High-Precision Portrait Classification Based on MTCNN and Its Application on Similarity Judgement

Juan Du
New Research and Development Center of Hisense, Qingdao 266071, China
dqxwpl@sina.com

Abstract. Portrait classification is a complex course including at least face detection, recognition and compare each of which contains multi-tasks, facing plenty of various challenging questions due to askew poses, illuminations, occlusions, image blurring and small scale face in the pictures. Though deep learning methods, such as Convolutional Neural Network (CNN) family and You Only Look Once (YOLO) series, had boomed a large number of areas on object detection and accelerated the solving of these difficulties on image processing, they are not specially designed for the image classification and may require a great deal of resource, expensive computation and taxing annotation. In 2016, an innovative face detection model named Multi-task convolutional neural network (MTCNN) arose and triggered viral and wide spread. Its high efficient and accurate performance on both face detection and face alignment tasks, real time effect based on lightweight CNN as well as effective conducting online hard sample mining, all contribute to significant improvement to the challenges above. This paper introduces the MTCNN algorithm and applies it to the similarity judgement with two industrial real problems together with FaceNet model. In addition, some effective practical methods on increase precision of classification are also proposed to gain better effect.

1. Introduction
The accuracy of portrait classification has long been a big challenge to artificial intelligence (AI) applications on image processing. Posture variations, extreme lightings, complex and high background noise, small proportional face area, low quality photos caused by weak camera, bad weather, shortage of light or shake of hands, and many other causes, can all deteriorate the detection and classification effect. As the first basic step and one of the leading directions of portrait classification, face detection has been explored and developed profoundly, and the improvement on it has remarkably facilitated many face applications, such as facial expression analysis. In 2004, Viola and Jones created cascade face detector with high-speed classifiers by training with Haar-Like features and AdaBoost [1]; meanwhile, they introduced the deformable part models (DPM) whose performance is excellent. Whereas, this model may degrade in real-world applications with larger visual variations of human faces even with more advanced features and classifiers (M. T. Pham [2] in 2010, B. Yang [3] in 2014). Besides, their high computational expense and expensive annotation in the training stage are also a headache.

In 2012, convolutional neural networks (CNNs) brought breakthrough to the precision in various computer vision tasks [4] and face recognition [5]. Later, Yang et al. [6] and Li et al. [7] respectively proposed improved CNN models for face detection, but their high computational requirement and ignorance to the interrelation between facial landmarks localization and bounding box regression...
seriously block their wide spreading.

There are also abundant researches on face alignment (T. F. Cootes [8] in 2001, X. Zhu [9] in 2012, X. P. Burgos Artizzu [10] in 2013). Zhang et al. use facial attribute recognition to enhance face alignment performance based on CNN [11]. But most algorithms on face detection and alignment didn’t consider the inherent relation between these two tasks. To solve the problem, Zhang et al. improves the accuracy of multi-view face detection with multi-task CNN [12], but the accuracy is limited by the initial detection windows produced by a weak face detector.

Additionally, mining hard samples during the course of training is essential to the power of detector. However, traditional hard sample mining demands laborious manual operations because of their offline manner. Therefore, an online hard sample mining method for face detection and alignment is imperative to the training process.

The problems mentioned above were generally well improved in the new model Multi-task convolutional neural network (MTCNN) proposed in 2016 [13]. It proposed a new lightweight cascaded CNNs based framework for joint face detection and alignment as well as effective method to conduct online hard sample mining. According to the experiment result, MTCNN produces very good real time and high-precision face detection result.

In this paper, two real applications closely associated with MTCNN are discussed after introducing their face detection model MTCNN and face recognition model FaceNet. The first practice of MTCNN is the character classification function, which automatically finds all the pictures containing the appointed person from a combined album composed of disparate people, and then generate all the personal albums belonging to each individual utilizing the classification to different people. Based on the solution to the first practical target, the second practice is to offer the beauty score by judging the similarity of the left half face and the right half face of a figure in the photo. During all these researches, some effective proposals on increasing the accuracy of the classification are proposed and analysed.

2. The Introduction of MTCNN
MTCNN is a reframed combined CNN model comprised of three layers of networks (As what shown in Figure 1) in following sequence: P-net→R-net→O-net. It utilizes the thought of candidate box and classifier to achieve fast and efficient face detection: P-net is used to produce candidate box rapidly; R-net serves as filter for picking up the candidate box with high accuracy; while O-net is for generating boundary box and key features of the face. The theme framework of MTCNN is similar with cascaded CNNs, but it deals with face area detection and facial feature detection together. Same with many other CNN models targeting to address the image issues, MTCNN also applies image pyramid, bounding box regression and Non Maximum Suppression (NMS) and series of CNN technology.

