Real Time Face and Hand Tracking With Correlation

Abstract

In this paper, we demonstrate an effective tracking system for moving face and hand objects in real time. Our approach contains of three sub-stages: adaptive background subtraction, face and hand detection using skin color and tracking them with pixels correlations respectively. There are famous robust tracking methods such as mean shift and active shape model in the literature. Despite providing robust object tracking, these methods have high computational burden. Taking into account the requirements of online processing, our approach can detect face and hands in image sequence using their skin color, geometrical shape information, and prior probabilities. The effectiveness of our approach comes from its simplicity. We demonstrate the performance of the proposed method by comparing it with the traditional tracking methods on real sequences.

1. Introduction

In computer vision, object tracking is the process of locating a moving object in consecutive video frames obtained using a camera. This basic problem has a wide range of applications from smart rooms to driver assistance and from perceptual user interface to augmented reality. Consequently, it has been an attractive field of investigation. There are numerous approaches proposed to tackle the problem of tracking objects in consecutive frames of a video. Mean shift based kernel tracking is a recently popularized method. The popularity of the mean shift is due its ease of implementation and robust tracking performance [1,2]. Despite its promising performance, it has two main limitations; its computational complexity is a major concern since this method relies on kernel density estimation of target and candidate objects. Alternatively, for tracking non-rigid shapes, the active shape model (ASM) utilizes a trained object model and attempts to find the best landmark translations between consecutive frames by allowing deformations from the mean shape [3]. The encoding ability of landmarks allowed tracking of non-rigid objects successfully using this approach including in medical imaging [4] and body pose estimation [5]. The original ASM, however, is not suitable for real-time tracking because of its high computational load and large number of iterations required for convergence [8].

For real-time deformable object tracking, computational complexity is of vital importance. In this article, we propose a real time face and hand detection method using skin color detection [6], adaptive background subtraction [7] and correlation tracking algorithm. The particular application we have in mind is medication adherence confirmation; therefore, it is assumed that the algorithm will be utilized in a relatively fixed location indoors with primarily artificial but non-varying lighting.

2. Proposed Method

As demonstrated in Figure 1, our proposed method consists of three sub-stages: adaptive background subtraction, face and hand detection using skin color (in foreground), and tracking them simply with cross-correlation maximization similar to matched filtering, respectively. In the following discussion, we provide details of the proposed method which is easily applicable to real time tracking situations due to its computational simplicity without losing robustness for the stated purpose.

2.1. Adaptive Background Subtraction

Background subtraction techniques can be classified into two broad categories [12]: non-recursive and recursive. In non-recursive background modeling techniques, a sliding-window approach is used for background estimation. A fixed number of frames is used and stored in a buffer. Recursive background modeling techniques do not keep track of a buffer containing a number of history frames. Instead, it recursively updates a single background model based on each incoming frame. In this article, we employ a recursive technique, since this approach generally requires less storage than non-recursive techniques [13]. Literature on recursive techniques is broad and propositions involving approximated median filter, Kalman filter, and mixture of Gaussians are found. Unlike Kalman filters that track the evolution of only a single Gaussian distribution model, the mixture of Gaussians method tracks multiple Gaussian distribution simultaneously for the color model of each pixel. In this article, we use a modified version of the usual pixel-level background subtraction method [7].

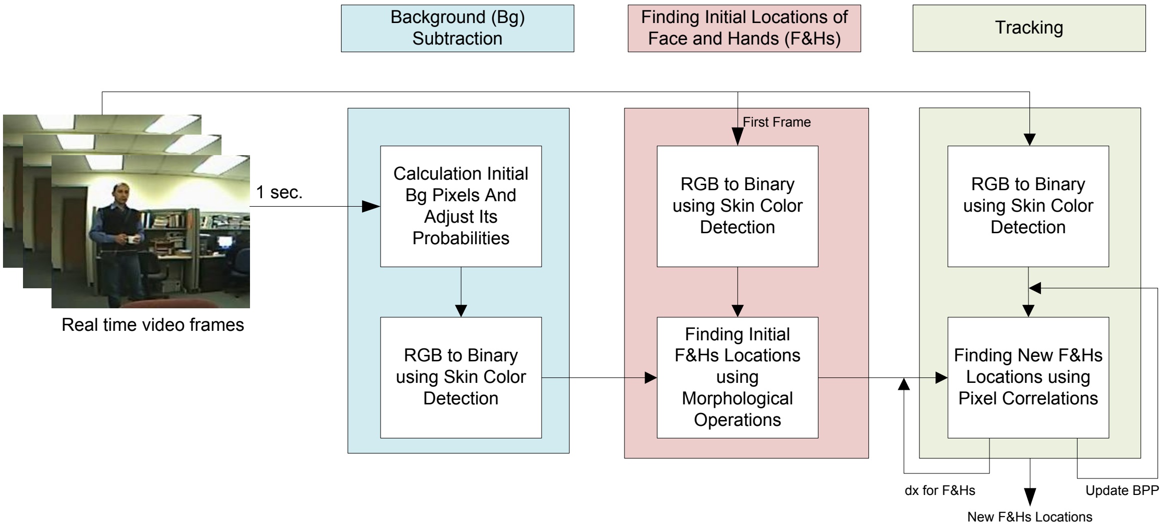
Background (Bg) Subtraction
Finding Initial Locations of Face and Hands (F&Hs)
Tracking
Real time video frames
1 sec.
Calculation Initial Bg Pixels And Adjust Its Probabilities
RGB to Binary using Skin Color Detection
First Frame
RGB to Binary using Skin Color Detection
Finding Initial F&Hs Locations using Morphological Operations
RGB to Binary using Skin Color Detection
Finding New F&Hs Locations using Pixel Correlations
dx for F&Hs
Update BPP
New F&Hs Locations

