LIPS DETECTION USING NEURAL NETWORKS

Jamal Ahmad Dargham[1] Ali Chekima[1] Sigeru Omatu[2*]

[1]*School of Engineering and Information Technology,University Malaysia Sabah,*

Locked Bag 2073, Teluk Likas,88999 Kota Kinabalu, Sabah, Malaysia

(jamalad, chekima}@webmail.ums.edu.my)

[2]*Computer and Systems Sciences, Graduate School of Engineering,*

Osaka Prefecture University, Sakai City, 599-8531, Japan

(omatu@cs.osakafu-u.ac.jp)

ABSTRACT

Lips detection is used in many applications such as face detection and lips reading. In this paper a method for lips detection in colour images in the normalised RGB colour scheme is presented. In this method, MLP neural networks are used to perform lips detection on segmented skin regions. Several combinations of chrominance components of the normalized RGB color space were used as the input to the neural networks. Two methods were used for obtaining the normalized RGB components from the RGB colour scheme. These are called maximum and intensity normalization methods respectively. The method was tested on two Asian databases. The number of neurons in the hidden layer was determined by using a modified network growing algorithm. It was found out that the pixel intensity normalisation method gave lower lips detection error than the maximum intensity normalisation method regardless of the database used and for most of the combinations of chrominance components. In addition, the combination of the g and r/g chrominance components gave the lowest lips detection error when pixel intensity normalisation method is used for both databases.

The effect of the scale and facial expression on the lips detection was also studied. It was found out that the lips detection error decreases as the scale factor increases. As for the facial expression, laughing facial expression gave the highest lips detection error followed by smiling and neutral expressions.

1. INTRODUCTION

Lips detection is used in many applications such as face detection and lips reading. Researchers have used either lips colour or lips shape or both for lips detection.

Gomez et al. [1] described an algorithm for lips detection in facial images. The algorithm begins by transforming the image by a linear combination of the red, green, and blue chrominance components of the RGB colour space. Each pixel O(x,y) in the transformed image is obtained as given in Equation 1.

$$O(x,y) = R(x,y) + B(x,y) - 6G(x,y) \quad (1)$$

A high pass filter is then applied to the transformed image to highlight the discriminative details of the lips envelope. Next, the original image and the high pass filtered image are averaged to obtain a new image. The new image is then thresholded to obtain a binary image. The largest object in the binary image is considered to be the lips. The algorithm only searches a fixed region of the image where the lips are assumed to be present.

Chang T. C. et al. [2] extracted facial features such as eyes, nostrils and lips in colour images by first identifying the skin regions in the images by chromaticity thresholding. Then, only the non-skin regions (holes) within the skin regions are searched for the other facial features. In addition, the face orientation, scale, and borders are first established before the search for the other facial features is carried out. The lips are found by thresholding the normalised red + blue – 2 * green chrominance component of a bounding box below the nostrils and above the chin. The thresholded bounding box is then searched line by line for a single region that meets certain criteria.

Sadeghi M. et al. [3] proposed a Gaussian mixture model of the RGB values of the pixels for lips detection. A modified version of the predictive validation technique that allows the use of the full covariance matrices is used to select the model parameters. A subsequent grouping of the mixture components are used as the basis for a Bayesian rule that labels each pixel as lips or non-lips.

Eveno, N. et al. [4] used a 'hybrid edge' for the mouth region localisation. The hybrid edges combine

Neutral at Scale Factor of 1 | Smiling at Scale Factor of 2 | Laughing at Scale Factor of 3

Figure 1. Sample images from the In-house database

pseudo hue and luminance information of the upper, middle and lower section of the lips. The mouth region is then found by projecting the hybrid edges along the X and Y axis.

Liew A. W. et al. [5] used a deformable geometric model to describe the lips shape. The model enables a priori knowledge about the lips expected shape while being flexible to describe different shape variations. They proposed a stochastic cost function that describes the joint probability of the lips and non-lips regions. The conjugate gradient routine was used to obtain the model parameters that minimise the non-linear cost function. It was reported that nearly all the lips contours from around 2000 lips taken from 20 speakers could be extracted.

The paper is organised as follows: The image databases are described next followed by a description of the proposed method. Then the results obtained are discussed before the paper concludes.