![Figure 1](image-url)  
**Figure 1.** The architectures of the three networks where “MP” means max pooling and “Conv” means convolution. The step size in convolution and pooling are 1 and 2 respectively.
2.1. The Process of MTCNN

2.1.1. Construction of the Image. Resizing the image with different scales and constructing image pyramid at first, so that the face detection can fit face with different sizes.

2.1.2. P-net. It is a shallow simple fully connected network (FCN) named Proposal Network (P-net), a network for proposing face area which deal with image in following steps:

   **Step1**: Extracting the initial facial features with FCN to decide the bounding box;
   **Step2**: Inputting the features into three convolution layers to identify whether it is a human face by face classifier.
   **Step3**: Filtering most of the candidate windows with Bounding-Box Regression and NMS, obtaining possible face locations with a locator for face features and generating the face area proposals.

2.1.3. R-net. It is a more complicated convolution network named Refine Network (R-net), a network for filtering predicted face windows got from P-net with high precision and optimizing them. For getting more credible and precise face area windows, R-net adds a 128 FCN after the last convolution layer to save more image features than P-net’s $1 \times 1 \times 31$ features. It uses stricter rules to select more carefully, and deletes massive candidate face area windows whose effect are not good enough. At last, R-net optimizes the output result with Bounding-Box Regression and NMS as well.

2.1.4. O-net. It is a relatively complicated convolution network named Output Network (O-net) with one more convolution layer than R-net, a network for outputting final five facial features by supervising the face areas and regressing the facial features. O-net gets more facial features than R-net and adds a 256 FCN in the end so that more image features can be preserved. Based on all these strategies for higher precision, O-net judges the face, regresses the face bounding box and locating the facial features again. After all these intricate steps, it outputs the top left corner coordinate and the lower right corner coordinate together with the five facial features.

2.2. The training method

Each stage of MTCNN networks is a multi-tasks network. The major tasks for each layer are face judgement, Bounding-Box Regression and Feature Location.

2.2.1. Face Judgement. The learning target is a bipartition problem. For each sample $x_i$, it uses cross-entropy loss function:

$$L_{i}^{\text{det}} = - \left( y_{i}^{\text{det}} \log(p_{i}) + (1 - y_{i}^{\text{det}})(1 - \log(p_{i})) \right)$$

$p_i$ is the probability that the face sample $x_i$ predicted by the MTCNN is a really face. $y_{i}^{\text{det}}$ stands for ground-truth, $y_{i}^{\text{det}} \in \{0,1\}$.

2.2.2. Bounding-Box Regression. For each candidate window, the offset (Such as the top left coordinate, the height and the width) between it and the nearest ground-truth is predicted. The learning target is a regression problem. The loss function is the square loss function:

$$L_{i}^{\text{box}} = \|y_{i}^{\text{box}} - y_{i}^{\text{box}}\|_2^2$$

$y_{i}^{\text{box}}$ is the regressed target from the network. $y_{i}^{\text{box}}$ is the ground-truth four dimensional coordinate, including the top left coordinate, the height and the width. The property of the Bounding-Box contains many kinds of relevant labelled information, such as blur, expression, illumination, invalid, occlusion, pose.

2.2.3. Feature Location. It is similar with Bounding-Box Regression. The loss function is as following:
\[ l_{\text{landmark}}^i = ||y_{\text{landmark}}^i - y_{\text{landmark}}^i||_2^2 \]  

Likewise, \( y_{\text{landmark}}^i \) is the regressed feature coordinate from the network. \( y_{\text{landmark}}^i \) is the ground-truth containing five coordinates: two eyes, two corners of the mouth and the nose.

### 2.2.4. Multi-source Training.

As the data set for training are different for disparate tasks during the learning course, when doing one task of training, the loss of other task’s training should be zero. Thus, the combination loss function should be as follows:

\[
\min \sum_{i=1}^{N} \sum_{j \in \{\text{det, box, landmark}\}} \alpha_j \beta_j^i L_i^j
\]  

\( N \) stands for the quantity of the training samples. \( \alpha_j \) is the importance of each task. In P-net and R-net, \( \alpha_{\text{det}} = 1, \alpha_{\text{box}} = 0.5, \alpha_{\text{landmark}} = 0.5 \). While in O-net, for gaining higher precise face coordinates, the parameters are \( \alpha_{\text{det}} = 1, \alpha_{\text{box}} = 0.5, \alpha_{\text{landmark}} = 1 \). \( \beta_j^i \) is the indicator of the sample type, in this case, stochastic gradient decent (SGD) can be used naturally to train these CNNs.

