ROBUST AND FLEXIBLE LANDMARKS DETECTION FOR UNCONTROLLED FRONTAL FACES IN THE WILD

A. Liang*, C. Wang, W. Liu and L. Li

Department of Computing
Curtin University, Perth, Western Australia

Abstract. In this paper, we propose a robust facial landmarking scheme for frontal faces which can be applied on both controlled and uncontrolled environment. This scheme is based on improvement/extension of the tree-structured facial landmarking scheme proposed by Zhu and Ramanan. The whole system is divided into two main parts: face detection and face landmarking. In the face detection part, we proposed a Tree-structured Filter Model (TFM) combined with Viola and Jones face detector to significantly reduce the false positives while maintaining high accuracy. For the facial landmarking step, we improve the accuracy and the amount of the facial landmarks by readjusting the face structure to provide better geometrical information. Furthermore, we expand the face models into Multi-Resolution (MR) models with the adaptive landmark approach via landmark reduction to train the face models to be able to detect facial landmarks on face images with resolutions as low as 30x30 pixels. Our experiments show that our proposed approaches can improve the accuracy of facial landmark detection on both controlled and uncontrolled environment. Furthermore, they also show that our MR models are more robust on detecting facial components (eyebrows, eyes, nose, and mouth) on very small faces.

1. Introduction. Detecting faces and plotting landmarks on each facial components (such as eyes, nose, and mouth) on a large set of images manually by human is a time consuming and tedious task. In Computer Vision field, this task is an important pre-processing before conducting face recognition, face tracking, and facial expression recognition on images or videos [2]. This is why developing systems for automatic face detection and its landmarks is an important research topic. However, even after a vast amount of recent developments, the problems still cannot be solved perfectly. The challenges include processing images with extreme illumination, occlusion, pose variation, image quality/resolution, and facial expressions.

Human face detection, feature extraction and analysis are three main problems which are foundational tasks in face recognition with computer vision [30]. A reliable and robust solution for face detection can speed up the progress on other high level research topic, for example, face recognition, tracking and 3D face reconstruction [2]. Although a lot of study have been conducted for face detection, it still remains a challenge to obtain accurate facial landmarking result to support higher level
research in computer vision. Therefore, a highly accurate, reliable face detection and landmarking approach is highly on demand [2].

In this paper we aim to propose a facial landmarking system for frontal faces in uncontrolled environment [18, 19]. Basically there are two parts in the system (Figure 1). The first part is the face detection. In this part we propose a Tree-structured Filter Model (TFM) in combination with Viola Jones face detector [26]. This will increase the time efficiency as well as reducing the false positive rate in detection of faces in the whole system. In the second part, we improve the face models proposed by Zhu and Ramanan [30] for face landmarks detection with more landmark points for better face representation. We improve the number of landmarks which leads to better accuracy and description of the facial components. Furthermore, we also enhance the capability to explore smaller faces with the proposed Multi Resolution (MR) model.

In this paper, we emphasize on facial images captured in the wild with various resolutions due to a fact that images are often captured in uncontrolled environment in practice. There is no guarantee that the human faces are in high quality/resolution. One of such examples is the Boston Marathon Bombings incident on 15 April 2013 [15]. Law enforcement has to investigate a significantly large amount of video footage and images from various devices such as surveillance cameras, onlookers cameras, and smartphones to find the possible suspects. The captured images of both suspects are among large amount of people on the street with low quality/resolution. This is why it is necessary to be able to detect the face and apply facial landmarks on various resolutions automatically to assist face recognition process.

The motivation of this paper is based on our study for current related works as discussed in section 2. The current face landmark detectors usually have insufficient number of landmark points and thus might not be sufficient to describe the facial components well. Furthermore, current face detectors are usually prone to high false positives when detecting multiple faces from the images in uncontrolled environment.

In summary, we make the following important contributions in this paper:

• With the proposed light-weight face model which acts as a ‘filter’ as well as in combination with the Viola Jones face detector [26], we can significantly reduce a lot of false positives on face detection in uncontrolled environment. This can be done with relatively high speed since we avoid redundant computations.
• We improve the current landmark detector with more landmark points for frontal faces and this improves the accuracy of face detection and facial feature description significantly.
• We develop an automatic adaptive landmark selection scheme for faces in various resolutions down to 30x30. This allows us to train face models which are capable of detecting facial landmarks from multiple frontal/near-frontal faces in uncontrolled environment where the size of the faces might vary greatly.

The rest of this paper is structured as follows. Section 2 briefly describes some approaches which are closely related to our research. Since facial landmark detection is a complex process, we break the research into two parts. The first is on improving face detection which is discussed in section 3. Enhancing face landmark detection is the second part which will be explained further in section 4. After successfully improving on both parts, we combine the proposed approaches in a single facial
Figure 1. Our proposed facial landmarking system.
landmark detection system in section 5. Section 6 summarizes all the discoveries and experiments. Furthermore, the limitations and future works will be briefly discussed.

2. Related Works. The foundation of the proposed approaches in this paper is based on the pictorial tree-structured face models proposed by Zhu and Ramanan [30, 8, 7]. Their approach is one of the best state-of-the-art facial landmarking techniques. It is able to perform three tasks, which are face detection, facial landmarks detection, and pose estimation. It can be applied on images in an uncontrolled environment.

The face models are defined as the connected facial landmarks on a tree-structure. Each landmark is represented as a node and the tree structure is essential to optimize the face models with dynamic programming and maintain the global elastic structure of the face. Each landmark is associated with HOG (Histogram of Oriented Gradients) features [5]. Furthermore, each face model represents a face structure in a particular expression and pose. Zhu and Ramanan provide 3 different set of face models [29] called Independent-1050, Share-146, and Share-99. All of them are trained on 900 face images (650 for Share-146 and Share-99) from CMU MultiPIE dataset [12] for positive images set and 1218 non-face images from INRIAPerson dataset [5] for negative images set. The difference between the Independent-1050 model with the other two models is the way these landmark features are trained. In Independent-1050, each landmark from each face model is trained independently regardless of the similarity among them. In total it has 1050 unique landmarks across all the face models in the set. Meanwhile, in both Share-146 and Share-99, there are some groups of facial landmarks which share the same feature description which means one landmark training is sufficient on multiple landmarks on the face models [25]. The reason for this sharing method is because some facial landmarks have a high feature similarity on different expressions or poses. For example, the corner of the eyebrow might not change too much whether the person is smiling or not (neutral). Zhu and Ramanan manually choose which landmarks to be shared with particular conditions (146 unique landmarks for Share-146 and 99 unique landmarks for Share-99). As a result, it improves the speed with a bit of sacrifice on accuracy. The Share-99 model is approximately 10 times faster than the Independent-1050 model.