Fig. 2. Samples images from the WWW database

2. THE DATABASES

Two databases were used. The first database is called the In-house database while the second database is called the WWW database. The In-house database comprises 15 subjects both males and females from the various races in Malaysia namely: Chinese, Malay, Indian and Indigenous people. Each subject has 12 images showing the frontal facial image of the subject at three distances from the camera. These distances are called scaling factors. Scale factor 1 represents a distance of 36 cm between the camera and the subject while scale factor 2 and 3 represent a distance of 72 cm and 108 cm respectively. For each scale factor, images for three facial expressions namely neutral, smiling, and laughing were taken as well as with glasses with neutral expression only. Thus, this database has a total of 180 images. These images were all taken indoor with a single digital camera under normal lighting conditions and with uniform background. Figure 1 shows a sample images from the database. The images are for neutral expression at scale factor of 1 followed by smiling expression at scale factor of 2 and finally laughing expression at scale factor of 3.

The WWW database has 45 images of Asian subjects collected from the World Wide Web. The subjects represent males and females of different ages. Some of the images were taken indoor while others were taken outdoor with varying backgrounds and lighting conditions. In addition, the pose of the face in the image varied widely from frontal to near portrait. The cameras used for taking these images as well the image processing techniques applied to them are unknown. The only restriction on these images is that they must show a face of an Asian person. Figure 2 shows sample images from the WWW database

3. PROPOSED METHOD

Figure 3 shows a block diagram of the proposed lips detection system. As can be seen in Figure 3, the intensity of the colour image is first normalised using the maximum intensity normalisation method as expressed by Equation 3. Then, the image is segmented into skin and non-skin regions using histogram thresholding on the combination of the r-g and r-b chrominance components. Since both pixel and maximum intensity normalisation methods will be used for lips detection on skin regions, then the intensity of the image is re-normalised using the pixel intensity method when pixel intensity normalisation is used for lips detection.

3.1 Intensity Normalisation

For an image having M by N pixels, the rgb components for pixel (x,y) are given by Equation 2. The set of equations given in Equation 2 perform pixel-by-pixel normalization of the RGB components of the RGB colour scheme. Thus, we call this method pixel intensity normalisation. We propose to normalise the RGB components by the maximum value of (R+G+B) over the entire image. We call this method maximum intensity normalisation.

$$\begin{cases} r(x,y) = \dfrac{R(x,y)}{R(x,y)+G(x,y)+B(x,y)} \\ g(x,y) = \dfrac{G(x,y)}{R(x,y)+G(x,y)+B(x,y)} \\ b(x,y) = \dfrac{B(x,y)}{R(x,y)+G(x,y)+B(x,y)} \end{cases} \quad (2)$$

Thus, the set of equations to perform the maximum

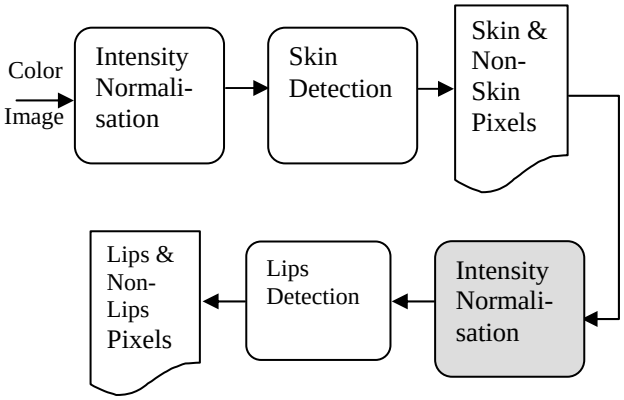

Figure 3. Block Diagram of the Lips Detection System

intensity normalisation are as shown in Equation 3.

$$\begin{cases} r(x,y) = \dfrac{R(x,y)}{Max(R+G+B)} \\ g(x,y) = \dfrac{G(x,y)}{Max(R+G+B)} \\ b(x,y) = \dfrac{B(x,y)}{Max(R+G+B)} \end{cases} \quad (3)$$

3.2 Determining the Network Structure

A Multi-layer Perceptron (MLP) neural network has an input layer, one or more hidden layers and an output layer. The number of neurons in the input layer is determined by the number of inputs (features) to the network while the number of required outputs determines the number of neurons in the output layer. The number of neurons in the hidden layers can be determined by trial and error, network growing or network pruning.

