Multi-face Detection System in Video Sequence

Phuong-Trinh Pham-Ngoc, Kang-Hyun Jo
Graduate School of Electrical Engineering, University of Ulsan
680-749 San 29, Mugeo 2-Dong, Nam-Ku, Ulsan, Korea.

Abstract – This paper presents an improved system for detecting multiple faces in a video sequence where detection is not limited to frontal view. We propose an adaptive selection of skin models from two different ones in RGB and ratio rgb space to overcome the illumination problem caused by automatic focus of camera. With the proposed solution, we receive more reasonable skin detection for different human races. We modify Local Binary Pattern (LBP) by adding a set of spatial templates. This LBP considers both principal local shapes and spatial textures of facial components. Human face is represented by a histogram of LBP coefficients. Moreover, the grayscale image of human face is changed to Discrete Cosine Transform (DCT) coefficients used in embedded Hidden Markov Models (eHMMs). A modified LBP (mLBP) histogram matching and eHMMs are composed to hierarchical classifier to determine whether skin regions are faces or not. The results of our system on different cameras are presented and discussed. We compare the performance capability of our system with other systems using separately eHMMs or LBP histogram matching. The proposed system can work well to detect attitudes of faces like frontal, rotated and profile faces.

Index Terms – Skin detection, LBP, eHMMs, profile faces.

I. INTRODUCTION

Face detection is required as the first step of the automatic face analysis system. This work has been widely investigated in recent years because it overlies many areas of application: face recognition, man-machine interaction systems, visual communication systems, video-surveillance, etc. However, face detection is a challenging task due to variation in illumination, variability in scale, location, orientation and pose. Facial expression, occlusion and lighting conditions also change the overall appearance of face. Many face detection methods in images [9] and in video sequence [2,3,7] have been published and have achieved some encouraging results.

For video sequence, face detection problem has been solved in two main approaches. The first way is to detect faces for every frame without using temporal information. The other is to detect a face in the first frame and then track the face through the sequence. This paper presents an improved face detection system in video sequence based on the first way. To reduce the effect of illumination changing caused by automatic focus of camera, we propose an adaptive selection of skin color models to receive more reasonable skin detection. Unreasonable skin regions are discarded by facial geometric conditions of human faces. Reasonable ones are prevented by replacing them with elliptic skin regions. Then we get the most potential ones considered as face candidates. We propose a modified LBP considering not only local spatial textures but also principal local shapes. A histogram of modified LBP coefficients is considered as facial representation. A combination of template matching and appearance-based methods is used for classification step. LBP histogram matching and eHMMs are combined into a hierarchical classifier to identify if face candidates are human faces. Our system is described in Fig. 1.

II. FACE CANDIDATE LOCALIZATION

A. An Adaptive Selection of Skin Color Models

The use of color information can reduce the search space for face detection. It allows fast processing and simplifies the task of face localization in complex backgrounds. It is also highly robust to geometric variations of the face patterns. Therefore, we use skin color detection as the first step in detecting faces.

In general, the automatic focus of camera or webcam causes the changing illumination. It affects the quality of captured images. To overcome this problem, we use an adaptive skin model selection shown in Fig. 2 to receive more reasonable skin detection. We define how to choose skin models of Lin [1] and Peer [5] to adapt with different illumination conditions. The model of Lin has a wide range of skin segmentation. It is sensitive to remain white and yellow color. However, it works well in really high or low illumination conditions. In contrast, the model of Peer has good skin detection result in normal illumination, but it often retains red color.

