

Combined Face-body Tracking in Indoor Environment

Xuefeng Song Ram Nevatia
University of Southern California
Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089-0273
xsong@usc.edu nevatia@usc.edu

Abstract

Background subtraction is commonly used for tracking objects in outdoor environment. But it doesn't work that well indoors, because of the problems caused by illumination change, shadows, occlusion and targets' changing appearances. In contrast, color tracking is relatively resistant to these problems, but suffers from the need of initialization. To complete the specific human tracking task in indoor environment, this paper utilizes the specific human face-body structure, and tracks face and body simultaneously and cooperatively. The advantage of this approach is that it can keep tracking in some bad situations, when one of the parts is missing, which makes it more robust than single-part tracking. Experimental tracking results on a meeting room video data are given.

1. Introduction

Human action analysis is one of the goals of computer vision. One crucial part of this task is tracking human through image sequences. Background subtraction [4] [5] [6] is a common method for discriminating a moving object from the static scene. However, it doesn't work indoors as well as outdoors. The problems are caused mostly by 4 factors. (1) Quick and big illumination changes: usually illumination changes in outdoor scene happen gradually along with the change of the sun's position or the cloud motion. But indoors, reflection and lighting changes happen swiftly, which create big blobs after background subtraction. (2) Appearance change: human appearances of indoor videos are more complex due to the proximity of the objects to the camera. (3) Occlusion: occurs more in indoor scenes, due to the presence of more scene objects and smaller space. (4) Unpredictable moving objects: The outdoor outlier moving objects are mostly vehicles or swinging branches. This problem becomes more complicated indoors as there are many kinds of indoor objects, such as doors, chairs, books, which all could move.

Compared to background subtraction based tracking methods, color tracking has some advantages. It is invariant to shape changes, and is not affected by shadow

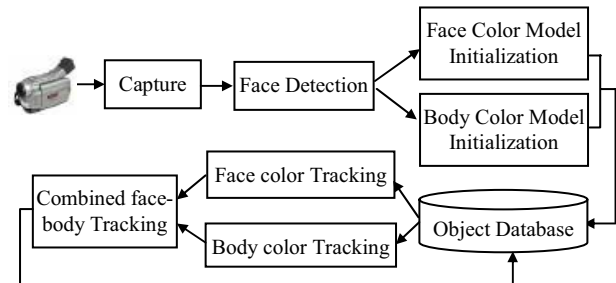


Figure 1: The flow chart of indoor human tracking

and the motion of other objects. Camshift [1] and Meanshift [2] are two successful color-tracking methods. Both of them search for the local optimal candidate, with a reasonable assumption that objects move continuously. However, they both have the shortcoming that they cannot initialize the motion regions. The Adaboost face detection method by Viola and Jones [3] is a good candidate for this purpose. It detects the frontal face in real time, with high detection rate and low false alarms.

Finally, humans have a clear body structure: head, upper body and legs, from top to bottom. By combing different evidence in a naïve Bayesian structure, this paper proposes a combined face-body tracking method, which tracks the face and body separately and cooperatively. It is robust in that it can keep tracking even when the evidence of one part is weak or missing. As shown in Fig.1, the proposed tracking method can be summarized in 3 parts: (1) Frontal face detection, which triggers the tracking of a new target. (2) Initialization of face and body color models. (3) Face-body tracking.

The paper is organized as follows. Section 2 reviews the Adaboost face detection method and the Camshift tracking methods. Section 3 elaborates the proposed body extraction method and the face-body tracking method. The experimental results are presented at Section 4.

2. Adaboost and Camshift

2.1. Adaboost Face Detection

Standard Adaboost algorithm learns a strong classifier $H_{ada}(x)$ by combining a set of weak 0/1 classifiers $\{h_i(x)\}$ using a set of weights $\{\alpha_i\}$:

$$H_{ada}(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^N \alpha_i \cdot h_i(x) \geq \frac{1}{2} \sum_{i=1}^N \alpha_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The selection of features and weights is learnt through supervised training. For face detection, the variant x is a square sub-window of the image. In Viola and Jones [3], the features are selected from four types of features (see Fig.2-a). The value is the difference between the sum of intensities in white regions and that in black regions. A feature of any type could be at different positions and of different sizes. So the total number of features is huge (180,000 for a 24x24 image patch). Adaboost training selects about 6,000 discriminating features for the face detection task. Fig.2-b shows the first two features.

To make the process faster, a cascade of classifiers is applied. The cascade can discard many easily determined non-face regions quickly, and focus the attention on the interesting regions. This makes the approach have a performance of 0.1 seconds for a 320x240 image on a current technology PC.