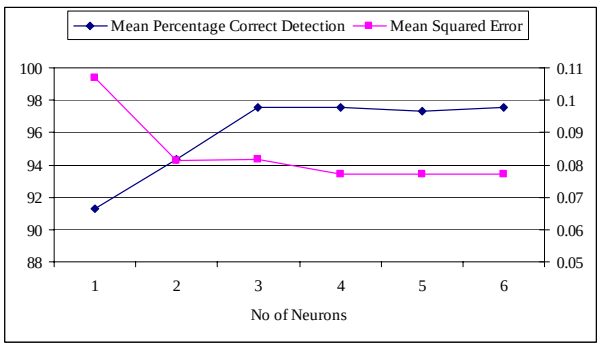

Figure 4. Relationship Between the Number of Neurons in the Hidden Layer and the Mean Percentage Correct Detection

In a previous study [6], it was found out that the histogram thresholding technique gives lower lips detection error when two chrominance components are used compared with a single chrominance. Thus, a the four chrominance components that gave the lowest lips detection error with the histogram thresholding technique will be combined in a group of two and three and will used as the input to the neural networks. Thus, two or three neurons will be used for the input layer. While one neuron will be used for the output layer decoded as 1 for lips and 0 for non-lips. Only one hidden layer will be used and the number of neurons in the hidden layer will be determined by using a modified network-growing technique.

In this method, the number of neuron in the hidden layer starts at one neuron and is incremented by one neuron between consecutive runs. For each run, thirty networks were trained using different weight, training and generalization sets. The average percentage correct lips detection for all thirty networks for all the images in the database is calculated. The method is terminated when there is no improvement in the average percentage correct detection for three consecutive runs. The number of neurons in the hidden layer is taken as the number of neuron in the hidden layer for the network structure that gave the highest percentage correct lips detection. Figure 4 shows an example of the relationship between the number of neuron in the hidden layer and the mean percentage correct lips detection as well as the mean squared error (MSE) of the thirty networks for each run when the combination of the r/g and r+b-2g chrominance components with maximum intensity normalisation is used on the In-house database. Tables 1 and 2 show the chrominance components used and the networks structures for both intensity normalisation methods for the In-house and WWW database respectively.

Table 2. Network Structures Used for Lips Detection on the WWW Database for Several Combinations of Chrominance Components

Maximum Intensity Normalisation		Pixel Intensity Normalisation	
Chrominance Components	Network Structure	Chrominance Components	Network Structure
r/g & r+b-2g	2-4-1	g & r-g	2-3-1
r+b-2g & r-g	2-4-1	g & r/g	2-1-1
r+b-6g & r-g	2-5-1	g & r+b-2g	2-5-1
r/g & r+b-6g	2-4-1	r-g & r/g	2-5-1
r/g & r-g	2-7-1	r-g & r+b-2g	2-1-1
r/g & r+b-6g & r-g	3-4-1	r/g & r+b-2g	2-2-1
		g & r-g & r+b-2g	2-9-1

Table 1. Network Structures for Several Combinations of Chrominance Components for the In-house Database

Chrominance Components	Network Structure	
	Maximum Intensity	Pixel Intensity
r/g and r+b-2g	2-3-1	2-4-1
r/g and r+b-6g	2-5-1	2-3-1
r/g and g	2-3-1	2-2-1
g and r+b-2g	2-5-1	2-2-1
g and r+b-6g	2-4-1	2-2-1
g and r/g and r+b-2g	3-3-1	3-7-1

3.3 Networks Training

For each combination of chrominance components, thirty networks were trained using all the lips pixels and a similar number of skin pixels. While sixty seven percents of lips pixels selected randomly and a similar number of skin pixels were used for validation.

4. LIPS DETECTION

The thirty networks trained on a given combination of chrominance components will be used to segment all images in the database into lips and non-lips regions. For each image, three performance metrics such as the percentage of correct detection, the false acceptance rate and the false rejection rate will be calculated. The performance of a given neural network is the average of the performance metrics for all images in the database. The performance for a given network structure for a given chrominance component is taken as the average of the performance metrics of the thirty networks trained on the given chrominance component. The percentage detection error is the sum of the average false acceptance rate and the average false rejection rate.

4.1 The Effect of Intensity Normalisation

The maximum intensity normalisation method will be used when segmenting all images in both databases into skin and non-skin regions. However, for lips detection on skin regions both maximum and pixel intensity normalisation methods will be used. Figure 5 shows the average percentage lips detection error for the 30 neural networks for a several combinations of chrominance components for the In-house database. As can be seen from Figure 5, the pixel intensity normalisation method gave lower lips detection error than the maximum intensity normalisation for almost all chrominance components used. However, for the two combinations of chrominance components that produced the lowest lips detection error there was no difference between both intensity normalisation methods.

