Face Detection in Thermal Image Based on Two-Stage CNNs

Yibo Guo[1,a)] Atsushi Shimada[1] Hideaki Uchiyama[1] Chao Ma[1] Hajime Nagahara[2]
Rin-ichiro Taniguchi[1]

Abstract: The IR images generated from an infrared camera show the temperature distribution of object. Because facial temperature is stable and independent of ambient lighting, an IR camera can be used in detecting facial regions in indoor facilities. The face detection algorithm uses Adaboost with local features, such as Haar-like, MB-LBP, and HOG features in the thermal image. However, these features can only be extracted by manual and few of them can be used. In this paper, we proposed a method that detects faces with thermal image by using CNN, which do not need to design the features and maintain a well performance. The face detector we provided has two stage: a CNN for facial region proposal and another CNN for calibration. The calibration network is used to adjust the detetcion window location in order to obtain a better recall. Finally, we compare the performance on thermal image dataset of our method with method that uses different designed features such HOG, and find that the face detector can maintain a similar or even better result, while being less complexity.

Keywords: thermal image, face detection, CNN, calibration network

1. Introduction

Research and development of face detection has been around in computer vision field for many decades. By using different categories of cameras, we can capture human faces in different spectral regions. A thermal camera can capture an image containing faces in the IR domain. By seeing the temperature of object, the thermal camera becomes a powerful tool for industrial inspection, medical imaging, chemical imaging, and surveillance. Since temperature of human faces is stable and independent of ambient lighting in indoor facilities such as meeting room, we can discriminate human faces in thermals image explicitly. Meanwhile the thermal images provide less detail information about human faces than color images captured by a normal camera, they can help for protecting personal privacy[1].

Variety face detection methods for thermal image has been provided by many research works. The segmentation-based method assumes that the facial temperature is within a fixed temperature range and use a threshold to find facial region from the thermal image. The projection-based method assumes that facial temperature is higher than the background and finds the facial region from vertical and horizontal projection profiles. The machine-learning-based method uses facial and non-facial patches as positive and negative samples and apply a machine-learning method such as SVM to build a classifier. Among the above mentioned methods, the machine-learning-based method such as using AdaBoost algorithm with Haar-like feature performed well and hence now widely used in the IR domain [1]. However, using such method for face detection with thermal images cannot work without designing the feature manual in advance. Meanwhile there is few feature about human face can be extracted from the thermal image. These may restrict the performance of face detection with thermal image. Recently many research work shows that the convolutional neural network (CNN) can be well-perform in object detection field, such as Alexnet[2]. By using detector based on CNN it can find a target without requiring for the feature to be designed in advance, which is efficient and convenient.

In this paper, we propose a method for face detection with thermal image by using CNN models. The method follows two step: facial region proposal and bounding-box calibration. In the first face detection step we train a CNN model with face/non-face samples cropped from thermal image, and then convert the architecture of network into fully-CNN. By doing this the face detector is enable to accept arbitrary size of a whole image as input and generate a heat-map corresponding to the original image. The second step calibration is to adjust facial bounding-boxes achieved from first step and make them to be better aligned to the ground truth by using another CNN. This can help our face detector to achieve a better recall.

2. Related work

The topic of face detection with thermal image has been around for many decades. One of a well-performed method proposed by Ma et al.[1] is that using Adaboost algorithm with mixed features for face detection. They found that using multiple type of features such as Haar-like and HOG for face detection, performs better than using just a single type. We believe that using the

[1] Kyushu University
[2] Osaka University
[a)] guoyibo@limu.ait.kyushu-u.ac.jp

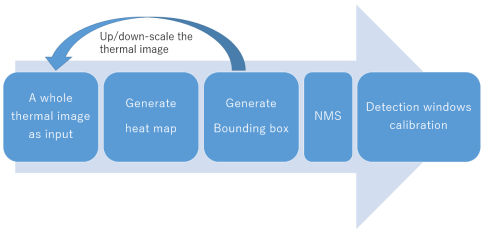

Fig. 1 The flow chart of CNN for face detection with thermal image.

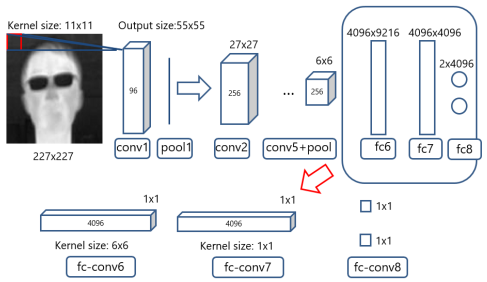

Fig. 2 We convert the fully-connected layers of trained CNN model into convolutional layers by reshaping the layer parameters. The CNN then can accept arbitrary size of thermal image and generate heat-map.

