Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 694321, 10 pages
http://dx.doi.org/10.1155/2013/694321

Research Article

Fast Facial Detection by Depth Map Analysis

Ming-Yuan Shieh and Tsung-Min Hsieh

Department of Electrical Engineering, Southern Taiwan University of Science and Technology, Tainan 710, Taiwan

Correspondence should be addressed to Ming-Yuan Shieh; myshieh@mail.stust.edu.tw

Received 15 September 2013; Accepted 17 October 2013

Academic Editor: Teen-Hang Meen

In order to obtain correct facial recognition results, one needs to adopt appropriate facial detection techniques. Moreover, the effects of facial detection are usually affected by the environmental conditions such as background, illumination, and complexity of objectives. In this paper, the proposed facial detection scheme, which is based on depth map analysis, aims to improve the effectiveness of facial detection and recognition under different environmental illumination conditions. The proposed procedures consist of scene depth determination, outline analysis, Haar-like classification, and related image processing operations. Since infrared light sources can be used to increase dark visibility, the active infrared visual images captured by a structured light sensory device such as Kinect will be less influenced by environmental lights. It benefits the accuracy of the facial detection. Therefore, the proposed system will detect the objective human and face firstly and obtain the relative position by structured light analysis. Next, the face can be determined by image processing operations. From the experimental results, it demonstrates that the proposed scheme not only improves facial detection under varying light conditions but also benefits facial recognition.

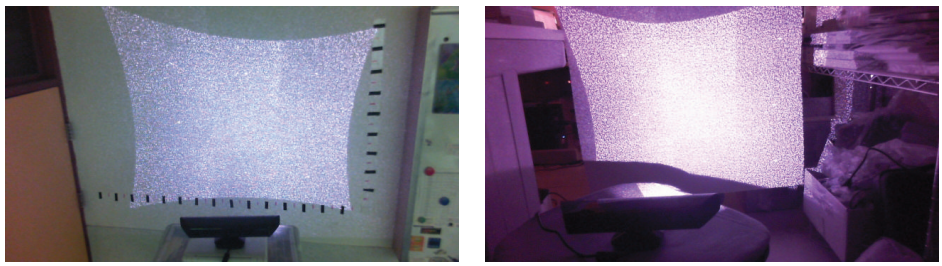
1. Introduction

The processes of digital image processing such as detection and recognition are similar to those of human vision. To enhance the effectiveness of digital image processing, numerous approaches focus on 3D image processing methodology especially depth map scan and related topics. To make closer interaction between human and device, the game console Wii released by Nintendo in 2006 had raised the studies on detections of pose, gesture, action and motion, and related topics. Further on 3D detection, the Kinect released by Microsoft is a motion sensing device as a game console for Xbox 360 and Windows PCs. By the Kinect, users just need to swing their hands, legs, or body and then can interactively control game role players. The new idea inspires numerous players and researchers to invest in 3D scanning, motion detection and interaction and related approaches.

The first success of Kinect is its depth map scan which let users easily determine the depth of every object from a screen. From the technical documents provided from the PrimeSense Ltd. [1], in the light coding solutions, the Kinect generates near-IR light to code the scene and then uses a standard off-the-shelf CMOS image sensor to read the coded light back from the scene. In which, the near-IR emitter diverges an infrared beam through a diverging lens and then the beam is projected on the surfaces in the form of uniform squares scattered as formed structured light planes. Then, the monochrome CMOS image sensor detects and recognizes the structured light map and then results in the depth map. Since near-infrared light is invisible and unaffected by ambient light, to diverge near-infrared light to detect distance is very suitable.

Besides, many similar studies [2–7] on structured light coding are proposed. In [2], Albitar et al. proposed a monochromatic pattern for a robust structured light coding which allows a high error rate characterized by an average Hamming distance higher than 6. Tong et al. present an up-to-date review and a new classification of the existing structured light techniques in [4].

Generally, to integrate two or more cameras/image sensors as a stereo vision device is a common technology for 3D capture. Another aspect, during these years numerous researchers focus on 3D scanning. Unlike 3D camera that collects color information about surfaces, 3D scanner collects depth/distance information about surfaces. These two aspects



(a) The speckle pattern projected on the surface from distance of 1.6 m (in a common room)

(b) The speckle pattern projected on the surface from distance of 0.5 m (in a dark room)

Figure 1: Test of the speckle patterns projected from different distances.

both provide useful information for stereo vision but also lack others [5]. That is why Kinect [6, 7] integrates an infrared projector and a monochrome CMOS camera as the depth sensor to collect distance information and adopts a RGB camera as the image sensor to collect color information for full 3D motion capture and facial recognition.

For advanced security, to detect and recognize biological characteristics such as fingerprint, face, voice, and iris, has become a commonly used technology. Among these biometric identification technologies, face identification is the most widely used. Since good recognition must follow good detection in face identification processes, how to detect the objective faces became a major topic, in which the depth map of image objects will be an important factor because if the object is far away from the camera then its image in size will be smaller than original without zooming. It means that if the depth map and the 2D image are considered simultaneously, then the facial detection and recognition will become easy.