### 2.2.5. Online mining difficulty sample.

MTCCN is different from the traditional way, mining the difficulty samples after training the original classifier to realize the online operation. For each small batch of samples, sort them according to their loss of forward transmission and select the first 70% of as “hard samples”, so that in the converse transmission, only the gradient of the hard samples needs to be calculated. This means those small weak samples which can contribute little to the enhancement of the model function will be ignored. The experiment shows that the method produces relatively better effect that manual selection.

### 2.3. Technology Details

#### 2.3.1. FCN

FCN deletes the full connection layer of the traditional CNN network framework, and then does the converse convolution and sample on the feature map of the last convolution layer or other suitable layer, so that the image can be restored to be the same size with the original picture. Besides, FCN can predict the type for each pixels of the converse convolution, and save the space information of the original image. In addition, during the converse convolution, FCN can make the prediction to the final image by extracting the converse convolution result of the other convolution layers, and proper selection of the extraction can offer better and more precise result.

#### 2.3.2. IoU

The relevance between the final calibrated prediction box of the sub-image and the nature box of the real sub-image (Normally calibrated by hand) is called IOU (Intersection over Union). The habitual standard of the calibration is the intersected area of the two boxes, or the area sum of their combined area.

#### 2.3.3. Bounding-Box Regression

When the value of IoU is smaller than a limitation, one dealing method is to give up the prediction result; while as the aim of the Bounding-Box regression is not to discard the prediction of the step, but to adjusting the result to make it approaches closer to the real value if the former prediction is too far from the real window. So, another way in reality is to address with a linear regression of the loss function, whose input and output are the converted result and the finally suitable result.

#### 2.3.4. NMS

NMS is to restrain the values that are not the maximum. In object detection, this algorithm can eliminate the prediction boxes with high coincidence and inaccuracy. What needs more attention is: this method may be not friendly enough to the coincided objects’ detection. Thus, Soft-NMS algorithm is designed to optimize the problem. It doesn’t delete the suppressed target directly, but
decrease its confidence level. In the end of the detection, it abandons the prediction boxes whose confidence level is lower than the final limitation.

2.3.5. PRelu. The activation function of MTCNN is PRelu, a type of Relu with parameters. PRelu adds parameters to negative value but doesn’t remove with filtration directly. This may lead to more computation and possible over-fitting, but offer better training result by saving more information.

2.4. Effect analysis of MTCNN

2.4.1. The effect of online mining hard sample. Using two O-net with same initial parameters, one addresses data with “Online mining hard sample” algorithm and another doesn’t use it. Their performance compare is as Figure 2 below:

![Figure 2. The Effect of Online Mining Hard Sample](image)

It shows that the O-net with “Online mining hard sample” is more accurate and sensible than the one without it. Therefore, MTCNN with “Online mining hard sample” algorithm produces lower loss.

2.4.2. The Effect of Face Detection. The following Figure 3 exhibits the compare between MTCNN and other algorithms on many data sets.

![Figure 3. MTCNN compares with other algorithms on face detection effect](image)
2.4.3. The Effect of the Joint Facial Landmarks Regression. Training two O-net with same initial parameters, one addresses data with “joint facial landmarks regression” and another doesn’t contain it. The performance of the O-net with “joint facial landmarks regression” is better than the others.

3. The Portrait Classification Based on MTCNN

The outstanding effect of MTCNN on face detection attracts the emergence of many applications. One hot direction is to classify figures according to their faces based on MTCNN and FaceNet. MTCNN is used for face detection and getting the exact face area during which MTCNN helps minimizing the background noise. FaceNet technology is to testify if the two faces are the same people. The general course can be shown as following Figure 4:

![Figure 4](image)

**Figure 4.** The example of portrait classification process based on MTCNN and FaceNet

3.1. Introduction of FaceNet

FaceNet judges by extracting the feature vectors of the two faces and comparing their difference. If the difference is small enough, it will regard them as the same people and classify them into the same type. Its key thought is to map the face image to the multi-dimensional space and present the similarity of the two faces by their space distances (differences). FaceNet utilizes the image mapping based on deep neural network and trains with the loss function based on triplets. The direct output of the network is a 128 dimensional vectors’ space. The structure of the FaceNet if as Figure 5 below:

![Figure 5](image)

**Figure 5.** The structure of the FaceNet network

The batch is the input face images for training. After the deep CNN, L2 is to normalize and get the feature vectors representation of the face images.