CNN model for face detection with thermal image could achieve the same good performance while no need for designing feature manual in advance. Many of CNN based face detection methods was provided since 1990s.

One of them proposed by Farfade et al.[3] is Deep Dense Face Detector (DDFD), a method that does not require additional annotations or components such as SVM and is able to accept a whole image as input by using a single CNN model. This can be comparatively efficient in face detection. We apply the CNN based method in thermal image which is gray image instead of color image captured by a normal camera.

3. Overall framework

3.1 Face detection pipeline

We proposed a method that detect facial regions with thermal image by using CNNs. As shown in the figure 1, our face detection method consists of two main stages: 1) a face detector for finding the location of faces in a thermal image and 2) a calibration network for adjusting the result. In the first stage we pre-trained a CNN model as face/non-face binary classifier. After converting the structure of the CNN it can take a whole image as input and generate a heat-map. The value of each point in the heat-map is the confidence of having a face, for its corresponding region in the input image. According to the heat-map we can obtain plural bounding-boxes represent for the almost location of a face. After NMS (non-maximum suppression) there remains only one bounding-box with highest confidence. In the second stage we train a CNN model for calibration by using abnormal pattern of bounding-box as training samples. The calibration network can help to adjust the location of window which is ensured by the face detector from the first stage, and make it better align to the center of the faces.

3.2 CNN for face detection

In this section, we provided the details of the CNN for face detection. The classifier has 5 convolutional and 3 fully connected layers, which is similar to AlexNet[2]. Because the thermal images contain few features, we start the fine-tuning from Annotated Facial Landmarks in the Wild (AFLW) dataset[4] in order to purchase more information about human faces for the CNN model. The AFLW dataset is consist of about 21K images with 24K facial landmarks. We also convert all of them into gray images in accordance with thermal images. To increase the samples, we do the data augmentation. We used sliding windows to crop subwindows from every image, and the size were the same as facial annotations. All samples will be resized into 113 $\times$ 113 the same as the input of the CNN. We then used our dataset, which contains 8,400 thermal images captured by IR camera, for fine-tuning the deep network. Through the two stages fine-tuning we got a facial classifier with nearly 99% accuracy.

The next step is to convert the CNN into fully-CNN. To do so we first translate the last 3 fully-connected layers of the CNN model into convolutional layers by using Caffe code[2], as shown in figure 2. We implemented the translation by changing the parameters of the fully-connected layers. Take the first fully-connected layers which is called fc6 as example: the input of original layer is 256 feature maps with size of 3 $\times$ 3, and the number of its weights is 2304 $\times$ 4096. Considering the weights as matrices, we can map these fully-connected weights to the convolution filters through the Caffe code. The new convolutional layer called fc-conv6 will have 4096 kernels with size of 3 $\times$ 3 $\times$ 256. Actually they are identical in memory so we can assign them directly. The biases are also identical to those of original layer. We then do the same work to all 3 fully-connected layers. Because the fully-connected layers is the only constraint for input size of the CNN model, after the network surgery the trained facial classifier is possible to accept arbitrary size of a whole image and obtain a heat-map. An example of the heat-map is shown in figure 3. Each point on the heat-map represent for the probability of having a face, which is corresponding 113 $\times$ 113 region in the original thermal image. By using the heat-map we can ensure the most confident region that contains a face in the input thermal image.

3.3 Calibration Network

In this section, we introduce the second stage of the face detector for thermal images, the calibration network, and how it works. The calibration network is a CNN model that can be used to adjust the bounding-boxes. The motivation of using the calibration network is shown in figure 4. Though the bounding-boxes generated from the last CNN represents the most confident region that could have a face in a thermal image, it may not be well aligned to the face[5]. We apply the calibration so that the face detector for thermal image can maintain a better recall.

The calibration network is a shallow CNN like NIN[6], which is consist of 4 convolutional layers and a global average pooling layers. Since we do not need to find the location of a face in the whole image, we choose a light network in this step to reduce the

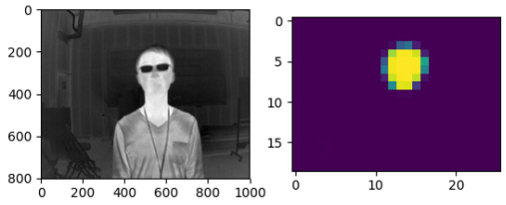

Fig. 3 An example image with its heat-map generated by CNN face detector. The yellow part means high scores for a face in that region.

computation. A detection window is cropped out according to the result of face detector in the first stage, and resized into 227 × 227.