We observed that there are some restrictions on their approaches which we want to overcome in this paper. Firstly, the faces are not detected before the facial landmarks detection, but simultaneously as the approach attempts to find well-corresponded facial landmarks from the whole image. Despite the fact that the face detection rate is satisfactory for normal resolution images, the computational cost could grow prohibitively large quickly as the image size increases (not efficient). So, there is a need to utilize a reliable approach that is able to detect faces quickly, so most non-faces regions can be discarded for efficiency purpose. Secondly, the amount of the landmarks are fixed with only 68 points for frontal faces and 39 points for side/profile faces. Even though these amounts might be sufficient for some face-related applications, we found that the information is not enough to define geometric information of the facial components. As an example, Conilione and Wang [3] propose a frontal face image retrieval method based on the semantic information of the facial components (e.g., shapes, size). 6 landmarks on each eye will not provide reliable semantic features. Furthermore, the accuracies of the landmarks
are occasionally not satisfactory. Lastly, the size of the faces must be at least 80 x 80 in resolution. We consider this as a limitation since it is often desirable to conduct face recognition on low-resolution faces [28].

As our scope is focused on frontal faces and one of our aims is to efficiently discover frontal faces prior to facial landmarking phase, we applied the technique proposed by Viola and Jones [26] for real-time detection. Their technique is well-known for its high speed and accuracy. The efficiency is the result of the following three features: 1) They implemented a novel and simple image representation called Integral Image to quickly compute the features; 2) The classifier is built with the idea of AdaBoost learning [10] which allows the system to select features from a large set of features. 3) The idea of combining the trained classifiers in a cascade structure to remove most of the non-faces region in the early phase, therefore reducing redundant computation. However, the false positive rate is very high when we use this detector for images with many faces with different resolutions. In our proposed system, we will use this approach at the first stage with the false positives rate significantly reduced using our proposed face models. The detail will be discussed in section 3.

There are three other state-of-the-art facial landmarking techniques which will be used for comparison in this paper.

The first one is CompASM approach proposed by Le et al [17]. It is an extension from the classic approach Active Shape Model (ASM) [4]. The improvement is based on the limitation of ASM where the facial landmarks model is constrained to each trained Gaussian model. This restriction makes it difficult to cover large amount of varieties on facial expressions, especially irregular expressions such as blinking (only one eye closed). Le et al overcome the problem by modifying the face models as a combination of 7 independent facial components (jawline, nose, lips, left eye, right eye, left brow, and eye brow) connected in a pictorial structure manner [8]. As a result, it is easier to handle various facial expressions since each facial component is trained separately and only combined for relative positions on the configuration model. Furthermore, they also applied Viterbi optimization [9] for the fitting step on facial contours.

The second one is STASM proposed by Milborrow and Nicolls [23]. This is another improvement of ASM [4]. The modifications are focused on the use of simplified form of SIFT features [20] and the feature matching approach of Multivariate Adaptive Regression Splines (MARS) [11]. These improvements made it possible to accurately detect 77 facial landmarks on multiple frontal/near-frontal faces in real time. However, we observe that this approach completely relies on the prior information of the face location from other approach (Viola and Jones face detector is used in their source code). As a result, if the face detector produces a false positive (non-face region), STASM will still impose the facial landmarks on it [22].

The last one is the Intraface approach proposed by Xiong and Torre [27]. They implemented the Supervised Descent Method (SDM) for the purpose of optimizing Non-linear Least Squares (NLS) function. This is applied on the face alignment for efficiency purpose. This adjustment enables Intraface to detect 49 facial landmarks on images or videos (face tracking). Besides, it is also able to handle various poses except the side/profile faces.

3. **Tree-structured Filter Model (TFM).** The process of face landmarking begins with locating the coordinates of the faces. As mentioned in section 2, we decided to use Viola and Jones [26] to detect the faces because of its efficiency.
Algorithm 1  Face Detection with Viola Jones (VJ) face detector and Tree-structured Filter Model TFM

1: procedure DETECT_FACES(img) \(\triangleright\) img is the image query

\[
\begin{align*}
VJ\_threshold & \leftarrow 3; \\
TFM\_threshold & \leftarrow -1.1; \\
VJ\_faces & \leftarrow []; \\
final\_faces & \leftarrow [];
\end{align*}
\]

2: \(VJ\_faces \leftarrow VJ(img, VJ\_threshold)\) \(\triangleright\) detect faces with VJ approach

3: if \(VJ\_faces\) is not empty then

4: \(n \leftarrow \text{length}(VJ\_faces)\) \(\triangleright\) \(n\) is the number of faces detected

5: for \(i = 1\) to \(n\) do

6: \(\text{face} \leftarrow \text{crop}(VJ\_faces(i))\) \(\triangleright\) crop only the face region

7: \(\text{score} \leftarrow \text{rescale}(\text{face}, 40, 40)\) \(\triangleright\) resize the face into 40x40 pixels

8: \(\text{score} \leftarrow \text{TFM}(\text{face})\) \(\triangleright\) verify the face with TFM

9: if \(\text{score} \geq \text{TFM\_threshold}\) then

10: \(\text{final\_faces} \leftarrow \text{final\_faces} + \text{face} \triangleright\) accumulate faces that pass the TFM threshold

11: end if

12: end for

13: else

14: print ”no face detected in this image.”

15: end if

16: return \(\text{final\_faces}\)

17: end procedure

However, even with its high accuracy, it is also prone to high number of false positives. That is why we propose a 'light-weight' Tree-structured Filter Model (TFM) based on the concept proposed by Zhu and Ramanan [30] to significantly reduce false positives. The purpose of this TFM is to 'filter' the face candidates passed by the Viola and Jones face detector. By combining Viola Jones face detector and the proposed TFM, we expect it to improve the face landmarking system in three aspects. Firstly, by knowing the faces location prior to landmarking process, it significantly reduces the region of interest which leads to lower computation. Secondly, we can exploit the information on the approximate face sizes to decide which face model is suitable for landmarking (more detail in section 4.2). The last is the significant decrease in the false positives which also leads to further speed improvement and higher accuracy.

The Pseudo Code is shown in algorithm 1. After conducting some experiments on various thresholds, we choose the merging threshold as 3 rectangles for Viola and Jones face detector and -1.10 for the TFM score threshold. We choose these thresholds because it significantly reduce false positives while still maintaining high rate of true positives.

3.1. Model Training. The TFM model is trained with the source code provided by Zhu and Ramanan [29] as seen in figure 2. The proposed TFM is trained on 10% scale level (face size approximately 30x30) of frontal face images from AR database [21]. Since TFM is developed only to ‘filter’ the results from Viola and
Figure 2. 2 face models on TFM. It includes neutral (left) and scream (right) expression.