3.1.1. Triplet loss. Triplet loss is innovated from the traditional method of loss functions that to mapping face image with a specific feature to the space, which try to differentiate the face image of a specific individual from the face images of other individuals. Triplet is such an example with (anchor, pos, neg), which gets the distances between the triplets and finds out the positive samples’ distance are smaller than the negative ones. By this way, triplets make the final judgement on whether two faces are the same people, and its mathematical expression is:

$$||f(x_i^a) - f(x_i^p)||^2_2 + \alpha < ||f(x_i^a) - f(x_i^n)||^2_2$$

3.1.2. The main steps of FaceNet. The process of FaceNet can be summarized as follows:

**Step1:** At the beginning of the mini-batch, getting face image samples from the training data set, deciding the quantity of samples in each batch and the number of the face images for each person.

**Step2:** Obtaining the the embedding of the sample face images from the CNNs, acquiring the triplets
by calculating the European distances between the embedding of the pictures.

**Step3:** Calculating the triplet-loss and optimizing the model, updating the embedding.

### 3.2 The methods of increasing the precision

In this part, many useful detailed solutions will be advised to raise the classification precision. Because during the real course of portrait classification based on MTCNN, there are many interfering factors that may worsen the result, such as image blurring, too small scale of face in the picture, and the abnormal positions, accordingly, auxiliary measures and algorithm are indispensible to optimize the accuracy.

**3.2.1. Low-Quality Images.** Pictures with low quality expends too much system resources, wastes time and computation, and even cause wrong classifications. Therefore, the first type of problems needs to solve is those which are blur, low resolution, over-exposed images. Laplace detection for blur pictures is suggested to use. Those face images whose fuzziness is larger than the limitation should be deleted in the beginning of MTCNN.

**3.2.2. Small Scale of Face Area.** If the scale of face in an image is too small, while the background contains many other people or things, the difficulty of face detection would be sharply increased, and these images are meaningless for the classification. So, the special step of scale supervision should be added into the detected face area output by MTCNN: if the scale is smaller than a threshold, the image should be discarded, or sent to be resized and detected by a new turn of MTCNN again until its scale fulfils the requirement only if its quality can pass the Laplace blurring test.

**3.2.3. Abnormal Facial Posture.** If the facial pose in the image is too extreme, for example, the face is raised to a too large angle that the distance between the eyes and mouth is too short, the image will bring error for the classification. So, requirement of normal face checking is necessary for increasing the accuracy, for instance, the distance of the two eyes should be larger than at least one third of the face width, or else the lateral angle of the face may be too big for classification and the image would be useless. In a word, filtering the abnormal posed face images is also necessary before MTCNN, and normalization algorithm should also be added before recognizing with FaceNet.

### 4. The Beauty Judgement Based on MTCNN

Based on the Model of portrait classification above, an extensive application based on similarity checking is inspired, naming Beauty Judgement.

There are many beautiful objects fulfilling people’s normal standard of beauty, and research shows that the more the object is symmetrical, the more beautiful it is. Thus, symmetry of an object is regarded as one common checking rule for judging its beauty degree. Based on this common sense, an application for reporting the beauty score of a face can be done by checking the similarity of its left half face and the right half face. This task seems simple and easy at the first glance, for its similarity with the target discussed in the fourth part of the paper. However, careful scrutiny to the topic reveals more challenges:

**Question1:** How to judge and cut the area of left half face and the right half face? Especially, how should this problem be dealt with when the face is not normal and upright?

**Question2:** Whether the MTCNN and FaceNet models can address the half of the face? If they can’t tackle the case, how should we use them to complete the face detection and recognition?

**Question3:** If the factors of the left half face is different from that of the right half face, for example, the light on the left half face is much brighter than that of the right face, how to eliminate the environmental effect?

