Experiments in analyzing the accuracy of facial feature detection

NICHOLAS ROEDER
roeder@cs.ualberta.ca

XIAOBO LI
li@cs.ualberta.ca

Department of Computing Science,
University of Alberta,
Edmonton, Alberta, Canada, T6G 2H1

Abstract

The accuracy of feature-based human face recognition is inherently dependent upon the correctness of the identification of the facial features. But the extent and details of this dependence are not known. In this paper, we attempt to develop a basic methodology that can be used to discover how sensitive the recognition process is to inaccuracies in facial feature detection from front-view id-type images.

1 Introduction

Automatic face recognition has many potential applications (e.g. authentication for security purposes or automatic teller machines) and has therefore generated much research interest. The two basic recognition approaches are face-based and feature-based. For face-based recognition, the entire image is analyzed as a two dimensional intensity function using some standard statistical techniques or a neural network [16, 1]. Feature-based recognition involves taking measurements of the individual constituents and contours of the face [14, 3, 10, 6, 13, 5, 17, 15]. While the ability to automatically detect facial features is crucial for feature-based recognition, it is also helpful in face-based recognition because the feature locations can be used to normalize size and orientation of the face.

In order for recognition to be successful, accurate and robust facial feature detection algorithms are needed. Although much research has been done in the areas of feature detection and recognition (using mostly manually detected features), the important connection between the two has received less attention. One question left unanswered is: How precisely do we need to locate each constituent and contour in order to correctly recognize an individual?

Our previous work involved a speculation / confirmation system of modules to detect facial constituents and contours [4, 11]. We subjectively judged our results by visually determining the accuracy of each feature detected. Although this provides a good starting point for determining the success of each feature detection module, it would be much more helpful to know how accurate each module needs to be for successful face recognition.

Therefore, the main purpose of the current experiment is to attempt to develop and apply a basic methodology based on our current feature detection modules (eyes, mouth, cheeks, and chin) involving a) extracting measurements from the detected features, b) gauging how tolerant the face recognition process is to inaccuracies in these measurements (and therefore, in facial feature detection), and c) judging the current automatic facial feature detection results based on this tolerance in order to build a basis for future improvements and directions.

2 Analysis of feature detection accuracy requirement

Here we begin the development of our methodology for exploring the accuracy requirements of facial feature detection. This section describes the input to the processes (including images and face measurements), the clustering and recognition techniques, and the observed effects of measurement inaccuracies.

2.1 Data preparation and methods

Image database

We expanded the face image database from our previous experiments [4, 11] (consisting of 84 images scanned from video) to bring the number of available grayscale images to 333 (all of a 256 × 256 resolution). The additional 249 images were scanned from a 1960 high school yearbook and consist entirely of young people without any facial hair or glasses. Different races are represented throughout the database and face orientation does vary, but the image backgrounds are consistently plain (except for a handful of cases) and the size of the faces is relatively similar.

In general, the quality of the yearbook images is lower than that of the original database images. The yearbook images are faded and often have less defined edges. The subjects in the yearbook images also have a wider variety of facial expressions (including many more open-mouthed smiles than our old set of images) and head tilt angles.

In the remainder of this paper, the "Book" image set will refer to the images scanned in from the yearbook. The 84 images used in our previous experiments will be referred to as "Lab" because the images were taken in our lab.

Features and measurements

In pattern recognition and computer vision literature, some terms have different meanings in different contexts. In order to avoid confusion, we define our use of the terms below.

face constituents The eyes and the mouth.

face contours The cheeks and the chin.

facial features Both the constituents and the contours combined.

face measurements Distances (and possibly angles) computed from facial features.

measurement selection Choosing a set of measurements to be used in recognition.

Manual identification of facial features in each image was performed in order to give a consistent measurement database for use with clustering and recognition. Based on the available features, we chose to take the measurements shown in Figure 1 for each face image. The numbers in the figure correspond to the following:

1. width of left eye
2. width of right eye
3. amount left eye is open

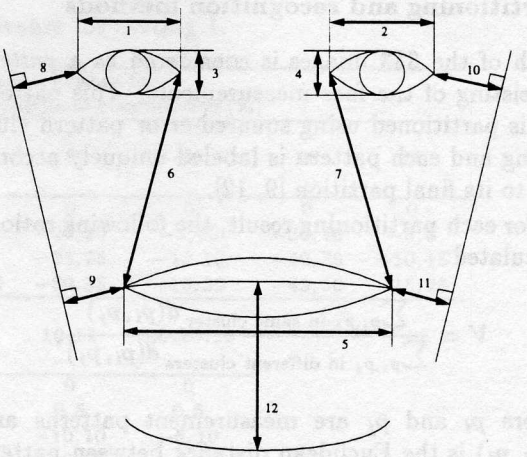


Figure 1: Depiction of face measurements used.

