Robust multi-clue face tracking system

Yujia Cao[1,2]
1. .Radio Engineering Dept. Southeast University Si Pai Lou 2#, 210096 Nanjing, Jiangsu, China
2 Human Media Interaction Lab University of Twent Electronics P.O. Box 217, 7500AE Enschede, Netherlands
y.cao@utwente.nl

Xin Wei[1], Li Zhao[1] , Riccardo Di Federico[3]
1.Radio Engineering Dept. Southeast University Si Pai Lou 2#, 210096 Nanjing, Jiangsu, China
3. Innovation Philips Consumer High Tech Campus 37, 5656AE High Tech Campus 37, 5656AE Eindhoven, Netherlands
nuptwx @163.com, zhaoli@seu.edu.cn
riccardo.di.federico@philips.com

***Abstract*—In this paper we present a multi-clue face tracking system, based on the combination of a face detector and two independent trackers. The detector, a variant of the Viola-Jones algorithm, is set to generate very low false positive error rate. It initiates the tracking system and updates its state. The trackers, based on 3DRS and optical flow respectively, have been chosen to complement each other in different conditions. The main focus of this work is the integration of the two trackers and the design of a closed loop detector-tracker system, aiming at achieving superior robustness at real-time operation on a PC platform. Tests were carried out to assess the actual performance of the system. With an average of about 95% correct face location rate and no significant false positives, the proposed approach appears to be particularly robust to complex backgrounds, ambient light variation, face orientation and scale changes, partial occlusions, different facial expressions and presence of other unwanted faces.**

Keyword-face tracking; multi-clue; face detector

I. INTRODUCTION

Face as the primary part of human communication has been a research target in computer vision for a long time. Large amount of monitoring/interaction applications in computer vision will benefit from an accurate and fast face tracking system. These include video conferencing, PUI (Perceptual User Interface), driver fatigue detection, eye-click pointer, human-robot interaction, video compression, lip reading and gaming control.

A wide range of computer vision algorithms have been applied to face tracking system. One category is skin color-based approaches. Probabilistic models in certain color space were built to distinguish the face area [1]. Color models are usually able to self-update. Color-based approaches are generally fast and light, but easily fail when other parts of the body or other skin color-like objects enter the camera view. They are also sensitive to lighting variation and the characteristics of the image input device.

Another category is shape-based approaches. They track the elliptic shape or contour of the face, such as elliptic template [2], deformable template [3] and active contour method [4]. They are not sensitive to background and lighting variation. However, their limitation lies in large angle out-plane head motion (tilt/pan), because the view of face shape and contour varies a lot.

The third category is related to motion-based approaches, such as the three-dimensional recursive search (3DRS) [5] and optical flow (OPFL) [6]. They calculate the velocity vector for a certain image area (pixel or block), and predict its location in the next frame in terms of the current location and the velocity.

In spite of great achievements made by previous research, robust face tracking remains as an elusive problem. One reason for this is that the face pattern greatly varies due to different head poses and facial expressions. Another reason is that most existing face tracking methods are not robust enough to deal with the presence of more than one face.

Our multi-clue face tracking approach aims at achieving both robustness and speed by integrating more than one tracker. Section 2 describes the three components and explains how they can be combined in a feedback loop. Design and results of thorough performance tests are discussed in section 3. Section 4 gives conclusion and several ideas for future improvement.

II. MULTI-CLUE FACE TRACKING SYSTEM

Each tracking algorithm has its advantages and limitations. However, due to different working principles, their performances complement each other in different situations. Aiming at overall robustness, our multi-clue face tracking approach integrates three components: a face detector and two object trackers. Effort has been put into two main designs: 1) Integration of two trackers: When the face cannot be reliably detected, the two face trackers work in parallel and provide two independent tracking results. We evaluate the reliability of the two results. Based on the evaluation, one single tracking face box (a rectangle on the face) is generated. 2) Interaction between detector and tracker: The face detector and integrated face tracker interact with each other by a in a feedback loop. When a face can be detected, the detector updates the trackers with the latest face position and size. When detector fails, the tracker keeps following the face area and therefore fills in the missing detections. When face can be detected again, the detector will use the tracking result to select the original face out of

978-0-7695-3571-5/09 $25.00 © 2009 IEEE
DOI 10.1109/GCIS.2009.207

all the face candidates. In this way, the system "locks" the face in a closed loop. When there is more than one face in the camera view, the system is able to keep following the same face (original face) until it goes out of the camera view.