#### 4.1 The General Process of the Beauty Judgement

First of all, this problem is greatly similar with the first application of MTCNN mentioned above in
the paper, the general process can be decided at first as following:

**Step1:** Image Filtration. Filtering and discarding the images with low quality, abnormal postures and nosy background. Especially, for the faces which turn left or right, limiting its deflecting angle to be less than 45 degrees, or else the face would be too difficult for comparing the similarity of the left and right half faces.

**Step2:** Face normalization.

**Step3:** Face detection with MTCNN. Getting the complete face and cutting off the nosy background the irrelevant people, by this way, the unique target face can be got without interferon.

**Step4:** Division of the left half face and the right half face and reconstruction of the compare target face images.

**Step5:** Input the two newly constructed faces into MTCNN models to detect the new complete faces made by hand again.

**Step6:** Judging the similarity of the two face images with FaceNet and outputting the result.

Specially, **Step2** and **Step5** are the additional steps added specially for the Beauty Judgment Model based on the portrait classification application.

4.2 Solutions to the Mentioned Difficulties

4.2.1 Division of the left half face and the right half face. The first policy to decrease the difficulty is the face angle limitation in the Step1 mentioned in 5.1. The limited deflecting angle can be adjusted according to the real need, and 45 degree is only an experience value for better getting the facial features. The second strategy is to add a step of normalization to the face images before MTCNN. The abnormal posture influence would be mitigated by the first and second measures. Based on these, the third method is to define correct way of cutting the left and right half faces. Using the coordinate of the cross point of the vertical axis of the nose coordinate and the horizontal axis of the two eyes as the top end point; at the same time, using the middle point of the two corners of the mouth as the bottom end point, then the segment of the top end point and the bottom end point could be the line for dividing the left and right half faces. Besides these three measures, the solution suggested in 5.2.2 can help increase the precision of the division as well.

5.2.2. The Reconstruction of the compare targets. After the division of the left and right half of the faces, the left half face can be copied to be its own right face, thus, a new complete face whose left and right half faces are absolutely same would be got and saved as Target Face Image One. Likewise, the right face of the original face image can be copied to be its own left face, and a new complete face image would be created manually and saved as the Target Face Image Two. In this way, the facial features of both the left half face and the right half face could be doubled and amplified, and then the question2 and question3 mentioned in 5.1 can be solved accordingly: because the input for both MTCNN and FaceNet would be complete faces but not half face.

Based on solutions advised above, the problems become easier to solve, and after the result of FaceNet comes out, the score of similarity of the person’s left half face and the right half face could be calculated and offered. If the similarity is higher than a normal experience standard, the result can be “Your beauty degree is 90!” or “You are a beauty!”

5. Conclusion

This paper introduces the important basic model MTCNN at first. It is one of the hottest model used most widely recently for its high precision and outstanding real time performance among the state-of-art algorithms for face detection. Then, the first basic application of portrait classification is researched based on MTCNN and FaceNet. Its direction is one of the most classical and popular area in nowadays AI visual research, and is also the base of many other industrial branches. In addition to the discussion to the basic models, several practical methods are also advised to improve the precision. At last, an interesting and creative research target of Beauty Judgement is discussed based on the
portrait classification model. Deep study reveals many brand new difficulties of the topic. Then solutions and suggestions are proposed with detailed analysis from multi-angles, and the general algorithm with six steps are presented and compared with the portrait classification.

Reference

[1] P. Viola and M. J. Jones, “Robust real-time face detection”. International journal of computer vision, vol.57, no. 2, pp. 137-154, 2004.
[2] M.T.Pham, Y.Gao, V.D.D. Hoang, and T.J. Cham, “Fast polygonal integration and its application in extending haar-like features to improve object detection,” in IEEE Conference, 2010, pp. 942-949.
[3] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features for multi-view face detection,” in IEEE International Joint Conference on Biometrics, 2014, pp. 1-8.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097-1105.
[5] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification verification” in Advances in Neural Information Processing Systems, 2014, pp. 1988-1996.
[6] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.
[7] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.
[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, 2001.
[9] X. Zhu, and D. Ramanan “Face detection, pose estimation, and landmark localization in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879-2886.
[10] X. P. Burgos Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in IEEE International Conference on Com-puter Vision, 2013, pp. 1513-1520.
[11] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision, 2014, pp. 94-108.
[12] C.Zhang, and Z. Zhang, “Improving multi-view face detection with multi-task deep convolutional neural networks” IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1036-1041.
[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao, “Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks”, IEEE Signal Process Lett. 23(10):1499-1503