Fig. 1. The proposed face detection system.
In the figure above, $I_p$ and $I_L$ are the same size binary images which are skin detection results from skin models of Peer and Lin, respectively. $I_{AND}$ is the result image from AND operator of $I_p$ and $I_L$. For each captured color image, these two images are calculated and compared together. They are considered as different images if the ratio $R_1$ given in function (1) is larger than 10%.

$$R_1 = \frac{\sum|I_p - I_L|}{\text{size}(I_p)}$$  \hspace{1cm} (1)

We also define whether the camera focus is near or not. The focus of camera is considered as near focus if the ratio $R_2$ given in function (2) is larger than 5%.

$$R_2 = \frac{\sum \text{skin pixel}}{\text{size}(I_{AND})}$$  \hspace{1cm} (2)

In general, the illumination is good if two skin detection results are similar. This is the best situation and we do not need to check the focus of camera. We use $I_{AND}$ image as the most adaptive skin detection result to reduce yellow and red color when they exist in two skin detection images.

In contrast, the illumination is too high or low if skin detection results are different. In addition, if the camera focus is far, the illumination is often low and the skin detection of Peer model usually loses many skin regions. In this case, we should use the skin detection of Lin to receive much more necessary information. If the camera focus is near, the illumination can be too bright. The adaptive selection for skin detection in this case is $I_{AND}$ image.

B. Face Candidate Localization
After skin detection, we label connected skin regions and erase the regions whose areas are smaller than a threshold. We call this step reducing small noise. Sometimes we maybe lose necessary face regions after skin segmentation. We will recover those regions by labeling connected non-skin regions in each skin region and change them into skin ones. This process ignores non-skin regions connecting directly to boundaries of their skin regions and non-skin regions whose areas are bigger than selected thresholds. With skin regions having properties of human faces, we preserve them by covering their areas with skin ellipses. Those works are very important for our face candidate localization step because we will use strong conditions of intensity histogram of skin pixels to separate connected faces. In some cases, human faces in images can be connected together or with other things such as hand, arm. This is one of challenges for face detection problem. Without those processes above, we may reduce or lose correct faces under strong separation. After separating connected objects, we reject non-face skin regions by some facial geometric conditions: the ratio between the width and height of skin region, the number of skin pixels in skin region, etc. Finally, we get the most reasonable skin regions considered as face candidates. Those candidates are changed to grayscale images used in hierarchical classifier which is described in next sections.

III. A MODIFIED LBP

Human face is a near-regular texture pattern generated by facial components and their configurations. Considering facial components such as eyebrow, eye, pupil, nose and face boundary, we select 8 main different spatial templates shown in Fig. 3 to preserve shape information of facial components. With only those spatial templates, we can describe all facial components; for example, eyebrow can be described by a union of templates d, b and c. However, we combine both those spatial and local texture information to improve the capacity of describing faces.

![Fig. 3. Eight main spatial templates.](Image)

In our method, instead of considering the central pixel $P_c$ only with its each neighborhood pixel as original LBP operator did [8], we use each pair of two neighborhood pixels ($P_{i1}, P_{i2}$) according to spatial templates to compare with the central pixel $P_c$. Eight spatial templates form 8 binary digits of mLBP number. So mLBP operator produces 256 different values. Equation (3) gives the computation of mLBP number.

$$mLBP = \sum_{i=0}^{7} S_i(x) \cdot 2^i$$  \hspace{1cm} (3)

where $S_i(x)$ is the $i^{th}$ binary digit of mLBP number; $S_i(x) = \begin{cases} 1, & (P_c > P_{i1}) \text{ and } (P_c > P_{i2}) \\ 0, & \text{otherwise} \end{cases}$

In fact, mLBP gives us information about both local shapes through 8 spatial templates and local textures. We retrieve more information to distinguish face and non-face objects.

IV. HIERARCHICAL CLASSIFIER OF FACE DETECTION SYSTEM

In each captured frame, our system defines face candidates as presented in section II. We apply a hierarchical classifier scheme shown in Fig. 4 for each face candidate to determine whether this is human face or not. We combine the template
matching and appearance-based methods into this classifier. To use those methods, we must create face and non-face database.

### A. Face and Non-face Database

Face detection will become an easy problem if we have clearly face and non-face class modeling. However, it is difficult to model non-face class because anything which is not a face belongs to non-face class. In our method, we collect 190 frontal and profile face images to create face samples. To create non-face class, we choose three main non-face objects: arm (10 samples), hand (31 samples) and noise (56 samples). All samples are 72x93 size color images. In our experiments, those samples are enough to represent face and non-face classes.

