An Accurate Extraction of Facial Meta-Information Using Selective Super Resolution from Crowd Images

Jieun Park[1], Yurim Kang[1], Yoo-Sung Kim[1,2*]

[1] Department of Information and Communication Engineering,
[2] Department of Artificial Intelligence,
Inha University
Incheon 22212, Korea

*yskim@inha.ac.kr

Abstract. An accurate extraction scheme of facial meta-information from low resolution crowd images is proposed. In order to detect crowd abnormal situations, extracting facial meta-information from crowd images is very helpful. However, since many crow images are of low resolutions, extracting facial meta-information is pretty difficult. To extract accurately facial meta-information from the low resolution crowd images, some face images which are not suitable to easily extract the meta-information should be classified and be improved the quality by a super resolution method. To confirm the feasibility of the proposed scheme, since gender of person can be regarded as very important facial meta-information, we compare the gender classification accuracies of using the proposed scheme and that of using the input crowd image itself. According to the experiment results, the proposed facial extraction scheme from crowd images using selective super resolution can improve the gender classification accuracy for crowd images than that for using the original crow images.

Keywords: Crowd abnormal detection, Low resolution crowd images, Facial meta-information, Selective super resolution, Gender classification

1 Introduction

Crowd analysis becomes important computer vision task because of its usability, including crowd movement analysis, abnormal behavior detection, and crowd size estimation[1]. As the urban population increases recently, crowd analysis and monitoring becomes more essential for public safety purposes. Crowd monitoring using intelligence video surveillance systems which can be encountered in daily life is increasing in many places such as subway stations, airports, and department stores even on the streets with a lot of people. In such a crowd situation, it is important to detect and prevent abnormal situations in advance through crowd monitoring. Therefore, accurate and rapid judgement of situation in crowd monitoring is required.

To recognize accurately and fast abnormal crow situations, using facial meta-information of person in crowd such as gender, wearing mask or sunglasses, and even

emotional expressions is very helpful. Of course, using these information can help to identify person in multi-camera environments. For example, if it is possible to analyze a face by accurately classifying gender, it will be helpful to understand a situation.

However, there are many difficulties to accurately extract facial meta-information from crowd images since they may be of low resolution and have heavy occlusions. In the case of face detection in a crowd with long distance from camera and many people, it is difficult to utilize it in future research because image resolution is not good as the size of the face decreases. Therefore, we propose a method to extract facial meta-information from low-resolution crowd images using a selective super resolution for improving the quality of the face images, so make them possible to use in future research. By solving low resolution problem occurring in crowd faces, the accuracy of gender classification can be improved, which can be contributed to crowd monitoring.

The rest of this paper is as follows. Section 2 describes the previous related works on the crowd dataset and on face detection model using convolutional neural network. In Section 3, the data set which is used for this study is introduced, and the proposed extraction scheme of facial meta-information from low-resolution images is discussed. In Section 4, the experiment to show the effectiveness of the proposed scheme is discussed. This paper is concluded with a short discussion of future studies in Section5.

2 Related Work

Crowd-11 dataset is widely used for crowd analysis researches. It classified crowd videos into 11 classes from considering various situations and places to classify abnormal behavior of crowds[1]. This includes crowd images in daily life and is close to the general CCTV images targeted in this paper. Therefore, some images from 8 classes suitable for this paper were selected except classes of 'Diverging Flow', 'Interacting Crowd', and 'No Crowd', which are difficult to identify crowd faces. Face detection tasks in crowd of real world have problems such as low resolution and heavy occlusion, unlike single person or refined situations. There is the widely used WIDER Face dataset for face detection in crowd[2]. In this dataset, considering various conditions such as pose, scale, and occlusion, it is classified into easy and hard situation according to the difficulty of face detection. In YOLO5Face([3]), they propose a method to improve the performance of face detection in hard situation close to the real world. It has high accuracy on very small faces, by adding a five-point landmark regression head into YOLOv5. Also, it belongs to the high real-time among object detection models. Therefore, we utilized YOLO5Face([3]) for face detection which is appropriate for the purpose of this study.