Figure 1 Overview diagram

This method consists of a likelihood classifier for each pixel given the background models for them. Specifically, it checks the current color of the pixel x^t in RGB coordinates at time t by:

$$p(x^t|BG) > c_{thr}$$

where c_{thr} is a preset threshold. In order to incorporate some level of robustness to lighting changes, the background models are typically designed as mixture models with B components for each pixel in image:

$$p(x|X_T, BG) \sim \sum_{m=1}^{B} \pi_m \eta(x, \mu_m, \sigma_m^2)$$

where μ_m and σ_m are respectively means and variances of GMM. The mixing weights denoted by π are non-negative. The GMM algorithm can select automatically the needed number of components per pixel and update π, μ and σ values by using a suitable recursive procedure. Hereby, undesired background information can be subtracted from the image leaving foreground objects for relatively static scenes.

2.2. Face and Hand Detection using Skin Color

After background subtraction, we can detect foreground objects in each frame (while the background object is still updated recursively). From the foreground segments, we need to extract face and hand locations (initialization in the first frame of real time video sequences and tracking in subsequent frames).

2.2.1. Skin Color Detection: We utilize skin color as a robust mask on the foreground region to extract visible uncovered body parts, including face and hands (for our target application of medication adherence it is presumed that these regions will be uncovered). In the literature, numerous techniques for skin color detection have been proposed. We use the method in [9] which can define explicitly (through a number of rules) the boundaries and the interior of the skin color distribution in RGB color space:

(R, G, B) is classified as skin if:
R > 95 & G > 40 & B > 20 &
Max{R, G, B} −min{R, G, B} > 15 &
|R−G| > 15 & R > G & R > B

The simplicity and robustness of this method have attracted many researchers [9, 10, 11].

2.2.2. Morphological operations to find locations of face and hands: After the skin color detection process, we are left with some binary objects which contain face and hands along with artifacts. To make a decision about which object is the face and which are hands, we measure a set of properties for each detected object. These properties are given as follows:

Area: The actual number of pixels in the object.

Centroid: The center position of mass of the object.

Diameter: The diameter of a circle with the same area as the object.

Convex Hull: The smallest convex polygon that can contain the object.

Solidity: The proportion of the pixels in the convex hull that are also in the object.

Eccentricity: The eccentricity of the ellipse that has the same second-moments as the object.

The discrimination between face and hands can be achieved by identifying the holes on the binary face object corresponding to the eyes and the mouth. Using these features, we can reliably classify each skin-color-foreground object into the face, hands, and artifact classes in the first frame. At initialization, the new detected face and hand objects are assigned a probability of zero (to be recursively updated in subsequent frames). If the same object is again determined as face or hand at subsequent frames, its probability is increased. This helps eliminate undesired artifacts detected alongside with face and hands.

2.3. Correlation Tracking

Tracking based on maximum cross-correlation is a simple technique that emerges from matched filter (or minimum Euclidean error-norm) approach to shift detection. In two consecutive frames containing the same object slightly shifted, the object's translation can be easily determined using correlation maximization. Given the binary skin-color objects on the previous frame and current frame, we can determine the matching objects by computing the cross-correlation of objects of interest from the previous frame with the current image and by selecting the nearest object that maximizes the correlation in the current frame. Spcifically, let the reference image be described by S and the segment of the current scene by I_{ij}. Then the correlation at each location (i,j) in the current frame is simply the dot product of two images, and is given by [14]: $C(i,j) = I_{i,j} \cdot S$

While correlation based tracking of objects in grayscale or color images is susceptible to many spurious local maxima, in our case, the binary images after background subtraction and skin color detection prevent such problems besides reducing the computational requirements of correlation calculation drastically. This process iteratively continues through consecutive frames.