4. amount right eye is open
5. width of mouth
6. right side of left eye to left side of mouth
7. left side of right eye to right side of mouth
8. left side of left eye to left cheek
9. left side of mouth to left cheek
10. right side of right eye to right cheek
11. right side of mouth to right cheek
12. middle of mouth to chin

All measurements are calculated relative to the distance between the two eye iris centers (which is often used in the literature [2]) and are simple Euclidean measures. Notice that the curvature of the chin was not used (although our automatic detection algorithm does calculate it) due to the difficulties involved in manually identifying it.

All measurements could exhibit some variety in a group of pictures taken of the same person. As more measurements are used, small changes in a subset of the measurements should have less of an effect on the recognition process. Nevertheless, it is important to use an image that is as normalized as possible in facial expression and head angle.

Measurements 6 through 12 should give some information about the relative shape of the face in certain areas. As well, measurements 6 and 7 together will indicate the relative angles of the eyes and mouth.

Since many of the images from the yearbook contained open-mouthed smiles, the thickness of the mouth and the lips was not used in our experiments. The current mouth detection routines are designed to process closed mouths only, but normally the endpoints of the mouth could still be found.

Partitioning and recognition methods

Each of the 333 images is considered as a *pattern* consisting of the face measurements. This pattern set is partitioned using squared-error pattern clustering and each pattern is labeled uniquely according to its final partition [9, 12].

For each partitioning result, the following ratio is calculated:

$$V = \frac{\sum_{p_i,p_j \text{ in same cluster}} d(p_i,p_j)}{\sum_{p_i,p_j \text{ in different clusters}} d(p_i,p_j)},$$

where p_i and p_j are measurement patterns and $d(p_i,p_j)$ is the Euclidean distance between pattern i and pattern j. The numerator of this ratio represents the tightness (or compactness) of the clusters and the denominator represents the cluster isolation. The V ratio quantifies the validity of the partitioning.

Each cluster was split evenly and randomly into two groups of patterns for recognition purposes: one for training and one for testing. A simple nearest neighbor rule is used in recognition, that is, each testing pattern takes the label of its nearest training pattern neighbor. At the end of the recognition phase, the *recognition rate* is defined as the percentage of testing patterns that are placed in the correct cluster.

Selection of suitable face measurements

Because of the relatively small size of the database and the limited measurement method, patterns defined by 16 different combinations of measurements, each using a subset of the original 12, were configured to obtain reasonable recognition results. In general, fewer measurements produced better recognition rates, but would also increase the likelihood of multiple images of the same person being placed in different clusters. In order to prepare the patterns for error analysis, we examined the clustering and recognition results for the 16 measurement combinations using from 4 to 8 clusters. In order for us to have considered using the data for subsequent analysis, it must have produced a) a low V ratio from clustering, b) relatively even-sized clusters, c) multiple images of the same person together in one cluster (4 people had multiple images), and d) a high recognition rate.

From these results, the three parameter settings described in Table 1 were chosen to be used in measurement error analysis. All gave even-sized clusters.

Table 1: The three chosen measurement combinations.

Setting	Measurements	V Ratio	Recog. Rate (%)
1	1,2,3,4,8,9,10,11	0.169	85.5
2	1,2,8,9,10,11	0.162	83.6
3	1,2,3,4,8,9,10,11,12	0.136	80.6

Table 2: Sensitivity of recognition to individual measurements.

Setting	Most to least sensitive with decrease in measurement	Most to least sensitive with increase in measurement
1	9,10,1,11,8,2,4,3	9,11,8,10,1,2,3,4
2	9,10,1,8,11,2	9,11,8,10,2,1
3	9,12,2,10,11,8,1,4,3	9,11,12,8,10,2,1,4,3

2.2 Effects of measurement inaccuracies on face recognition

This section describes the development of the methodology for analysing the effects of measurement inaccuracies on recognition success. Note that all 333 images were involved in this analysis.