Nowadays, there are numerous approaches on facial detection and recognition. The common technologies of facial recognition consist of Eigenface, Fisherface, waveletface, EGM (Elastic Graph Matching), PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), Haar wavelet transform, and so on. It is worthy noted that most approaches develop the theory and algorithms on 2D image processing. There are many approaches [6–11] that focus on 3D integrated face reconstruction and recognition, in which the depth map becomes as an important factor. For instance, in the approach [8], Burgin et al. extend the classic Viola-Jones face detection algorithm [9] which considers depth and color information simultaneously while detecting faces in an image. The studies proposed by Rodrigues et al. in [10] discuss an efficient 2D to 3D facial reconstruction and recognition scheme.

In this paper, the human facial features of configuration and movement are estimated by using Haar wavelet transform. The features will be considered as patterns for facial detection and recognition. To determine skin color range, geometric relationships of features, and eigenfaces as patterns, the system will conclude the facial appearance and features to the most similar pattern. Then, the interface will show the meaning of the facial expression.

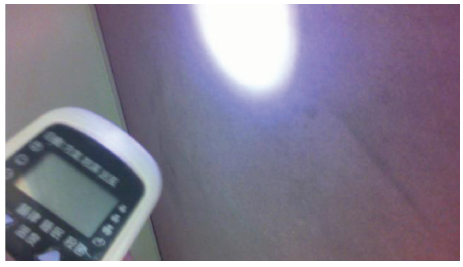
2. Structured Light Based Depth Map Analysis

In Kinect system, there are two speckle patterns that could appear on the camera: one is the primary speckle coming from the diffuser (projected on the object) and the other is the secondary speckle formed during the imaging due to the lens aperture and object material roughness. It only concentrates in primary speckle. The primary speckle pattern, produced by the diffuser and diffractive optical element (DOE) and then projected on the surface, varies with z-axis, in which the PrimeSense Ltd. calls the speckle pattern structured light. Based on the extended depth of field (EDof), DOE is the embodiment of astigmatic optical element, which has different focus for different angle direction. Besides, DOE is designed to reduce the divergence angle so that light intensity would vary slower with distance.

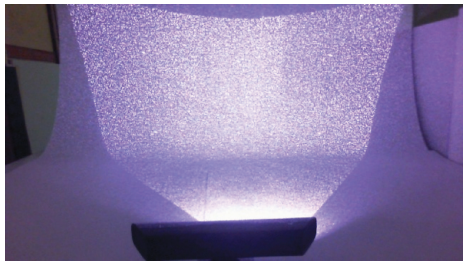
Figure 1(a) shows the speckle pattern projected when the distance between the surface and the device is 1.60 m; Figure 1(b) displays the speckle pattern projected in a dark room from the distance of 0.5 m. Moreover, in order to test how the speckle pattern is generated by an infrared light source, one can modify the webcam: LifeCam Cinema (Microsoft, as shown in Figure 2(a)) by removing the filter of infrared light. As shown in Figures 2(b) and 2(c), there indicate the speckle patterns of the projection surfaces captured by the modified webcam where the infrared light is emitted from a remote control and the Kinect respectively. Notice that the infrared light source of the Kinect, emits a pyramidal speckle pattern.

In this paper, the detectable range of objects is the distances from 60 cm to 10 m and in front of the Kinect. By integrating a Kinect and its depth map analysis, the proposed system aims to improve the effectiveness of facial detection. The diagram of proposed scheme is as shown in Figure 3. At first, the Kinect emits a near-IR light beam which is then projected on surfaces and forms a speckle pattern. The speckle pattern is captured as a grayscale image by a monochrome CMOS camera of the Kinect. The grayscale image is then processed by histogram thresholding and transferred to a binarized image containing the contours of objectives. Next, by using median filter operation, the minor image blocks will be eliminated. Finally, after the following steps of edge detection and ellipse detection, the objective facial blocks will be determined.

(a) The webcam of infrared filter removed



(b) The speckle pattern from a remote control



(c) The speckle pattern from a Kinect

Figure 2: The test of speckle patterns emitting from different infrared light sources.

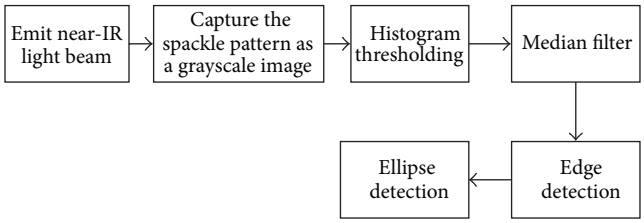


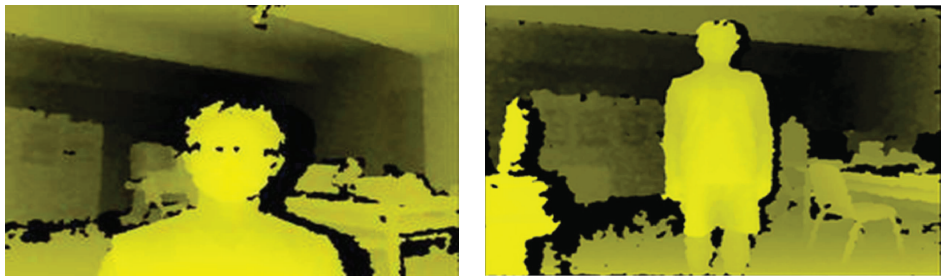
Figure 3: The diagram of facial detection by the depth map analysis of Kinect.