Jones face detector [26], the variety of the facial expressions is reduced. We only include neutral and scream expression from 224 face images for efficiency purpose. We choose 12 landmarks for the training. These landmarks are 2 eyebrow centres, 2 eye centres, 1 nose tip, 2 mouth corners, and 5 landmarks on jawline. These landmarks are chosen because we believe they are sufficient to closely 'describe' a human face. Furthermore, in addition to the 1218 images from INRIA dataset [5], we also add 1650 random non-faces images as the negative image set for training session in order to improve its capability to distinguish between faces and non-faces.

Since the Viola and Jones face detector is able to detect multiple frontal/near-frontal faces quickly, we also have to avoid high processing overhead from TFM. In order to achieve that, we scaled all the face images from Viola Jones face detector down to 40x40 for efficient computation. Since TFM is also trained on small faces anyway, it makes sense for it to work well on detecting the scaled-down face images regardless of the original face sizes.

3.2. FDDB Database. Face Detection Data Set and Benchmark (FDDB) [13] is an excellent dataset to test the performance of face detector approaches (not facial landmarks detector). It provides a large set of faces in uncontrolled environment. It contains 5171 faces derived from 2845 images on Faces in the Wild dataset [1]. For experimental purposes, all the images are classified into ten subsets for cross-validation evaluation. The ground truth and source code to produce the ROC curve are also provided.

Since the scope of this research is emphasized on frontal faces, we manually select 1535 images which contain 2130 frontal/near-frontal faces for our testing set. Some examples can be seen in figure 3.

3.3. Performance Evaluation. In order to prove the effectiveness of the TFM on reducing false positives while maintaining high accuracy, we conducted the performance evaluation based on the Receiver Operating Characteristic (ROC) curve. This curve represents the number of both true positive and false positive at various matching scores. A good approach will need to produce true positives as high as possible while keeping false positives as low as possible.

We compare our proposed TFM with SHARE-146 model [30] and Viola Jones face detector itself [26]. SHARE-146 is chosen because this face model is able to detect faces with the lowest resolution compared to all face models provided by Zhu
Figure 3. Some chosen frontal faces samples from FDDB.

and Ramanan, so it is more likely to detect faces in uncontrolled environment where there are various sizes of the faces. Share-146 contains 13 face models on various poses. In order to compare fairly, we only test the frontal face model.

3.4. Comparison with VJ and Share-146. As shown in Figure 4, Viola and Jones (VJ) face detector (green/dashed line) manages to achieve the highest accuracy but also produces the highest rate of false positives. After combining it with TFM (blue/solid line), the amount of false positives can be reduced significantly while only losing a small number of true positives. Some examples can be seen in figure 5. Share-146 (red/dotted line) is able to acquire a high level of true positives with relatively low false positives. However, it is still slightly worse than our proposed TFM. Since SHARE-146 has to explore the whole image to find the faces, the computational cost is significantly higher compared to TFM which emphasizes on particular region of interest detected by Viola and Jones face detector. Table 1 shows the average computational cost to process 1535 images.

|      | VJ  | VJ+TFM | Share-146 |
|------|-----|--------|-----------|
| Time | 0.074 | 0.237  | 3.804     |

4. Face Landmarks Detection.

4.1. Accuracy Improvement.
4.1.1. Tree Reconstruction. As proposed by Millborrow and Nicolls [24], a large amount of facial landmarks involved in a single face model is able to increase the likelihood of better fitting of the landmarks on the face. In other words, it is possible to improve accuracy of facial landmarks detection by increasing the density of the landmarks. Based on this information, we revised the tree structure of the face models proposed by Zhu and Ramanan. The revision is emphasized on the primary facial components such as eyebrows, eyes, nose and mouth. In order to provide better geometric information (e.g. shape and size), the additional landmarks have to cover facial components as a region instead of a line (See figure 6 and 7).

4.1.2. AR Database. We choose frontal face images from AR Database [21] for training the new set of face landmarks. Among various face databases, AR database provides facial landmark ground truth [6] that meets our expectation. It contains 130 landmarks circumscribing all facial components which describe the contour and shape (Figure 8). Moreover, this database is free for research purpose. In total, there are 896 set of ground truth for 896 frontal face images. It consists of 4 facial expressions (neutral, smile, angry, scream) from 112 people (58 men and 54 women) taken on 2 different period/sessions (2 weeks apart). (896 = 112 people x 4 facial expressions x 2 sessions).

From the source code provided by Zhu and Ramanan [29], we trained 4 face models from 4 facial expressions with 112 images each on the first period/session. Figure 9 visualizes the proposed model. We also used the same negative images set which are 1218 images from INRIAperson Database [5].
4.1.3. Performance Evaluation. To evaluate the performance of the proposed tree-structured face model, we compare the landmark accuracy with two other approaches. The first experiment is compared with the Independent-1050 face model proposed by Zhu and Ramanan. In the second experiment, CompASM approach proposed by Le et al [17] is chosen. Both experiments will be conducted on 448 images from the second session of AR Database.

As each approach deals with different number of facial landmarks, we cannot compare the entire landmarks directly. Therefore, particular common landmarks are chosen which are also the primary/fiducial landmarks as suggested in [2]. Some examples of primary landmarks are the nose tip, eye corners and mouth corners. This type of landmarks are significant due to its reliability on face recognition and tracking with a large variety of facial expressions.
After conducting intensive observation on all approaches, we decided to choose 15 and 17 primary points for Independent-1050 and CompASM respectively as seen in figure 10. The measurement will based on the relative error from the ground truth. Let $G$ as the landmark ground truth set and $R$ as the result of automatic facial landmark detection. The relative error of particular landmark $j$ on a single image $i$ is defined as:

$$\text{Relative Error} = \frac{|G_j - R_j|}{|G_j|}$$
This measures the Euclidean distance between both common landmark $j$ from $G$ and $R$. The lower the value, the better it is. The distance has to be further normalized by dividing with the Inter-Ocular Distance (IOD) of the face in image $i$. IOD is the distance between two eye centres which can be calculated from the provided landmark ground truth coordinates. As stated in [2], this normalization is necessary to avoid dependency between the relative error and the size of the faces.