For each set of measurements chosen in Table 1, we perturbed each individual measurement and certain groups of measurements by adding errors to examine the effects on recognition success. Using the original clusters for each measurement set, inaccuracies ranging from -100% to $+100\%$ (in increments of 5%) of the measurement were added only to the testing patterns before recognition.

Table 2 shows the relative sensitivity of recognition to errors in individual measurements in each of the chosen settings. In general, the perturbed eye measurements (1, 2, 3, and 4) appear to have less of an effect on recognition than do the measurements involving the cheeks (8, 9, 10, and 11).

To better understand the effects of measurement inaccuracies, we record some experimental results in Tables 3 and 4. Because of space limitations, Setting 3 is not detailed. Each table shows the ranges (from negative to positive) in measurement inaccuracies in percent that would be allowable in order to preserve a certain recognition rate. In many cases, the positive range is larger than the negative range, meaning that recognition is impacted to a lesser degree with expanded measurements than with shrunken ones. Note that each column in the tables contains results that are independent of the other columns; that is, only the measurement(s) in

Table 3: Measurement error tolerance for Setting 1.

Recog. Rate Desired (%)	Allowable inaccuracy range for each measurement or measurement combination (%)							
Single measurement	**1**	**2**	**3**	**4**	**8**	**9**	**10**	**11**
85	0,10	0	0,40	0	0	0	0	0
80	−15,35	−10,15	−50,65	−35,100	−20,20	−5,10	−35,10	−5,5
75	−25,55	−40,50	−80,100	−75,100	−25,25	−15,10	−40,35	−10,15
70	−30,65	−50,75	−100,100	−100,100	−50,35	−15,20	−45,50	−15,25
Measurement combination	**1,2**	**3,4**	**1,2,3,4**	**8,9**	**10,11**	**8,9,10,11**		
85	0	0,30	0,5	0	0	0		
80	−5,5	−15,65	−5,20	−5,5	0,5	0,5		
75	−15,25	−50,100	−15,25	−5,10	−10,10	−5,10		
70	−20,35	−75,100	−20,25	−10,15	−10,15	−5,10		

the column headings were perturbed in each test.

These tables also give data which specify the sensitivity of individual measurements as shown in Table 2. In particular, the measurements involving the cheeks (8, 9, 10, and 11) are much more sensitive than those involving the eyes (1, 2, 3, and 4), while measurements 8 and 10 have greater allowable inaccuracy ranges than 9 and 11.

From this information, any efforts in feature detection improvement should be concentrated on the features which involve measurements that have the smallest allowable inaccuracy ranges for successful recognition.

Tables 3 and 4 will be used extensively in the methodology developed for objective judgment of automatic facial feature detection presented in Section 3.3. In that section the allowable inaccuracy ranges for individual *measurements* will be used to determine similar ranges for individual *facial feature* detection modules, which will lead to a classification of the feature detection quality.

3 Feature detection error analysis

This section completes the development of our basic methodology by using the effects of measurement inaccuracies discussed in the previous section to judge the automatic feature detection. In preparation, we first summarize the detection techniques we use [4, 11].

3.1 Summary of feature detection techniques

We use four distinct modules to detect eyes, mouths, cheeks, and chins in front-view id-type face images. A context module is used to find approximate locations of the eyes and mouth before closer examination [4]. The cheek and chin modules employ the locations of the eyes and mouth in determining the relevant subimages for detection [11]. Morphological operations and deformable templates [17, 7] are applied in the eye and mouth modules, while the cheek and chin modules apply a simplified Adaptive Hough Transform (AHT) technique [8].

3.2 Subjective analysis of feature detection results

We present our subjective analysis here mainly to serve as a comparison with the objective results that are developed below. Our subjective analysis decides the detection quality of individual face features based on human judgment of how "close" the detection is to the true feature. Note that only those images in which correct constituent contexts were found could be further processed in this experiment. The combined results are shown along with a breakdown of the results for the two different image sets. The "Lab" images were involved in previous experiments [4, 11]; the "Book" set refers to the images scanned in from the yearbook. The subjective results are shown in Figure 2.

3.3 Objective analysis of feature detection results

In order to objectively judge the results, the *measurements* found from the manually identified features can be compared with those calculated from the automatic detections for the same image, but this will not give a direct link to the accuracy of the individual *feature detection modules* (which is our main objective). Therefore, depending on the facial

Table 4: Measurement error tolerance for Setting 2.