Facial image preprocessing is an important aspect of the facial detection and facial recognition. The images are sensitive to ambient conditions such as the brightness of ambient light, resolution and characteristics of the image device, and signal noises. There will be noise, distortion, low contrast, and other defects occurring during facial detection processes. In addition, the captured distance and direction of objectives, focal size, and so forth might make face blocks be with different sizes and locations in the image.

To ensure that all the objective faces in the image can be detected with consistent features such as size, aspect, contour, and position of facial blocks, it needs to do suitable image preprocessing. The common image preprocessing methods include facial position correction (rotation, cropping, and scaling), facial image enhancements, geometric normalization, grayscale normalization, and so forth, in which in order to get good facial recognition must follow to get upright positions of facial images; the facial image enhancement is to improve the facial images and then results in clearer images and the images in uniform size and conditions are more conducive to the image processing for facial detection and recognition. The necessary image preprocessing of facial detection will be discussed below.

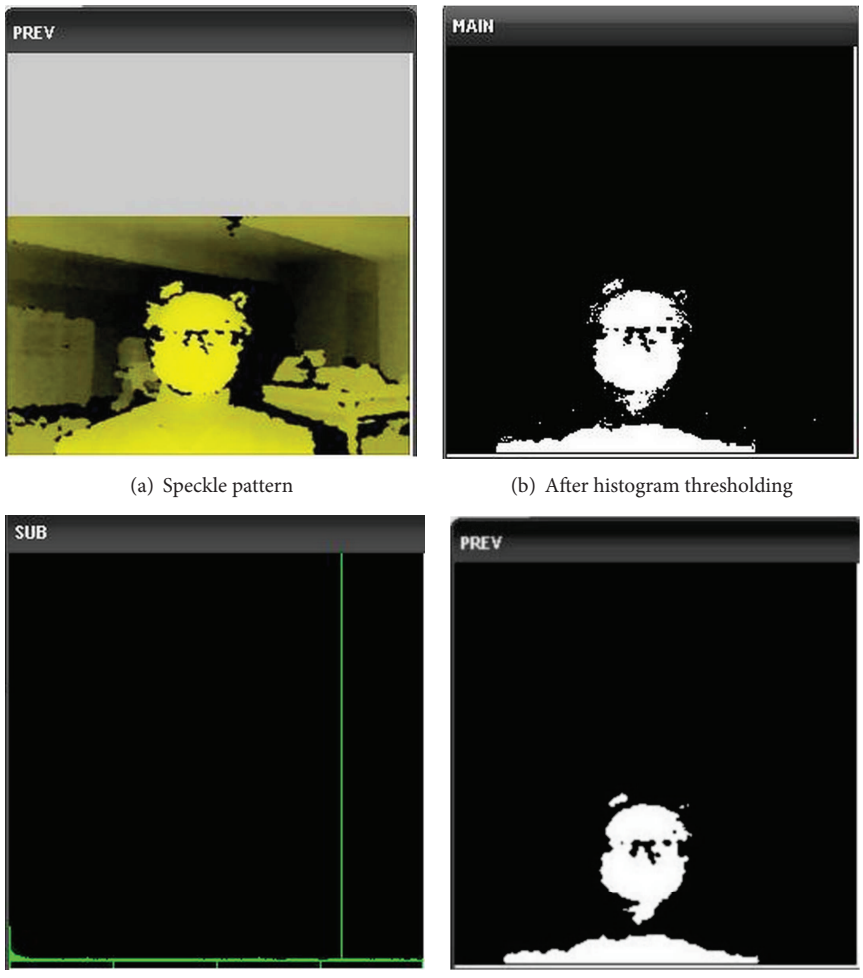
2.1. Light Coding. A typical structured light measurement method is to project a known light pattern into the 3D scene viewed by camera(s) and/or by means of the triangle measurement and geometry relations computation and then can determine the contours of objective surfaces. Similarly, the PrimeSense Ltd. calls the above technology in Kinect "light coding," which means the speckle spots on the projection are able to be coding to represent the depths of surfaces. That is, the objects will be marked in the same light code because of their similar depth through structured light measurement and determination. Such processing results in a depth map as shown in Figure 4. It is noted that the original grayscale images have been transferred into yellow grayscale ones in this paper in order to display more significantly.

2.2. Histogram Thresholding and Median Filter. There are many noises or unexpected blocks in the grayscale image and



(a) The depth image at 0.8 m (b) The depth image at 2.0 m

Figure 4: The depth maps determined after light coding.



