ESTIMATION OF SPEAKING SPEED FOR FASTER FACE DETECTION IN VIDEO-FOOTAGE

Osamu Ikeda
Faculty of Engineering, Takushoku University
815 Tate-machi, Hachioji, Tokyo, 193-0985 Japan

ABSTRACT

We previously reported a face detection system based on color segmentation using HSV. It was shown that the color is more effective than other colors not only in accurate segmentation but also in effective extraction of facial features. The first is crucial for detection and the latter for recognition. When it comes to video footages of news program, sound often accompanies the video and persons express themselves by moving facial parts while speaking. In this paper we improve the face detection in speed using both sound and video in a combined way. First, the rate of syllables spoken is estimated from the sound. Next, for a beginning short video clip of each new scene, a differential image is formed with the frame distance corresponding to the rate to find mouth and eyes. This enables us to reduce the number of sampling points for segmentation to a great degree and to enhance the reliability of the detection. Also music is discriminated from speaking by the estimation. These contribute to much faster detection of face.

1. INTRODUCTION

Research on face detection or recognition has extensively been made in recent years in such fields as image processing and computer vision [1], [2]. For example, Rowley, Baluja and Kanade proposed a neural network- based algorithm [3]. Schneiderman and Kanade developed a Naïve Bayes classifier [4]. Osuna, Freund and Girosi presented an algorithm to train Support Vector Machines [5]. And Turk and Pentland proposed to use eigenfaces [6]. Those algorithms, however, are not so fast. Viola and Jones reported a rapid object detection method [7]. Fröba, Ernst, and Küblbeck presented a real-time face detection system [8], using AdaBoost [9]. But the detection rate for the first method is below 95%. In the field of multimedia, on the other hand, sound and texts as well as images have been used to better understand the semantic meanings of multimedia documents [10], [11]. Several improvements have been reported to enhance the accuracy of face segmentation for a short face video. They combine temporal segmentation or tracking with spatial segmentation [12], or they adopt manual segmentation [13] as a last resort. In an emerging field of surveillance, Wang and Kankanhalli detect faces based on the dynamically changing number of attention samples [14], where AdaBoost is used for face detection and cues of movement, hues and speech are used to adaptively correct the samples. As for speech its existence alone is their interest and no analysis is made for it.

We reported a segmentation method based on HSV color [15]. It was shown that HSV is more advantageous than RGB or YCbCr both in segmentation and in feature extraction. A face detection and image retrieval system was constructed using an extended region-growing method based on the color. It was shown that the system could achieve the detection rate of more than 95% but that it took time of the order of 0.1 sec, on average, for a 304 x 232 pixel image using a 1.2 GHz Pentium III processor.

In this paper, we improve the system in speed for the case of video footages of news program without sacrificing the detection rate. We analyze the accompanying sound to estimate the rate of syllables spoken, and using the rate we form a differential image to find where the mouth and eyes are in the image. This enables us to focus the sampling point around those parts, thereby making the face detection faster and reliable.

2. FACE DETECTION SYSTEM

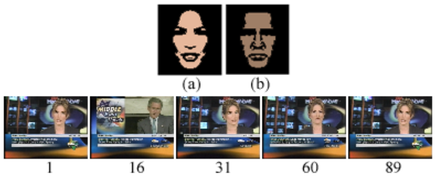

Fig. 1 Face detection and image retrieval [15].

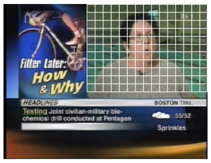
Filter Later: How & Why
HEADLINES
Testing Joint civilian-military bio-chemical drill conducted at Pentagon
BOSTON THU.
33/32
Sprinkles

Fig. 2 Sampling points for segmentation over a pre-determined window.