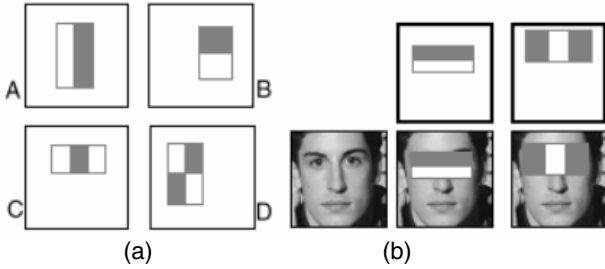


Figure 2: (a) Example rectangle features. (b) The first two feature selected by Adaboost training.

2.2 Camshift Color Tracking Method

Camshift [1] is an efficient color tracking method, which climbs the probability density gradients to find the local peak of the distribution. The first step is to compute the probability density image $P(x, y)$. Based on it, the search procedure could be described as 4 steps:

1. Set the search window at the location in the previous frame (\bar{x}_0, \bar{y}_0) .
2. Calculate the mean location in the search window by
$$\bar{x}_1 = \frac{\sum_{i=1}^n x_i \cdot P(x_i, y_i)}{\sum_{i=1}^n P(x_i, y_i)}, \bar{y}_1 = \frac{\sum_{i=1}^n y_i \cdot P(x_i, y_i)}{\sum_{i=1}^n P(x_i, y_i)}$$
3. Center the search window at the mean location computed in Step2, $\bar{x}_0 = \bar{x}_1, \bar{y}_0 = \bar{y}_1$.
4. Repeat Steps 2 and 3 until convergence

How to calculate the probability density image is an open part. Instead of the hue color model used in [1], we create a 32x32x32 RGB color histogram model for both the tracking target and the background. For each pixel, the probability of its belonging to a target is evaluated in this way:

$$P(\text{target} | \text{pixel}) = \frac{P(\text{pixel} | \text{target}) \cdot P(\text{target})}{P(\text{pixel} | \text{target}) \cdot P(\text{target}) + P(\text{pixel} | \text{bg}) \cdot P(\text{bg})} \quad (2)$$

where “bg” means “background”, $P(\text{target}) + P(\text{bg}) = 1.0$, and $P(\text{target})$ is set as the ratio of the target’s area to the area of the rest of the image. $P(\text{pixel} | \text{target})$ and $P(\text{pixel} | \text{bg})$ can be derived directly from the color histograms of the target and the background. Compared to the computing probability method from only the target color model, this target-background color models put higher weights on the more distinguished pixels. The target and background color models could be and should be updated from time to time, but this topic is not addressed in this paper.

Fig. 3 shows an example. On the probability image, the left person (in black) and the middle person (in yellow) are very obvious, while the right person (in white) is not clear since white is very similar to the background.



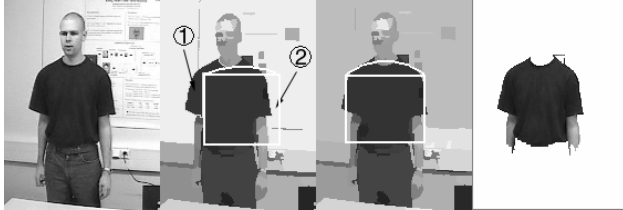
Figure 3: Target detection. (a) Original image. (b) Image intensity proportional to probability of the targets.

3. Upper Body Extraction and Face-body Tracking

3.1 Upper Body Extraction

In our tracking approach, both the face and upper body are tracked. Face detection part initializes the color model of the face. Intuitively, the body is below the face, and of a reasonable size. To get an accurate body patch, we model human body as a rectangle with an arch on top, which is described as a 5-parameter vector $(x, y, \text{width}, \text{height}, \text{arch_height})$. Then we segment the image into regions by Meanshift Segmentation method [8] and search for the regions belonging to body. For a specific body model (see Fig. 4-b), a region belongs to body if half of its pixels are inside this model. The match error of this model is the number of outlier pixels. There are two types of outliers. One type (labeled (1) in Fig.4-b)

consists of the pixels that belong to body regions, but are outside the model. The other (labeled (2) in Fig.4-b) consists of the pixels that are inside the model, but don't belong to any body regions. The best model, which has the minimal error, is searched for in the 5D parameter space. Small regions could be wrongly included, but the consequence is trivial because it only involves little number of pixels. Fig.4 shows an example.



(a) sample (b) bad match (c) good match (d) body

Figure 4: Illustration of human body analysis. Original image (a) is segmented into (b)(c). A body model, a rectangle with an arch on top, is matched with the regions. Finally, body region is extracted at (d).