(a) Speckle pattern (b) After histogram thresholding

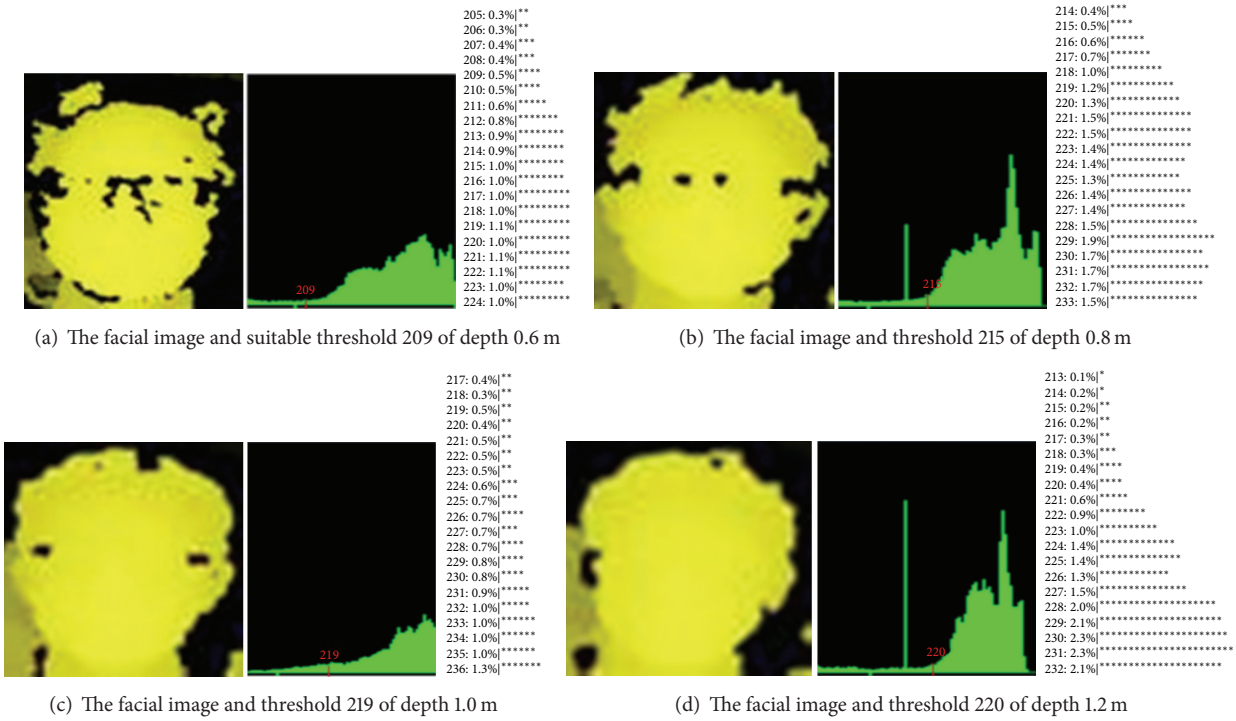
(c) Histogram (d) After median filter

Figure 5: The image preprocessing by filters.

need to be removed. Firstly, the image can be in binarization by histogram thresholding operation. Then, the noises in the binary image can be filtered by median filter operation. After these two steps of image preprocessing as shown in Figure 5, the objective facial blocks will be split from the original image successfully.

However, the threshold of binarization is difficultly decided by a constant. From Figures 6(a) to 6(d), it is seen that there need to be different thresholds in different depths. One can observe these figures and if to look at the point of 0.5% height in histogram, the possible thresholds 209, 215, 219, and 220 in the depths 0.6 m, 0.8 m, 1.0 m, and 1.2 m can be approximated by the formula of their respective depth as the follows:

$$\text{Threshold} = 219 + (\text{depth} - 1) * 2.5. \quad (1)$$



(a) The facial image and suitable threshold 209 of depth 0.6 m

(b) The facial image and threshold 215 of depth 0.8 m

(c) The facial image and threshold 219 of depth 1.0 m

(d) The facial image and threshold 220 of depth 1.2 m

Figure 6: Different thresholds need to be given for different depth maps.

The estimated thresholds by (1) will become 209, 214, 219, and 224.

2.3. Edge Detection and Ellipse Detection. The resultant image after median filter shows the objective block including face and part of body. To execute gradient computation for edge detection (Figure 7(a)) and then to match the block to an ellipse model in axis ratio of 1.2 (a common face) for ellipse detection (Figure 7(b)), the objective facial block is determined as shown in Figure 7(c).

2.4. The Advantages of Adopting Structured Light Analysis. The facial detection based on structured light analysis starts from the grayscale image which is a monochrome image of the speckle pattern. Besides, the foundation processes of the speckle pattern are almost unaffected from ambient light which results in more reliable detection. It benefits the objective detection be superior to being influenced in dusky or bright or inconstant illumination. Moreover, the computations in such monochrome way also cost lower than those while dealing with color image detection. Thus, the proposed scheme is suitably adopted for fast facial detection facing different even bad illumination conditions.

3. Haar-Like Facial Detection

The concept of Haar-like features was firstly proposed by Papageorgiou et al. in 1998 [12] and then widely used in object recognition [12–15]. They intended to adopt Haar wavelet transfer algorithms to deal with the facial detection of upright faces but found there were certain limitations existing in the application. In order to obtain the best spatial resolution, they proposed 3 kinds and 3 types of characteristics. In [13], Viola and Jones have made an expansion based on these foundations, who propose 2 kinds and 4 types of characteristics defined as 3-rectangle features and 4-rectangle features.

