TRACKING AND SEGMENTING PEOPLE WITH OCCLUSIONS BY A SAMPLE CONSENSUS BASED METHOD

Hanzi Wang and David Suter

Department of. Electrical. and Computer Systems Engineering
Monash University, Clayton 3800, Victoria, Australia
{hanzi.wang; d.suter}@eng.monash.edu.au

ABSTRACT

One of the most difficult issues in visual tracking is to track people in groups, especially under occlusions. In this paper, we present a novel sample consensus based method, which utilizes both color and spatial information of human bodies, to model the appearance of people. We use this appearance model to segment and track people through occlusions. We show experimental results in several video sequences to validate the effectiveness of the proposed method.

1. INTRODUCTION

Tracking people is one of the most challenging tasks in computer vision. Human motion is non-rigid because when people walk towards or away video camera, both the shape and the size of images of people change. One simple approach is to use the connected component method [1] to extract several isolated regions and label each region as a person. However, this method is far from being efficient in complicated scenes. For example, a person can separate into several regions when the person is occluded by background objects such as trees or a pole; or when several people meet together and form a group, their connected components merge into a single one.

It is clear that an appearance model of people should be used in more successful approaches.

A lot of work has done in modeling human appearance [2, 3, 4, 5, 6, 7, 8]. In [7] and [4], the authors employed a mixture of Gaussians to estimate the color distributions of human bodies. Although some appealing results have been obtained, choosing the right number of Gaussians is a challenging problem. The authors of [5] employed color histograms to model the color distributions of human bodies. However the spatial information of human bodies is neglected. Indeed two people may have the same color histograms though they dress in different ways. Another problem of color histograms is that they require the sample size to be large enough to ensure statistical efficiency. In [2], a probability mask (i.e., an appearance template) is used to model each pixel of human bodies. Because this method records the appearance at each pixel, it requires a large memory to store the template. Non-parametric methods using Gaussian kernel density is employed in [6, 8] to model the human appearance. Although such is an improvement over [5] and [4], the biggest problem of the method is its high computational cost.

In [9], we have successfully applied the SAmple CONsensus (SACON) to model backgrounds involving dynamic scenes. In this paper, we extend that work and use sample consensus to model the appearance of human bodies. We also use the obtained appearance model to segment and track people with occlusions. We exploit both spatial and color information of the human bodies in our method. Since we use sample consensus, the computational burden is greatly reduced.

2. TRACKING AND SEGMENTING PEOPLE

Our main aim is to track people with occlusion. In order to simplify the question, we consider the occlusion between only two people (A and B). However, our method can be easily extended to segment and track multiple people. We use two tracking modules in the framework of the proposed method: box-based tracking and appearance-based tracking. When people are separated and are not occluded each other, we use the box-based tracking module; when people are occluded, or meet and form a group, we use the appearance-based tracking module.

2.1. Box-Based Tracking

In order to track people, we need an effective model to approximate the shape of human. Usually, there are two widely-used human shape models in the literature: ellipse [6] [3] and box (or rectangle) [2, 5]. Although an ellipse is

0-7803-9134-9/05/$20.00 ©2005 IEEE

more accurate than a box in modelling the shape of a human body, we find the box is effective in most cases and it is also easy to perform. We take the state of the box as: $S=(c_x,c_y,h,w)$, where c_x, and c_y is the center of the box, h and w is the half height and half width of the box. We update the state of all tracked boxes at each frame. When we detect two people occlude each other, we only update the centers of each corresponding box.

For each frame, we obtain several boxes corresponding to foreground regions. The system tries to associate each box with one of the detected tracks. This is done by the following steps: (a) merge small boxes, whose region size is less than a threshold T_{person}, with the closest large box; (b) merge the boxes whose projections onto x-axis are overlapped larger than a threshold T_x, but the projections onto y-axis are not overlapped; (c) merge the boxes whose distance (to the other box/boxes) is less than a threshold T_d; (d) associate each box to the closest track, whose distance to the box is small.