3 Extracting Facial Information by a Selective Super Resolution

3.1 Overall procedure for the extraction of facial meta-information

Fig. 1 shows the overall procedure for the extraction of facial meta-information from the crowd images. First, when crowd image comes in, face detection is performed by YOLO5Face([3]) as described in the previous section. Cropped face image from the crowd has very low resolution as the number of people increases or the size of the face decreases due to the distance from the camera. The lower resolution, the more difficult it is to identify the face, so it will be difficult to utilize in future research. Therefore, when the resolution of the face images is smaller than the threshold, we regard the face image as the hard case for extracting facial meta-information from the original image so that super resolution step is applied to improve the quality of the input. As the threshold value, after careful consideration of input data succeed by experiment we decided to use 48 * 48. At the last step we can extract the facial meta-information. In this paper, we try to extract gender information from the person's face.

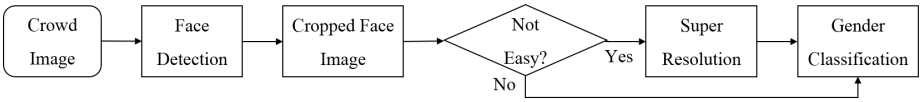

Fig. 1. Overall procedure for the extraction

3.2 Dataset

We need crowd dataset in real world for face detection and subsequent research in crowd analysis. Crowd data in real world means most like real situation, rather than a refined one. The most similar Crowd-11 dataset([1]) was used as our experiment, and some of them are selected in consideration of the angle and distance of camera. In other words, we included the distance enough to reveal the whole upper body of person and excluded the angle that interferes with face recognition due to heavy occlusion between people.

From Crowd-11 dataset([1]), we extracted total of 551 images by sampling 111 videos with an average length of 4 to 5 seconds at one frame per second as shown as Fig 2. To use the selected images in face detection and gender classification, ground truth data is required. We make annotation files by labeling the x and y coordinate values of face region and gender of the face in the images. There are 6,146 male faces and 4,984 female faces in total 11,130 faces.

Fig. 2. Examples of selected images for this study from Crowd-11 dataset

Within the selected crowd images, the performance of face detection and subsequent research is different because the size of the face. Since this paper focused on crowd face and analyze the experimental results according to the resolution of the face images, the selected images were classified according to the resolution of the image extracted from the face, not the resolution of the entire image.

A total of 11,130 faces were sorted according to the number of pixels, which means resolution. They are classified into three cases as shown in Table 1 according to the section where the number of images decreases sharply depending on the number of pixels. The case where the resolution of the face image itself is relatively large is easy case, the middle resolution is medium case, and the smallest is hard case. It is possible to identify the face in easy case to confirm with the naked eyes, while medium and hard cases have difficulties. The number of faces in hard case is the largest because there are more cases of small faces compared to large faces in crowd.

Case	Number of Face Images	Resolution
Easy Case	596	48x48 ~
Medium Case	5,102	29x29 ~ 48x48
Hard Case	5,432	~ 29x29

Table 1. Classification of the difficulty level of face images based on the resolutions

3.3 Face Detection

In this paper, face detection in crowd situations like CCTV video is required. Therefore, face detection was performed on the finally selected Crowd-11 [1] dataset using YOLO5Face([3]) model, which has excellent performance in the WIDER face dataset([2]) close to the real world.

Next processing step is conducted using face images extracted from the Crowd-11 dataset through face detection. In the case of the detected and enlarged face images of crowd, the resolution is not good. So, it is difficult to directly utilize original face images in future research. Therefore, among the studies to improve resolution, ESRGAN([4]) which has excellent perceptual improvement ability was applied to obtain better results in subsequent studies.

3.4 Super Resolution

Next step is conducted using face images extracted from the Crowd-11 dataset through face detection. In the case of the detected and enlarged face images of crowd, the resolution is not good as shown on the left of Fig 3, because there are many people, and the distance from camera is so long. So, it is difficult to directly utilize original face images in future research. Therefore, among the studies to improve resolution, ESRGAN([4]) which has excellent perceptual improvement ability was applied to obtain better results in subsequent studies. We got better resolution of face images by applying ESRGAN([4]) to hard and medium cases as shown on the right side of Fig 3.