The rapid object detection based on Haar-like features [11, 13] is proposed by Viola and Jones in which there are three characteristics as follows:

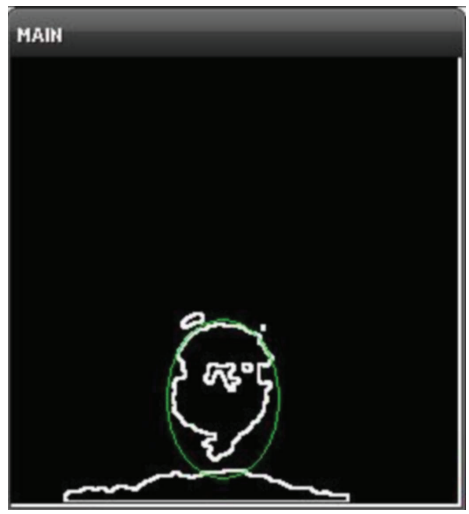
(1) the use of integral images achieves the fast characteristic computation;

(2) constructing a classifier by the method of AdaBoost [14] to collect few important characteristics;

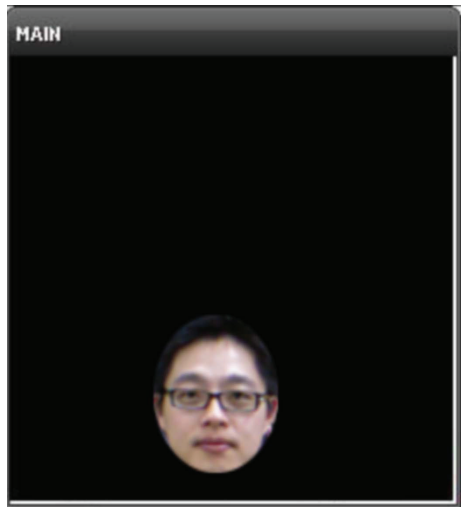
(3) Using a boosted cascade of simple features, it enhances the detection by focusing on useful features.

In the studies in [13], Viola and Jones proposed the concept of integral images and the theory based on the AdaBoost real-time facial detection. They construct an upright facial classifier which is based on 200 characteristics concluded after classifying 4,916 artificial faces in the size of 24×24 and 3,500,000 inhuman faces. From these two examples of rectangular characteristic model, the AdaBoost facial classifier can achieve 95% detection rate; moreover in 14804 inhuman face examinations, the proposed scheme achieved 100% false positive rate. To adopt a boosted cascade of classifiers, it improves the effectiveness of facial detection and reduces the computation time because the inhuman faces will be passed in real-time human facial detection.

(a) After edge detection by gradient computations



(b) After ellipse detection



(c) The resultant facial detection

Figure 7: The image processing for facial detection.

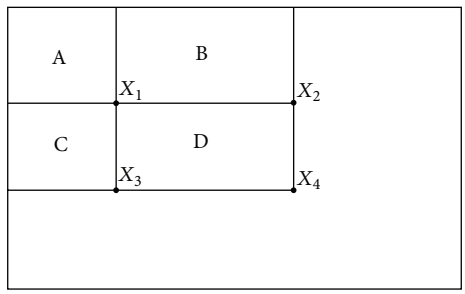
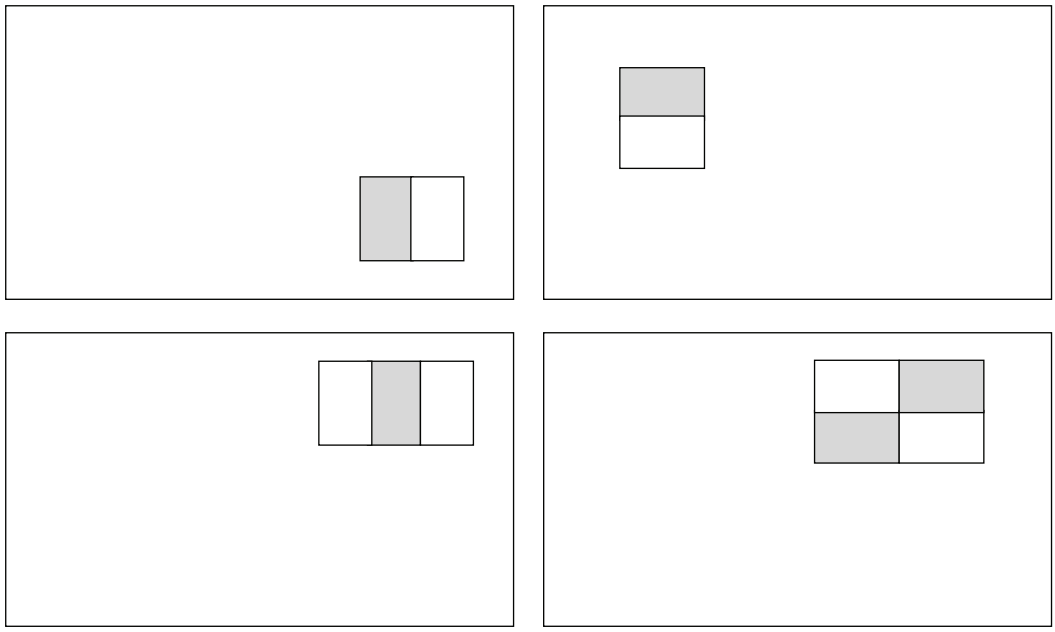