At the step (d), there are several possible cases to be considered: (1) if a box only corresponds to one track, and vice versa, the box is associated with the track; (2) if a box corresponds to two or more tracks, it means that several tracks merge at this frame. We use the human appearance model to handle this case. (3) if two or more boxes correspond to only one track, it means the track separates. We delete the track and create two or more child tracks which succeed the state of the track; (4) if no track is found to correspond to a box, we create a new track for the box. However, we label the new track as a temporary track until the track appears several consecutive frames, when we validate it as a new track; (5) if no any box corresponds to a track for several consecutive frames, it means that track is missing. We delete the track from the track list.

To compute the distance between two boxes A and B, we use:

$$D_{box}=\max(0,d_x)+\max(0,d_y) \tag{1}$$

where:

$$d_x=\begin{cases} c_{xA}\square w_A\square c_{xB}\square w_B, when\, c_{xA}\square c_{xB} \\ c_{xB}\square w_B\square c_{xA}\square w_A, when\, c_{xA}<c_{xB} \end{cases}$$

and similar for d_y.

From the above equation, we can see that: when $d_x < 0$ (or $d_y < 0$), it means that the projections of the two boxes onto x-axis (or y-axis) are overlapped; when $d_x > 0$ (or $d_y > 0$), i.e., no overlapping area for the projections of the two boxes onto x-axis (or y-axis); when the two boxes are overlapped for the projections of the two boxes onto both x-axis and y-axis, the value of D_{box} is equal to zero.

2.2. Modeling the Foreground

It is necessary to initialize the appearance of people when they enter scene or separate from a group, to differentiate each person from the others when they meet to form a group or occlude each other. We assume that the poses are upright and their appearance models do not change dramatically when they group or occlude each other.

2.2.1. Choose Color Space

The first step to model the human appearance is to choose the color space. Because RGB color space is sensitive to the change of illumination, employing RGB color space to model human appearance is less effective when the human is masked by shadows (of himself or other objects). Similar to [6], we employ normalized chromaticity r, g (where r = R/(R+G+B) and g = G/(R+G+B)) and intensity I (where I = (R+G+B)/3). Thus, each sample at location j: $S(j)$ can be expressed by three channels of r, g, I; $S(j)=(S^{c_1}(j),S^{c_2}(j),S^{c_3}(j))=(r,g,I)$, where $S^{c_i}(j)$ is the value of the sample in the ith channel (C_i).

2.2.2. Sample Consensus

Our work is inspired by RANSAC [10]. RANSAC is robust to multi-modal distributed data and can find a mode even if the data has multiple structures and a high percentage of outliers. If we partition a person into several blobs (usually three blobs: head, torso, leg), and we treat each blob as one mode, we can treat a person as a set of multi-modal distributed data. We now define a foreground model Sample Consensus.

When k persons $P_{j\ (j=1,..,k)}$ form a group or occlude each other, we need to classify each pixel $p_{i\ (i=1,..,n)}$ within the binary mask for the group into one of $\{P_j\}$. If we let A_j be the number of pixels of the jth person mask P_j, $\{S_j(m)\}_{m=1,..,Aj}$ be the samples from the jth person, a pixel p_i within the mask of the group can be classified to the kth person by:

$$p_i\square P_k, k=\arg\max_k \square_{c=r,g,I}\frac{\sum_m \square_k^c(p_i)}{A_k} \tag{2}$$

where $\square_k^c(p_i)=\begin{cases}1 & if\left|p_i^c\square S_k^c(m)\right|\square T_r \\ 0 & otherwise\end{cases}$, T_r is a threshold value.

2.2.3. Considering Spatial Information

Spatial information of data is important and provides the spatial structure of the data. However, spatial information is not considered in Equation (2). Now, we will take spatial information into account.

We consider the vertical direction (y-axis) and the horizontal direction (x-axis) separately. For the vertical direction, because we assume that the poses are upright, we do not want samples of the feet of a person, which have similar color, to contribute to the classification of a pixel at the head location of the person. Thus, Equation (2) is revised as:

$$p_i \square P_k, k = \arg\max_{k} \square_{c=r,g,I} \frac{\sum_{m'} \square^{c}_{k}(p_i)}{A'_k} \qquad (3)$$

where m' is a set of samples of P_k, whose y values are close to that of p_i; A'_k is the number of the set of samples $\{m'\}$.