The calibration network takes the cropped window as input. We prepared N pattern for calibration. Each pattern are predefined as a 3-dimensional vector $[s_n, x_n, y_n]_{n=1}^{N}$. In these set of vectors the x_n, y_n are used for axis offsets and the s_n is used for scale changes. Make (x, y) to be the top-left corner and (w, h) to be the size of a detection window, the calibration pattern adjusts the window to be

$$(x - \frac{x_n w}{s_n}, y - \frac{y_n w}{s_n}, \frac{w}{s_n}, \frac{h}{s_n})$$

In our work, we prepared $N = 45$ patterns, which is formed by

$$s_n \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$$

$$x_n \in \{-0.15, 0, 0.15\}$$

$$y_n \in \{-0.15, 0, 0.15\}$$

The output of the calibration network is a vector of confidence scores $[c_1, c_2, c_3 \ldots, c_N]$. In figure 5, we compare the face detector with and without calibration. The test result shows that the face detector performs with calibration apparently better than that without calibration, especially in recall.

4. Experiments

4.1 Training process

In training the CNN for face detection, we first used the AFLW dataset for fine-tuning as described in section3.2. We then use our thermal image dataset[1] captured by IR camera for the second fine-tuning. The thermal image dataset includes totally 8,400 thermal images with human face and background, which is consist of 14 males and 6 females, with 10 variations in camera distance, 21 poses, and people with and without glasses. Considering the infrared camera is usually apply in indoor facilities and thermal sensitive, the scenario is settled to be in the meeting room and the environmental temperature and light are stable. Since we choose the leave-one-out cross validation, we divide the dataset into two parts: 7980 thermal images of 19 people for training, and the rest 420 images of last people for validation. After data augmentation we obtain about 70k positive and 1.3 million negative training samples. For fine-tuning we used 60k iterations and batch size of 128 images, where each batch contained 16 positive and 112 negative examples. This result in 20 CNN model for face detection. After training the face detection CNN models we converted the networks structure as shown in section3.2. By doing this they were possible to accept a whole image with arbitrary of size and to generate a heat-map. The heat-map shows the score for every 113 × 113 window with a stride of 32 pixels.

Fig. 4 The calibration network adjusts the detection window to be better aligned to the face region. The bounding-box (red) is generated from the face detector in the first stage, the bounding-box (blue) is the adjusted result.

In the next step we trained another CNN for calibration. In collecting the training data for it we only use the thermal image dataset. We perturb the face annotations with the 45 calibration patterns, which are introduced in section3.3. For example, for the n-th pattern $[s_n, x_n, y_n]$, we use $[\frac{1}{s_n}, -x_n, -y_n]$ to deform the bounding box, then crop and resize it into size of 227 × 227 as input. Refer to the face detection CNN we also divide the dataset into two parts, for training and validation respectively. After 120k iterations the accuracy of the calibration nets are all between 80% and 90%.

4.2 Face detection for thermal images

After training the models and doing network surgery we tested our face detection approach on the thermal image dataset. As mentioned in section4.1, we used the leave-one-out cross validation. There are 420 thermal images for each participant and we tested on all of the 20 participants. We first scaled the thermal images up and down to detect faces of smaller or larger than 113 × 113 respectively, like scale pyramid. We defined the minimum scale and the maximum scale of scale pyramid to be 200 and 3200 respectively. The image is first up-scaled to height of 3200. We then apply a factor $f_s = \sqrt[3]{0.5}$ (about 0.79)[3] to down-scaled the thermal image until the height of it is less than 200. By doing this work we can obtain plural heat-maps and extracted facial bounding-boxes according to them. The next step is to reduce the number of bounding-boxes by using NMS. In this work

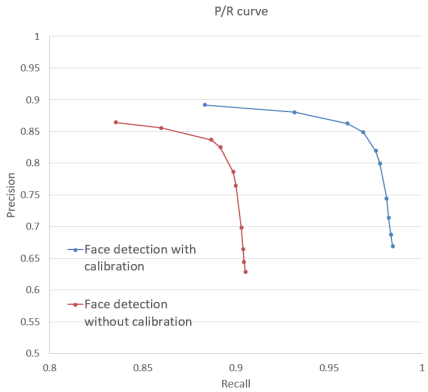

Fig. 5 We compared the performance of face detector with and without calibration on our thermal image dataset. The precision/recall curves show that the calibration work can provide big promotion.

we choose a method named NMS-average. We first filter out the bounding boxes with confidence less than 0.5. Next we cluster the bounding-boxes according to an overlap threshold, which is defined as 0.2[3]. We averaged location of all bounding-boxes and give a highest weight to the bounding box who has highest score in each cluster. Finally, we use the maximum score of the bounding-boxes as the final score of a detection window.