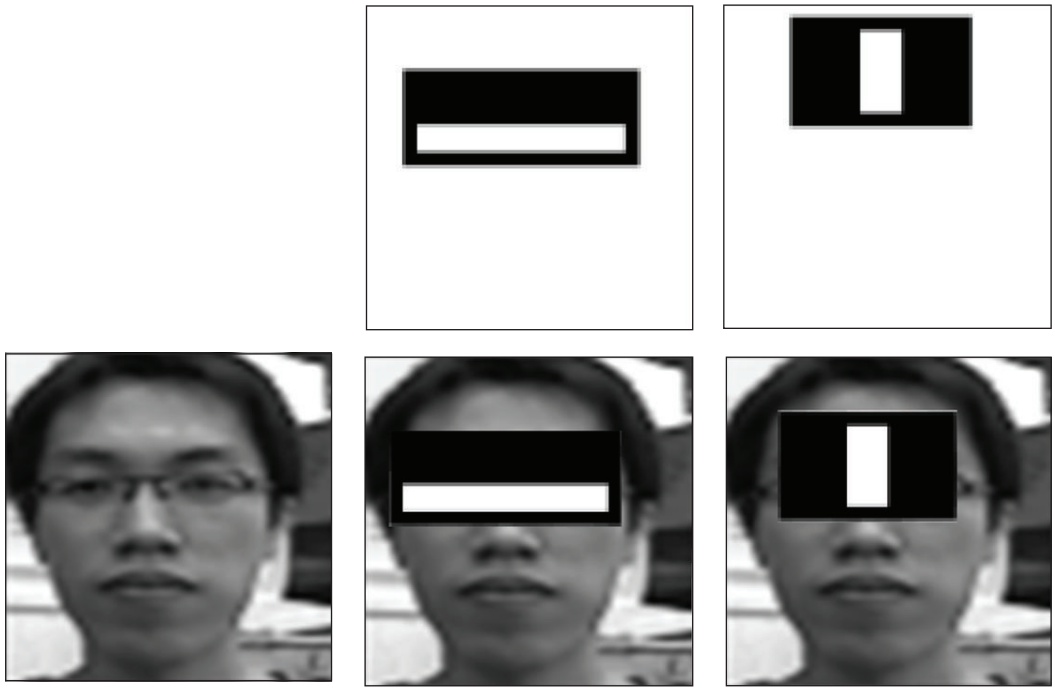
Figure 8: The diagram of determining an integral image.

3.1. Integral Image. Because there are usually more than ten thousand training samples in a rectangular image to represent the features, for instance, if one needs to count the total of pixels in any rectangle, the computation will be huge and time-consuming. The concept of integral image is to count the sum of features in the rectangle, which is then defined as the new image value of respective pixel.

For instance, in Figure 8, the value X_1 represents the total of the pixels in the rectangular A as a feature and the values X_2, X_3, and X_4 indicate the total of the pixels in



(a) Four types of basic rectangle features proposed by Viola and Jones.



(b) To find out a face by Haar-like features

Figure 9: The concepts of the Haar-like rectangle features.

the rectangular A + B, A + C, and A + B + C + D, respectively. Then one can easily get the total of the pixels in rectangular D by $X_4 + X_1 - (X_2 + X_3)$. It indicates that if the integral image could be determined firstly then the computation cost of features will be reduced through integral image than from original image.

3.2. Rectangular Feature, Weak Classifier, and AdaBoost Algorithm. The AdaBoost algorithm is one iteration generation method, which aims to combine those with meaningful classified features among numerous weak classifiers as a new strong classifier, in which, a weak classifier refers to whose performance is better than ones of the stochastic classifiers. In the definitions proposed by Viola and Jones, the Haar-like features are mainly formed by the basic rectangle blocks of 2~3 white-black sections as shown in Figure 9(a). For facial detection, one can adopt the obvious differences on illumination existing in facial features as Haar-like features,

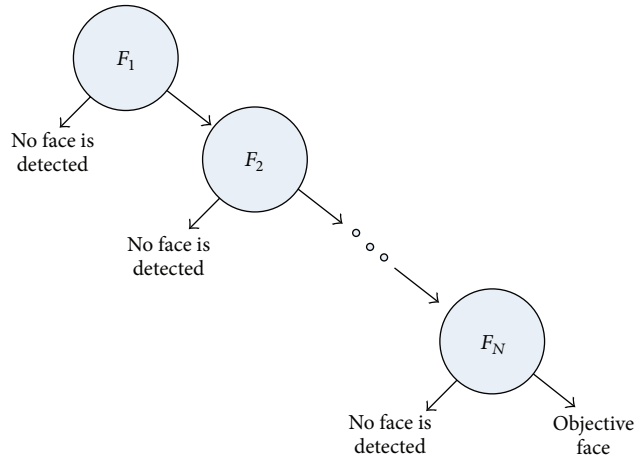


Figure 10: The diagram of a cascade of boosted classifiers.

such as in Figure 9(b), to find out the location of the objective face in a facial image.