Recog. Rate Desired (%)	Allowable inaccuracy range for each measurement or measurement combination (%)					
Single measurement	**1**	**2**	**8**	**9**	**10**	**11**
85	0, 30	0, 30	−15, 20	−5, 5	−30, 0	−5, 0
80	−10, 55	−25, 40	−25, 20	−5, 5	−35, 10	−5, 10
75	−15, 60	−35, 55	−35, 30	−10, 15	−40, 30	−10, 15
70	−25, 85	−45, 80	−60, 45	−15, 20	−45, 50	−10, 20
Measurement combination	**1,2**	**8,9**	**10,11**	**8,9,10,11**		
85	0, 20	0, 5	0, 5	0		
80	−5, 25	−5, 5	−5, 5	0, 5		
75	−10, 30	−10, 10	−10, 10	−5, 5		
70	−15, 40	−10, 15	−10, 15	−5, 10		

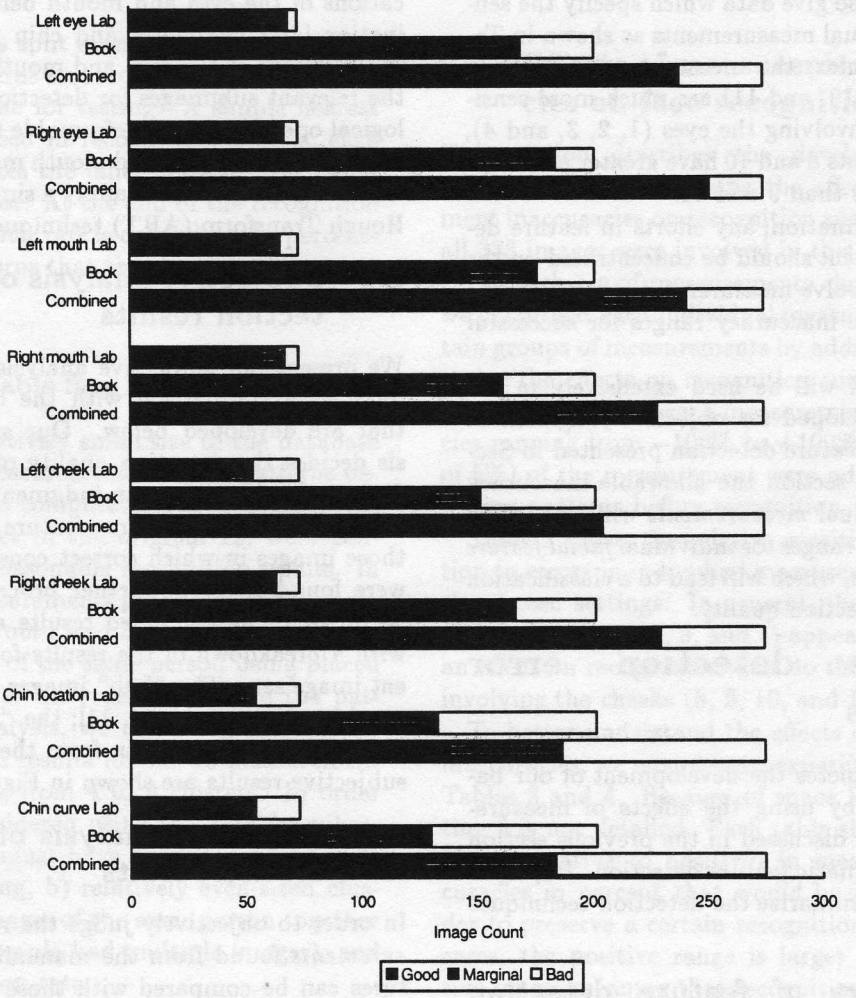


Figure 2: Subjective judgment of facial feature detection.

feature involved, we use certain representative measurements to calculate the accuracy of the feature detection. The chosen measurements are discussed with each individual module below.

In order for the objective judgment to be consistent, we use the allowable inaccuracy ranges for selected measurements found in Tables 3 and 4, to determine whether the feature detection was **good**, **marginal**, or **bad** as follows:

1. Choose the face measurements involved in the current facial feature detection module being judged.
2. Set up ranges of error for the individual measurements involved that will be tolerated for the feature detection to be considered **good**. For the allowable inaccuracy ranges for **good** detection, use the median of all negative and positive extremes of ranges for the individual measurements corresponding to a recognition rate of 80% in Tables 3 and 4. The error of all measurements must be within their individual allowable inaccuracy ranges in order for the feature detection to qualify as **good**.
3. A similar technique is used for finding the ranges for **marginal** detection, although the recognition rate is now 70%.
4. Any feature detection whose corresponding measurements do not fall in either the **good** or **marginal** ranges is considered **bad**.

