Abstract—Previous studies on face detection in video footages show that segmenting faces accurately and reliably is often hard to succeed, leading to interactive manipulation. This paper presents a simple and practical face detecting and identifying system in video footages from input key face images with focus on segmentation. Each time the scene change is detected the system tries to find a similar face on the beginning frame to each of the key faces. We present an improved color-based segmentation that uses RGB and HSV systems and includes isolated but enclosed face parts, to generate more informative binary face patterns with representative color information. A few variations of the face image are synthesized from the input key face images, to more effectively detect faces in the video footage that may slightly be tilted or rotated from the frontal one. Some experimental results are given.

Index Terms: face detection, pattern matching, segmentation, video retrieval

I. INTRODUCTION

Researches on face detection in images and its related areas have extensively been made in recent years especially in the fields of image processing and computer vision [1]. Among them, Rowley, Baluja and Kanade [2] and Feraud et al. [3] used neural network-based methods, Schneiderman and Kanade developed a Naïve Bayes classifier [4], Osuna, Freund, and Girosi proposed a method called Support Vector Machine classifier [5], and Turk and Pentland presented a method that uses eigenfaces [6]. Those algorithms are aimed at detecting the existence of face and its location in image accurately and in real time [7], [8].

In the field of multimedia, on the other hand, the focus of research has been on not just detection but also identification of faces, people, or some specific objects in video images or video footages [9], [10]. Satoh, Nakamura and Kanade, for example, tried to retrieve the name from the face or the face from the name using the video, video caption and the transcripts, where they detect faces with the neural network-based method and track the face [9]. Since the accuracy of segmentation affects to the identification, several improvements have been reported, which combines temporal segmentation or tracking with spatial segmentation [11] or manual segmentation[12]. Long et al. presented a method that uses three consecutive frames to take into account motion and user interaction to cope with situations in which automatic detection fails. [10]

This paper presents a simple and practical face detecting and identifying system in video footages from input key face images; that is, it is aimed at finding similar faces in the video footages to the input ones. In our system each time scene change is detected, a sampling point is scanned on the beginning frame and color-based spatial segmentation is carried out at each point to get segmented face images, which are then transformed to binary patterns with representative color information to correlate with those of input face images.

To achieve accurate segmentation in this system, we use both RGB and HSV color systems, discrimination by skin color, a few different segmentation errors, and identification of eyes and mouth. Also we include spatially segmented isolated but enclosed face parts to make the segmented image more informative. As for the key face images, a few variations of the segmented face image are synthesized to detect slightly tilted or rotated faces effectively. The parameters can be controlled depending on the properties of the key images and the intention of the retrieval. For example, the parameters could be set restrictedly close to those derived from the key image, which may enable us to efficiently retrieve the relevant scenes in the footage. In the other extreme case, when as many similar faces as possible are desired to detect, then those restrictive parameters can loosely be set.

II. FACE DETECTION AND IDENTIFICATION SYSTEM

One of the two video footages used is an edited one of CNN headline news, consisting of 100 frames, where an anchorperson appears four times and three interviewing scenes are sandwiched as shown in Fig. 1. The other also consists of 100 frames edited from a variety of news programs, which may correspond to 100 different scenes, as shown in Fig. 2. Figure 3 demonstrates how our system works. If we input one of the anchorperson’s images to find where she appears in the video footage, the system outputs the beginning frames of the four scenes among seven. If we add the second image to find where he is, the system adds the beginning frame of the first interview scene, where he appears. Shown on the right in the figure are binary patterns of their frontal faces synthesized from the input face images.

Figure 4 shows a flowchart of the system. Given a key face image, the face part is segmented for each of a number of combinations of sampling point, color system and segmentation error. The most appropriate result is selected among the segmented images, and the average color of (r, g, b) or (h, s, v) and their standard deviations are computed. Then, two variations of the segmented face image, for example, are synthesized and they are made into binary patterns with the common color information.
Fig. 1 One of the two short video footages used in experiments, which consists of 100 frames of 304x232 pixel image. This footage consists of four scenes where the same anchorperson appears and three different interviewing scenes. The numbers are those of the beginning frames of the seven scenes.