If there are $L$ landmarks to be tested ($L = 15$ for Independent-1050 and $L = 17$ for CompASM) on image testing set as many as $N$ ($N = 448$), then the average of the relative error can be defined as:

$$\text{dist}_{\text{IOD}}(G_i^j, R_i^j) = \frac{||G_i^j - R_i^j||_2}{\text{IOD}_i}$$

(1)
Even though the relative error is a simple and robust way of accuracy measurement, it is prone to outliers which might skew the result significantly. To further improve the validity of the results, we also measure the detection rate of the chosen landmark \( \mathbf{L} \). The concept behind this evaluation is to check whether the detected landmarks are reasonably close to the ground truth (e.g., below particular threshold). If the distance is too large, it will be considered "not detected." We use the same evaluation as in [2]. Three thresholds are defined: 20%, 10%, and 5% of the IOD. The visualization for the thresholds can be seen in figure 11.

Since the purpose of increasing the landmarks is also related to geometric information, we also measure some basic information. We measure the width and height difference with the ground truth \( \mathbf{G} \) on both eyes and mouth. Let \( \mathbf{W} \) as the width of facial component \( c \) from facial components set \( C \) (\( C = \{ \text{left eye, right eye, mouth} \} \), \( |C| = 3 \)). The evaluation will be defined as:

\[
\text{width}_{\text{IOD}}(G^c_i, W^c_i) = \frac{\text{width}_{\text{diff}}(G^c_i, W^c_i)}{\text{IOD}_i}
\] (3)
It is quite similar to relative error measurement where we measure the difference (the width in this case) normalized with the IOD on the corresponding image. For each facial component $c$, the average difference is defined as:

$$\text{Width diff.}^c = \frac{\sum_{i=1}^{N} \text{width}_\text{IOD}(G^c_i, W^c_i)}{N}$$  \hspace{1cm} (4)$$

The measurement on height difference is also similar to the width difference. Let $H$ be the height of facial component $c$. The evaluations for height difference are defined on both formulas as follows:

$$\text{height}_\text{IOD}(G^c_i, H^c_i) = \frac{\text{height}_\text{diff}(G^c_i, H^c_i)}{\text{IOD}_i}$$  \hspace{1cm} (5)$$

$$\text{Height Diff.}^c = \frac{\sum_{i=1}^{N} \text{height}_\text{IOD}(G^c_i, H^c_i)}{N}$$  \hspace{1cm} (6)$$

In summary, the evaluations we conducted are as follows:

- Average of relative error normalized with Inter-Ocular Distance (IOD).
- Detection rate on 3 different thresholds.
- Measure the difference in some basic geometric information: width and height on eyes and mouth.

4.1.4. Comparisons with Independent-1050. We summarize the results into two tables. The first table contains the relative error and detection rate, while the second one contains the width and height difference. As shown in Table 2, Independent-1050 produces almost as twice relative error as our proposed model. Furthermore, our detection rate is significantly higher. With these metrics, we can claim that we have successfully improved the accuracy of detecting the primary landmarks on frontal faces. Similar improvement can also be seen in Table 3. We have achieved smaller difference in comparison with ground truth especially on the eyes. Some examples can be seen in figure 12.

**Table 2.** 15 Fiducial Points Error and Detection Rate from the Independent-1050 model and Our Proposed Model.

| Model            | Rel. Error | 5% IOD | 10% IOD | 20% IOD |
|------------------|------------|--------|---------|---------|
| Independent-1050 | 0.0726     | 41.29% | 77.26%  | 96.73%  |
| Proposed model   | **0.0365** | **78.48%** | **96.35%** | **99.75%** |

**Table 3.** Width and Height Error Rate from the Independent-1050 model and Our Proposed Model

| Component   | Model            | Width Error | Height Error |
|-------------|------------------|-------------|--------------|
| Right Eye   | Independent-1050 | 0.0956      | 0.0561       |
|             | Proposed model   | **0.0267**  | **0.0233**   |
| Left Eye    | Independent-1050 | 0.0941      | 0.0515       |
|             | Proposed model   | **0.0254**  | **0.0214**   |
| Mouth       | Independent-1050 | 0.0438      | 0.0522       |
|             | Proposed model   | **0.0361**  | **0.0403**   |
4.1.5. Comparisons with CompASM. The result is shown in the same manner as section 4.1.4 with additional 2 primary points (2 sides of the nose). From table 4 and 5, our proposed model outperforms CompASM in all of the evaluation metrics. In fact, we were surprised by the large improvement on the mouth area (Table 5). We investigated the results thoroughly in order to find the reason behind this gap and we discovered that CompASM is not suitable for images with screaming facial expressions (figure 13). Such cases increase their error rate significantly. We therefore re-evaluate their approach with only 336 face images excluding screaming expressions. The results are summarized in table 6 for each expression. One can find that our proposed model still achieves much better performances on most evaluations except the height differences on eyes where CompASM can compete with our approach closely and it even has smaller differences on neutral expression.

| Model          | Mean | 5% IOD  | 10% IOD | 20% IOD |
|----------------|------|---------|---------|---------|
| CompASM        | 0.0769 | 46.33% | 77.52% | 94.49% |
| Proposed model | 0.0353 | 79.93% | 96.85% | 99.80% |

4.2. Multi-Resolution (MR) Models. Although we have improved the facial landmarks detection accuracy by retraining the model in section 4.1, such model is still limited to face images with large resolutions (approximately 240x240 or larger). This poses a significant limitation since the method can only apply to some applications where the face images are taken with high quality camera on controlled environment. Unfortunately, such requirement is often not satisfied in general especially in uncontrolled environment (e.g outdoor, crowd, party).
Table 5. Width and Height Error Rate from CompASM and Our Proposed model.

| Component   | Model       | Width Error | Height Error |
|-------------|-------------|-------------|--------------|
| Right Eye   | CompASM     | 0.0256      | 0.0233       |
|             | Proposed model | 0.0265      | 0.0233       |
| Left Eye    | CompASM     | 0.0515      | 0.0300       |
|             | Proposed model | 0.0252      | 0.0216       |
| Mouth       | CompASM     | 0.0979      | 0.1516       |
|             | Proposed model | 0.0359      | 0.0399       |

Figure 13. CompASM is not suitable for scream expression.