The allowable inaccuracy ranges for **good** and **marginal** feature detection are shown in Table 5.

Table 5: Allowable inaccuracy ranges for measurements.

Feature	Quality of detection	Allowable inaccuracy range for each measurement (%)	
Eyes		**1 or 2**	**3 or 4**
	Good	$-10, 35$	$-35, 65$
	Marginal	$-35, 75$	$-100, 100$
Mouth		**5**	
	Good	$-15, 20$	
	Marginal	$-25, 30$	
Cheeks		**8 or 10**	**9 or 11**
	Good	$-35, 20$	$-5, 10$
	Marginal	$-45, 45$	$-15, 20$
Chin		**12**	
	Good	$-15, 20$	
	Marginal	$-25, 30$	

Eye module

Measurements 1 and 3 for the left eye and 2 and 4 for the right were used in the objective analysis of the eye detection module. These measurements only involve the size of the eye and not the position and therefore cannot be entirely dependable. However, adding measurements 6, 7, 8, and 10 would reflect the eye position but make the results less dependable because inaccuracies in cheek and mouth detection would be involved.

As can be seen by comparing Figures 2 and 3, using these objective criteria qualifies less **good**, more **marginal**, and less **bad** detections for eyes compared to the subjective results.

Mouth module

For similar reasons as those involved with the eye module, only the *size* of the mouth was used for objectively judging the results. Therefore, only measurement 5 was involved, but this measurement was not in any of the error tolerance analysis because it was removed during the feature selection described in Section 2.1.

In this case the objective results for the mouth width are very similar to the combined subjective results for the left and right mouth points (which can be seen by comparing Figures 2 and 3).

Cheek module

For objective analysis of the cheek module, we used measurements 8 and 9 for the left cheek and 10 and 11 for the right.

Comparing the subjective and objective results (Figures 2 and 3, respectively) shows a huge disparity in cheek detection accuracy. The most probable contributor is how the cheek measurements depend so highly on the position accuracy of both the eyes and the mouth. The cheeks may be in the absolute correct location, but if the eye and/or the mouth is not correct, the cheeks will be judged as being incorrect as well. So the objective results of the cheek detection are not very informative, considering their dependence on the accuracy of the face constituents. The allowable inaccuracy ranges for measurements 9 and 11 are quite restrictive as well, but they are consistent with our discoveries in Section 2.2 about the sensitivity of these measurements.

Chin module

Measurement 12 was involved in only one setting in Section 2, and the inaccuracy ranges were far too restrictive. Using the ranges shown in Table 5, the chin detection resulted in more **good**, less