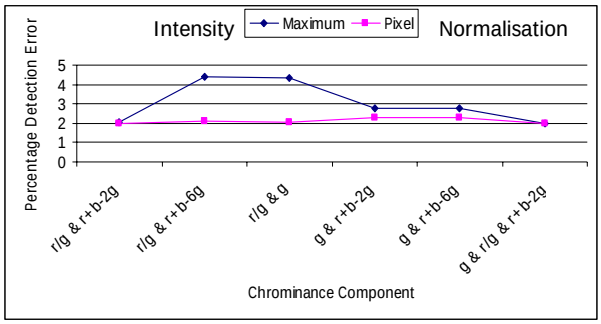

Figure 5. The Effect of Intensity Normalisation on the Percentage Lips Detection Error for the In-house Database

For the WWW database, different combination of chrominance components were used for the maximum and pixel intensity normalisation methods. However, there are 3 combinations which were used for both intensity normalisation methods. Thus, only these 3 combinations will be used for comparison. As can be seen from Figure 6, the pixel intensity normalisation method gives lower lips detection error by almost 2% than the maximum intensity normalisation method.

Table 3. The Combination of Chrominance Components that Gave the Lowest Percentage Lips Detection Error for Each Database and for both Intensity Normalisation Methods.

In-House Database

Intensity Normalisation	Chrominance Combination	Value
Maximum	g & r/g & r+b-2g	1.9271
Pixel	r/g & g	1.9825

WWW Database

Intensity Normalisation	Chrominance Combination	Value
Maximum	r+b-6g & r-g	7.6923
Pixel	g & r/g	5.7248

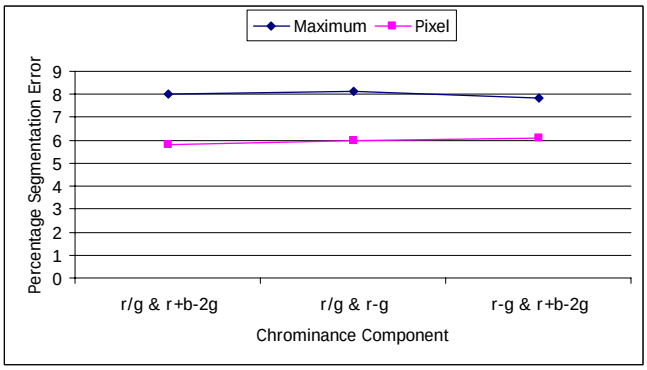

Figure 6 The Effects of Intensity Normalisation on the Mean Percentage lips Detection Error for the WWW Database

4.2 The Effect of the Chrominance Component

Table 3 shows the combination of chrominance component that gave the lowest lips detection error for each database and for each intensity normalisation method. As can be seen from Table 3, the combination of g and r/g gave the lowest lips detection error for the pixel intensity normalisation method regardless of the database. However, for the maximum intensity normalisation method, the lowest lips detection error for each database was achieved using a different combination of chrominance components.

4.3 The Effect of Scaling

As can be seen from Figure 7, the percentage of lips detection error decreases as the scale factor increases regardless of the chrominance components or the intensity normalisation method used. The average percentage decreases of the percentage detection error between the scale factors vary from 34% to 45%. This decrease can be attributed to the decrease in the number of skin and lips pixels in the image as the scale factor increases while the image size remains constant (see Figure 1).

4.4 The Effect of the Facial Expression

Figure 8 shows the relationship between the percentage of lips detection error and the facial expressions for several combinations of chrominance components for both intensity normalisation methods. As can be seen from Figure 8, when the pixel intensity normalisation is used, the percentage of lips detection error for images with laughing facial expression is higher than that given by images with smiling expression which is in turn higher than that given by images with neural expression regardless of the combination of chrominance components used.

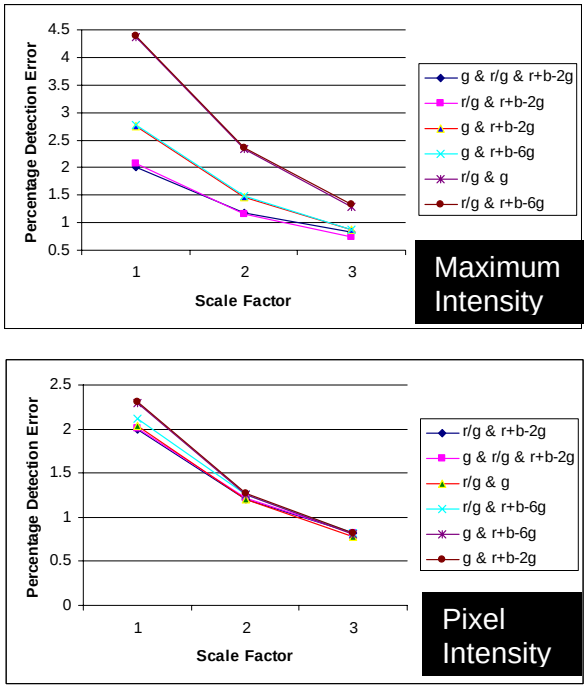

Figure 7. The Effect of Scaling for Maximum Intensity and Pixel Intensity Normalisation