3. Experimental Result

The proposed algorithm has been tested in various images sequences and several experimental results are shown in this paper to compare the performance with other methods. In Figure 2, tracking results of the proposed method for the Walking sequence which has 200 frames of size 320x240 pixel^2 are displayed. Small red circles at the first frame at Figure (2) shows the results of the initial face and hand detection and classification using the process described above. Big yellow circles in the image demonstrate designated search areas to track the detected face and hand objects in the next frame. As seen on other frames in Figure (2), despite the very small sizes and rapid displacements of the hands, the proposed approach is able to detect and track accurately throughout the walking scene. Table 1 shows elapsed times of both the proposed method and competing tracking methods. As seen in this table, the proposed method can obtain desired results not only robustly but also faster.

Table 1 Comparison of Tracking Algorithms

Walking sequence (200 Frame)	*Elapsed Time (sec.)*
PDF based MS	24.34353
Histogram based MS	8.361277
Active Shape Method	32.23432
Proposed method	5.768125

We also utilized the proposed tracking method in a setting that simulates medication taking by a person while sitting. Subjects are asked to repeatedly eat candy out of a pillbox. A classifier that detects the hand-face merging event is implemented to run in parallel with the proposed tracking algorithm and the frames in which the subjects ate the candy were successfully determined. In Figure 3, the labels fp and hp indicate the probability that the corresponding object is the face or is a hand (recursively incremented as each object is tracked through consecutive frames). In both scenarios, the algorithm can process about 30 frames per second on a PC with 1GB RAM and 3.2GHz processor speed, operating in Matlab (implemented efficiently using matrix-vector elementwise operators).

4. Conclusion

We proposed a simple yet robust real time face and hand tracking scheme using adaptive background subtraction, skin color detection and correlation based tracking. The performance of the proposition is verified in various tracking scenarios and results are shown here for two applications.

5. References

[1] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, 2002, pp. 603-619.

[2] Yilmaz, A., "Object Tracking by Asymmetric Kernel Mean Shift with Automatic Scale and Orientation Selection". Computer Vision and Pattern Recognition, IEEE Conference, 2007, pp.1-6.

[3] Cootes, T.J., Taylor, C.J., Cooper, D.H., and Gragam, J.: "Training models of shape form sets of examples". British Machine Vision Conf.,1992,pp. 9–18

[4] Q. Xu, R. Hamilton, R.Schowengerdt, S. Jiang. "A deformable lung tumor tracking method in fluoroscopic video using active shape models: a feasibility study". Phys. Med. Biol, 2007, pp. 5277–5293.

[5] J.F. Vasconcelos, R. Cunha, C. Silvestre and P. Oliveira. "Landmark Based Nonlinear Observer for Rigid Body Attitude and Position Estimation". Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, USA, 2007, pp.12-14.

[6] V. Vezhnevets, V. Sazonov and A. Andreeva. "A Survey on Pixel-Based Skin Color Detection Techniques". In Proc. Graphicon, 2003, pp. 85-92.

[7] Z. Zirkovic, F. V. Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction",Pattern recognition letters, 2005, Elsevier.

[8] D. Seghers, P. Slagmolen, Y. Lambelin, J. Hermans, D. Loeckx, F. Maes, and P. Suetens. "Landmark based liver segmentation using local shape and local intensity models", A Grand Challenge, 2007, pp. 135-142.

[9] P. Peer, J. Kovac, and F. Solina. " Human skin colour clustering for face detection". In submitted to EUROCON 2003, International Conference on Computer as a Tool.

[10] J. Ahlberg, "A system for face localization and facial feature extraction". Tech. Rep. LiTH-ISY-R-2172, Linkoping University.1999.

[11] L. Jordao, M. Perrone, J. Costeria, and J. Santos-Victor. "Active face and feature tracking". In Proceedings of the 10th International Conference on Image Analysis and Processing, 1999, pp. 572–577.

[12] S-C.S. Cheung, C. Kamath. "Robust Techniques For Background Subtraction In Urban Traffic Video", In Proceedings of the SPIE, Volume 5308, 2004, pp. 881–892.

[13] M. Smith, "Background Subtraction for Urban Traffic Monitoring using Webcams", Master Thesis, University of Amsterdam, 2006.