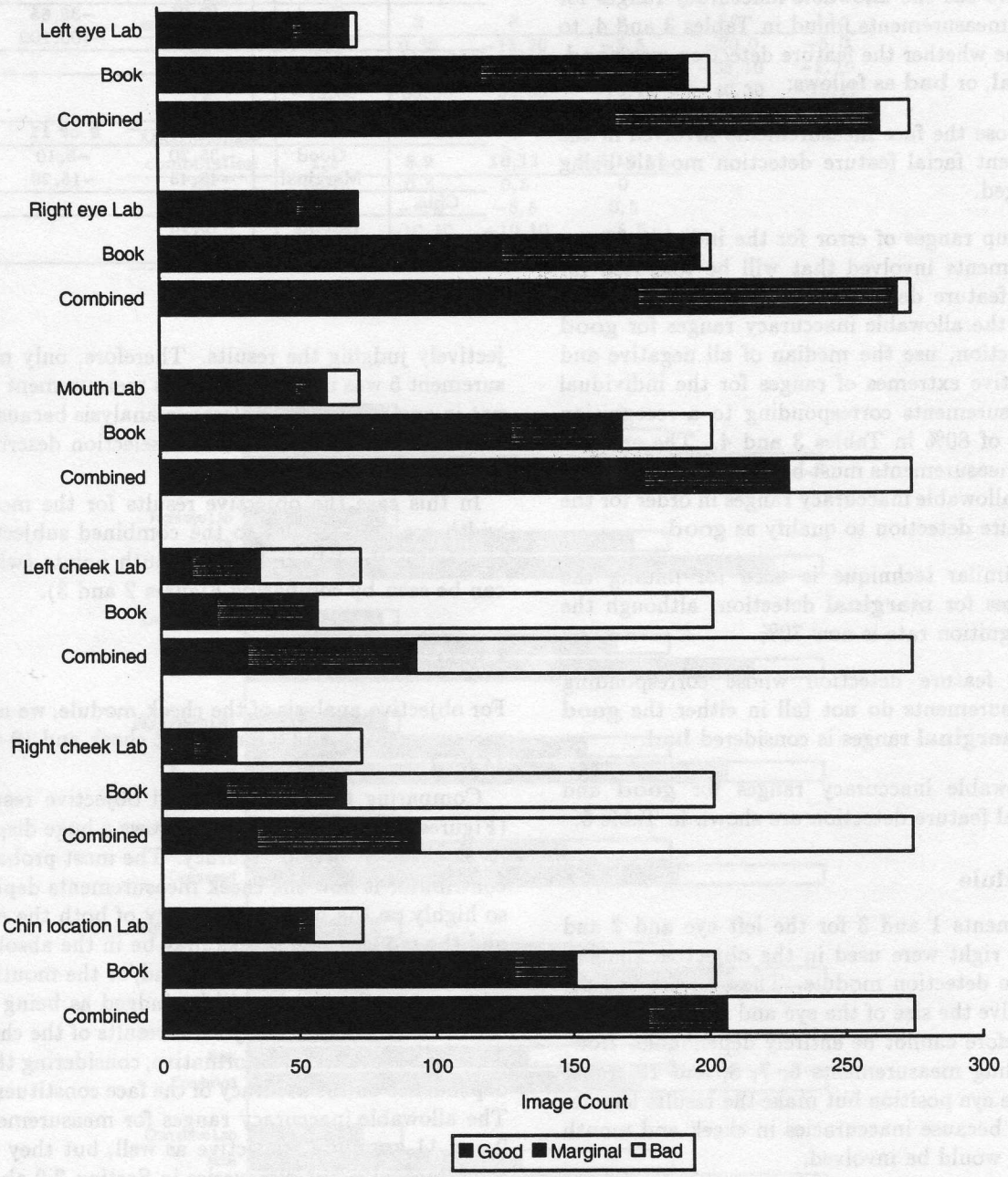
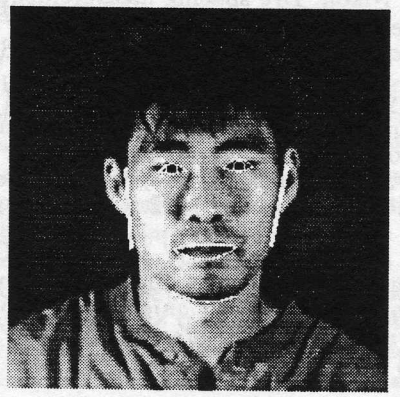


Figure 3: Objective judgment of facial feature detection.

(a)



(b)

Figure 4: Automatic feature detection results for "Book" image (a) and "Lab" (b).

marginal, and less **bad** detections than according to our subjective judgments (as seen in Figure 3).

3.4 Comparison of subjective versus objective feature detection results

Figure 4 contains two examples of facial feature detection. Image (a) is taken from the yearbook, whereas (c) is a "Lab" image. In order to compare the differences between subjective and objective judgments of the results, Table 6 shows both for the images in the figure. (Each x/x combination in the table represents the subjective/objective results.) The column containing the data for mouth detection has two ratings for subjective results corresponding to the left and right mouth corners. In general, the subjective and objective results are fairly close, except for the cheeks, for which the objective rating is typically worse than the subjective (for the probable reasons discussed above).

Notice the differences in image quality between the "Book" and the "Lab" images. In particular, the "Book" images have a thatched or striped effect that the "Lab" ones do not. As well, the "Lab" images are generally clearer and have more contrast.

4 Conclusion

In summary, these experiments revealed the following major points of interest:

1. The sensitivity of individual measurements can be applied in directing improvements in the detection modules involving those measurements.