For the horizontal direction, we take the median of the x values of the samples of P_k as c_x when P_k is occluded by other persons or in a group. Because c_x of P_k may change when the person is in the group, or with occlusion, it may cause error if we use c_x from the previous frame (i.e., at t-1 frame) to express the horizontal location of P_k at current frame (i.e., at t frame). This is especially noticeable when, for example, two persons exchange position in the horizontal direction. In order to use the horizontal spatial information, we use a two-step method:

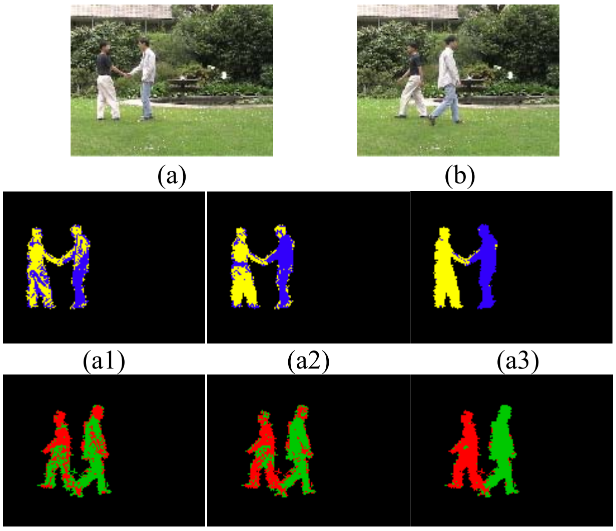
(1) We use Equation **(3)** to get the initial classification of pixels within the mask of the group;

(2) We compute $c_x(P_k)$ of the pixels classified as P_k. Then, we use Equation (4) to reclassify each pixel within the group mask:

$$p_i \square P_k, k = \arg\max_{k} \frac{\square_{c=r,g,I} \frac{\sum_{m'} \square^{c}_{k}(p_i)}{A'_k}}{|x(p_i) \square c_x(P_k)|} \qquad (4)$$

where $x(p_i)$ is the x value of p_i. (We note that $|x(p_i) \square c_x(P_k)|$ can be zero. Thus, we set a downward threshold to avoid dividing zero).

Figure 1 show the results obtained by sample consensus model without and with using spatial information. From Figure 1 we can see that, compared with the results obtained by the proposed method without using spatial information or using only vertical information, the results using both vertical and horizontal spatial information are the most accurate.



(b1) (b2) (b3)

Figure 1. (a) and (b) Original images; (a1) and (b1) segmentation results without using spatial information; (a2) and (b2) segmentation results using only vertical spatial information; (a3) and (b3) segmentation results using both vertical and horizontal spatial information.

3. EXPERIMENTAL RESULTS

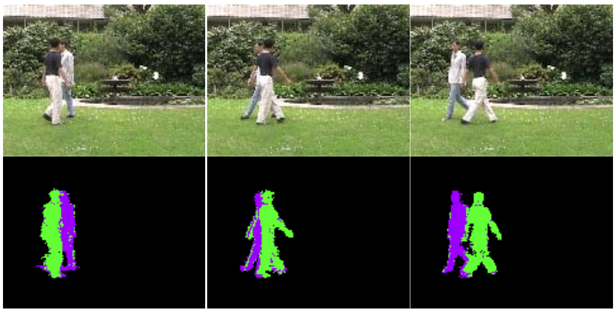


Figure 2. Segmentation results according to Equation (4). The first row: original images (frame numbers from the left column to the right column are 227, 228 and 229); The second row: human segmentation results;

To segment and track people in video sequences, first, we apply our recently proposed SACON background modelling method [9] to extract the foreground regions from the video sequences. Then, we input the foreground regions which correspond to humans, to the algorithm proposed in this paper. We initialize the human appearance models after people have been detected for several frames (i.e., when they are stable) and afterwards, we update the human appearance models at each frame until they occlude each other or merge to a group.

Figure 2 shows the segmentation results of humans by the proposed method. From Figure 2, we can see that, although there are some pixels misclassified, most pixels of the two persons are correctly classified even when the persons are subject to occlusions.

Figure 3 shows more segmentation results and tracking results by the proposed method. The proposed method successfully tracked both persons when they pass across and occluded each other.