The system structure is shown in figure 1. Detailed discussions are given in the following sub-sections.

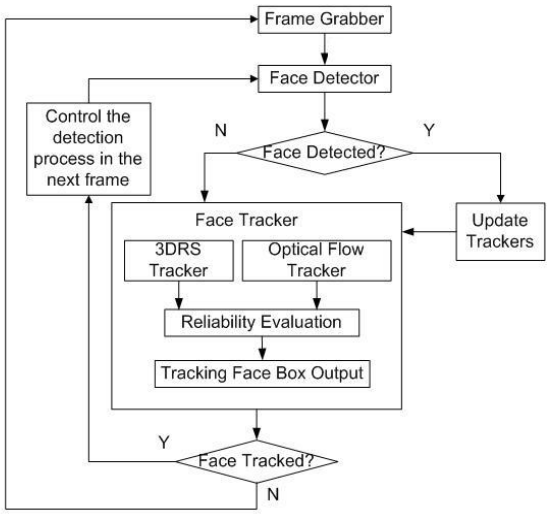

Figure 1. System Structure

A. Detector

The face detector provides fast and reliable detection on nearly frontal and upright faces. The core function of this detector is based on the Viola-Jones's algorithm [5], which has fairly robust and very fast performance.

In our face tracking system, the face detector enables automatic initiation and updates the states of the two trackers. Therefore, it is required to have low false positive rate. Wrong detection will lead the trackers to track the wrong object. Therefore, a grouping procedure is applied to all the detection candidate rectangles. Only if a group contains more than certain number of candidates, it is output as face candidate. Furthermore, a detection result identification process compares all the output face candidates with the current tracking result (details see section 2.5). In this way, false positive detections and background faces are effectively rejected. The detector only outputs the original face unless it is no longer in the frame.

B. Integration of two trackers

The choice of the tracker(s) was less simple. Several major object tracking algorithms have been tested. However, only few possess the necessary reliability and speed required for our application. The final choice was to combine the operation of two lightweight trackers, 3DRS motion estimation [6] and Optical Flow [7]. We found that the two trackers have complementary performance, i.e. when one fails or becomes not reliable under certain conditions; it is likely that the other one is still working correctly and precisely.

In the ideal situation, the two trackers follow the face together, thus generating largely overlapping tracking boxes. However, it often happens that the two tracking boxes diverge in position and size and sometimes even completely separate from each other. In order to provide a clue for evaluation, we record the center point location (x and y coordinate) of two tracking outputs in the latest 5 seconds. Furthermore, two measures have been defined:

a) Fast motion: Optical flow tracker has good performance on following fast motion, while 3DRS shows its main limitation. Therefore, we detect fast motion by comparing the velocity of the two tracking boxes in the last 5 seconds. When the following condition is true, we assume fast motion happened.

$$\frac{\sum_{i=2}^{N}|x_{3DRS}^{i}-x_{3DRS}^{i-1}|}{\sum_{i=2}^{N}|x_{OPFL}^{i}-x_{OPFL}^{i-1}|}+\frac{\sum_{i=2}^{N}|y_{3DRS}^{i}-y_{3DRS}^{i-1}|}{\sum_{i=2}^{N}|y_{3DRS}^{i}-y_{3DRS}^{i-1}|}\geq T \quad (1)$$

where $N=5\times framerate$, threshold T is set to 3 according to experiment, $(x_{3DRS}^{i},y_{3DRS}^{i})$ and $(x_{OPFL}^{i},y_{OPFL}^{i})$ are the center point of 3DRS and optical flow face box in the *i*th frame, respectively.