2. The information gained regarding the accuracy requirements of facial feature detection can be used to criticize current detection techniques with a more objective basis. For example, from this study we found that face contour detection needs to be more precise than face constituent detection (to preserve successful recognition).

3. The connection between face *measurements* and face *features* is important and not trivial. To reasonably judge the accuracy of each detection module, the measurements chosen should reflect the size and position of the facial feature. But involving too many measurements may cause the module to be judged too harshly due to the interdependence of the face measurements and features.

4. Detecting face contours without the need for face constituent information would noticeably improve contour detection results. A simpler improvement that would help in the cases with noticeable head tilt, would be to use the angle of an imaginary line between the eyes in setting the face contour subimages.

The *methodology* developed in this paper should provide a basic framework that could be applied to future projects of larger magnitude, perhaps involving more features, measurements, and images. Knowing the precision requirements of feature detection necessary to maintain successful recognition facilitates an objective judgment of facial feature detection, and provides a solid basis for future developments.

Acknowledgement: This research is supported in part by the Canadian Natural Sciences and Engineering Research Council under Grant OGP9198 and a STEP grant from the provincial government of Alberta.

Table 6: Comparison of subjective/objective results for feature detection for images in Figure 4. "g"=good, "m"=marginal, "b"=bad.

Image	Facial feature					
	Left eye	Right eye	Mouth	Left cheek	Right cheek	Chin location
a (Book)	g/m	g/m	g,m/g	m/b	g/m	g/g
b (Lab)	m/m	g/m	g,g/g	g/b	g/m	g/g

References

[1] Robert J. Baron. Strengths and weaknesses of computer recognition systems. In Andrew W. Young and Hadyn D. Ellis, editors, *Handbook of Research on Face Processing*, chapter 10, pages 507–18. Elsevier Science Publishers B.V., P.O. Box 1991, 1000 BZ Amsterdam, The Netherlands, 1989.

[2] Vicki Bruce and Mike Burton. Computer recognition of faces. In Andrew W. Young and Hadyn D. Ellis, editors, *Handbook of Research on Face Processing*, chapter 10, pages 487–506. Elsevier Science Publishers B.V., P.O. Box 1991, 1000 BZ Amsterdam, The Netherlands, 1989.

[3] S.R. Cannon, G.W. Jones, R. Campell, and N.W. Morgan. A computer vision system for identification of individuals. *IEEE IECON'86 Proceedings*, 1:347–51, 1986.

[4] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[5] I. Craw, D. Tock, and A. Bennett. Finding face features. *ECCV92*, May 1992.

[6] V. Govindaraju, S.N. Srihari, and D.B. Sher. A computational model for face location. *Proceedings of the 3rd Int'l Conference on Computer Vision*, pages 718–721, 1990.

[7] C-L. Huang and C.-W. Chen. Human facial feature extraction for face interpretation and recognition. *Pattern Recognition*, 25(12):1435–1444, 1992.

[8] J. Illingworth and J. Kittler. The adaptive Hough transform. *IEEE Trans. on Patt. Anal. and Machine Intell.*, PAMI-9(5):690–698, 1987.

[9] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, New Jersey, U.S.A., 07632, 1988.

[10] L.C. Lambert. Evaluation and enhancement of the AFIT autonomous face recognition machine. Master's thesis, Air Force Institute of Technology, 1987.

[11] X. Li and N. Roeder. Experiments in detecting face contours. *Proc. Vision Interface '94*, pages 96–103, 1994.

[12] Lionel M. Ni and Anil K. Jain. A VLSI systolic architecture for pattern clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 7:80–89, 1985.

[13] Mark Nixon. Eye spacing measurement for facial recognition. *SPIE Applications for Digital Image Processing VIII*, 575:279–85, 1985.

[14] Toshiyuki Sakai, Makoto Nagao, and Takeo Kanade. Computer analysis and classification of phototgraphs of human faces. *First USA-Japan Computer Conference Proceedings*, pages 55–62, October 1972.

[15] M.A. Shackleton and W.J. Welsh. Classification of facial features for recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVIP-91)*, pages 573–79, 1991.

[16] Matthew Turk and Alex Pentland. Face processing: models for recognition. *SPIE Intelligent Robots and Computer Vision VIII: Algorithms and Techniques*, 1192:22–32, 1989.

[17] Alan Yuille, David Cohen, and Peter Hallinan. Facial feature extraction by deformable templates. Technical Report 88-2, Harvard Robotics Laboratory, 1988.