3.2 Face-body Tracking

Fig.5 shows a generic naïve Bayesian structure that describes human with some quantitative measurements of face, body and the positional relation between face and body. $X_f = (x, y)$ and $X_b = (x, y)$ are the positions of the face and body rectangles¹. In the context of color tracking, $\rho(X_f)$ and $\rho(X_b)$ measure the similarity of a candidate rectangle to the original face or body color model. $(X_f - X_b)$ measures the relation between the positions of face and body. Apply the Bayesian rule to get:

$$P(H | \rho(X_f), \rho(X_b), X_f - X_b) \propto P(\rho(X_f) | H) \cdot P(\rho(X_b) | H) \cdot P(X_f - X_b | H) \cdot P(H) \quad (3)$$

We analyze the components on the right side of the equation (3) from right to left. First, $P(H)$ can be treated as a constant. Second, We assume that face is above the body, and they are not far away from each other in both X and Y directions. We present this face-body structure with the product of two Gaussian functions below.

$$P((X_f - X_b) | H) = G(\mu_x, \sigma_x, \frac{x_f - x_b}{width}) \cdot G(\mu_y, \sigma_y, \frac{y_f - y_b}{height}) \quad (4)$$

where the distances are normalized by window size, (μ_x, σ_x) and (μ_y, σ_y) are learnt offline from some samples. Finally, function $\rho(X)$ is within the range [0,1]. We propose an intuitive way to represent the probability $P(\rho(X_f) | H)$ and $P(\rho(X_b) | H)$ linearly.

¹ The size of the tracking window is fixed by initial face detection.

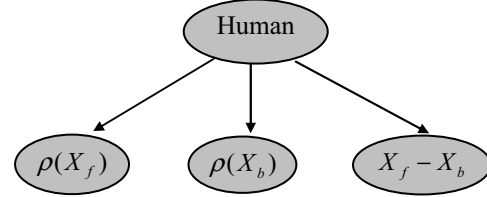


Figure 5: Naïve Bayesian structure of human representation by face and body

$$\begin{cases} P(\rho(X_f) | H) = \frac{2}{1+\alpha} (\alpha + (1-\alpha) \cdot \rho(X_f)) \\ P(\rho(X_b) | H) = \frac{2}{1+\beta} (\beta + (1-\beta) \cdot \rho(X_b)) \end{cases}, 0 \leq \alpha, \beta \leq 1 \quad (5)$$

where α, β represent the probability $P(\rho(X)=0 | H)$ for face and body separately, $2/(1+\alpha)$ is a normalization factor to keep $\int P(\rho(X_f) | H) = 1$. Finally, tracking is to search for the local maximum:

$$(X_{face}^*, X_{body}^*) = \arg \max (\alpha + (1-\alpha) \cdot \rho(X_f) \cdot (\beta + (1-\beta) \cdot \rho(X_b)) \cdot G(\mu_x, \sigma_x, \frac{x_f - x_b}{w_f}) \cdot G(\mu_y, \sigma_y, \frac{y_f - y_b}{h_f}) \quad (6)$$

Instead of a time-consuming search in the 4D space (x_f, y_f, x_b, y_b) , we first track face and body separately by applying Camshift, which maximizes $\rho(X)$. Then based on the best face candidate (x_f^1, y_f^1) and following the criterion of Equation 6, we search the corresponding body (x_b^1, y_b^1) in a local region. We repeat starting from the best body position (x_b^2, y_b^2) and get (x_f^2, y_f^2) . Finally, we pick the winner between $(x_f^1, y_f^1, x_b^1, y_b^1)$ and $(x_f^2, y_f^2, x_b^2, y_b^2)$.

The advantage of this method is that when the evidence of one part (for example, face) is not strong, instead of drifting away, the other part (body) can hold it to a proper position. Then the missing part has the chance to recover later. Experiments show this approach is more robust than either the face-only or the body-only tracking.

4. Experiment

Our approach is tested on the two ICVS-PETS-03 meeting videos (see Fig. 6). The participants are labeled as Person-1 to Person-6 separately. From Table 1, we can see that the detection rate of Person-1 jumps dramatically from 62.49% to 97.46%, and the detection rate of Person-

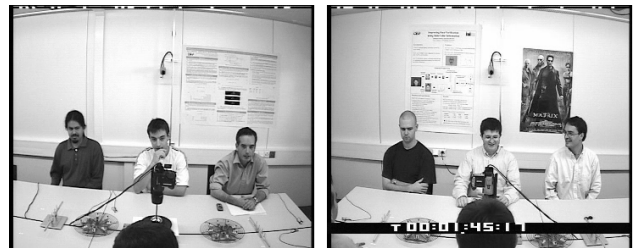


Figure 6: ICVS-PETS-03 meeting videos

3 jumps from 73.95% to 89.35%, when comparing face only tracking to face-body tracking. For other people, the detection rates don't increase much because the face evidence is strong enough to finish most of the tracking on its own. In addition, since we use frontal face detection for initialization, no person can be tracked until his/her face becomes frontal. This causes more than half of the missing detections. The other missing detections are caused by partial or full occlusions.