However, when maximum intensity is used there is a difference between the chrominance components used. For those chrominance components that yielded low lips detection error, the effect of facial expression on the percentage of lips detection is similar to effects obtained when pixels intensity normalisation is used. However, for those chrominance components that gave high lips detection error such as r/g & g or r/g & r+b-6g the facial expression has almost no effects on the percentage of lips detection.

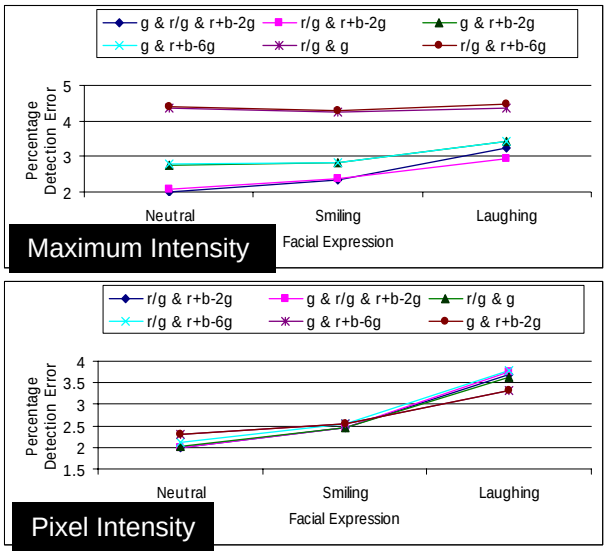

Figure 8. The Effect of Facial Expression on the Percentage of Lips Detection

5. CONCLUSIONS

In this paper a method for lips detection using MLP neural networks was presented. The proposed method first segments the image into skin and non-skin regions using histogram thresholding technique. Then the MLP neural network is used to segment the skin regions into lips and non-lips regions. Several combinations of the four chrominance components that gave the lowest lips detection error using the histogram thresholding technique were used as the input to the MLP. The number of neurons in the hidden layer was determined by using a modified network growing algorithm. It was found out that the pixel intensity normalisation method gave lower lips detection error than the maximum intensity normalisation method regardless of the database used and for most of the combinations of chrominance components. In addition, the combination of the g and r/g chrominance components gave the lowest lips detection error when pixel intensity normalisation method is used for both databases.

The effect of the scale and facial expression on the lips detection was also studied. It was found out that the lips detection error decreases as the scale factor increases. As for the facial expression, laughing facial expression gave the highest lips detection error followed by smiling and neutral expressions.

REFERENCES

[1] Gomez, E., Travieso, C.M., Briceno, J.C., Ferrer, M.A. "Biometric identification system by lip shape". Proceedings of the 36th Annual International Carnahan Conference on Security Technology. 2002. pp. 39 – 42

[2] Chang, T.C., Huang, T.S., Novak, C. "Facial feature extraction from color images". Proceedings of the 12th IAPR International. Conference on Pattern Recognition, Computer Vision & Image Processing. 1994. Vol. 2. pp. 39 - 43

[3] Sadeghi, M. Kittler, J. Messer, K. "Modelling and segmentation of lip area in face images" Proceedings of IEE Conference on Vision, Image and Signal Processing. 2002. Vol. 149, pp. 179 – 184

[4] Eveno, N., Caplier, A., Coulon, P.-Y. "A parametric model for realistic lip segmentation". Proceedings of the 7th International Conference on Control, Automation, Robotics and Vision, 2002. vol.3 pp. 1426 - 1431

[5] Liew, A.W.C., Leung, S.H., Lau, W.H. "Lip contour extraction using a deformable model". Proceedings of the International Conference on Image Processing. 2000. Vol.2. pp. 255 – 258

[6] Dargham J. A., Chekima A., "Lips Detection in the Normalised rgb Colour Space". Proceeding of the 2nd International Conference on Information & Communication Technologies: From Theory to Applications, ICTTA'06, Damascus, Syria.