#### B. Mixed mLBP Histogram and Difference $D_f$ Matchings

We use the histogram of mLBP coefficients to represent a face. If we only use single mLBP histogram for the whole face candidate image, occlusion will affect template matching algorithm seriously. To reduce the impact of occlusion, in general, we divide human face into two parts: the upper part from nose up to forehead and the lower one from nose down to neck. We calculate individual histogram for each part and connect them sequentially to create one mixed 255x2 bin histogram representing to face candidate image. By this way, we reduce effectively the influence of occlusion.

Given an image $I$, one mixed mLBP histogram is denoted by $H_{mLBP}^{mix}(I)$. We adopt error measurement because of simple and fast computation. A distance measurement is defined as:

$$D(H(I_1), H(I_2)) = \sum_{i=1}^{n} H_{i}^{mLBP_{mix}}(I_1) - H_{i}^{mLBP_{mix}}(I_2)$$

where $H_{mLBP}^{mix}(I_1)$ and $H_{mLBP}^{mix}(I_2)$ are two mixed mLBP 255x2 bin histograms, and $n$ is the number of bins.

Given a face database with $m$ samples, for any sample $P$, we change it from color image to grayscale one and define its histogram-matching feature as the average distance to face training samples as follows:

$$f_{face}(P) = \frac{1}{m} \sum_{i=1}^{m} D(H(P), H(X_i))$$

where $X_i$ is a face training sample.

In fact, this histogram-matching feature has the discriminating ability between face and non-face patterns. To demonstrate this property, we can see Fig. 5a which shows the positive and negative distance measure distribution over 156 face samples and 121 non-face samples. With this feature $f_{face}$, we use thresholds called $T_{face}$ to classify face and non-face objects. In our experiment, if $f_{face}$ is smaller than 1800, face candidate can be considered as face with 99% of correct detection. If $f_{face}$ is bigger than 3500, the face candidate is almost not a human face. Only in the range [1800,3500] of $T_{face}$, it is still hard to say if the face candidate is a face. We improve face detection in this range by the following feature.

With non-face database, for any sample $P$, we also define its histogram-matching feature as the minimum of three average distances to three non-face object training samples, given by

$$f_{nonface}(P) = \min(f_{arm}(P), f_{hand}(P), f_{noise}(P))$$

where $f_{arm}$, $f_{hand}$ and $f_{noise}$ are calculated following to (5). The difference between $f_{nonface}$ and $f_{face}$ shown in Fig. 5b also has the discriminating ability between face and non-face patterns. We call it difference $D_f$ given by

$$D_f(P) = f_{nonface}(P) - f_{face}(P)$$

We use $D_f$ to improve the face detection rate when $T_{face}$ is in [1800,3500]. We define the explicit thresholds $T_{Df}$ for $D_f$ to distinguish face and non-face patterns.

We specify matching conditions for both $f_{face}$ and $D_f$ and use them jointly for the two first matching stages in hierarchical classifier. The decision rules are as follows:

$$\delta(f_{face}, D_f) = \begin{cases} 1, & \text{if } (f_{face} \leq 1800) \text{ or } (f_{face} \text{ and } D_f \text{ are as listed in table 1}) \\ 0, & \text{otherwise} \end{cases}$$

<table>
<thead>
<tr>
<th>$f_{face}$</th>
<th>$D_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect ratio $\geq 0.85$</td>
<td>Aspect ratio $&lt; 0.85$</td>
</tr>
<tr>
<td>[1800,2500]</td>
<td>0–</td>
</tr>
<tr>
<td>[2501,2900]</td>
<td>300–</td>
</tr>
<tr>
<td>[2901,3500]</td>
<td>300–</td>
</tr>
</tbody>
</table>

where aspect ratio is a ratio between the height and the width of one face candidate.

