Fast face detection and localization from multi-views using statistical approach

Teoh, Eam Khwang; Anvar, Seyed Mohammad Hassan; Yau, Wei-Yun

2011

Anvar, S. M. H., Yau, W., & Teoh, E. K. (2011). Fast face detection and localization from multi-views using statistical approach. International Conference on Information, Communications & Signal Processing, 1-5.

<https://hdl.handle.net/10356/100453>

<https://doi.org/10.1109/ICICS.2011.6173610>

© 2011 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/ICICS.2011.6173610>].

Downloaded on 26 Apr 2023 22:43:02 SGT

Fast Face Detection and Localization from Multi-views using Statistical Approach

Seyed Mohammad Hassan ANVAR ^{1,2},

School of Electrical and Electronic Engineering,
Nanyang Technological University ¹,
Singapore

Wei-Yun YAU ²,

Institute for Infocomm Research ²,
A*STAR, Singapore

Eam Khwang TEOH ¹

School of EEE,
Nanyang Technological University ¹,
Singapore

Abstract—Window-based face detection methods are fast. However their results are coarse, pose dependent and require fine face alignment for face analysis. Recently a statistical approach is introduced by Toews and Arbel [1], which is able to detect faces in multiple poses and does not require face alignment. However, their method is slow compared to the window-based method. In this paper, we proposed a method, which capable of detecting faces in multiple poses in near real time and also does not require face alignment. Experimental results show that our proposed method has comparable accuracy with the Toews and Arbel's method but has significantly lower processing time.

Index Terms—Face detection; localization; scale invariant feature; statistical method; real-time.

I. INTRODUCTION

HUMANS can recognize face images without significant effort but this problem is still a challenging topic in the area of computer vision [2]. Automatic face recognition problem has attracted much attention during the past two decades. One of the crucial issues in face recognition challenge is automated detection and localization of faces. This stage is needed in order for the computer to extract appropriate face images to the face recognition engine to do the identity verification. The result of this stage can affect the overall result of other stages significantly as wrongly detected or un-detected faces will cause errors. Other than the performance, the speed of this stage is also important to ensure that the overall processing time is within an acceptable duration. Obviously, it is crucial for face detection and localization to be completed as fast as possible, i.e. in real time. Recently, many approaches based on sliding window have been proposed for face detection such as Viola and Jones method [3], Li and Zhenqiu [4], Huang et al.[5] and Schneiderman and Kanade [6]. They used a sliding window in multi scales, then searches every pixel of the given image to determine the existence of a face. For multi-view face detection, the entire processing has to be repeated for different poses. There are some issues with the window-based methods. First of all, they yield coarse face detection result, which still require fine face alignment. However, the fine alignment solution, such as proposed by Batur and Hayes [7], is not able to be processed in real-time (require a few seconds to align a single face). Face alignment typically uses facial landmarks. Thus, it will fail for profile faces where the landmarks are not all available, making the approach not workable for scenarios requiring multi-pose faces.

Recently a statistical method has been proposed by Toews and Arbel [1], which does not need face alignment. This method is able to detect faces in the given image and estimate a scale and position for them. However, their method is quite slow compared to the window-based methods.

In this paper, a statistical approach is proposed. It is similar to the method proposed by Toews and Arbel [1] where it can detect faces in multiple pose and does not require fine face alignment. However, unlike [1], it requires only manual intervention in a single face image and the proposed approach is more than 4 times faster compared to [1]. Experimental results show that the performance of our method seen in the Receiver Operating Characteristic (ROC) curve is almost similar or in some cases better than the method in [1].

II. METHOD

In the proposed method, first a face model is shaped based on scale invariant features extracted from training images. This model is then used to detect and localize faces in the given images.