Fig. 2 The other video footage used in experiments, which consists of 100 different scenes of 304x232 pixel image frames. The faces vary in color, position within the frame, size, number and in the extent of how frontal. Also the backgrounds vary, and the lighting conditions appear to be different for different images.

Fig. 3 Finding and identifying faces from key face images. If we input the anchorperson’s image, the system outputs the beginning frames of the four scenes among seven. If we add the second picture as key image, then the system adds the beginning frame of the first interviewing scene.

Fig. 4 Flowchart of our system. The parameters relevant to each processing can be set interactively.

Then, the video stream is analyzed to detect the scene change. If there exists a significant change, a sampling point is scanned on the beginning frame of the detected scene. At each point, color-based segmentation is carried out for each of a number of combinations of HSV and RGB colors and multiple segmentation errors from the color values at each sampled point. The segmented images are then transformed into binary patterns with the information of average color and standard deviations. Each pattern is correlated with each of the key face patterns. The frame with the maximal correlation value is obtained for each scene. Finally the frames are selectively displayed with their numbers according to the correlation value.

III. SEGMENTATION

We use the following simple evaluation to detect the scene change as

\[
D_{\text{scene}}(n) = \frac{1}{3X_I Y_I} \sum_{x=1}^{X_I} \sum_{y=1}^{Y_I} \sum_{c=r,g,b} |i(x, y, c; n) - i(x, y, c; n - \delta n)|
\]  

where \( n \) is the order of the frame, \( i \) is the image with size \( X_I \) by \( Y_I \) in pixel, \( c \) is \( r, g \) or \( b \) component, and \( \delta n \geq 1 \), which depends on how quickly one scene moves to another. The scene is judged to have changed if \( D_{\text{scene}} > D_{\text{th}} \) holds.

Which color system to be used may depend on the image. To cope with a variety of images, we use either RGB or HSV or both. Letting the sampled image value be \((r_s, g_s, b_s)\) for RGB or \((h_s, s_s, v_s)\) for HSV, the image region continuous with the sampled point within an error is segmented as:
knowledge of the ratio of the standard deviation to mean of the face image, it may be possible to decide the spacing with the uniformity of the face in RGB or HSV values. In this case, however, may not be usual. The spacing may also be affected by the uniformity of the face image, which, available from the size of the key face image, which, to cope with a wider range of faces or lighting conditions.

The error in Eq. (2) or (3) may be set roughly to the ratio of the standard deviation to mean value of the face image, which, statistically the ratio ranges roughly between 10% and 20% for each of $r, g, b, s, v, r, s, v, h, b, g, r, g$. We may be able to select an appropriate color system based on the properties of the key image, or we can use both to cope with a variety of different conditions. The weights may be applied, for example, to compensate for lighting conditions or to optimize the segmentation for some specific image.

The image region which satisfies the condition in Eq. (2) or (3) but is not continuous with the sampled point is included in the segmented image, if even a slightest part of the region is longitudinally sandwiched by the already segmented part within a certain distance, as shown in Fig. 5. This extension may be useful in making the segmented image more informative, leading to enhanced matching with the key images.

The error in Eq. (2) or (3) may be set roughly to the ratio of the standard deviation to mean value of the face image part. The ratio may be different for a different face and under a different lighting condition. Statistically the ratio ranges roughly between 10% and 20% for each of $r, g, b, s, v, r, s, v, h, b, g, r, g$. We may be able to select an appropriate color system based on the properties of the key image, or we can use both to cope with a variety of different conditions. The weights may be applied, for example, to compensate for lighting conditions or to optimize the segmentation for some specific image.

The appropriate sampling spacing on the frame might be selected to cope with a wider range of faces or lighting conditions.

The color window is used to accept or reject the color at a segmented part within a certain distance, as shown in Fig. 5. This extension may be useful in making the segmented image more informative, leading to enhanced matching with the key images.