b) 3DRS active: The activity of the 3DRS tracker is measured by calculating the average velocity in the last 5 seconds. We assume the 3DRS is active when the following statement is true. Threshold T1 is set to 10% of frame width.

$$\frac{1}{N-1}\sum_{i=2}^{N}|x_{3DRS}^{i}-x_{3DRS}^{i-1}|\geq T_1 \text{ or } \frac{1}{N-1}\sum_{i=2}^{N}|y_{3DRS}^{i}-y_{3DRS}^{i-1}|\geq T_1 \quad (2)$$

The entire evaluation process is shown in figure 2. The two tracking results match when their overlap area is larger than certain percentage (T_a) of their total area. Based on the assumption that the remaining optical flow feature points are reliable enough to be on the face, optical flow output is given higher confidence weight when the two trackers do not match. When most of the points have failed, no fast motion has happened in the latest 5 second and 3DRS is still active, we assume the 3DRS tracking box provides a valid output. Otherwise, we assume that the original face is already out of the camera view.

The result of the evaluation process plays an important role. It determines the output strategy, which means how to integrate the two separated tracking results into one. As shown in figure 3, when the two trackers match, the average of two tracking boxes are output. When optical flow tracker is more reliable, the system outputs the center point of optical flow tracker and average size. This is because the size of the optical flow tracking box is usually smaller than the real face size. When 3DRS is more reliable, we still take the remaining optical flow feature points (if any) into account for center point calculation. When both trackers fail, there is no tracking output.

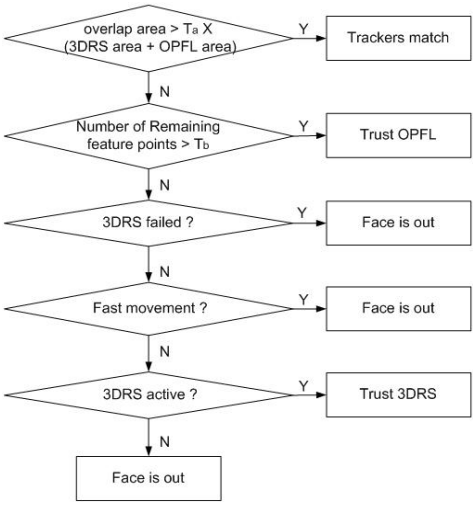

Figure 2. Evaluation process of two trackers' reliability

C. Detector-Tracker interaction

1) Feedback detector to tracker

Because the face detector is set to have very low false positive, we always use the detection result when a face can be detected. The detector updates the trackers with the last face position and size each time it finds a face. More specifically, it updates the position and size of the 3DRS tracking box and resets the optical feature points inside the face area.

2) Feedback tracker to detector

When the detector fails, the tracker keeps following the face area and therefore fills in the missing detections. When the face can be detected again, the detector will compare the face candidates with the tracking result. If the tracker indicates that the original face is not in the camera view, the detector outputs the face candidate with the biggest size. If the tracker is still tracking the original face, the detector only outputs the original face and updates the trackers. All the background faces and false positives will then be rejected. In this way, the integrated detector-tracker system "locks" the original face by a closed loop. The detection result identification process is described in figure 4. When two trackers match or at least one of them is reliable, the detector outputs the face candidate which has the largest overlap area with the tracking box. Otherwise it outputs the face candidate with the largest area as the original face.

III. EXPERIMENT

A. Setup

In order to shorten the development cycle and improve the effectiveness, the system implementation was based on the OpenCV library [8]. The library sources are highly optimized and suitable for real-time applications.

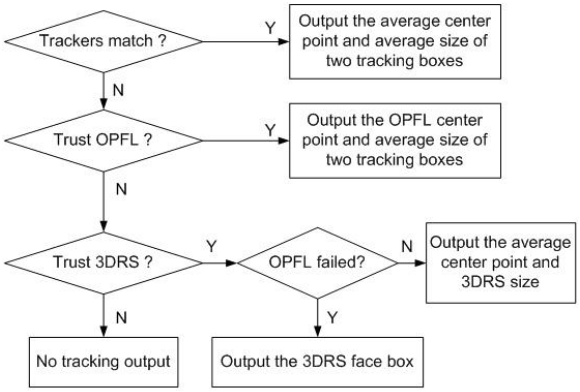

Figure 3. Tracking face box output strategy

The system runs on a normal PC equipped with Pentium IV 2.8GHz CPU and 1G Ram. A consumer webcam is used as image input device. Input frame rate is 30frame/s and image resolution is 320 × 240. The system is able to render about 15 frames per second.