The model construction starts with determining two reference points in every image of the training set. These two reference points should be selected such that they would be visible in the all poses of a rotated face. For example, if the head has up and down rotation (elevation angle), the best reference points invariant to the rotation are the center of two eyes. Alternatively, if the head has left and right rotation (azimuth angle), the best reference points are the points between the two eyes aligning to the nose and immediately below the nose. Once the reference points are determined in all the images of the training set, SIFT features are extracted using the method proposed by Lowe [8]. Each feature obtained from this method contains geometric and appearance descriptors. Geometric descriptor for feature f_i includes (x_i, y_i) position, scale σ_i and orientation ϕ_i . The appearance descriptor for feature f_i is an array of length 128 bins from the gradient magnitudes of pixels in a region with scale σ_i around the position (x_i, y_i) . The model construction proceeds with normalizing the geometric descriptors of each extracted feature in image k with the position (x_r^k, y_r^k) , scale σ_r^k and orientation ϕ_r^k of the reference points in that image. Since we have two reference points, it is easy to assign them a position, scale and orientation. One of the points are selected as reference position (x_r^k, y_r^k) . For instance, the

position of left eye or the position of the nose base, and the length of the connection line between two reference points in pixel unit is considered as the reference scale σ_r^k and the angle between this line and the horizontal line is selected as the reference orientation ϕ_r^k . Whenever all features are normalized to their related reference points, they all shape a *primitive* model with hundreds of thousands of features, depending on the number of training images and the number of features extracted per image. Not all these features are informative. Therefore, a mechanism should be considered to remove redundant features. We used likelihood ratio test in (1) and evaluated it for every feature in the primitive model to find the most informative features.

$$\rho_i = \frac{p(f_i|face)}{p(f_i|\overline{face})}, \quad (1)$$

where the term $p(f_i|face)$ means the probability that feature f_i comes from a real face image and term $p(f_i|\overline{face})$ tells the probability that feature f_i comes from the other object and of the background areas of the face. These two terms are not easy to determine in an exact manner. An approximation to this ratio can be obtained by comparing the appearance descriptor of feature f_i to the appearance descriptors of the other features in the *primitive* model in the following way: The appearance descriptors of all other features which are similar to the appearance descriptor of feature f_i within a threshold Thr_i^a are determined and those features, which present in a vicinity of feature f_i in the normalized space, are considered as agreement group and others as non-agreement group. The ratio between the number of members in these two groups can be taken as an approximation of the likelihood ratio test for feature f_i . The value of threshold Thr_i^a for feature f_i is determined such that it maximizes the likelihood ratio in (1) as it is shown in (2).

$$Thr_i^a = \operatorname{argmax}(\rho_i) \quad (2)$$

After determining the likelihood ratio for every feature in the *primitive* model, *intermediate* model is formed by applying a threshold to the ratio in (1) and pruning the features with low probability. This procedure can reduce the number of features by approximately a factor of ten to twenty. However, this amount of features in the *intermediate* model still contains many redundant features. Therefore, a clustering approach is applied to reduce the model features further. For every feature f_i in the *intermediate* model, the binary similar appearance features matrix χ in the neighbourhood area of the features are determined. Cell $\chi(i, j)$ denotes the similarity value between feature f_i and f_j as illustrated in (3).

$$\chi(i, j) = \begin{cases} 0, & f_i \text{ is not similar to } f_j \\ 1, & f_i \text{ is similar to } f_j \end{cases} \quad (3)$$

The number of similar features to the feature f_i are computed by summation of the rows in the matrix χ . These

values indicate the similarity value $\pi(f_i)$ for feature f_i as illustrated in (4).

$$\pi(f_i) = \sum_{j=1}^N \chi(i, j) \quad (4)$$

where N is the number of features in the model. The similarity members values $\pi(f_i)$ are sorted in descending order and from the top of the list downward ($i = (1, 2, \dots, N)$). The median for each group is taken to represent the descriptor of that group. Each feature in the *intermediate* model can only be a member of one group. Therefore, if a feature was a member in multiple groups, that feature is considered as a member in the group with the maximum number of members and its membership to the other groups is discarded.

Applying the clustering algorithm to the *intermediate* model reduces the number of features by approximately another factor of ten to twenty, resulting in about a thousand features in the *final* model. This model is used to detect faces from arbitrary views in the given images. This stage is very crucial because removing the redundant features not only speeds up the algorithm but also helps better detection.

