Face Detection in Videos Using Skin Color Segmentation and Saliency Model
Chao Wang, Yunhong Wang, Zhaoxiang Zhang
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
E-mail: ableblaze@163.com, yhwang@buaa.edu.cn, zxzhang@buaa.edu.cn

Abstract: Given a video sequence containing face candidates, detecting faces is a challenging problem, which can be attributed to the difficulty in handling the appearance variability of the face. Based on skin color segmentation combined with the saliency model, a novel method is proposed to detect human faces in videos. Firstly, a skin color model in the YCbCr chrominance space is built to segment skin-color pixels from background in frames. Then, the candidate regions of human face can be extracted using mathematical morphological operators. Finally, a saliency model is proposed to detect eyes and mouths in face regions. Experimental results demonstrate successful face detection performance over a wide range of facial variations in color, illumination and scale in videos.

Key Words: Face detection, skin color segmentation, saliency model

1. INTRODUCTION
Visual detection of objects, such as faces, has received significant attentions in the computer vision community. Detecting face is a crucial step in the identification applications like recognition or identification of faces, video surveillance, intelligent human-computer interface, face image database management and querying image databases. Detecting and locating human faces and facial features in an image or image sequences are important tasks in dynamic environments, such as videos, where noise conditions, illuminations, locations of subjects and pose can vary significantly from frame to frame [1][2].

There have been various approaches proposed for face detection, which could be generally classified into four categories [3]: template matching based methods, feature-based methods, knowledge-based methods, and learning based methods. Template matching based method means the final decision comes from the similarity measurement between input image and the template. It is scale-dependent, rotation-dependent and computational expensive. Feature-based methods use low-level features such as intensity [4], color [5], edge, shape [6], and texture to locate facial features, and further, find out the face location. Knowledge-based methods [7] detected an isosceles triangle (for frontal view) or a right triangle (for side view). Learning based methods use a lot of training samples to make the classifier to be capable of judging face from non-face. Despite of the notable successes achieved in the past decades, making a trade-off between computational complexity and detection efficiency is the main challenge. Among the face detection algorithms, skin color based detection information is an important category [8].

This paper is organized as follows. In Section 2, skin color segmentation and morphological operations are applied to detect candidate face regions. These regions are processed by the saliency model in Section 3. Section 4 calculates the positions of eyes and mouths. Experimental result are given in Section 5. Finally, the paper is discussed and summarized in section 6.

2. SKIN COLOR SEGMENTATION
Skin color segmentation is a very fast approach in face detection because faces can be detected directly from segmented skin areas. The final goal of skin color segmentation is to build a decision rule, which will discriminate between skin and non-skin pixels. This is usually accomplished by introducing a metric, which measures distance (in general sense) of the pixel color to skin tone. The type of this metric is defined by the skin color modeling method.

RGB is a colorspace originated from CRT (or similar) display applications, when it was convenient to describe color as a combination of three colored rays (red, green and blue). It is one of the most widely used colorspaces for processing and storing of digital image data. One method to build a skin classifier is to define explicitly (through a number of rules) the boundaries skin cluster in some colorspace. For example [9]:

\[
\begin{align*}
R &> 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and } \\
\max\{R; G; B\} - \min\{R; G; B\} &> 15 \text{ and } \\
|R - G| &> 15 \text{ and } R > G \text{ and } R > B.
\end{align*}
\]

Modeling skin color requires to choose an appropriate color space and identifying a cluster associated with skin color in this space. Based on Terrillon et al.’s [10] comparison of nine different color space for face detection, it has been observed that the RGB colorspace is not the best choice for face detection. In order to improve the performance of skin color clustering, we use YCbCr space to build a skin color model since it is perceptually uniform, and the chrominance components are almost independent of luminance component in the space. There are non-linear relations between chrominance (Cb,Cr) and luminance (Y) of
The skin-color segmentation and the face region using dilation and erosion operations: (a) original image, (b) the result of skin color segment, (c) the result of skin color segment after using dilation and erosion operations, (d) face candidates, (e) the face candidate after using dilation and erosion operations.

Skin color in the high and low luminance region. YCrCb is an encoded nonlinear RGB signal, commonly used by European television studios and for image compression work. The equation of transformation between RGB to YCbCr is given by the following equation [11]:

\[
\begin{bmatrix}
Y \\
C_b \\
C_r \\
1
\end{bmatrix} =
\begin{bmatrix}
0.257 & 0.507 & 0.078 & 16 \\
0.148 & -0.291 & 0.439 & 128 \\
0.439 & -0.369 & -0.071 & 128 \\
0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B \\
1
\end{bmatrix}
\] (2)

where R, G, B take the typical values from 0 to 255 (8-bit precision), Y is in the same range (0-255), and Cb, Cr components are into the range (16-240). The transformation simplicity and explicit separation of luminance and chrominance components makes this colorspace attractive for skin color modelling.

In the color detection process, each pixel is classified as either skin or non-skin based on its color components. Although skin color varies from person to person, they tend to get clustered into a compact region in YCbCr space.

After obtained skin color regions, it is effective to use morphological operations as a combination of dilation and erosion for face detection. The operation of dilation and erosion are fundamental to morphological processing and are used most often in various combinations in practical processing. Morphological operation can simplify image data while preserving their essential shape characteristics and eliminate irrelevancies. So it can effectively help to derive a more accurate contour of skin segment. Skin color segmentation results was shown in Fig. 1: (a) is the original image from “Economic Half-Hour” and (b) is the result of skin color segment after employing the YCbCr model. Using (b) as the mask we can get the face candidates (d). While (c) is the result of skin color segment after (b) using dilation and erosion operations. (e) is the face candidates chosen from (d) which is employed as a mask.