Table 6. Error Rate and Detection Rate from CompASM and Our Proposed model on Each Expression.

| Evaluation Metric | Neutral | Smile | Angry |
|-------------------|---------|-------|-------|
| 17 Fiducial Points | CompASM | 0.0541 | 0.0342 | 0.0340 |
|                   | Proposed | 0.0342 | 0.0231 | 0.0232 |
| Right Eye Width   | CompASM | 0.0401 | 0.0218 | 0.0405 |
|                   | Proposed | 0.0318 | 0.0459 | 0.0321 |
| Left Eye Width    | CompASM | 0.0435 | 0.0248 | 0.0253 |
|                   | Proposed | 0.0218 | 0.0244 | 0.0226 |
| Mouth Width       | CompASM | 0.0414 | 0.0221 | 0.0242 |
|                   | Proposed | 0.0429 | 0.0210 | 0.0234 |
| Right Eye Height  | CompASM | 0.0202 | 0.0244 | 0.0236 |
|                   | Proposed | 0.0218 | 0.0244 | 0.0226 |
| Left Eye Height   | CompASM | 0.0184 | 0.0219 | 0.0242 |
|                   | Proposed | 0.0429 | 0.0210 | 0.0234 |
| Mouth Height      | CompASM | 0.0444 | 0.0332 | 0.0359 |
|                   | Proposed | 0.0429 | 0.0332 | 0.0386 |
| 5% IOD            |          | 60.50%  | 80.83%  | 66.36% |
| 10% IOD           |          | 87.55%  | 96.69%  | 87.31% |
| 20% IOD           |          | 98.74%  | 99.84%  | 98.90% |

Motivated by this limitation, we plan to develop some face models based on the previous proposed face model (from section 4.1) to handle images with smaller
face resolutions. Initially, we attempted to train the model with the same number of landmarks (130 points) and tree structure on various level of smaller images scale. We still conduct the model training on AR database [21]. Unfortunately, such training succeeds only with slightly smaller scale from original. Based on this experiment, we speculate that the reason behind the failed training is due to the high level of density of the landmarks on small image. For example, 130 landmark points on a face with size 90x90 pixels will be too dense and barely distinguishable between neighbouring landmarks. As a result, most landmarks lose its uniqueness on features and become difficult to distinguish.

4.2.1. Landmarks Reduction. We discovered two important facts from our investigation. First, the model training is still feasible with 80% image size scale. Second, the high density of landmarks causes the training to fail on low-scale images. Based on these observations, we develop a systematic framework to reduce the landmarks accordingly.

There are three important aspects for the framework. 1) We have to preserve the important landmarks such as eyes corners, mouth corners and nose tip. 2) The reduction has to be done in a uniform-distributed manner for the whole face to conserve the symmetrical property of the tree structure. 3) When a landmark is removed, other landmarks which are closely connected to it have to be adjusted to avoid having a big “gap” on the line of landmarks (Figure 14).

In order to fulfill the first requirement, we choose some particular landmarks as the VIP (Very Important Points). These landmarks will always be in the face models regardless of the resolution of the faces. We develop an effective face structure with the chosen VIP in figure 15. In total, we choose 18 VIP including 4 eyebrow corners, 4 eye corners, 1 nose tip, 3 landmarks on nose edge, 1 landmark above nose tip between both eyes, 2 mouth corners, and 3 landmarks along jawline. All these VIP are used to divide the facial landmarks into 17 parts as seen in figure 16. This division is necessary to fulfill the second requirement. As the face image gets
smaller, the landmarks between 2 VIP on all 17 parts will be reduced accordingly. This is to ensure the reduction occur uniformly on the whole face.

The third requirement is a bit more difficult since we need to develop a schema where the connected landmarks have to be adjusted properly after some landmarks are removed. In order to visualize this, please refer to figure 17. The red landmarks represent the initial landmarks before reduction. After removing one landmark, we distribute the remaining landmarks (green landmarks) along the line of initial landmarks evenly. These steps are repeated until the distance between neighboring landmarks are not too close.

4.2.2. Model Training. By utilising the proposed automatic landmark reduction, we can train the face model on any resolution. The training set used in Section 4.1.2 is also used in this section. The face images will be scaled from the original high resolution down to particular resolution and the corresponding landmarks ground truth will be scaled down and reduced accordingly. In this case, one should remember that the ground truth landmarks for low resolution images are actually obtained automatically from high resolution images and they are not truly ground truth. In this experiment, we train 4 sets of face models on 4 different scale scale levels: 70%, 50%, 30%, and 10%. On each set, there are 4 facial expressions: neutral, smile, angry, and screaming.

For the rest of this paper, we will refer to the proposed face models (including model from section 4.1) as MR (Multi Resolution) model followed by the number of landmarks associated with that resolution. For example, the first proposed face model will be referred as MR-130, which is the original model. The remaining will be referred as MR-103, MR-70, MR-36, and MR-14 for scale level 70%, 50%, 30%,

Figure 15. Location of the Very Important Points (VIP).
Figure 16. Division of the face into 17 parts.

Figure 17. Adjustment of the 'new' landmarks after reduction.

and 10\% respectively (refer to figure 18). Table 7 briefly summarizes the information on MR models.
Algorithm 2 Landmark Reduction on Each Part between 2 Very Important Points (VIP) (Figure 16)

1: **procedure** Reduce_Landmarks(coord, percentage)

2:  
3:  coord is a set of coordinates (x,y) of landmarks.  
4:  the first and last landmarks are the VIPs.  
5:  percentage is the scale of the landmarks based on the corresponding face image.  
6:  the assumption is that the percentage is less than 80% image scale.

7:  final_coord ← []
8:  row ← size(coord, 1)  
9:  landmarks ← row − 1  
10:  total_percentage ← percentage * 100.0 * (landmarks)  
11:  dist_per_pair ← total_percentage/landmarks  
12:  **while** dist_per_pair < 80 and landmarks ≠ 1  
13:  reduce the landmarks one by one
14:    landmarks ← landmarks − 1  
15:    dist_per_pair ← total_percentage/landmarks  
16:  **end while**
17:  if landmarks = 1  
18:    final_coord ← [coord(1)  coord(row)]  
19:    **end if**
20:  else
21:    final_coord ← [coord(1)]
22:    for i = 1 to landmarks do
23:      position ← (i/landmarks) * (row − 1)
24:      int ← floor(position)
25:      frac ← position − int
26:      if int + 1 < row then
27:        new_landmark ← coord(int + 1) + (coord(int + 2) − coord(int + 1)) * frac
28:      else
29:        new_landmark ← coord(int + 1)
30:        **end if**
31:      end if
32:    **end for**
33:  **end if**
34:  return final_coord
35: **end procedure**

Recall that we define 18 Very Important Points (VIPs) in section 4.2.1, i.e., these landmarks will always be preserved on any face models. This applied to all MR models except the MR-14. We make a small exception here by keeping only one landmark (which is nose tip) instead of five in the nose area (figure 19). This decision was made because of our observation on small faces which reveals that the features around human nose are too subtle.

4.2.3. PUT Database. In this Multi Resolution section, we first test the capability of the MR models on various face resolutions in the controlled environment. We also
conduct an experiment on uncontrolled environment in section 5 after we combine MR models with TFM from section 3 and Viola Jones face detector [26] as the complete facial landmarking system.

We use PUT database [14] for this experiment. It contains 9971 face images of 100 people on white and plain background with a large variety of poses. The faces are clearly visible without any extreme brightness since the illumination is partially
A small exception is made on MR-14 for the number of VIP (Very Important Points). We only keep the nose tip instead of five landmarks for the area around the nose.