IV. KEY IMAGE PATTERNS FROM KEY FACE IMAGES

We synthesize a few variations from each of the input key face images to enhance the detection for near-frontal faces, as shown in Fig. 4. For example, when we stand on the assumption that the original image is near-frontal, three patterns are generated as follows. The simplest variation is the horizontally symmetric version of the original face image.

In the same way as for the video stream, those variations of key face images are segmented using RGB and HSV colors and a few segmentation errors at each of some sampled points. Among them, the most appropriate segmented face images are chosen, and their properties of size and means and standard deviations of color components are computed. These images are made into binary patterns.

V. PATTERN MATCHING

Using the properties of the key face image, we may decide the color system, the sampling spacing, the color window, the segmentation error, and the window of segmented image size. The color window is used to accept or reject the color at a sampled point. The window of segmented image size uses such loose restrictions as $3X < Y$ and $X < 2Y$ and may also use more strict ones relevant to the key face image. After obtaining the segmented images, they are made into binary patterns, which are then judged whether they are patterns of face. We impose on the binary patterns of the segmented images that the numbers of null-value pixels in the three regions in Fig. 6 exceed given values, which are proportional to the size $XxY$. Then, the segmented binary pattern is correlated with one of the key face patterns, where the former is reduced or enlarged independently in $x$ and $y$ directions so that the resulting pattern is equal to the key face pattern in size. In the case of $Y > 2X$, however, the fitting ratio in $y$ is made same as in $x$, and after the segmented pattern is fitted to the key face pattern the top part of the segmented image is used for the correlation.
VI. EXPERIMENTS

Table I shows statistics of the 21 segmented faces from the news video footages, which include a variety of skin colors and ages. It is seen that means of $r$, $g$ and $b$ components are not necessarily the same but tend to decrease in that order, while the ratios of the standard deviation to mean tend to slightly increase in that order. They are very different for different persons as well as for different lighting conditions and color adjustments of the video camera.

Figure 7 shows the inclusion of isolated image parts as part of the face to make the resulting segmented patterns more informative, where eyelids and part of lips and teeth are included. Figure 8 shows the three binary patterns generated from the single key face image. If the original image is frontal enough, the synthesis may be easy. Figure 9 shows that the spacing of the sampling point, $dx$ ($= dy$), larger than $X/dx = 2$ is enough to attain more than 0.8 as the correlation value for the key face pattern $T1$ and that the threshold value increases to 4 for $T16$ and 7 for $T75$, where $T1$, $T16$, and $T75$ are shown in Fig. 10 and their sizes are denoted by $X_x \times Y_y$. The ratio of standard deviation to mean of the segmented image, averaged over the three color components, is 0.098, 0.120, 0.147 for $T1$, $T16$, and $T75$, respectively. Hence the spacing has to be smaller as the ratio increases. In the result, HSV color system and multiple error values 0.1, 0.15, 0.2 and 0.25 were used for segmentation. Fig. 11 shows that the correlation peak becomes narrower as the ratio of standard deviation to mean increases, where the spacing used is small enough. Then, the segmentation error has to be selected more carefully as the ratio is larger. Figures 12 and 13 show that for the case of $T75$ RGB color system cannot correctly segment the face part while HSV system can reasonably achieve it.

Using variations of key face image is effective in enhancing the detection capability, as shown in Table II. In the case of using a single key face pattern $T1$ given in Fig. 10 for the first video footage shown in Fig. 1, the correlation values for the frame numbers 1, 31 and 60 were 0.90, 0.57 and 0.64, respectively, while the other frames had the maximum value of 0.54. Using the two variables in addition to $T1$, shown in Fig. 8, enhances those values significantly to 0.90, 0.72 and 0.76, while the maximum value of the other frames remain the same.