B. Experiment design and measurements

Our target application is in the area of ergonomics for PC workers. Therefore, the experiments were carried out in an office environment. We set up three sessions to assess the system performance with respect to a) user behavior, b) background and lighting variation and c) presence of unwanted faces.

User behavior test aims at evaluating the performance on different face views and head movements. To this purpose, we recorded 9 sequences from subjects with different race, gender, age and hair style. We asked them to perform simple tasks in front of the monitor which involved 1) head rotation (pan, tilt and roll) 2) occlusion (hands or cups) 3) changing size (distance change between the user and the monitor) 4) fast motion (e.g. subjects leave their desk). The average number of frames for each sequence is around 1200. All sequences have the same background and lighting condition. Ground truth data of face location and size are manually marked in each frame.

We defined performance metric, dealing respectively with reliability of the tracked face box. With the reliability metric we want to evaluate the ability to find a face when it is in the camera view. Several indicators are considered including detection rate, tracking rate, locating rate, false positive rate and false negative rate. Detection rate measures the reliability of the face detector. False positive happens when the detector outputs non-face area and leads the tracker to track wrong object. The tracking result is deemed correct when the center point of the ground truth face box lies inside the tracking box. Tracking rate, defined as the percentage of correct results over the total amount of frames, indicates the reliability of the whole system.

Background and lighting variation test aims at measuring the reliability of our approach towards background complexity and lighting variation. Video clips were recorded

at 15 different locations with complex backgrounds and different lighting conditions. The system is also tested on a facial video database with time varying lighting [9]. Head movement is relatively simple and slow.

Multi-face test assesses the ability tracking the original face and rejecting the background faces. 4 video sequences with more than one faces were recorded for the multi-face testing. Each of them has the original face in every frame. Background face either comes in later or has a farther initial distance from the camera.

C. Results and Discussion

1) User behavior test

Three metrics are calculated for all the 9 test sequences on different user behaviors. Table 1 shows the average reliability result. Though the detector shows good performance on nearly-frontal face, the average detection rate is only 43.2%. Such low detection rates are mostly due to the diverse face views caused by head motion. However, there is no false positive in any of the 9 sequences, which means that the detector always initiates the tracker with the correct object. Since the tracker has good performance on tracking non-frontal faces, the (combined detector + tracker) average output tracking rate increases to 96.8%. Among the 9 sequence, the highest result reaches 99.4%. False negatives occur in all sequences when face moves into the camera view from outside. In these circumstances the detector can fail, either because it moves in too quickly and is then blurred, or the position of the head is non frontal.

TABLE 1: AVERAGE RELIABILITY MEASUREMENT RESULT OVER 9 SEQUENCES

Number of Frames	1227
Number of Frames with face	1139
Detection rate (%)	43.2
Tracking rate (%)	96.8
False positive rate (%)	0
False negative rate (%)	2.8

2) Background and lighting variation test

Complex background tends to increase the chance of false positives, as some luminance patterns can resemble actual faces. However, thanks to the conservative settings adopted for the Viola-Jones detector and the feedback between trackers and detector, the system was able to reject all possible incorrect outputs (see figure 4).

The reliability to different lighting intensity and direction is implicitly shown by the performance on the 15 video sequences with different backgrounds. Test on the time varying lighting video database further proves that our face tracking system is robust wide range of lighting variations (see figure 5). With relatively simple and slow head motion, the face tracking rate reaches 100% on all sequences in this test section. Face is correctly and precisely located in every frame.

