ELectronics EXpress

IEICE Electronics Express, Vol.7, No.15, 1106–1111

A novel method of 2D articulated body tracking under self-occlusion and ambiguity

Kittiya Khongkraphan[1] **and Pakorn Kaewtrakulpong**[2,a)]

[1] *Department of Computer Engineering*

[2] *Department of Control System and Instrumentation Engineering,*
King Mongkut's University of Technology Thonburi,
126 Prachautid, Bangmod, Toongkru, Bangkok, Thailand, 10140

a) *pakorn.kae@kmutt.ac.th*

Abstract: A novel masker-less based method to track a 2D articulated body under self-occlusion and ambiguity in monocular image sequences is proposed. The proposed method applies SMC to both color and motion features. We employ a self-updated binary occlusion mask to increase accuracy in tracking. To alleviate the effect of illumination, each color body part is formed by a Gaussian mixture model in HS space, and the distribution intersection is used in distance measures of two probability distributions. Moreover, a motion cue is used to prune spurious solutions. Our technique can track the target reliably, especially in occlusion and ambiguous cases.

Keywords: tracking, occlusion, ambiguity

Classification: Science and engineering for electronics

References

[1] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," *Proc. CVPR*, pp. 126–133, 2000.

[2] W. Qu and D. Schonfeld, "Real-time decentralized articulated motion analysis and objecting from videos," *IEEE Trans. Image Process.*, pp. 2129–2138, 2007.

[3] A. Gritai, A. Basharat, and M. Shah, "Geometric constraints on 2D action models for tracking human body," *Proc. ICPR*, 2008.

[4] D. Ramanan, D. Forsyth, and A. Zisserman, "Strike a pose: Tracking people by finding stylized poses," *Proc. CVPR*, pp. 271–278, 2005.

[5] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Distributed occlusion reasoning for tracking with nonparametric belief propagation," *Proc. NIPS*, 2004.

[6] G. Hua, M. H. Yang, and Y. Wu, "Learning to estimate human pose with data driven belief propagation," *Proc. CVPR*, pp. 747–754, 2005.

[7] K. Khongkraphan and P. Kaewtrakulpong, "Robust contour tracking in cluttered background using snake on weighted-gradient image," *Proc. IWAIT*, 2007.

[8] C. Ciaran, O. E. Noel, and S. F. Alan, "Detector adaptation by maximizing agreement between independent data sources," *Proc. IEEE In-*

DOI: 10.1587/elex.7.1106
Received March 29, 2010
Accepted July 02, 2010
Published August 10, 2010

ternational Workshop on Object Tracking and Classification Beyond the Visible Spectrum, 2007.

[9] S. J. McKenna, Y. Raja, and S. Gong, "Tracking color objects using adaptive mixture models," *Image and vision computing*, pp. 225–231, 1999.

[10] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

1 Introduction

Tracking human body parts or poses has attracted the attention of many computer vision researchers. A number of algorithms have been proposed to address human body pose tracking, and one of the initial approaches was based on image-based approach. These approaches often fail under self-occlusion, where multiple human body parts occupy the same region of the input image. Moreover, several approaches generally suffer incorrect tracking due to ambiguity problems, since some symmetric body parts such as arms and legs normally have a similar appearance. Solving the self-occlusion and ambiguity problems is the focus of our paper.