Fig.7 gives a few examples, showing the cases where the face-only tracker fails, but face-body tracker succeeds. In case-1 and case-2, face tracker can't track the left person because the faces are mostly or totally unseen. But taking the advantage of clearly observable body, face-body tracker can still track him successfully. In case-3 and case-4, the same thing happens on the person in the middle.

On an ordinary Pentium III 1.0G HZ computer, we achieved 3.3frames/second detection speed for 360x288 image sequence.

Table 1: Detection Rates Comparison

ID	Show Time (frames)	Face Tracking		Face-Body Tracking	
		Detected	Percent	Detected	Percent
1	7678	4798	62.49%	7483	97.46%
2	9260	8822	95.27%	8888	95.98%
3	8308	6144	73.95%	7424	89.35%
4	8176	7984	97.65%	7988	97.70%
5	6482	6280	96.88%	6360	98.12%
6	6590	6378	96.78%	6454	97.94%

5. Conclusion

This paper focuses on simultaneous tracking of multiple parts, i.e. face and body for human tracking. Experiments show it could enhance the robustness of the tracking system. Two other important issues, illumination change and occlusion, are not mentioned here. How to deal with them will be part of our future work.

Acknowledgements

This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract MDA-904-03-C1786.

References

[1] G. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, vol. 2nd Quarter, 1998

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift", *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 142-149, 2000

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *IEEE Conf. on Computer Vision and Pattern Recognition 2001*, vol. 1, pp. 511-518, 2001

[4] T. Horprasert, D. Harwood, and L. Davis, "A statistical approach for real-time robust background sub. and shadow detection", *Proc. of IEEE ICCV Frame-rate Workshop*, 1999

[5] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body". *IEEE trans. on PAMI*, Vol 19, 1998

[6] T. Zhao, R. Nevatia, F. Lv, "Segmentation and tracking of multiple humans in complex situations", *IEEE Conf. on Computer Vision and Pattern Recognition*, 2001

[7] Yoav Freund and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". *Computational Learning Theory: Eurocolt '95*, pages 23-37. Springer-Verlag, 1995

[8] D. Comaniciu, P. Meer: "Mean shift: A robust approach toward feature space analysis." *IEEE Trans. Pattern Anal. Machine Intell.*, 24, 603-619, May 2002

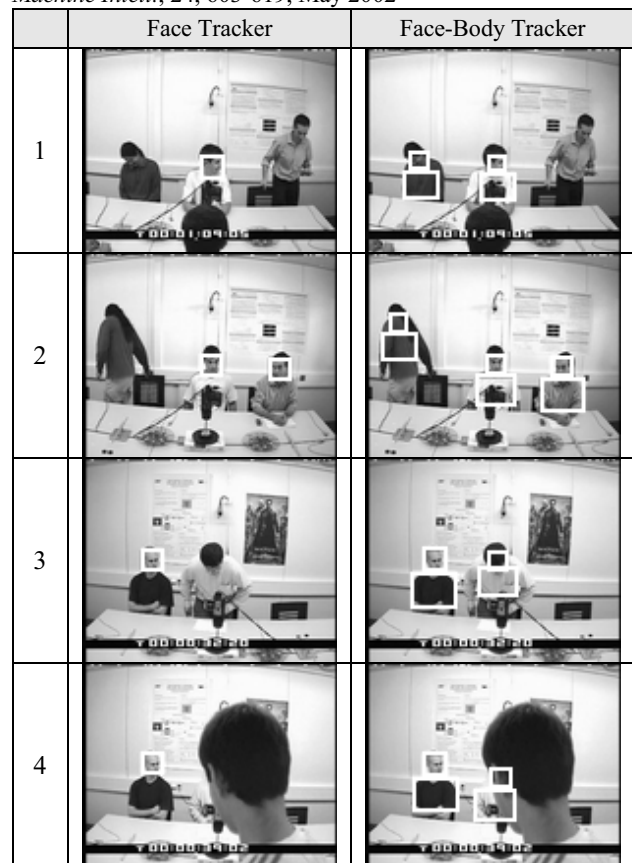


Figure 7: Cases when face-body tracker exceeds face tracker