Table 7. The proposed 5 Multi Resolution models. This lists the resolution range they can handle, the number of landmarks, and the size of the faces on training dataset.

| MR Models | Target Face Sizes | Landmark | Trained Face Sizes (Approximate) |
|-----------|-------------------|----------|---------------------------------|
| MR-130    | Above 255         | 130      | 300x300                         |
| MR-103    | 180 - 255         | 103      | 210x210                         |
| MR-70     | 120 - 180         | 70       | 150x150                         |
| MR-36     | 60 - 120          | 36       | 90x90                           |
| MR-14     | Below 60          | 14       | 30x30                           |

controlled. All the images are available in high resolution and the ground truth are also provided for evaluation purpose.

We choose 196 frontal faces from 98 people as the testing image set. This set contains faces with size approximately 750x750. To evaluate the performance of the face models on various resolutions, we scale the images down to 7 different levels. We choose particular scale level to reduce the face sizes into 600x600, 450x450, 300x300, 210x210, 150x150, 90x90, and 30x30.

4.2.4. Performance Evaluation. We use the same evaluation index as discussed in section 4.1.3 where we calculate the average relative error and detection rate compared to the ground truth on PUT database. The comparisons are between our MR model and the Share-146 model provided by Zhu and Ramanan [30] (As a reminder, this model can achieve the lowest resolution among all provided face models by Zhu and Ramanan). Instead of comparing 15/17 primary landmarks, we only choose 11 landmarks to be compared as we have less common landmarks among all MR models and SHARE-146 (Figure 20). Geometrical information testing is not included since low resolution images do not often provide sufficient details for the shape of facial components.

4.2.5. Comparison with Share-146. The relative error and detection rate can be seen in table 8 and 9. As reported by Zhu and Ramanan, SHARE-146 model is
unable to detect landmarks on faces as small as 30x30. Our proposed MR models are still able to detect facial landmarks with less error on all resolutions compared to SHARE-146 model. Furthermore, we also improved the detection rate significantly with different resolutions.

Table 8. 11 Facial Landmarks Relative Error from the SHARE-146 model and our Multi Resolution Models.

| Face Sizes | SHARE-146 | MR model |
|------------|-----------|----------|
| 750x750    | 0.0911    | 0.0543   |
| 600x600    | 0.0914    | 0.0542   |
| 450x450    | 0.0912    | 0.0537   |
| 300x300    | 0.0922    | 0.0548   |
| 210x210    | 0.0902    | 0.0596   |
| 150x150    | 0.0930    | 0.0669   |
| 90x90      | 0.0920    | 0.0832   |
| 30x30      | not detected | 0.1225 |

Table 9. Detection Rate (%) from the SHARE-146 Model and MR Model.

| Face Sizes | 5% IOD   | 10% IOD  | 20% IOD  |
|------------|----------|----------|----------|
| S-146      | MR       | S-146    | MR       |
| 750x750    | 22.26    | 53.06    | 61.73    |
| 600x600    | 21.10    | 52.32    | 61.60    |
| 450x450    | 22.96    | 54.36    | 61.32    |
| 300x300    | 21.10    | 52.08    | 60.62    |
| 210x210    | 22.87    | 45.55    | 61.87    |
| 150x150    | 20.83    | 38.68    | 59.69    |
| 90x90      | 20.22    | 25.42    | 61.46    |
| 30x30      | -        | 12.76    | -        |

Table 9. Detection Rate (%) from the SHARE-146 Model and MR Model.
5. Facial Landmarking System. After the successful experiments in section 3.4 (TFM), 4.1.4 (MR-130), and 4.2.5 (MR models), we now integrate all these components into one complete facial landmarking system, as shown in figure 1. This proposed system is expected to be able to handle multiple frontal faces in various sizes in uncontrolled condition. We test the proposed system with the AFLW database next.

5.1. AFLW Database (Uncontrolled). In this section, we choose the Annotated Facial Landmarks in the Wild (AFLW) database [16] for experiment. AFLW contains 25,993 faces gathered from 21,997 images without control over the facial expression, illumination, occlusion, pose, face size and amount of the faces. It provides an ideal case for testing the proposed model for application in real-life images. Because our scope is restricted to frontal faces, we manually selected 200 images containing 687 frontal/near-frontal faces. Most of these chosen images contain multiple faces on various resolutions taken on uncontrolled environment in one image. Figure 21 shows some examples.

5.2. Performance Evaluation. We planned to conduct evaluations based on the relative error of the landmark accuracy. Unfortunately, the ground truth provided in AFLW database is not sufficient, since some of the ground truth landmarks are actually inaccurate and some faces do not have ground truth (Figure 22). We choose an alternative metric by counting the number of detected faces (true positive) and non-faces (false positive). For visualization, we also provides some images to compare in figure 23.
As discussed in section 3, the proposed TFM improves time efficiency while reducing false positive. This is why we conduct another simple comparison to measure the speed. We choose two test images and scale them up to 500% with increment of 100%. The purpose is to discover how the image resolution would affect the growth of time consumption when detecting facial landmarks.

5.3. Comparison With Share-146. As AFLW database [16] has various face sizes, we once again compare our proposed system to SHARE-146 [30] face model which achieves the highest true positive among all the models provided by Zhu and Ramanan. Share-146 is a set of 13 face models with different poses ranging from -90° to 90° with 15° increment. As our main concern in this paper is about frontal faces, to make a fair comparison, we execute the Share-146 model with only 1 model with 0° (frontal faces only) and the summary of the results is listed on table 10. For comparison, we also put the experimental results with all 13 models and one can find that our proposed combination of TFM and MR models can achieve much higher detection rate with only 2 false positives.

| True Positive | False Positive |
|---------------|----------------|
| SHARE-146     | 473 (68.85%)   | 139            |
| SHARE-146 (Frontal Only) | 442 (64.34%) | 7              |
| MR Model      | 597 (86.90%)   | 2              |

5.4. Speed Comparison. The face models proposed by Zhu and Ramanan relies on a full-scale image scanning. As a consequence, the size of the image has a significant impact on the computation complexity. On the other hand, our proposed system utilizes the prior information about the location of face regions (From Viola Jones detector + TFM) to avoid redundant computations. Even though the speed of our proposed system is still far from real-time, it still shows a significant improvement.

Before comparing the speed, we need to decide which models to use for comparison. Firstly, we choose the SHARE-146 model and will only test one frontal face model (0° face pose). Since SHARE-146 has only 146 unique landmarks, we need
Figure 23. Some examples from AFLW database detected with Share-146 model (frontal only) (first column) and MR models (second column). On the second row, Share-146 has a false positive on the trees in the background. On the last row, Share-146 is unable to detect the small faces. Furthermore, the right eyebrow is not aligned properly by Share-146.