3.3. Boosted Classifiers Cascade. The cascade of boosted classifiers, as shown in Figure 10, is working with Haar-like features. It needs to be trained with a few hundred sample views of a particular object such as a face, called positive examples, which are scaled to the same size (maybe 20×20), and negative examples, arbitrary images of the same size.

After a classifier is trained, it can be applied to a region of interest (of the same size as used during the training) in an input image. The classifier outputs a "1" if the region is likely to show the face and "0" otherwise. To search for the object in the whole image one can move the search window across the image and check every location using the classifier. The classifier is designed so that it can be easily "resized" in order to be able to find the objects of interest at different sizes, which is more efficient than resizing the image itself. So, to find an object of an unknown size in the image the scan procedure should be done several times at different scales.

3.4. Haar-Like Feature-Based Cascade Classifier. The cascade in the classifier means that the resultant classifier consists of several simpler classifiers (stages) that are applied subsequently to a region of interest until at some stage the candidate is rejected or all the stages are passed. The word "boosted" means that the classifiers at every stage of the cascade are complex themselves and they are built out of basic classifiers using one of four different boosting techniques (weighed voting). The basic classifiers are decision tree classifiers with at least 2 leaves.

The feature used in a particular classifier is specified by its shape, position within the region of interest and the scale (this scale is not the same as the scale used at the detection stage, though these two scales are multiplied). For example, in the case of the third line feature the response is calculated as the difference between the sum of image pixels under the rectangle covering the whole feature (including the two white stripes and the black stripe in the middle) and the sum of the image pixels under the black stripe multiplied by 3 in order to compensate for the differences in the size of areas. The sums of pixel values over a rectangular region are calculated rapidly using integral images.

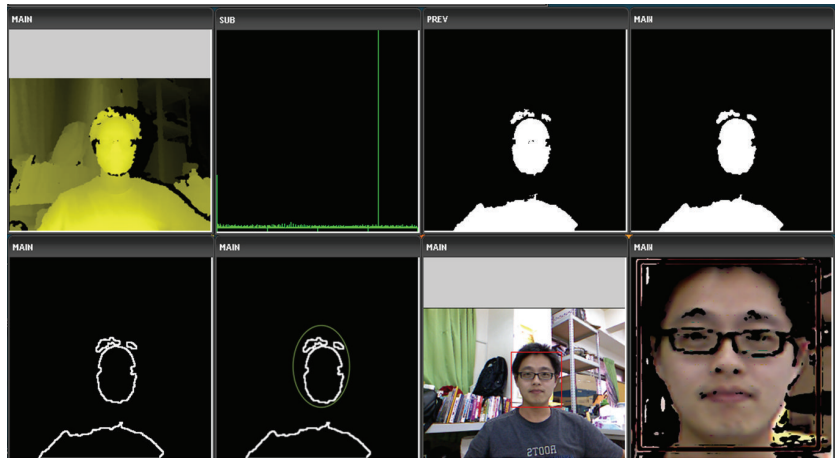
4. Results and Discussion

The detection experiments are executed by using a Kinect, the depth sensing device produced by Microsoft Corp. It emits invisible infrared light beams through a diffuser to be scattered on detected surface. The speckles projected on the surface are detected by the CMOS camera of Kinect as a depth image which could display 3D scene and be used to determine 3D poses and motions.

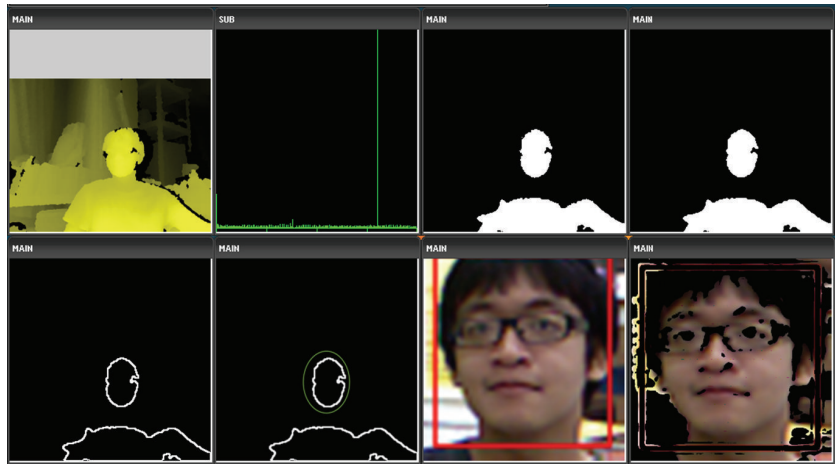
Figure 11 displays three cases of real facial detections. Each subfigure contains 8 experimental results, from top to bottom and left to right, and there are the speckle pattern, the histogram, the binarized image after thresholding, the image after median filter, the image after edge detection, the facial block after ellipse detection, the image after Haar-like facial detection, and the resultant image after skin segmentation. Even in complex background, the experimental results still demonstrate the feasibility of proposed scheme. From the data after over 5,000 pattern (inhuman face included) tests, the average of successful facial detection rate is 95.3%. If only counts the human face tests, the success rate will be reduced into 85.7%. It is necessary to recover in skin segmentation.