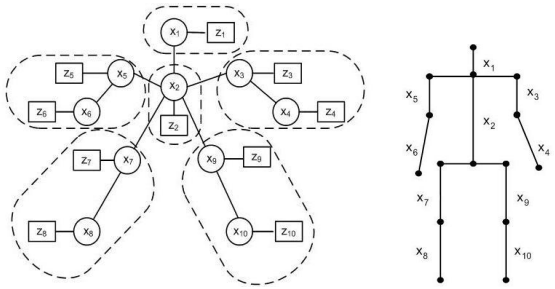
1.1 Previous work

To deal with the self-occlusion and ambiguity problems, a number of approaches utilize multiple cameras [1]. The main drawback of these approaches is the camera system setup. Several approaches focus on using strong prior knowledge of body poses; however, these are limited to normal poses in known activities such as walking [2] or aerobics [3]. Some approaches apply a binary occlusion mask to resolve the self-occlusion problem [4, 5]. Such approaches require an assumption that the order of the limbs is known a priori. Some approaches use detected part candidates to help estimate postures approximately, even under self-occlusion [6]. Alternatively, some researches [2] form a human model using a collection of trees instead of a single tree [1]. The configuration of each tree can be obtained sequentially. Like the binary occlusion mask approach, these approaches require knowledge of limb order.

2 Proposed method

The main concept of our approach is to apply a binary occlusion mask and to decompose a human model into several subtrees to resolve the self-occlusion problem. Additionally, we combine both spatial and temporal information in the observation computation to deal with ambiguity problems. Firstly, an order of human body parts is determined for hierarchical tracking. Then the human body parts in the subtree closest to the camera are evaluated. The next step involves a binary occlusion mask being generated from that solution and then used in subsequent tracking of the other body parts according to their orders.

DOI: 10.1587/elex.7.1106
Received March 29, 2010
Accepted July 02, 2010
Published August 10, 2010



(a) (b)

Fig. 1. (a) The collection of subtrees of an articulated human body, (b) human body parts represented in the model.

2.1 Human body model

Similar to [2], we model a 2D view of a human body as a collection of M subtrees, which is denoted by $\mathbf{T} = \{T_k | k = 1, 2, \ldots, M\}$, where T_k is the k^{th} subtree. Each tree consists of hidden nodes and corresponding potentials, as shown by the circles and rectangles, respectively, in Fig. 1 (a). The human body parts represented by the nodes in the model are shown in Fig. 1 (b). The hidden nodes are represented by $X = \{\mathbf{x}_i | i = 1, 2, \ldots, N\}$, where $\mathbf{x}_i$ is the body part. Each part consists of five states: (x, y) position, orientation, width and length. A corresponding observation set is denoted by $Z = \{\mathbf{z}_i | i = 1, 2, \ldots, N\}$, where $\mathbf{z}_i$ is the image observation node for the i^{th} body part. The relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is represented by the observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$.

2.2 Human body order determination

We base our method on the assumption that orientation of the human face signifies the order of human body parts. To obtain the human facing, the human head is first detected [7] and then the facial skin is extracted from the head region [8]. A line separating the head region into two symmetrical parts is formed. The facing is then assigned to the side that has the maximum percentage of detected facial areas. In case of non-overlapping region, the facing is set to that obtained in the previous frame. A frame of a walking sequence and detected skin regions are displayed in Fig. 2 (a) and (b), respectively. The head contour and separating line are shown in Fig. 2 (c) in purple and black, respectively. The overlapping region between head and skin is shown in green in Fig. 2 (c).

3 Tracking

The head position is obtained from Section 2.2, whilst the remaining parts of the tree (torso, left arm, right arm, left leg and right leg) are individually tracked based on their orders. It is based on hierarchical tracking, and we apply a binary occlusion mask to alleviate detection of spurious features. Our approach uses the Sequential Monte Carlo (SMC) method [10] to track the human body. Each tree of the human model is considered a state space in

DOI: 10.1587/elex.7.1106
Received March 29, 2010
Accepted July 02, 2010
Published August 10, 2010

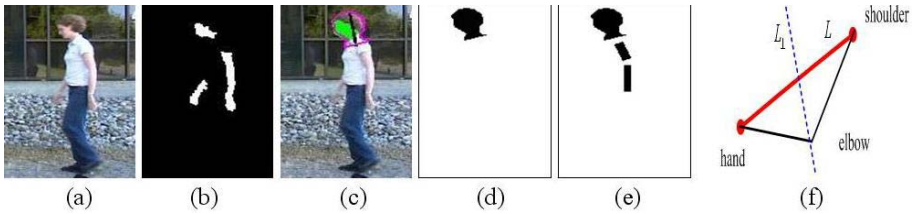


Fig. 2. (a) input image, (b) human skin, (c) human facing, (d) binary occlusion mask before left arm tracking, (e) binary occlusion mask after left arm tracking, (f) kinematics constraint.