The test stage begins with extracting SIFT features from the test image. The appearance descriptors of the extracted features are compared to the appearance descriptors of the model features to find the possible matches. This comparison is performed based on an Euclidean distance between features' appearance descriptors that should be less than a threshold Thr_i^a obtained in the training stage. Each possible match to the model suggests the existence of a face in the given image. Considering all the face positions suggested by possible matches, the probability of presenting a face typically concentrates in some regions in the given image. In order to find the center of mass of these clusters and remove sparse points, all the possible or candidate locations found are clustered using mean-shift [9]. For M cluster centers determined in the mentioned method, the existence of a true face in the place of the cluster center r , ($r \in [1, 2, \dots, M]$) is checked using the likelihood ratio test in (5).

$$\delta(r) = \prod_{i=1}^N \frac{p(f_i|face)p(face)}{p(f_i|\overline{face})p(\overline{face})} \quad (5)$$

where N is the number of features in the model and f_i 's are the features underlying the cluster r . This function includes two parts: the likelihood ratio and the prior ratio. The first part was evaluated before in (1). The prior ratio is a constant and can be considered in the threshold term. A threshold on $\delta(r)$ determines the existence of a true face in the location of cluster r .

III. EXPERIMENTAL RESULTS

In this section, the face detection and localization method proposed by Toews and Arbel [1] is implemented by the

Figure.1 Compares the processing time between our proposed method and the method proposed by Toews and Arbel [1]. The processing time indicated is based on the average time taken for each method to process a test image of resolution 320×240 pixels in a total of 500 images. As shown in this figure, our method is more than 4 times faster than the method in unlike [1]. We also implemented our method in C++ programming platform.

author and their results are compared to the results of the method proposed by the author in this paper in the same training and test data and conditions (same PC and programming platform).

A. Data description

Images and data, which were used in this experiment, come from the standard color FERET database [10]. This dataset contains single face of different persons in the resolution of 768×512 pixels. Since the image scales are very large, they are resized to the resolution of 320×240 pixels and converted to grayscale. It includes 1209 persons with various age, gender and ethnicity in different scales and illumination conditions. An average of 10 images per person were captured and covered a wide range of face rotation from left profile to oblique, frontal and right profile.

The other dataset, which was used in this experiment, is the CMU profile face images [11]. This dataset contains faces taken in high clutter background and one image may include more than one faces and the faces could be in different views. The size of the images in this dataset is different and the resolution of faces is poor. Among the images in this dataset, only the images containing faces where the length of the nose from the mid-point between the two eyes to the bottom of the tip of the nose were of at least 15 pixels or more, were selected. We found that below this size, the resolution is not high enough for the SIFT features to be reliably extracted.

B. Face Detection and Localization Results

1000 persons out of a total of 1209 were randomly chosen and separated equally into either the test or training groups. Among all the images of each person, one image from random view is selected. Therefore each training and test group contains 500 images of random person in random pose. Both the method proposed by the author and the method from Toews and Arbel [1] (implemented by the author) are applied to the image from the training and test groups. The average processing time of an image computed from a total of 500 images are compared and illustrated in

Figure.2 Compares the processing time between our method and the methods proposed by Huang et al. [5] and Li and Zhenqiu [4]. These methods are designed to detect faces from multi-views. However, they only yield coarse face detection in the given images and an alignment algorithm is required. Alignment algorithm is time consuming. It is seen in this figure that our method has an acceptable processing time considering that it also provides the localization and pose estimation. Although the other methods are faster than our method, the average fine alignment time of about 1 sec has to be added to them. Thus, our method will be faster if the total time together with fine alignment is taken into account.

Fig. 1. The implementation is done on Matlab programming platform. The results show that our method is more than 4 times faster than the Toews and Arbel's method in a typical image of 320×240 pixels. We also have implemented our algorithm in C++ programming platform to compare the speed of this method to the speed of the window-based method. As it is demonstrated in Fig. 2, the time consumed by our method for a typical image of size 320×240 is less than one second and is about twice the time of the window-based face detection methods proposed by Li and Zhenqiu [4] and Huang et al. [5]. However, the time measured for our method is for face detection including geometric localization and alignment. However, the time measured for the window-based methods is only for face detection and does not include fine alignment. If we add the alignment algorithms which will take more than 1 sec, then their running time would be more than our method. Our code also has not been optimized and it is possible to further reduce the processing time. In addition, we found out that more than half of the reported time for our method (about 500ms) is taken by the SIFT feature extraction algorithm [8]. We could investigate using other faster feature extraction algorithm in the future.