Let us review the face detection and image retrieval system reported in a previous paper [15], using a video footage consisting of seven scenes and 100 frames, as shown in Fig. 1, where an anchorperson appears four times and three interviewing scenes are sandwiched. If we input her face image in Fig. 1(a), the system outputs the beginning frames of the four scenes where she appears. If we add the second

0-7803-9332-5/05/$20.00 ©2005 IEEE

image in (b), then it adds the beginning frame of the scene, where he appears. The system can also detect faces rather than recognize them if we replace specific face images with a typical face image and use different values of the relevant parameters.

In this system, first, the video stream is analyzed to detect scene change through the differential image operation between neighboring two frames. Next, a point is scanned over the pre-determined window of the beginning frame of a new scene, as shown in Fig. 2. The color at each point is judged whether or not it belongs to the class of face. If that is the case, segmentation is carried out using the region-growing method, which is based on the continuity of eight neighborhoods with the sampled point and is subject to a given error:

$$\frac{w_h|h-h_s|+w_s|s-s_s|+w_v|v-v_s|}{w_h+w_s+w_v}<e \qquad (1)$$

where (h,s,v) are the components of HSV color at a point in the image, (h_s,s_s,v_s) are those at the sampled point, w's are weights, and e is the maximal error. In this case, image regions that satisfy Eq. (1) but are not continuous with the sampled point are included in the segmented image if they are close enough to the original segmented image, to make the resulting segmented image more informative.

Then, the segmented image is checked whether or not its dimensions, X and Y, satisfy the facial restriction in Eq. (2) and whether or not its binary pattern satisfies Eq. (3):

$$Y<2.5X \text{ and } X<1.25Y \qquad (2)$$

$$N > N_e \text{ for eyes,} \quad N > N_m \text{ for mouth} \qquad (3)$$

where N is the number of vacancies in the regions assigned for eyes or mouth as shown in Fig. 3, and N_e and N_m are pre-determined numbers.

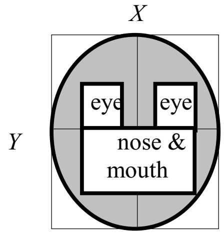

Fig. 3 Eyes and mouth regions for the segmented image.

Further, the binary pattern is correlated with the input binary face pattern, where the former is fitted in size to the latter. In the case of $Y > 2X$, where a neck part may often be included in the segmented image, however, the fitting ratio in y is made the same as in x, and the top part of the segmented face pattern is used for the correlation.

The segmentation is repeated as many times as the one in Eq. (4), and highly correlated single or multiple facial images are output.

$$\begin{gathered}\{\text{number of segmentation errors}\}\\ \text{x } \{\text{number of sets of the weights on } HSV\} \\ \text{x } \{\text{number of sampling points}\}\end{gathered} \qquad (4)$$

In the correlation we use an average face image pattern for face detection while a specific face one for recognition. The color window used to check the sampled color is wide so as to cover the face color range for detection while it is narrow depending on the specific image for face recognition. In the recognition, the degrees of matching in color and in pattern are compared to select the maximally correlated image with the input one.

3. FORMATION OF DIFFERENTIAL IMAGE

One of the methods to make the system fast is to reduce the number of sampling points. In Fig. 2, the number of sampling points on the face is only 14 out of the total number of 192. For video footages of news programs, it is possible to know a rough location of the face by forming differential images, specifically for such facial parts as the mouth, eyes and head.

The differential image with m frame distance is given as

$$i_{dif}(x,y,c;m)=\sum_k \left|i(x,y,c,(k+1)m)-i(x,y,c,km)\right| \qquad (5)$$

where $i(x,y,c,n)$ is the image intensity at the coordinates (x,y) for the frame n and the color component c, and the factor of normalization is omitted. The movements of such facial parts take time so that there may exist an optimal value for the frame distance; when it is too small the resulting image may be noisier and when it is too large other parts such as head may be more emphasized. The formation of the image in Eq. (5) over a relatively short time may be enough for the purpose. The optimal value of m is estimated from the accompanying sound data.