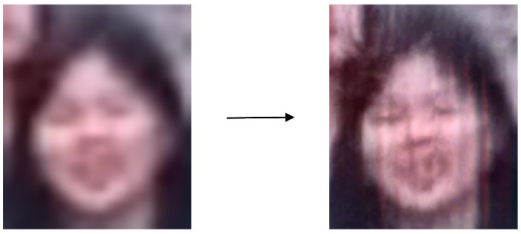

Fig. 3. Examples of images before and after ESRGAN application

3.5 Extraction of Gender Information as Facial Meta-Information

In post processing using detected crowd face images, increasing the accuracy of gender classification may help in detecting crowd anomaly behaviors. Therefore, we conducted a simple gender classification experiment to confirm an improvement of performance. In this case, Real-time Convolutional Neural Networks Gender Classification([5]) with good real-time performance was used in our experiment for application in video-surveillance.

4 Experiments

4.1 Face Detection

We conducted face detection using YOLO5Face Detector([3]) on the selected Crowd-11 dataset. To evaluate the accuracy of detection, if the detections over IOU(Intersection Over Union) > 0.5 is considered as the correct ones. Through this, we obtained about 80.6% of Recall and 71.1% of Accuracy. As the result of evaluation in Table 2, the performance of the easy case with the high resolution of the face images was the best, while the performance of the hard case with the low resolution was the worst.

Case	Recall	Accuracy
Easy Case	87.67%	84.41%
Medium Case	85.04%	80.32%
Hard Case	74.79%	59.99%
Total	80.6%	71.1%

Table 2. Performance evaluation by Face detection

4.2 Gender Classification

To improve the performance in gender classification, ESRGAN([4]) was applied to the face images of medium and hard cases, except for the easy case. The image after application has 4 times the resolution of the original face images. As the result of total evaluation in Table 3, the performance after application improved by 1%.

Case	Before ESRGAN	After ESRGAN
Easy Case	72.15%	-
Medium Case	68.86%	69.78%
Hard Case	65.24%	65.24%
Total	67.27%	68.20%

Table 3. Performance evaluation by gender classification before and after applying ESRGAN

5 Conclusions

In this paper, an accurate extraction scheme of facial meta-information from low resolution crowd images is proposed. In the proposed scheme, to accurately extract the facial meta-information even from the low-resolution crowd images, YOLO5Face detector which is well known as the good detector against small faces is used, and the

small images which are of lower than the threshold resolution determined by experiments are enhanced the quality by using ESRGAN. After that processing, the facial meta-information can be extracted more well. To confirm the feasibility of the proposed scheme, since gender of person can be regarded as very important facial meta-information, we compare the gender classification accuracies of using the proposed scheme and that of using the input crowd image itself. According to the experiment results, the proposed facial extraction scheme from crowd images using selective super resolution can improve the gender classification accuracy for crowd images than that for using the original crow images. This result shows the improvement of about 1% in gender classification. It is meaningful in that it tried to improve the performance in related studies by applying super resolution algorithm after face detection. Therefore, if an additional learning process in ESRGAN using train data optimize for face images is carried out, the possibility of development in future research is expected.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00203, Development of 5G-based Predictive Visual Security Technology for Preemptive Threat Response). This work was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A1A01041815).

References

1. Crowd-11 : A Dataset for Fine Grained Crowd Behaviour Analysis, IEEE Conference on CVPRW (2017)
2. Yang, Shuo and Luo, Ping and Loy, Chen change and Tang, Xiaoou, WIDER FACE: A Face Detection Benchmark, IEEE conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Delong Qi, Weijun tan, Qi Yao, Jingfeng Liu, YOLO5Face: Why Reinventing a Face Detector, arXiv preprint arXiv:2105.12931 (2021)
4. Xingtao Wang and others, ESRGAN: Enhanced Super-Resolution Generative Adversarial Network, arXiv preprint arXiv:1809.00219v2 (2018)
5. Octavio Arriaga and Matias Valdenegro-Toro and Paul Plöger, Real-time Convolutional Neural Networks for Emotion and Gender Classification. arXiv preprint arXiv:1710.07557 (2017)