To choose one of our proposed MR models which has similar number of unique landmarks. The closest one is MR-36 because it contains 4 different facial expressions on frontal faces with the total of 144 unique landmarks.
We compare the speed of the two models on two different scenarios (refer to figure 24). The top picture (original size 579x389) is the first scenario where the face(s) occupies only a small portion of the image. This is a good example where acquiring face location prior to facial landmarking leads to high efficiency. The other image (original size 500x335) represents the second scenario where the face(s) occupies a large portion of the image.

We first observe the speed change in the first scenario. As shown in figure 25, the growth of time consumption of our proposed system is much slower as the image size grows larger. Meanwhile, if we apply SHARE-146 directly, the time cost escalates quickly from 7.7 seconds to 202.5 seconds just by scaling the image by 500% (25x larger size).

The second scenario is where the speed improvement becomes less obvious. After analyzing figure 26, even though our proposed system is still faster, the time consumption still grows significantly along with the increasing image size. This is to be expected since the face regions occupy approximately 40% of the image. However, this problem can be solved by exploiting the prior information on face sizes. Even though MR-36 was trained to handle small resolution (check table 7), it still works on large faces. However, this just leads to a lot of redundant computation. We can
scale the face regions down to a particularly suitable size if they are too large. For this experiment, we choose the size of 150x150. As shown in figure 27, this simple approach significantly reduces the growth of the computational time and make it more stable regardless of the size of the image.

5.5. Result Analysis. Observation of both results reminds us once again how significant prior knowledges are for facial landmarking process. In section 2, we already mention about the importance of knowing location of the faces prior to facial landmarking for speed efficiency. This has been shown in section 5.4 where our proposed approach avoids a large number of redundant computation. However,
there is one more lesson we can learn from the results in section 5.3. Prior information on face poses is also necessary to avoid false positives. Since SHARE-146 contains 13 various face pose models, no prior information on the face poses will force us to test on all 13 face models. As a result, it produces high false positives as many as 139 false positives. After we focus only on a single frontal face model, the false positives are significantly reduced to 7. Even though the accuracy drop by approximately 4.51%, this is still reasonable considering there are some near-frontal faces which might be undetected by only 1 frontal face model on SHARE-146. Such a concern on face poses should be considered as our future research work.

Table 11. Relative Error (PUT database) from the SHARE-146 model, Our Proposed MR Model, STASM, and Intraface.

| Resolution | SHARE-146 | MR model | STASM | Intraface |
|------------|-----------|----------|-------|----------|
| 750x750    | 0.0911    | 0.0543   | 0.4817| 0.0546   |
| 600x600    | 0.0914    | 0.0542   | 0.3450| 0.0547   |
| 450x450    | 0.0912    | 0.0537   | 0.2218| 0.0534   |
| 300x300    | 0.0922    | 0.0548   | 0.0872| 0.0544   |
| 210x210    | 0.0902    | 0.0596   | 0.0436| 0.0535   |
| 150x150    | 0.0930    | 0.0669   | 0.0419| 0.0566   |
| 90x90      | 0.0920    | 0.0832   | 0.0423| 0.0699   |
| 30x30      | not detected | 0.1225 | 0.0727| 0.1079   |

5.6. Comparison with other State-of-the-art Approaches. In the last experiment, we compare our proposed approach with some other state-of-the-art facial landmarking approaches. We choose Intraface [27] and STASM [23]. This experiment will be focused on the accuracy of the landmarks detection. We evaluate the relative error as used in our early experiment on section 4.1.3 on PUT database [14] with the same testing dataset as in section 4.2.3. The summaries are shown in table 11 and 12.
Table 12. Detection Rate (%) from the SHARE-146 Model, MR Model, STASM, and Intraface.

| Res. | SHARE-146 | MR | STASM | Intraface |
|------|-----------|----|-------|-----------|
| 750  | 22.26     | 61.73 | 97.26 | 61.73     |
| 600  | 21.10     | 61.60 | 96.94 | 61.73     |
| 450  | 22.96     | 61.32 | 96.94 | 61.91     |
| 300  | 21.10     | 60.62 | 96.75 | 61.41     |
| 210  | 22.87     | 61.87 | 97.31 | 67.35     |
| 150  | 20.83     | 60.69 | 96.85 | 70.32     |
| 90   | 20.22     | 61.64 | 97.36 | 69.44     |
| 30   | -         | -    | -     | -         |

Table 13. Landmark Points Amount from the SHARE-146 model, Our Proposed MR Model, STASM, and Intraface.

| Resolution | SHARE-146 | MR model | STASM | Intraface |
|------------|-----------|----------|-------|-----------|
| 750x750    | 68        | 130      | 77    | 49        |
| 600x600    | 68        | 130      | 77    | 49        |
| 450x450    | 68        | 130      | 77    | 49        |
| 300x300    | 68        | 130      | 77    | 49        |
| 210x210    | 68        | 103      | 77    | 49        |
| 150x150    | 68        | 70       | 77    | 49        |
| 90x90      | 68        | 36       | 77    | 49        |
| 30x30      | not detected | 14     | 77    | 49        |

We analyze the results in two resolution ranges. Firstly, we observe the high resolution (from 300 x 300 or above). Even though it is not the best facial landmarks detector yet, our proposed approach still achieves a satisfying performance and on a par with Intraface. We believe that the high number of landmark points in MR-130 plays a significant role in its accuracy. Furthermore, since high resolution faces contain more highly-detailed facial information, the extracted features (HOG features) are more accurate in the landmarks matching process. It might be considered peculiar why STASM’s performance decreased significantly in this resolution range. This is not caused by the dropping accuracy of the landmarks detection, but because of its lack of approach to distinguish faces and non-faces region [22] (even in controlled environment) as mentioned in section 2. When the images are in high resolution, it is more prone to false positives even with the plain background. That is why some of the detected faces by STASM are completely missed (See figure 28). Our proposed TFM model helps to avoid these false positives.

For the lower resolution range (below 300x300), our proposed approach achieves slightly lower performance compared with the other two approaches. According to our observation, we discover a possibility which might cause the small drop in landmark accuracy. Since the training images for MR models are scaled down involving cubic interpolation and the "false" ground truth values are obtained with approximation as discussed in section 4.2.1, these might create a small margin of error when training MR models which has a negative impact on detecting landmark accurately.

