*Advances in Computer Science Research (ACSR), volume 61*
7th International Conference on Education, Management, Computer and Society (EMCS 2017)

# A Research Method of Identifying a Speaker Based on Human Motion

Jiping Liu[1, a], Hong Gao[1, b] and Chao Chen[2, c]

[1]Center for experiment Management, Bohai University, Jinzhou 121013, China

[2]College of Information Science and Technology, Bohai University, Jinzhou 121013, China

[a]bhdxdljp@126.com, [b]bhdxgh@126.com, [c]chch56@126.com

**Keywords:** Speaker identification; Image processing; Face detection; Hand detection; Movement detection; F1 score

**Abstract.** In this paper, a novel method is purposed to identify a speaker in a video. Instead of using techniques of audio processing and lip movement, head motion and hand waving are adopted as a criterion to identify a speaker as those two kinds of movements are always accompanied when a person is speaking. It is obvious that a speaker is hard to identified accurately when the person is standing too far from the observing point to identify his/her lip movement or the surrounding is too noisy to identify his/her sound. Therefore, this method is purposed to identify a visual speaker based on the high level movement which avoids the two disadvantages mentioned before. Several different image processing algorithms are employed to detect the movement of face and hand. Moreover, the three-frame difference algorithm is modified to improve the accuracy and efficiency when detecting movements. Any other moving objects beyond the regions of face and hand will not be detected. F1 score is used to evaluate this system. After testing on 1,973 frames containing different occasions and characters, the average value of F1 score reaches 91.91% which proves the feasibility of this project. Furthermore, a conclusion can be drawn that the nearer the speaker is, the higher the F1 score is.

## Introduction

In recent years, the technologies of image processing are developing with high speed and are widely used in daily life. For instance, the digital camera, which employs the technique of face detection, is able to identify the human face automatically. A lot of researches and algorithms are proposed in areas related to this project, and most of them have gain great successes.

The aim of this project is to build a system that can identify a speaker in a recorded video based on his or her consecutive body movement which is made up of head motion and hand waving. Instead of using techniques of audio processing and lip movement, head motion and hand waving are adopted as a criterion to identify a speaker as those two kinds of movements are always accompanied when a person is speaking. As all of the technologies adopted in this project are related to image processing, MATLAB is chosen as the programming language which performs much better than C/C++ or Java in terms of efficiency.

In this paper, Chapter 2 illustrates the system design and algorithms utilized for each step when implementing this project. Experimenting, evaluation

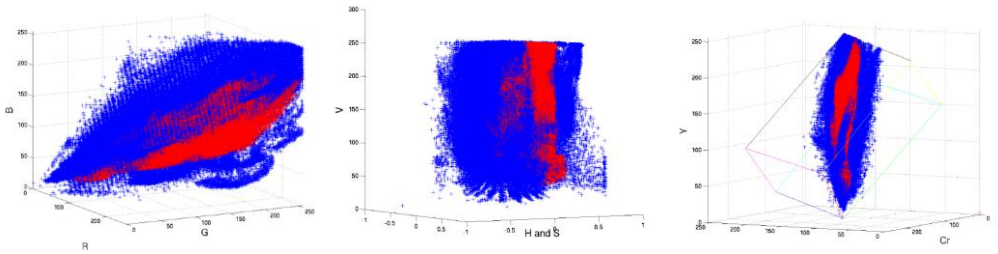



Figure 1. The distributions of skin colours in different colour spaces [1]

And testing for separate stages are given in Chapter 3. Finally, Chapter 4 draws a conclusion on visual speaker identification.

## Algorithms

**Skin Regions Detection.** Detecting face and hand among skin regions is able to promote efficiency and accuracy. A skin colour model should be built to distinguish skin regions from non-skin regions in a digital colour image. YCbCr colour space is chosen as a suitable one to build a skin model. Fig. 1 shows the compactness of distributions of skin colours in RGB colour pace, HSV colour space and YCbCr colour space respectively. A conclusion can be drawn that the distributions of skin colours in YCbCr colour space is similar to Gaussian model which can be adopted to calculate the probability of a pixel belonging to skin colour. Therefore, a digital image can be turned into a skin likelihood image whose skin regions are much brighter than non-skin patches. Skin regions are segmented by using an adaptive threshold Tto turn the image into a binary one. Morphological processingis adopted for removing noises.