Figure 4. Sample output of background test. False positive face candidates in the background are rejected with a diagonal

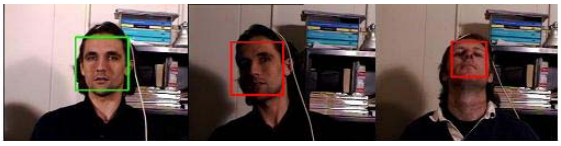

Figure 5. Sample output of time varying lighting test

3) Multi-face test

Figure 6 shows several sample snapshots of the multi-face test output. All faces candidates output by the viola-Jones algorithm are displayed. However, the detection result identification process (see chapter 2.3) selects the original face and rejects the background face. Tracker only follows the original face. In the 4 sequences, all the background faces are successfully discarded, even when the two face areas are partly overlapped or the background face is closer to the camera.

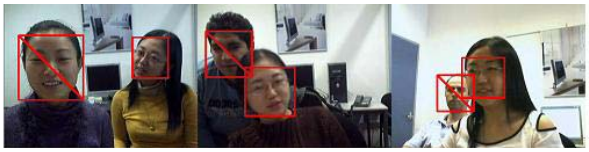

Figure 6. Sample output of multi-face test

IV. CONCLUSION AND FUTURE WORK

In this research we designed, implemented and tested a real-time multi-clue face tracking system. The three main components are a face detector, based on Viola-Jones algorithm, and two trackers, respectively based on 3DRS motion estimation and optical flow. The main goal was to achieve superior robustness by integrating the three blocks, taking advantage of the specific strength of each component. The detector, the only block specific to face detection, initializes the tracking system and continuously updates the trackers with the latest detection. Whenever detector fails, typically on non-frontal faces, the two trackers take over, both follow the face. The final output is then computed from the two tracking results on the basis of an internal reliability evaluation. Conversely, when the detector gives again a detection result, the current tracking output is used to select the original face among all detection candidates. The system has been tested in office environment. Three test sessions were carried out with focus on user behavior, background

and lighting variations, and presence of more faces, respectively. Under complex user behavior, the face tracking rate reaches 96.8%. Test of the robustness against complex background and varying lighting conditions also showed good performance.

Although this study was conducted in the context of ergonomics for office workers, it could be used for a number of other applications, including gaming interfaces, video conferencing and driver fatigue detection.

Future improvements can be considered on following aspects: 1) Better performance on false negatives, which mainly happen when a non-frontal face comes into the camera view. This limitation could be overcome by extending the training database of the face detector with multi-view faces. 2) Webcams with wider viewing angle could improve the system performance when the tracked face comes close to the monitor. In such situation, only part of the face would be visible in the camera view, making it impossible to detect it.

REFERENCES

[1] H.Stern and B.Efros, "Adaptive color space switching for tracking under varying illumination", Image and Vision Computing, 2005, 23(3):353–364.

[2] S.Birchfield. "Elliptical head tracking using intensity gradients and colorhistograms", In Proc. CVPR, 1998: 232–237.

[3] R.Kjeldsen and A.Aner, "Improving face tracking with 2d template warping", In Proc. IEEE Int'l Conference on Automatic Face and Gesture Recognition, 2000: 129–135.

[4] X.Bing, Y.Wei, and C.Charoensak, "Face contour tracking in video using active contour model", In Proc. International Conference on Image Processing, 2004, 4: 1021–1024.

[5] P.Viola and M.Jones, "Robust real-time object detection", International Journal of Computer Vision, 2002,1(2).

[6] T.Tang, J.Wang and Y.Liu, "Adaptive frame recovery based on motion activity", IEEE Workshop on Signal Processing Systems, 17-19 Oct. 2007: 692-697.

[7] C.K.Hsieh, S.H.Lai, and Y.C.Chen, "Expressional face image analysis with constrained optical flow", IEEE International Conference on Multimedia and Expo. 23-26 Jun. 2008: 1553-1556.

[8] G. Bradski. The OpenCV Library. Dr. Dobbs Journal of Software Tools, 2000.

[9] http://www.cs.bu.edu/groups/ivc/data.php.