in: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 16–, 2002.

[4] Shan, C., Learning local binary patterns for gender classification on real-world face images. Pattern Recognition Letters 33, 431 – 437, 2012.

[5] Leng, X., Wang, Y., Improving generalization for gender classification.,in: ICIP, pp. 1656–1659, 2008.

[6] Eidinger, E., Enbar, R., Hassner, T. , Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security 9,2170 – 2179, 2014.

[7] Zhang K, Tan L, Li Z, et al. Gender and smile classification using deep convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 34-38, 2016.

[8] O. M. Parkhi, A. Vedaldi, A. Zisserman, “Deep Face Recognition,” in British Machine Vision Conference, 2015

[9] Levi, G., Hassner, T., Age and gender classification using convolutional neural networks, in: IEEE Conf. on CVPR workshops, 2015.

[10] King D E. Dlib-ml: A machine learning toolkit[J]. Journal of Machine Learning Research, pp. 1755-1758. 2009, 10(Jul)

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[12] K. Zuiderveld, “Contrast Limited Adaptive Histogram Equalization”, Academic Press Inc., 1994.

[13] Hadid, Abdenour, and M. Pietikainen. "From still image to video-based face recognition: an experimental analysis." Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on. IEEE, 2004.

[14] Parhami B. Voting algorithms[J]. IEEE Transactions on reliability, 43(4): 617-629, 1994.

[15] http://tensorflow.org/

[16] Huang G B, Learned-Miller E. Labeled faces in the wild: Updates and new reporting procedures[J]. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep, 14-003, 2014.

[17] Rothe, Rasmus, Radu Timofte, and Luc Van Gool. "Dex: Deep expectation of apparent age from a single image." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015.