To compare the performance of our method with that of Toews and Arbel [1], we plot the Receiver Operating Characteristic (ROC) curves based on 500 randomly selected test images from FERET dataset. The performance comparison for is shown in Fig. 3. It is shown the performance is as slightly better if not as good as [1]. Finally, we evaluated the proposed method with the method from Toews and Arbel using some images from the CMU dataset which contain faces satisfying the minimum size criteria. As shown in Fig. 4 (Left column), our method is able to detect and localize the face images of arbitrary pose accurately, so long as the faces are not severely occluded or smaller than the required minimum size. The estimated pose is quite similar to what we labeled in the training stage (Two reference points, one between the two eyes and the other below the tip of the nose). Comparing the results of our

Figure.3 Illustrates the Receiver Operating Characteristic (ROC) curves obtained from 500 randomly selected test images taken from FERET dataset. Although our method is much faster than Toews and Arbel's method, the performance of our method is slightly better than the performance obtained from the method of Toews and Arbel.

method (Fig. 4 (left column)) with the results of Toews and Arbel's method (Fig. 4 (right column)), we are able to detect all the cases detected by their method and also able to detect cases which they missed as shown in Fig. 4 (e) and (f).

IV. CONCLUSION

A probabilistic method has been proposed to detect and localize face images from arbitrary view. This method is faster than the current state-of-the-arts for multiple pose face detection and localization proposed by Toews and Arbel [1]. This is achieved without degrading the accuracy in the detection. In addition, half of the time consumed by our method is required by the SIFT feature extraction algorithm to extract the scale invariant features from the given image. However, our algorithm does not depend on SIFT features and any type of scale invariant features could be used, such as SURF [12] and SCARF [13] that are faster than SIFT. As a future work we intend to improve our code by replacing the SIFT features with other scale invariant features and optimizing the code to obtain a completely real time implementation of our method.

ACKNOWLEDGMENT

This research was supported by the A*STAR agency and the Institute for Infocomm Research of Singapore.

REFERENCES

- [1] M. Toews, and T. Arbel, "Detection, Localization, and Sex Classification of Faces from Arbitrary Viewpoints and under Occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1567-1581, Sep, 2009.
- [2] R. Chellappa, P. Sinha, and P. J. Phillips, "Face Recognition by Computers and Humans," *Computer*, vol. 43, no. 2, pp. 46-55, 2010.
- [3] P. Viola, and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, May, 2004.
- [4] S. Z. Li, and Z. Zhenqiu, "FloatBoost learning and statistical face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1112-1123, 2004.
- [5] C. Huang, H. Z. Ai, Y. Li et al., "High-performance rotation invariant multiview face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 671-686, Apr, 2007.
- [6] H. Schneiderman, and T. Kanade, "Object detection using the statistics of parts," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 151-177, Feb-Mar, 2004.

Figure.4 Compares the results of applying our method (Left column) to some images taken from the CMU dataset with the results of Toews and Arbel's method (Right column). It is seen that whenever the face images have sufficient resolution or not severely occluded, our method is able to detect and localize them. The obtained results in most of the cases are similar or better than the results of Toews and Arbel's method. However, our method is able to detect and localize faces more than 4 times faster than their method.

- [7] A. U. Batur, and M. H. Hayes, "Adaptive active appearance models," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1707-1721, 2005.
- [8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, Nov, 2004.
- [9] D. Comaniciu, and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, May, 2002.
- [10] FERET. "The Color FERET Database," available at : <http://face.nist.gov/colorferet/>, 2011.
- [11] CMU. "Frontal and profile face databases," available at : <http://vasc.ri.cmu.edu/idb/html/face/>, 2011.
- [12] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Computer Vision - Eccv 2006, Pt 1, Proceedings, Lecture Notes in Computer Science* A. Leonardis, H. Bischof and A. Pinz, eds., pp. 404-417, Berlin: Springer-Verlag Berlin, 2006.
- [13] J. T. Stephen, A. M. Bruce, and A. S. Karl, "Real-Time Robust Image Feature Description and Matching," in *Computer Vision - Accv 2010*, Queenstown, 2010.