3. SALIENCY MODEL

Fig. 2: General architecture of the saliency model.

Computational models based on saliency detection have been widely used in computer vision and robotics to predict important areas in the visual field. In the saliency model [12], visual features are calculated by using linear filtering at eight spatial scales, followed by center-surround differences, which compute local spatial contrast in each feature dimension (color saturation, intensity and edge information) for a total of 42 maps. An iterative lateral inhibition scheme instantiates competition for salience within each feature map. After competition, feature maps are combined into a single conspicuity map for each.
feature type. The three conspicuous maps are then are summed into the unique topographic saliency map. The saliency map is implemented as a 2-D sheet of Integrate-and-Fire neurons. The WTA, also implemented using Integrate-and-Fire neurons, detects the most salient location and directs the attention towards it. An inhibition-of-return mechanism transiently suppresses this location in the saliency map, such that the attention is autonomously directed to the next most salient image location. The details of the model used in the present study have been briefly schematized in Fig. 2.

Motivated by the visual attention model, which has been successfully extended to some applications, we will employ the similar idea to indicate the face location in the video frames by constructing a facial saliency map. Studies from developmental vision show that infants demonstrate an innate predisposition towards face-like configurations (not necessarily a face) involving shapes resembling eyes, nose and mouth without the use of texture or color even within days of when they were born [13]. This suggests that those parts may hold a special key to face detection. As we can see in Fig. 3: (a) is an original face image cut from a hole scenes using the method in Section 2, and (b) is the saliency map of (a) while 60 percent saliency part of (a) is shown in (c).

4. LOCALIZATION OF FACE FEATURES

Among the various facial features, eyes and mouth are the most prominent features for recognition and estimation of head pose. In this paper, we locate eyes and mouth positions based on the vertical and horizontal projection derived from an eyes and mouth region map. Employing two separate face masks, one from 60 percent saliency part of a face image and the other from the skin color segment of a face image(using dilation and erosion operations and the structuring element of morphological operations in this part need to be changed due to difference resolution of frames) we can build an eyes and mouth region map. Fig. 4 shows an example of eyes and mouth region map.

Integral projection method was firstly applied to face recognition by T.Kanad [14], it is effective to use this method in extracting the location of facial features. This method assumes that a binary image $g(x,y)$ with the height of $h$(in pixel) and width of $w$, then the vertical integral projection is defined as:

$$V(x) = \sum_{y=0}^{y=h-1} g(x,y)$$

Similarly, the horizontal integral projection is defined as:

$$H(x) = \sum_{x=0}^{x=w-1} g(x,y)$$

The top half of eyes and mouth region map’s vertical and horizontal projection is shown in Fig. 5(a), while Fig. 5(b) is the projection of the bottom half of eyes and mouth region map. Through the derivation of edge curve on integral projection we can certain the position of peak points. At the same time, we can see the location of the eyes and mouth corresponding to the
position of peak point as shown in Fig. 5(a) and Fig. 5(b). In the vertical projection, not only the peak point’s value of the eyes is higher than the eyebrow’s peak point but also the position of eye’s peak point is under the position of eyebrow’s peak point on ordinate. Mouth and nose with a similar relationship with the above. While in the horizontal projection, the eyes and the mouths’ locations are just the the positions of peak point on ordinate. Afterward we locate the face features by certain the peak point as shown in Fig. 5(c).

5. EXPERIMENT

The algorithm of this paper is realized using Matlab R2008a on a PC configured with Pentium(R) Dual-core E5300 CPU(2.60GHz) and 2GB memory. Fig. 6 shows illustration of face detection results in video sequence of the approach. The video are chosen from Internet, and consist of both indoor scenes and outdoor scenes under fluorescent or sunlight.

In the experiment, in order to measure the results, we count the false detection (count the false detections in the correct detections of faces) in the entire frames. There are about 300 frames in every video, and the frame size varied from $288 \times 384$ to $360 \times 640$ while the face size varied from $30 \times 20$ pixels to $100 \times 150$.

Table 1 shows the accuracy of face detection in experimental videos. The accuracy of eyes detection in videos is shown in Table 2: Video.1 is a video which has low resolution, but only one face in a frame and the face region take a large part of a frame. In these videos the false detections happen when the eyes closed, and the percentage of saliency is 55% the accuracy reach the maximal value. Video.2 is a video which has a higher resolution, and only one face in a frame besides the face region take large part of a frame. These videos always have a high accuracy of detection, and the accuracy reach the maximal value when percentage of saliency is 70%. Video.3 is a video which has low resolution, not only have more than two face in a frame but also the face region is very small. These videos can not reach high accuracy, because of the small face region and the low resolution, the morphological operations destroy the face features while the features are not salient enough to the saliency model. The accuracy only reach 70% when percentage of saliency is 65%. As we can see in Table 3, the accuracy of mouth detection is almost all 100%. This is because mouth is the most outstand face feature in the bottom half of a face region. So it is not easy to be destroyed by morphological operations and salient sufficient to the saliency model.

6 CONCLUSION

Experiments show that this method has strong robustness. It can detect the face features in various conditions such as backgrounds movement and environment illumination variation. The algorithm of this paper can used for some face feature detection system. In the next work, the solution of raise the accuracy of detection in low resolution videos will be considered.
ACKNOWLEDGEMENT

This work is funded by the National Natural Science Foundation of China (No. 60873158), the National Basic Research Program of China (No. 2010CB327902), the Fundamental Research Funds for the Central Universities, and the Opening Funding of the State Key Laboratory of Virtual Reality Technology and Systems.

References