SMC. The key idea of SMC is the approximation of the posterior density function $p(\mathbf{s}|\mathbf{z})$ by a set of samples (or particles) and correlation weight pair, $\{\mathbf{s}^i, w^i\}$; $i = 1, 2, \ldots, N$, where $\mathbf{s}^i$ is the i^{th} sample of the state vector, w^i its corresponding weight, $\mathbf{z}$ the observation vector and N the number of samples. The probability of each sample is computed from

$$w^i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(1-w^i_{c,m})^2}{2\sigma^2}}. \tag{1}$$

where w^i is the corresponding weight of the i^{th} sample, σ the standard deviation and $w^i_{c,m}$ the observation value obtained by combining color and motion distance measures. It can be computed from

$$w^i_{c,m} = \alpha w^i_c + (1-\alpha) w^i_m. \tag{2}$$

where w^i_c, w^i_m are color appearance and motion scores, respectively. (The scores are normalized to be between 0 and 1.) α is the weight of blending between color and motion information. The sample with maximum weight is selected to be the solution

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}^i} p(\mathbf{x}|\mathbf{z}) \approx \arg\max_{\mathbf{s}_i} w^i; i = 1, 2, 3, \ldots, N. \tag{3}$$

where $\hat{\mathbf{s}}$ is the state estimate of the object, $\mathbf{s}^i$ the sample of the state vector and w^i the corresponding weight. After tracking, the binary occlusion mask is updated for tracking subsequent body parts. The binary occlusion masks of before and after tracking the left arm are shown in Fig. 2 (d) and (e), respectively.

3.1 Sample generation

The concept of sample generation is to create appropriate candidates for each joint position and evaluating them according to body-part model to obtain the best solution. The samples are generated based on body-part ratio of Leonardo Da Vinci's anatomy concept, prior knowledge of potential positions on thinning lines of detected body-part silhouette and of detected skin regions, and estimated body-part order.

Starting from neck and shoulder joints, an upside-down-T shape is fitted on the head-separated line (see Fig. 2 (c)) and samples are generated around

DOI: 10.1587/elex.7.1106
Received March 29, 2010
Accepted July 02, 2010
Published August 10, 2010

its end and intersection points. Position of the hand closer to camera is estimated from potential end positions of the remaining thinning lines. For elbow, samples are limited by a kinematics constraint on the line $L_{\perp}$, as depicted in Fig. 2 (f). Its samples are generated around $L_{\perp}$. Another upside-down T-shape is fixed using the neck joint and the estimated torso location. Similarly, the samples of hip and abdomen are generated. Foot and knee positions are generated similarly to those of the arm.

4 Observation function

Generally, the observation function is used to measure the observation likelihood of the samples. A color feature is widely used in human tracking thank to their robustness to rotation in depth, shape changing and scale changing. However, it suffers color changing over time due to changes in scene illumination and visual angle. Another major problem is ambiguity due to the similarity of limb features. To alleviate the above problems, we integrate spatial and temporal information into the observation computation for color and motion features. For the color feature, we use hue (H) and saturation (S) components of the HSI space [9]. Lightness (I) is not used in the observation function so that it is robust against global illumination changes. Each human body is formed by a Gaussian mixture model $m(\mathbf{s})$ based on HS components.

$$m(\mathbf{s}) = \sum_{i=1}^{M} w^i \eta(\mathbf{s}; \mu^i, \wedge^i). \tag{4}$$

where $\eta(\mathbf{s}; \mu^i, \wedge^i)$ denotes the i^{th} Gaussian component with mean μ^i and covariance $\wedge^i$. w^i is the i^{th} mixing weight satisfying $\sum_{i=1}^{M} w^i = 1$. From the color models of the sample $m(\mathbf{s})$ and the template $m(\mathbf{t})$, the similarity measure is defined by the distribution intersection between the color model of the sample and the template as

$$w_c = m(\mathbf{s}) \cap m(\mathbf{t}). \tag{5}$$

where w_c is the color score. (A discrete approximation of the intersection is implemented in our work.)