5. Conclusion

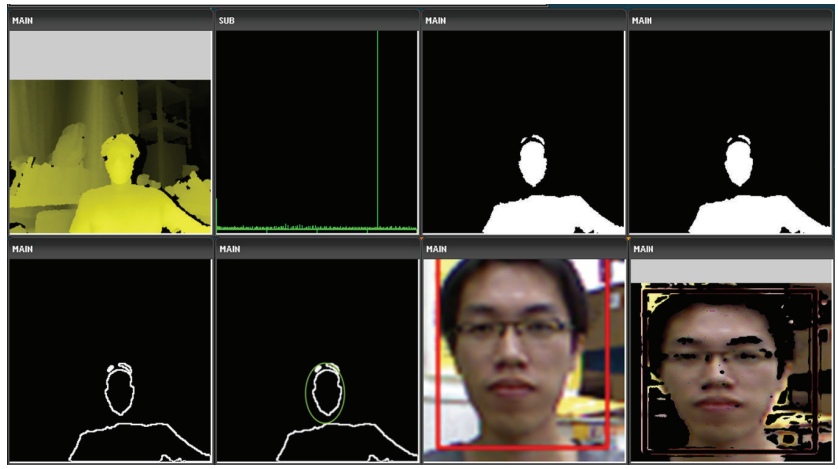
The proposed scheme consists of three subsystems, the first part is the structured light based depth sensing system, the second is the depth map analysis system, and the third is the Haar-like feature based cascade classifier. The structured light device provides the depth maps and helps the system to detect the human face by proposed fast facial detection. The Haar-like feature-based cascade classifier then makes good and fast facial detection. The proposed facial detection scheme based on depth map analysis is proven to obtain



(a) Case 1



(b) Case 2



(c) Case 3

Figure 11: The experimental results of real facial detection based on Haar-like features.

better effectiveness of facial detection and recognition under different environmental illumination conditions. From the experimental results, even in complex background, it still demonstrates the feasibility of proposed scheme.

Acknowledgment

This work is supported by the National Science Council, Taiwan, under Grant nos. NSC101-2221-E-218-029 and NSC100-2632-E-218-001-MY3.

References

[1] PrimeSense/Solution, http://www.primesense.com/solutions/technology/.

[2] C. Albitar, P. Graebling, and C. Doignon, "Robust structured light coding for 3D reconstruction," in *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV '07)*, pp. 1–6, October 2007.

[3] C. Je, K. H. Lee, and S. W. Lee, "Multi-projector color structured-light vision," *Signal Processing*, vol. 28, no. 9, pp. 1046–1058, 2013.

[4] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3D full human bodies using kinects," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 4, pp. 643–650, 2012.

[5] G. Vogiatzis and C. Hernández, "Self-calibrated, multi-spectral photometric stereo for 3d face capture," *International Journal of Computer Vision*, vol. 97, no. 1, pp. 91–103, 2012.

[6] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. V. Gool, "Random forests for real time 3D face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[7] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo, "Kinect identity: technology and experience," *Computer*, vol. 44, no. 4, Article ID 5742015, pp. 94–96, 2011.

[8] W. Burgin, C. Pantofaru, and W. D. Smart, "Using depth information to improve face detection," in *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI '11)*, pp. 119–120, March 2011.

[9] Z. Ren, J. Meng, and J. Yuan, "Depth camera based hand gesture recognition and its applications in Human-Computer-Interaction," in *Proceedings of the 8th International Conference on Information, Communications and Signal Processing (ICICS '11)*, pp. 1–5, December 2011.

[10] J. Rodrigues, R. Lam, and H. du Buf, "Cortical 3D face and object recognition using 2D projections," *International Journal of Creative Interfaces and Computer Graphics*, vol. 3, no. 1, pp. 45–62, 2012.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[12] C. P. Papageorgiou, M. Oren, and T. Poggio, "A General framework for object detection," in *Proceedings of the 6th IEEE International Conference on Computer Vision*, pp. 555–562, January 1998.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I511–I518, December 2001.

[14] T. Susnjak, A. L. C. Barczak, and K. A. Hawick, "Adaptive cascade of boosted ensembles for face detection in concept drift," *Neural Computing and Applications*, vol. 21, no. 4, pp. 671–682, 2012.

[15] A. Ehlers, F. Baumann, and B. Rosenhahn, "Exploiting object characteristics using custom features for boosting-based classification," in *Image Analysis*, vol. 7944 of *Lecture Notes in Computer Science Volume*, pp. 420–431, 2013.