4. SOUND ANALYSIS FOR FRAME DISTANCE

The mouth movements may be closely correlated with the pronunciation of syllables, in view of which we estimate the rate of syllables spoken. First, the envelope of the sound waveform is derived from the sound waveform, which is then made binary with an appropriate value so that the wavelets that roughly may correspond to syllables be countable:

$$w_{let}(t)=\begin{cases}1 \text{ for } e(t)>c_w \mathrm{Max}\{e(t)\}\\ 0 \quad \text{otherwise}\end{cases} \qquad (6)$$

where $w_{let}(t)$ is the on-off pattern of the wavelets, $e(t)$ is the envelope, and c_w is a threshold value. Then, the average rate of the wavelets are given as

$$r_w = N_w / T_w \qquad (7)$$

where N_w is the number of wavelets over the time T_w. And the mean frequency for each wavelet is calculated as

$$f_a=\frac{\int \left|f\mathrm{FT}\{a(t)\}\right| df}{\int \left|\mathrm{FT}\{a(t)\}\right| df} \qquad (8)$$

where $a(t)$ is the waveform. We impose the minimal time duration of wavelet to be 512/48000 sec, where 512 means the minimal number of data for FFT and 48000 is the sampling frequency of the sound data; since, too short a wavelet may not necessarily mean the movement of the mouth. Aiming at discriminating music from speaking, we also use the duration of 256/48000.

Repeating the calculation for r_w and f_a by changing the value of c_w, we obtain the wavelet pattern $\hat{w}_{let}$ that has the largest number of high frequency components of f_a for $\hat{c}_w$. The optimal rate of r_w may include pauses between the sentences, and faster mouth movements than the average rate are desired to be reflected in the resulting differential image. So twice the rate $\hat{r}_w$ is used for the formation of differential image. Then, the corresponding m value in Eq. (5) is given by

$$\hat{m} = 30/(2\hat{r}_w) \qquad (9)$$

The differential image obtained with the optimal frame distance is averaged in each sub-window, which is then made binary using a threshold value to obtain a pattern. The pattern may show mostly the mouth and/or eyes, so that the sampling is made on and around the pattern.

5. EXPERIMENTS

An example showing the process of finding the optimal wavelets is shown in Fig. 4 for the sound data of speaker 4 in Table 1, where the parameter c_w was actually changed with a step of 0.05. It is seen that the mean frequency profile for c_w=0.3 has largest number of high frequency components. In this case the number of wavelets found agrees with the number of syllables spoken as shown in Table 1. The differential image for m=3 for this speaker is shown on the left in Fig. 5 and its averaged image within sub-windows is shown on the right, where each sub-window has 12x12 pixels. In Fig. 6, ten of 121 images of this video clip on the top show a typical case of movements. Averaged differential images with m = 1 to 10 in the middle and their binary patterns obtained for the threshold value 179 of 255 at the bottom show that the binary pattern most focuses on the face with the optimal frame distance 2.6. Similar results are also shown in Fig. 7 for another speaker 5 in Table 1, who gives much larger movements as shown in the figure. Fig. 8 shows that those binary patterns are on the mouth and eyes. We can also observe by using different threshold values that as the value of m is different from the optimal one, the binary pattern tends to have more elements outside of the face part.

In the case of music the rate of wavelets estimated tends to be unusually high or low and the rates obtained for the minimal FFT numbers 256 and 512 tend to be very different from each other, as shown in Table 1. These enable us to easily discriminate a music scene from a speaking one. Table 1 and Fig. 9 show that estimated numbers of the wavelets roughly agree with those of syllables spoken, where the sound data vary from 1.6 sec to 20 sec. As a result those methods combined can improve the speed of face detection by a factor of more than 10, so that we can expect to process more than 20 frames per second with this system.