**Face Detection.** Appearance-based method, which is based on classification, using ensemble methods, of an over complete set of simple image features, can be viewed as the supreme promising approach for face detection . This algorithm was proposed by Paul Viola and Michael Jones in 2001 and developed by Lienhart. Commonly, Viola and Jones algorithm for face detection consists of three main steps including feature extraction, boosting and multi-scale detection.

In terms of feature extraction, the simple features adopted in Viola and Jones algorithm are reminiscent of Haar basis functions that are employed by Papageorgiouer al. Moreover, every feature is gained by subtracting the addition of pixels contained in white areas from the summation of pixels in grey areas. Integral image, which is viewed as an intermediate representation, is applying in computing rectangle features. The integral image locates at point(x, y) covers the total amount of pixels above and to the left of x, y inclusive (see Eq. 1) .

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x', y') \tag{1}$$

Whereii (x, y) represents the integral image while I (x, y) stands for the original image. A pair of recurrences below is used in Eq. 2:

$$\begin{aligned} s(x,y) &= s(x,y-1) + i(x,y) \\ ii(x,y) &= ii(x-1,y) + s(x,y) \end{aligned} \tag{2}$$

Where$s(x,y)$ is the sum of a cumulative row, *s(x-1) =0* and *ii (-1, y) =0*. Every rectangular sum is able to be calculated in four array references by employing an integral image.

The second step is boosting. Some combine weak classifiers define the boosting, which promotes learning to be a simpler and more efficient process. Negative sub-windows can be rejected mostly when most of the positive examples are able to be detected by boosted classifiers. Simpler classifiers are called to discard most of sub-windows following by further complex classifiers to get low incorrect positive rates. The boosting works as follow details:

1) Learn a simple and singe classifier from a given dataset and then check errors.

2) Provide the error data a higher weight.

3) Learn a second classifier that is simple and easy from the re-weighted dataset.

4) Make a combination of the first and the second classifiers, and then reweight error data in both of them.

5) Repeat above steps until T classifiers are achieved.

6) The last classifier is the combination of all above T classifiers.

Multi-scales detection is employed in Viola and Jones face detection method as the third stage to make sure faces of any size can be detected. As learning and testing are based on the rectangle features, different scales of faces should be calculated. Therefore, it is necessary to set an appropriate scale factor.

A geometric test, which is based on the shape features of a face, shall be adopted to verify the candidates of faces after Viola and Jones face detection since some non-face regions may be included. Three connected component operators are adopted: Compactness, Solidity and Orientation. The criteria of judgement are a combination of the area, A, the perimeter, P, and the size, Dx and Dy of the min-max box of connected components.

## Hand Detection

In recent years, a lot of algorithms have been proposed in detecting hands and are employed into various applications. Concerning colour-based method, it can be viewed as an effective and fast approach as features of hands are hard to be obtained. The flexible variations of both the form and the angle of hands promote the choice of colour-based method. Moreover, computational cost is able to be saved when a same method is adopted for both the detection of face and hands. By establishing a skin colour model, hands and face regions which are obtained in skin patches are able to be distinguished from non-skin areas. The hand detection technology described later is based on colour.

In this stage, some criteria concerning size and position are used to detect hands. For instance, the hand is assumed to be smaller than a face. Moreover, during the processes of identifying a hand, the adjacent frames are observed to make a further verification. Using those assumptions shall make the detection easier than using hand features as they are rather complicated which will increase the computational cost. In order to find hand regions among skin patches, components are sorted in decreasing order in lists:

1) List of components leftmost (short for Ll): helpful to detect the left hand
2) List of components rightmost (short for Lr): helpful to detect the right hand

Criteria for temporal tracking are:

3) List of components closest to the previous left hand position (short for Lcl)
4) List of components closets to the previous right hand position (short for Lcr)

The left hand selection involves (Ll, Lcl) and the right hand selection involves (Lr, Lcr). In each list, the top element is viewed as the likely candidate. The next component in a list will be taken into consideration only when elements at the top of all lists among a selection are not the same.