After those two stages, the face detection rate can reach over 85%. In order to increase the performance of our system, eHMMs is used as the last step to check face candidates which are not satisfied the two first template matching steps to give the final conclusion.

#### C. Embedded Hidden Markov Models (eHMMs)

In our algorithm, we define non-face class as three different sub-classes: arm, hand and noise. It means our face detection is changed to four class pattern classification problem. EHMMS classifier [4] performs pattern recognition for a four-class problem by determining the maximum likelihood to find the most similar class for candidate object. Given training sets of positive and negative samples, we will have four eHMM models corresponding to four classes: face, arm, hand and

![Fig. 4. Hierarchical classifier scheme.](image)

![Fig. 5. Distribution of measurement: (a) Distribution of distance measure $f_{face}$, (b) Distribution of difference feature $D_f$.](image)
noise. A face candidate which was ignored by the two first stages of face detection system is checked by eHMMs. Finally this is not human face if the result of this face candidate under eHMM stage is non-face.

After hierarchical classifier, human faces in a current frame are detected. Our system will continue with new captured frame from video sequence unless the stop alert is announced.

V. EXPERIMENTAL RESULTS

We tested the performance of our system with the database forming from different sources: NRC-IIT facial video database, our own facial video database and movies. From NRC-IIT database, we checked 23 single-face video sequences with 11 persons in different pose and orientation of faces. For our own database, we used different kinds of camera such as Genius videocam series V4, Cosy net PC camera-PC590, Chinese PC camera and Canon powershot A95 to record video sequences and use them to check our system, especially the performance of proposed skin detection scheme. This database includes 20 video sequences with multi-face appearance from 15 different persons in few days. The result from those video sequences proved a good multi-face detection capability of our system in various poses and orientations of faces. Especially, because of using mLBP histogram matching, our system can also detect faces with 90 degree pose as shown in Fig. 6(a).

![Fig. 6. Several face detection results: (a) Video sequence from our database with Genius webcam, (b) Video sequence from the movie “Harry Porter”.](image)

In addition, we tried to check our system with 10 video sequences taken from several films such as “Harry Potter”, “The Myth”, “The Chronicles of Nania”, etc. In those video sequences, there are multi-face appearance, complex human actions and backgrounds. Our system also can work well in those databases as shown in Fig. 6(b). However, because color effect is often used in films, the skin detection step was influenced to cause much noise to make false detections.

The performance of our system depends on connected component step in face candidate localization. If the number of face candidates increases, the speed of our system decreases. Approximately, the speed of our system is 5fps for 160x120 image sequences; 2.57 fps for 320x240 image sequences and 1.33 fps for 640x272 image sequences with Pentium 4 CPU, 2.6GHz, 512MB of RAM in Visual C++ environment. All face detection results of our system are shown in table 2. The comparison between our method and others is given in table 3.

<table>
<thead>
<tr>
<th>Database</th>
<th>Detection rate</th>
<th>Missing rate</th>
<th>False rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRC-IIT and our own database</td>
<td>93%</td>
<td>7%</td>
<td>3%</td>
</tr>
<tr>
<td>Film sequence</td>
<td>82%</td>
<td>18%</td>
<td>14%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our algorithm</th>
<th>LBP hist. matching</th>
<th>eHMMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection rate</td>
<td>93%</td>
<td>80%</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

We presented an improved frame-based face detection system. We proposed an adaptive selection of skin models to reduce effectively the effect of automatic camera focus for different cameras. Our system work well to detect multiple faces in various positions, scale, orientation and pose. Especially, our system is improved to detect profile faces with the angle of pose about 90 degrees. Comparing to use only eHMMs or LBP histogram matching, our system has better capacity of detecting multi-faces. Our future work is improving the speed of this system to at least 10fps to apply in real time application.

ACKNOWLEDGMENT

The authors would like to thank to Ulsan Metropolitan City and MOCIE and MOE of Korean Government which partly supported this research through the NARC and post BK21 project at University of Ulsan.

REFERENCES