6. CONCLUSIONS

We improved our previous face detection method in speed and in reliability using sound and video in a combined way for video footages of news program. First, the rate of syllables spoken is estimated from the sound. Next, a differential image is formed with the frame distance corresponding to the rate for a short video clip of each new scene, to find mouth and eyes. This reduces the number of sampling points to a great degree and also makes the detection more reliable. Also music is discriminated from speaking by the estimation, which further contributes to faster detection. An extensive evaluation of the system is now under way.

REFERENCES

[1] R. Chellappa, C.L. Wilson, and S. Sirohey, “Human and Machine Recognition of Faces: A Survey,” Proc. IEEE, vol. 83, pp. 705-740, 1995.

[2] M-H. Yang, D.J. Kriegman, N. Ahuja, “Detecting Faces in Images: A Survey,” IEEE Trans. PAMI, vol. 24, pp. 34-58, 2002.

[3] H.A. Rowley, S. Baluja, T. Kanade, “Neural Network-Based Face Detection,” IEEE Trans. PAMI, vol. 20, pp. 23-38, 1998.

[4] H. Schneiderman and T. Kanade, “Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition,” Proc. CVPR, pp. 45-51, 1998.

[5] E. Osuna, R. Freund, and G. Girosi, “Training Support Vector Machines: An Application to Face Detection,” Proc. CVPR, pp. 130-136, 1997.

[6] M.A. Turk and A.P. Pentland, “Eigenfaces for pattern recognition,” J.Cognitive Neuroscience, vol. 3, pp. 71-96, 1991.

[7] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” Proc. CVPR, pp. 511-518, 2001.

[8] B. Fröba, A. Ernst, and C. Küblbeck, “Real-Time Face Detection,” Proc. 4th IASTED SIP, pp. 497-502, 2002.

[9] Y. Freund and R. E. Schapire, “A Short Introduction to Boosting,” J.Japan.Soc.Artificial Intelligence, no. 14, pp. 771-780, 1999.

[10] Y. Wang, Z. Liu and J. Huang, “Multimedia Content Analysis,” IEEE Signal Processing Magazine, vol. 17, pp. 12-36, 2000.

[11] S. Satoh, Y. Nakamura, and T. Kanade, “Name-It: Naming and Detecting Faces in News Videos,” IEEE Multimedia, vol. 6, pp. 22-35, 1999.

[12] D. Wang, “Unsupervised Video Segmentation Based on Watersheds and Temporal Tracking,” IEEE Trans. Circuits and Systems for Video Technology, vol. 8, pp. 539-545, 1998.

[13] C. Toklu et al., “Simultaneous Alpha Map Generation and 2D Mesh Tracking for Multimedia Applications,” Proc. ICIP, vol. 1, pp. 113-116, 1997.

[14] J.Wang and M.S.Kankanhalli, “Experience based Sampling Technique for Multimedia Analysis,” Proc. ACM Multimedia, pp.319-322, 2003.

[15] O. Ikeda, “Segmentation of Faces in Video Footage Using HSV Color for Face Detection and Image Retrieval,” Proc. ICIP, III, pp. 913-916, 2003.

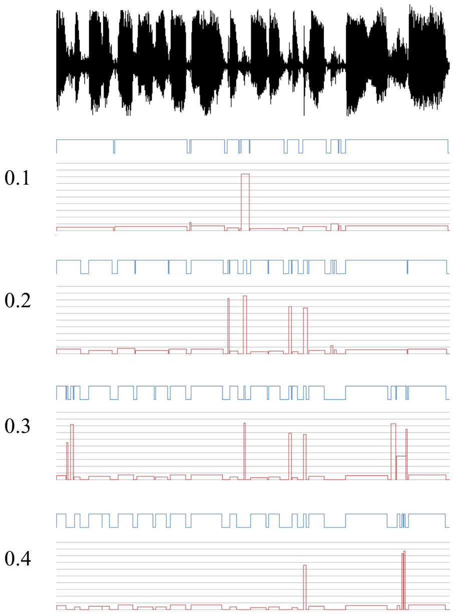

Fig. 4 Profiles of wavelets and their mean frequency for four threshold values of c_w for the sound of the speaker 4 in Table 1.