## Motion Detection

The method employed is based on a widely used motion detection algorithm named three-frame difference which calculates the different pixels among every three adjacent frames to find out the moving objects. In this project, the algorithm is modified to focus on detected object regions instead of a whole frame which makes it more efficiency and accurate. Assuming g1(x, y) represents the motion variation image of a k – 1 frame and a k frame in specific regions, while g2(x, y) stands for that of the k frame and a k + 1 frame. They are able to be obtained by Eq. 3:

$$g_1(x,y) = |f_k(x,y) - f_{k-1}(x,y)| \ (k = 2,3,4,\dots)$$
$$g_2(x,y) = |f_{k+1}(x,y) - f_k(x,y)| \ (k = 2,3,4,\dots) \quad (3)$$

where $f_{k-1}(x,y)$ is the k–1 frame while $f_k(x,y)$ represents the k frame and $f_{k+1}(x,y)$ stands for the k+1 frame.

In order to convert the difference images into binary images, a suitable threshold value T will be selected to operate on these images based on grey feature. Four steps are required in the selection of a suitable threshold.

1) Divide the regions of detected targets in a current frame and in the previous frame into 2*2 blocks respectively. Calculate the number of pixels $a_i$ in every block in the region of a current frame and that of bi in a certain region of the previous frame. Assume that there are kblocks,m and nstand for the length and width of the image respectively (seeEq.4).

$$S = \sqrt{\frac{1}{mn} \times \sum_{i=1}^{k} (a_i - b_i)^2} \quad (4)$$

2) Generally, the threshold is set as T=1.1S roughly. The frame difference image of the current frame and the previous frame is binarized using this T.

3) The mean of pixels that are less than T in frame difference images is computed and marked as M. Assume that q is the amount of pixels that are less than T, and d(i) represents for the pixel (see Eq. 5). Set M as the threshold for the frame difference image and binarize that image.

$$M = \frac{1}{q}\sum_{i=1}^{q} d(i) \tag{5}$$

4) Obtain the frame difference image of the third frame and the current frame. Then binarize it with M as the threshold. Repeat the third step.

The AND operation is performed on difference images and Equation 6 gives the definition:

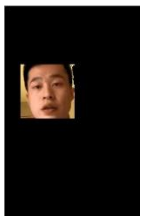
In order to get the coordinates of a moving region in AND operation images by an object clustering approach, two steps are needed

1) Label every object pixel. First of all, the AND operation images should be scanned from left to right and then from top to the bottom. When meeting a pixel that belongs to object region, detect the other eight pixels around it.

2) Label eight pixels with a new number which should begin from 1 and add 1 each time if they are not labelled. Otherwise, the minimum number among eight neighbours of a current pixel is labelled with it.

3) Cluster every object region. The motion object image is scanned line by line from top to the bottom with both directions. When meeting a pixel belongs to an object region, detect eight neighbours around it. If the number of the current pixel a is smaller than the smallest number b among neighbouring pixels, replace a with b. The image is

$$p(x,y) = g_1(x,y) \otimes g_2(x,y) = \begin{cases} 1, & g_1(x,y) \neq 0 g_2(x,y) \neq 0 \\ 0, & otherwise \end{cases} \tag{6}$$



(a)Input image (b) Result

Figure 2. Skin regions segmentation Figure 3. Face detection Figure 4. Hand detection

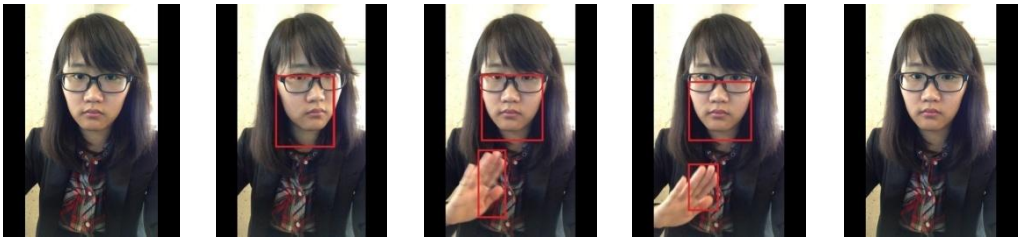


Figure 5. Examples of detecting the face and hand movement by the system

Scanned again from the bottom to top after the whole image has been detected, and caparisoning label numbers is repeated. This process will not stop until no number will be changed.