After finding the almost location of face we then apply the calibration. We crop the detection windows given by face detection CNN and resize them into 227 $\times$ 227 as the input of calibration network. The calibration network will output a vector of confidence score $[c_1, c_2, c_3 \ldots, c_N]$. We then use the average results of the patterns of high confidence score as the adjustment $[s, x, y]$,

$$[s, x, y] = \frac{1}{Z}\sum_{n=1}^{N}[x_n, y_n, s_n]I(c_n > t) \tag{1}$$

$$Z = \sum_{n=1}^{N} I(c_n > t) \tag{2}$$

$$I(c_n > t) = \begin{cases} 1, & if\ c_n > t \\ 0, & otherwise \end{cases} \tag{3}$$

The t is a threshold to filter out patterns with low score. Here we set the value of it as 0.08. We verify our two stages face detector for on the thermal image dataset and compare its performance with the face detection with thermal method that using HOG, LBP, HAAR and mixed features[1]. All of the validations of these method apply the leave-one-out cross validation. As shown in figure 6, we can see from the precision and recall curves that the face detection with thermal image by using CNN can maintain a good performance and even better than some of the designed feature, such as HOG. Meanwhile we find that our method still not perform good enough on the precision. Some of detection windows located in incorrect place, this because when we apply data augmentation some of positive samples shift from original ground truth. Since our thermal image data only contain 8400 images while the CNN model own enormous number of kernels, We believe that the CNN model hold potential to achieve better performance if we increase the content of our dataset.

5. Conclusions and future work

In this paper we proposed a method using CNN for face detection with thermal image, which consists of two stage: facial region proposal and detection window calibration. It can accept arbitrary size of faces in thermal image as input and generate heatmap. We verified our method in dataset under an indoor scenario with and without calibration stage. The result shows that the calibration network can make large progression in recall and precision performance of face detection with thermal images. We also compare our method with face detection method with thermal images by using different features[1]. The face detector maintains a good performance in recall but still not good enough on precision. This is because the amount of dataset used for training is still very limited, and the data augmentation method produces some inappropriate positive samples. In future we are planning to increase our dataset such as capturing more thermal images that

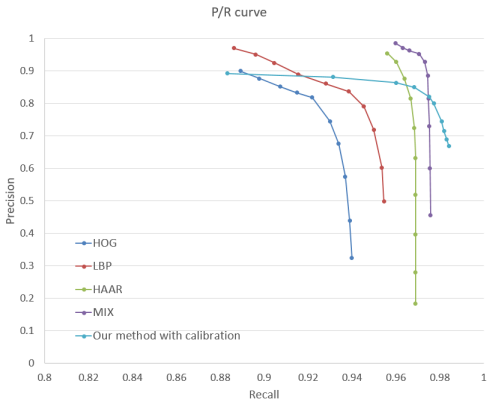

Fig. 6 We compare the performance of our method, CNNs for face detection and calibration with thermal image, with methods which use HOG, LBP, HAAR and mixed features respectively. For calculating the precision and recall we take the detection windows have IOU more than 50% with ground truth as true positive.

contain human face in different conditions and with variety poses. Meanwhile, we also prepare to use better sampling strategies in order to purchase better performance on face detection with thermal image by using CNN.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grand Number JP15K12066, Japan.

References

[1] Chao Ma, Ngo Thanh Trung, Hideaki Uchiyama, Hajime Nagahara, Atsushi Shimada, and Rinichiro Taniguchi. Mixed features for face detection in thermal image. In The International Conference on Quality Control by Articial Vision 2017, pp. 103380E103380E. International Society for Optics and Photonics, (2017).

[2] Alex Krizhevsky, Ilya Sutskever, and Georey E Hinton. Imagenet clasication with deep convolutional neural networks. In Advances in neural information processing systems, pp. 10971105, (2012).

[3] Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li. Multi-view face detection using deep convolutional neural networks. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 643650. ACM, (2015).

[4] Kstinger, Martin, Wohlhart P, Roth P M, et al. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. IEEE International Conference on Computer Vision Workshops IEEE, 2012:2144-2151(2012).

[5] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 53255334(2015).

[6] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400(2013).