Utilizing the properties of the key face images makes the detection easier and faster. The segmented image of $T12$ in Fig. 14 has unusual average $r$, $g$ and $b$ values, 145, 103, 28 of 255, respectively, due to the lighting condition. If we use the color window of 0.61 $< g/r < 0.81$ and 0.09 $< b/r < 0.29$ the twelfth frame had the correlation value 0.98 while most of the frames has been rejected by the color window and a few went through it, resulting in the maximum correlation value of 0.06. The key face image on the 45th frame, on the other hand, has a normal skin color with average values (198, 148, 129) of 255. In this case, using the color window of 0.70 $< g/r < 0.85$ and 0.55 $< b/r < 0.70$ made us obtain the result that the 45th frame has 0.99 while other frames had the maximum value 0.68. If we take into account the correlation in color, the latter value reduces to 0.53, where a Gaussian function with the standard deviation of 15% for each color component is used.

Our present algorithm can detect more than 90 faces among 100 in Fig. 2 without restrictions such as color window, although the number depends on how complete the segmented image is. We also found in the experiments that HSV tends to be much more useful than RGB for the segmentation.

VII. CONCLUSIONS

A simple and practical face-detecting and identifying system in video-footages from key face images was presented with focus on the segmentation algorithm. We confirmed our system works well. However, the system needs to be improved in terms of both speed and accuracy. As for the speed, it may be useful to know where the eyes and mouth are on each scene by obtaining a differential image. This may localize the sampling points on each frame. And utilizing the audio information may also be useful to distinguish the scenes worth to detect faces from those of transient scenes. As for the accuracy, combining the techniques developed for face detection with our system may be useful.

REFERENCES

Table I Statistics of the 21 segmented face images in the video footages, where images taken under unusual lighting condition or having significant intensity variations on the face are excluded, and the mean value of $r$ is normalized to unity.

<table>
<thead>
<tr>
<th></th>
<th>$r$</th>
<th>$g$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>1.00</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>standard deviation/mean</td>
<td>14.2%</td>
<td>14.5%</td>
<td>16.1%</td>
</tr>
</tbody>
</table>

Fig. 7 Segmentation with and without the inclusion of isolated face parts, where (a) part of a frame, (b) segmented image without the inclusion, (c) its binary pattern with the color information, (d) segmented image with the inclusion, (e) its binary pattern with the color information.

Fig. 8 Synthesis of variations from the key face image, where (a) part of the input image, (b) segmented image, (c) and (e) binary patterns horizontally symmetrical, and (d) frontal binary pattern synthesized from (c) and (e).

Fig. 9 Profiles of correlation value as function of $X_t/dx$ for each of the three key face patterns given in Fig. 10, where the same video footage as in Fig. 1 was used. In the figure T1(others), for example, shows the maximum correlation value profile that the six frame images other than the first one give.

Fig. 10 Key face patterns of T1, T16, and T75 with the sizes $X_t \times Y_t$ of 40x67, 36x57, and 36x70, respectively.

Fig. 11 Profiles of the correlation value as function of segmentation error for three key face patterns.
Fig. 12 Segmentation for a face with hair of a similar color, where (a) part of the frame, (b) segmented face using RGB color system, (c) its binary pattern, (d) segmented face using HSV color system, (e) its binary pattern.

Fig. 13 Profiles of correlation value as function of $X_t/dx$ for two cases using RGB and HSV color systems for the key face pattern T75, where others(HSV) and others(RGB) are profiles of the maximum correlation value for the first five frames and multiple error values 0.1, 0.15, 0.2, and 0.25 are used.

<table>
<thead>
<tr>
<th>key face patterns used</th>
<th>Frame number and Correlation value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>T1</td>
<td>0.90</td>
</tr>
<tr>
<td>T1 and its two variations</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table II. The effect of using variations of key face image on enhancing the detection of the anchorperson, where the same video footage as in Fig. 1 was used and HSV color system, $e_{HSV} = 0.1, 0.15, 0.2, 0.25$, and $dx = dy = 10$ are used.

Fig. 14 Two key face images and their binary patterns T12 (top) and T45 (bottom) in the second video footage in Fig. 2.