Some irregular regions may be found after the object clustering technique. These regions should be denoised and considered as motion regions

## Results and Discussion

Fig. 2 below shows the outcomes of segmenting skin regions from two different input images which lay a foundation for face and hand detection. After invoking the Computer Vision System Toolbox, Fig. 3 illustrates the results of face detection where face is detected correctly in the first image while the crossed-

Hands are detected as face candidates in the second one. Therefore, it is necessary to adopt three connected components to verify the result and remove non-face patches (see Fig. 4). The next stage is about hand detection. Fig. 4 demonstrates the result of Fig. 2(b) which is used as the input image. The last step is about movement tacking based on the results of face detection and hand detection. Fig. 5 illustrates the consecutive frames in two test videos where different characters have different head and hand movement. By adding a bounding box around the moving head and hand, a speaker is able to be identified.

After testing on 1,973 frames containing different character wearing different clothes and different background, the average value of F1 score is 91.91%. And the scenario varies from one person to two. Furthermore, a conclusion can be drawn that the nearer the character is from the observing point, the higher the value of F1 score is. More data should be recorded for testing where the character is standing far from the camera and a modified threshold should be adopted in the system to determine whether amovement happens.Furthermore, according to the test results, the variations of background and illumination have little impact

On face and hand detection which means using skin colour segmentation as the basis of face detection and hand detection is reliable.

Besides, assumptions used for hand detection narrows the limitation of clothes characters could wear and lower the accuracy even they save the computational cost. For instance, the exposed small part of upper arm in the scenario will be detected as a hand region incorrectly by the system. Hence, the methods of detecting hands should be improved in the future.

## Conclusions

This project proposed a method of identifying the speaker based on a high level movement detection which consists of head motion and hand motion. First of all, the face and hand are detected respectively on the basis of skin colour. The Viola and Jones algorithm and three connected components are employed to detect face. This combination generates a further accurate result. Hands are identified by some assumptions: the size of the component should be smaller than that of a detected face; the position of this region in the next frame should be closed to it and so on. Therefore, the hand regions are determined. The last step is about detecting movements of detected face and hand. A modified algorithm named three-frame difference, which is widely used in movement detecting, is employed in this stage. The movement is detected within specific areas instead of a whole frame. By doing so, other moving objects will not be detected. A judgment about whether there is a movement will be made by setting an adaptive threshold which will be calculated for each run. A bounding box is added once the movement is detected.

## References

[1] Xing G, Qi WY. (2006), “Skin Color Segmentation Based on Color Space and its Application in Face Detection [J]”, Television Technology, Vol. 30, Iss. 7, 91-93.

[2] Qin LF, He DJ. (2009). “Research on Face Detection Algorithm Based on Skin Color Segmentation [J]”, Computer Engineering and Design, Vol. 30, Iss. 19, pp. 4461-4463.

[3] Nagapriya, K.K., Ashwini, H., and Subramanya, B. (2012), "A Novel Face Detection and Tracking Algorithm in Real-Time Video Sequences [J]", International Journal of Electronics Signals and Systems, Vol. 2, Iss. 1, pp. 25-28.

[4] Ahonen, T., Hadid, A., and Pietikainen, M. (2006), "Face Description with Local Binary Patterns: Application to Face Recognition [C]", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, Iss. 12, pp. 2037-2041.

[5] Zhao LH, Liu FH. and Xu X.(2004), "Survey of face detection methods [J]", Computer Application Research, Vol. 9, Iss. 12, pp. 73-77.

[6] Otsu, N. (1979), "A threshold Selection Method from Gray-Level Histograms [C]", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-9, No. 1, pp. 62-66.

[7] Li LL., Deng XS (2003), "Application of MATLAB in image processing technology [J]", Microcomputer Information, Vol. 19, Iss. 2, pp. 65 - 66.

[8] Masip, D., Bressan, M., and Vitria, J. (2005), "Feature Extraction Methods for Real-Time Face Detection and Classification [J]", EURASIP Journal on Applied Signal Processing, Vol. 13, pp. 2061-2071.

[9] Viola, P. and Jones, M. (2004), "Robust Real-Time Face Detection [J]", International Journal of Computer Vision, Vol. 57, Iss. 2, pp. 137-154.

[10] Lienhart, R., Kuranov, A., and Pisarevsky, V. (2003), "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection [J]", Pattern Recognition, Vol. 2781, pp. 297-304.