4. CONCLUSIONS

In this paper, we present a new sample consensus based method for modeling human appearance and handling occlusion problem in human segmentation and tracking. Both color and spatial information are considered in human appearance model. The proposed method can successfully segment and track people through occlusion. We have applied our method to both outdoors and indoors video sequences and we have achieved promising results.

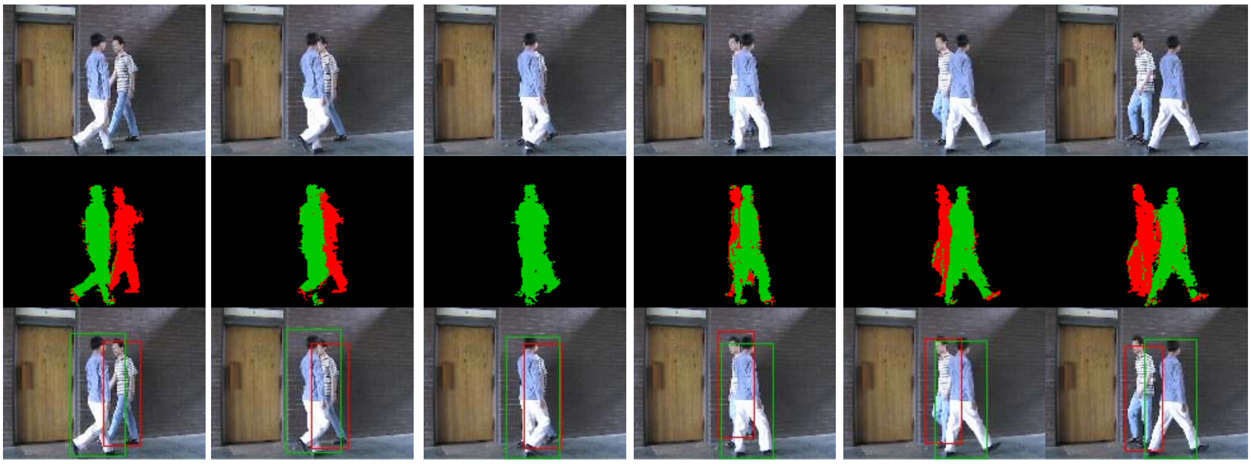


Figure 3. Segmentation and tracking results. The first row: original images (frame numbers from the left column to the right column are 103 to 108); The second row: human segmentation results; The third row: tracking results.

5. REFERENCES

1. R. Lumia, L. Shapiro and O. Zungia, "*A New Connected Components Algorithm for Virtual Memory Computer,"* Computer Vision, Graphics, and Image Processing. **22**(2): p. 287-300, 1983.
2. A. Senior. "*Tracking People with Probabilistic Appearance Models,"* in *ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems*: p. 48-55, 2002.
3. M. Hu, W. Hu and T. Tan. "*Tracking People Through Occlusions,"* in *International Conference on Pattern Recognition*: p. 724 - 727, 2004.
4. S. Khan and M. Shah. "*Tracking People in Presence of Occlusion,"* in *Asian Conference on Computer Vision*: p. 263-266, 2000.
5. S.J. McKenna, et al., "*Tracking Groups of People,"* Computer Vision and Image Understanding. **80**: p. 42-56, 2000.
6. A. Elgammal, et al., "*Background and Foreground Modeling using Non-parametric Kernel Density Estimation for Visual Surveillance,"* Proceedings of the IEEE. **90**(7): p. 1151-1163, 2002.
7. S.J. McKenna, Y. Raja and S. Gong, "*Tracking Color Objects Using Adaptive Mixture Models,"* Image Vision Computing. **17**(3): p. 225-231, 1999.
8. A. Elgammal and L.S. Davis. "*Probabilistic Framework for Segmenting People Under Occlusion,"* in *Proc. IEEE 8th Int. Conf. Computer Vision*: p. 145–152, 2001.
9. H. Wang and D. Suter. "*SACON: A Consensus Based Model for Background Subtraction,"* MECSE-15-2005, Monash University, Australia, 2005.
10. M.A. Fischler and R.C. Rolles, "*Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,"* Commun. ACM. **24**(6): p. 381-395, 1981.