[14] S. Wong, "Advanced Correlation Tracking of Objects in Cluttered Imagery", The Proceedings of SPIE: Acquisition, Tracking and Pointing XIX, Orlando USA, 2005.

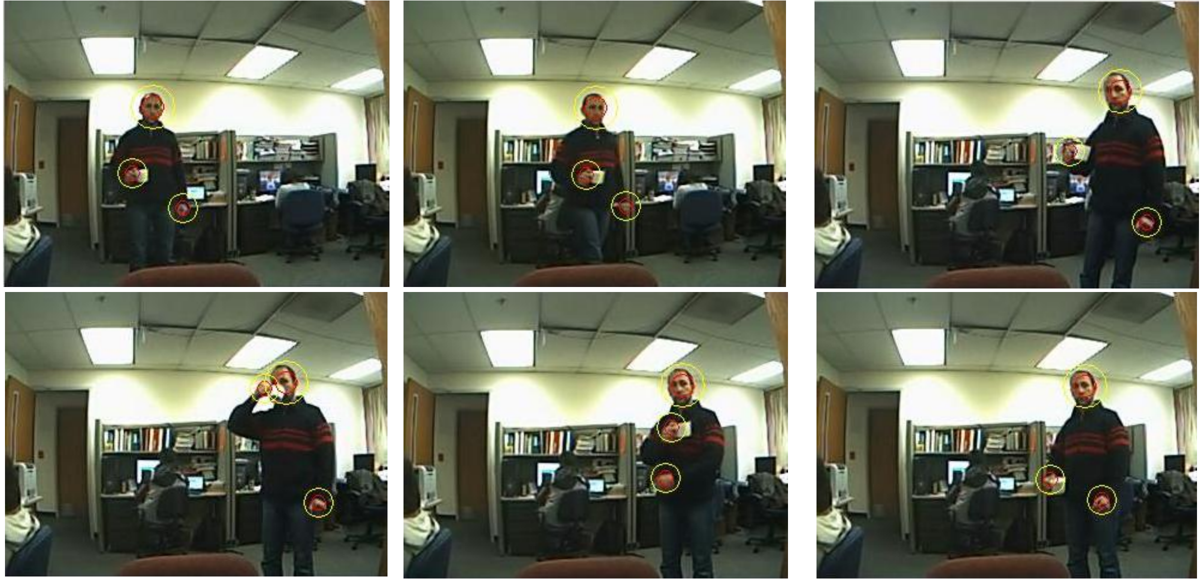

Figure 2 Tracking results of the Walking sequence. Frames 1, 40, 122, 139, 157 and 200 are displayed.

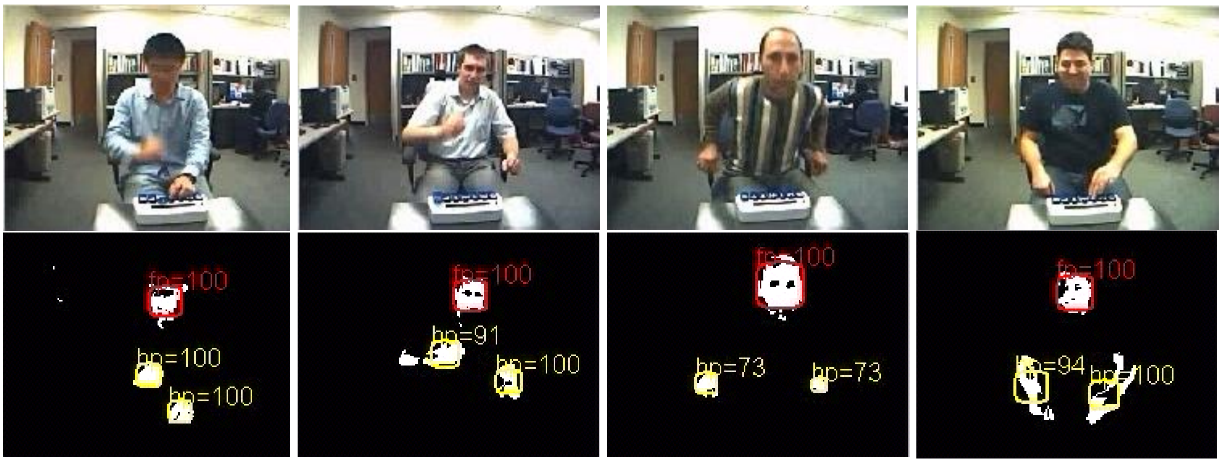

Figure 3 Tracking results of our approach at the EatCandy application which has been designed for a specific purpose are displayed.