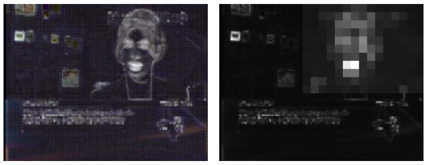

Fig. 5 Left: differential image for m = 3 for the video clip of the speaker 4 in Table 1 and right: averaged one in the sub-windows.

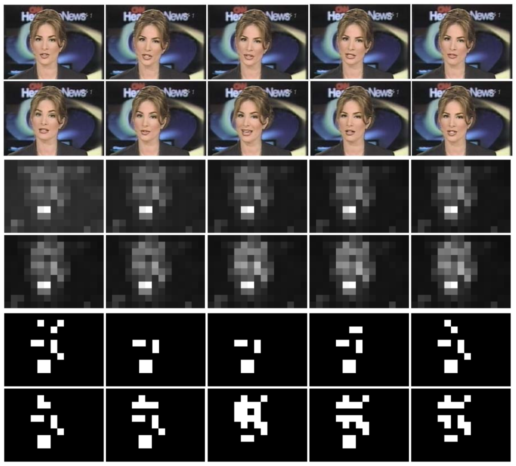

Fig. 6 Top: 10 images for the video clip of the speaker 4 in Table 1; middle: differential images averaged for m = 1 to 10 from top left to bottom right; and bottom: binary patterns obtained using a threshold value.

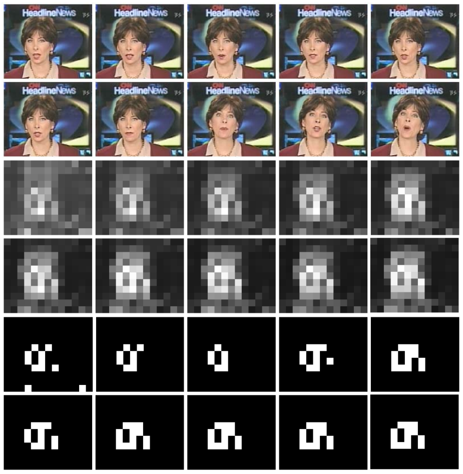

Fig. 7 Top: 10 images for the video clip of the speaker 5 in Table 1; middle: differential images averaged for m = 1 to 10 from top left to bottom right; and bottom: binary patterns obtained using a threshold value.

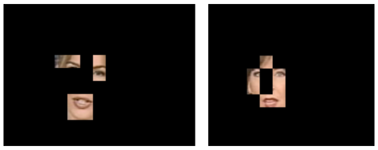

Fig. 8 Face parts for the binary patterns obtained for the two video clips in Figs. 6 and 7.

Table 1 Examples of estimated number of wavelets for several speaking and music scenes.

speaker/music	Time (sec)	syllables	wavelets N_{fft}=512	wavelets N_{fft}=256	rate 1/sec
1. cnn, male	20.00	103	78	104	3.9
2. abc, male	17.68	82	106	172	6.0
3. bbc, female	10.24	51	55	64	5.4
4. cnn, female	4.04	23	23	33	5.7
5. cnn, female	2.80	15	15	19	5.4
6. cnn, music	7.00	-	78	133	11.3
7. cnn, music	3.42	-	11	38	2.4

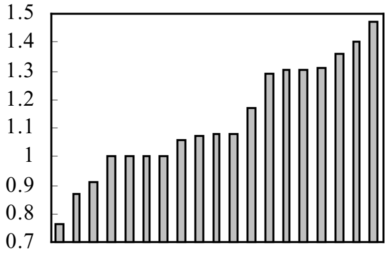

Fig. 9 Ratios of the number of wavelets estimated to that of syllables spoken for various short video clips ranging from 1.6 to 20 sec.