In motion feature, we also model the motion feature by a Gaussian to represent the distribution of the distance between predicted point and sample. The position of each body part is first predicted by a first-order motion and then used to compare with positions of samples.

$$w_m = \frac{1}{\sqrt{2\pi\omega^2}} e^{-\frac{d^2}{2\omega^2}}. \tag{6}$$

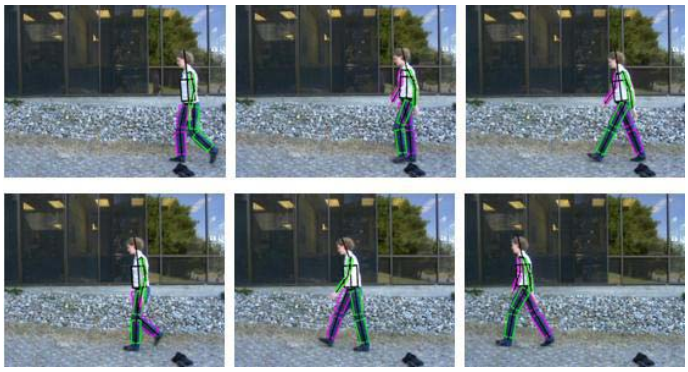
where d is the distance between the positions of predicted and sampled positions, w_m the motion score and ω the standard deviation.

5 Experiments

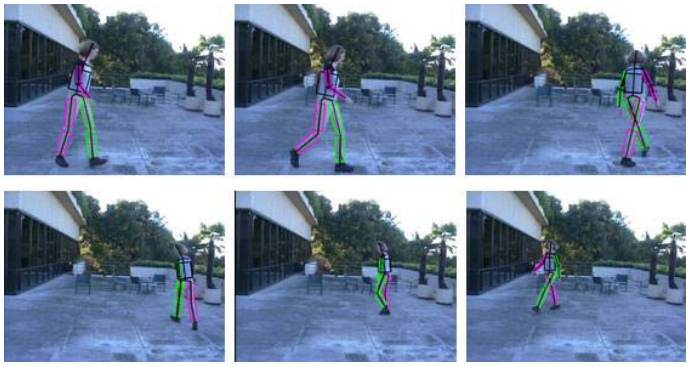
To evaluate the performance of our proposed method, we tested it on two image sequences, namely, "walk straight" and "walk in a circle", containing

DOI: 10.1587/elex.7.1106
Received March 29, 2010
Accepted July 02, 2010
Published August 10, 2010

60 and 145 frames, respectively. Figs. 3 (a) and (b) show the performance of the two sequences. The right limbs are shown in purple, whilst the left limbs are shown in green. The accuracy is evaluated from an averaged error distance compared with manually labeled positions, as 2.03 and 2.35 pixels/joint in "walk straight" and "walk in a circle", respectively.



(a)



(b)

Fig. 3. Some results using our approach (a) "walk straight" (b) "walk in a circle".

6 Conclusion

We propose a masker-less based approach to track the human body. The proposed method applies SMC to both color and motion features. By employing a self-updated binary occlusion mask and a model of the observation computation, the method can track objects reliably in cases of self-occlusion and ambiguity. The measurement is based on a color cue modeled by the Gaussian mixture model in HS space, and the distribution intersection is used in distance measures of two probability distributions. To increase the performance of tracking in cases of ambiguity, a motion cue is used to prune spurious solutions. From our experiments, our proposed method performed very well, especially in occlusion and ambiguous cases.

DOI: 10.1587/elex.7.1106
Received March 29, 2010
Accepted July 02, 2010
Published August 10, 2010