However, our proposed MR models are more stable and consistent on detecting approximate location of each facial components in low resolution especially when facial hair is present (e.g beard) or the eyebrows are very close to the hair (less distinguishable). Some examples are shown in figure 29 with face size approximately 30x30. As shown in figure 30, MR models are more robust against misalignment on facial components. Other than a landmark on the chin, all facial components are...
sufficiently detected. We believe that our adaptive landmarks scheme (via landmark reduction from section 4.2.1) plays a significant role since we emphasize only on the important/primary landmarks on small faces. On the other hand, Intraface might regard the beard as the mouth itself and produces misaligned landmarks. STASM
might also detect the landmarks inaccurately (third row). Furthermore, since the number of the landmarks from STASM is always fixed at 77 points (table 13), all the landmarks are crumpled on small faces.

6. Conclusion. We proposed a facial landmarking framework for frontal faces for both controlled and uncontrolled environment. The approaches are extended from the concept of tree-structured face models proposed recently by Zhu and Ramanan [30]. The framework consists of two fundamental components which are face detection and face landmarks detection. For the first component, we proposed a Tree-structured Filter Model (TFM) to assist Viola and Jones face detector [26]. Experiments show that TFM is able to significantly reduce false positives while maintaining high rate of true positives efficiently. In the second component, we rearrange the tree structure of the face model which leads to improvement of the accuracy and increase the amount of landmarks which provides better geometrical information of the facial components. Furthermore, in order to improve the capability of the face models to detect landmarks on small faces, we trained the MR models with adaptive landmark scheme via landmark reduction. Our experiments show that the proposed MR models are able to detect facial landmarks on faces as small as 30 x 30 pixels. Lastly, we compare the MR models with other two state-of-the-art facial landmarking approaches: STASM [23] and Intraface [27] in various
resolutions. Our approach is on par with Intraface and better than STASM on large faces. Furthermore, MR models provide more landmarks on facial components. On the other hand, even though MR models are slightly less accurate on small faces, it is more stable to localize facial components on small faces due to the adaptive landmark scheme.

Even with all these improvements, there are still many aspects to be improved. For the future research, we intended to improve the facial landmarking system to handle more variations of facial expressions (including irregular expressions), non-frontal poses, and improve the computation speed.

Acknowledgments. We appreciate the constructive comments from both reviewers which help improving the quality of this paper significantly. This work is also supported by the National Science Fund of China with grant number 61363066.

REFERENCES

[1] T. L. Berg, A. C. Berg, J. Edwards and D. A. Forsyth, Who’s in the picture? in NIPS, 2004.
[2] O. Çeliktutan, S. Ulukaya and B. Sankur, A comparative study of face landmarking techniques, EURASIP Journal on Image and Video Processing, 2013 (2013), 13.
[3] P. Conillione and D. Wang, Fuzzy approach for semantic face image retrieval, The Computer Journal, 55 (2012), 1130–1145.
[4] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, Active shape models-their training and application, Computer Vision and Image Understanding, 61 (1995), 38–59.
[5] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, IEEE, (2005), 886–893.
[6] L. Ding and A. M. Martinez, Features versus context: An approach for precise and detailed detection and delineation of faces and facial features, IEEE Trans. Pattern Anal. Mach. Intell., 32 (2010), 2022–2038.
[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, Object detection with discriminatively trained part-based models, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32 (2010), 1627–1645.
[8] P. F. Felzenszwalb and D. P. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision, 61 (2005), 55–79.
[9] G. D. Forney Jr, The viterbi algorithm, Proceedings of the IEEE, 61 (1973), 268–278.
[10] Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in Computational Learning Theory, Springer, (1995), 23–37.
[11] J. H. Friedman, Multivariate adaptive regression splines, The Annals of Statistics, 19 (1991), 1–67.
[12] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker, Multi-pie, Image Vision Computing, 28 (2010), 807–813.
[13] V. Jain and E. Learned-Miller, FDDB: A Benchmark for Face Detection in Unconstrained Settings, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.
[14] A. Kasinski, A. Florek and A. Schmidt, The PUT face database, Image Processing and Communications, 13 (2008), 59–64.
[15] J. C. Klontz and A. K. Jain, A case study of automated face recognition: The boston marathon bombings suspects, IEEE Computer, 46 (2013), 91–94.
[16] M. Koestinger, P. Wohlhart, P. M. Roth and H. Bischof, Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.
[17] V. Le, J. Brandt, Z. Lin, L. Bourdev and T. S. Huang, Interactive facial feature localization, in Proceedings of the 12th European conference on Computer Vision - Volume Part III, ECCV'12, Springer-Verlag, Berlin, Heidelberg, (2012), 679–692.
[18] A. Liang, W. Liu, L. Li, M. R. Farid and V. Le, Accurate facial landmarks detection for frontal faces with extended tree-structured models, in Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, (2014), 538–543.
[19] A. Liang, C. Wang, W. Liu and L. Li, A novel landmark detector system for multi resolution frontal faces, in Digital Image Computing: Techniques and Applications (DICTA), 2014 International Conference on, IEEE, (2014), 1–8.
[20] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60 (2004), 91–110.
[21] A. M. Martínez and R. Benavente, The AR Face Database, Technical Report 24, Computer Vision Center, Bellatera, 1998, http://www.cat.uab.cat/Public/Publications/1998/MaB1998/CVCReport24.pdf.
[22] S. Milborrow, Stasm User Manual, http://www.milbo.org/stasm-files/stasm4.pdf, 2013.
[23] S. Milborrow and F. Nicolls, Active Shape Models with SIFT Descriptors and MARS, VISAPP, http://www.milbo.users.sonic.net/stasm.
[24] S. Milborrow and F. Nicolls, Locating facial features with an extended active shape model, in Computer Vision--ECCV 2008, Springer, (2008), 504–513.
[25] A. Torralba, K. P. Murphy and W. T. Freeman, Sharing visual features for multiclass and multiview object detection, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20 (2007), 854–869.
[26] P. Viola and M. J. Jones, Robust real-time face detection, International journal of computer vision, 57 (2004), 137–154.
[27] X. Xiong and F. De la Torre, Supervised descent method and its applications to face alignment, in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, (2013), 532–539.
[28] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys (CSUR), 35 (2003), 399–458.
[29] X. Zhu and D. Ramanan, Face detection, pose estimation and landmark localization in the wild, http://www.ics.uci.edu/~xzhu/face/.
[30] X. Zhu and D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR ’12, IEEE Computer Society, Washington, DC, USA, (2012), 2879–2886.

Received June 2015; 1st revision March 2016; final revision August 2016.

E-mail address: antoni.liang@postgrad.curtin.edu.au
E-mail address: chenyu.wang@postgrad.curtin.edu.au
E-mail address: w.liu@curtin.edu.au
E-mail address: l.li@curtin.edu.au