Multiple People Labeling and Tracking Using Stereo

Nurul Arif Setiawan, Seok Ju Hong, Chil Woo Lee*
*Dept. of Computer Engineering, Chonnam National University

Abstract In this paper, we propose a system for multiple people tracking using fragment based histogram matching. Appearance model is based on IHLS color histogram which can be calculated efficiently using integral histogram representation. Since histograms will loss all spatial information, we define a fragment based region representation which retain spatial information, robust against occlusion and scale issue by using disparity information. Multiple people labeling is maintained by creating online appearance representation for each people detected in scene and calculating fragment vote map. Initialization is performed automatically from background segmentation step.

Keywords: Integral Histogram, Tracking, Multiple People.

1. Introduction

Detection and tracking of people in dynamic scene is one of the important modules for many vision application systems including surveillance, activity analysis, robot vision and human computer interaction. For such systems, the challenge is to determine people trajectory and maintain correct association of the tracked objects in natural situation which poses several problems such as occlusion, feature similarity and dynamic environment.

Mean-shift is popular tracking algorithm which based on histogram region with kernel model and its localization by performing minimization toward local basin of convergence [5]. Other trackers focused on filtering method to evaluate the objects dynamics and estimate trajectory hypotheses by means of particle filter [6,7] or Kalman filter [8]. In most of the trackers, objects are represented by its color histogram as it robust to view changes, noise and partial occlusion by assuming that objects has constant appearance throughout the scene. But histogram has several issues mainly the loss of spatial information and color histogram by itself is not robust against illumination changes and color similarity with background. To handle histogram limitation, [7] add Edge Orientation Histogram feature for their particle filter implementation. Several improvements on kernel and histogram definition for mean-shift tracker also have been proposed to maintain spatial information by using spatiogram [9] or correlogram [10].

In this paper, we extend the method developed by [2] by utilizing disparity information from stereo camera and provide multiple tracking framework. In [2], an object is represented by a template modeled with several patches. Tracking is performed by comparing histogram of patches of object's template to a new hypothesis for position and scale. The main tool for this method is the integral histogram structure proposed in [1] which enables us to extract histograms of multiple rectangular regions in efficient way. The author [1] has shown that exhaustive search over image region can be performed efficiently for object localization and with better tracking result than local search by the mean-shift algorithm. In fragment based tracking, the lost of spatial information is compensated by multiple patches and spatial relationship between patches retain spatial information.

In this paper, we propose some improvement for fragment based tracking mainly by adding disparity information and multiple people tracking framework. As note in [2], one issue in intensity only tracker is scale space search. To match over different range scale hypotheses, they enlarge and shrink the template by 10% and performed histogram matching in that scale. This exhaustive search over scale space can be reduced by using disparity.

Inclusion of the disparity information to add robustness of the tracker has been proposed by [3][11][13]. Plan view construction is used in [11] to track people with camera placement above head level. In [3], the authors use depth information to segment background and foreground and to provide scale space estimation for

template matching. Disparity as weight for mean-shift kernel tracking is used in [13] since by using disparity information, tracker will robust against similar background appearance. Disparity information it self is not always reliable because in case objects have a little visual texture resulting invalid region in disparity image. Our algorithm utilizes disparity information as weight in histogram representation and to estimate scale space search directly. This adds information to the objects representation resulting in better separation in case of partial occlusion and reducing computational cost.

Another improvement over [2] is to use Improved HLS color information over grayscale image. Although only with grayscale images, the algorithm gives good results with low computational costs, in more complex situation, color information will perform better since it holds more information that grayscale. By using HLS color space, we can separate illumination and use only chrominance component to model the object's histogram. This means less computational cost over RGB which require 3D histogram definition to capture all color information. Improved HLS color space is extension of HLS color space with better defined model [4] and has been implemented for background segmentation under Gaussian Mixture Model [12].

In next section, we will describe our algorithm for initialization and building appearance model including adding of depth information to add trackers robustness against occlusion and to provide estimation of scale space search. In section 3, fragment based tracking and multiple people tracking framework will be outlined. Some experimental results will be discussed in section 4, and finally section 5 concludes this paper and discusses some future works.

2. Segmentation and Appearance Model

2.1 System Overview

Our approach is built on IHLS color space [4]. In [12], IHLS color space has implemented for background segmentation with good result. Appearance model will be based on the IHLS color histogram which can be calculated efficiently using integral histogram structure [1]. Since histograms will loss all spatial information, we define a novel fragment based region representation which robust against occlusion and scale issue by using color and disparity information. Multiple people labeling is maintained by creating online appearance representation for each people detected in scene and calculating fragment vote map in each frame. Initialization is performed automatically from background segmentation step. Our system architecture is shown in Fig. 1.

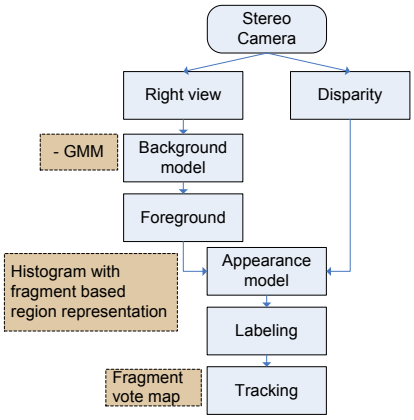

Figure 1. System architecture

2.2 Background segmentation

For automatic initialization, segmentation provide simple and quite reliable way to segment moving objects assuming that every moving objects in scene are actually people. More advanced method can be incorporated such as face or people detector. We use Gaussian Mixture Model using Improved HLS color space [12] to segment the foreground objects from a fixed stereo camera. RGB color obtained by camera is converted into Improved HLS color space using following equation [4]:

$$
\begin{aligned}
&s = \max(R,G,B) - \min(R,G,B) \\
&y = 0.2125R + 0.7154G + 0.0721B \\
&c_{r1} = R - \frac{G+B}{2}, c_{r2} = \frac{\sqrt{3}}{2}(B-G) \\
&c_r = \sqrt{{c_{r1}}^2 + {c_{r2}}^2} \\
&\theta^H = \begin{cases} \text{undefined} & \text{if} \quad c_r = 0 \\ \arccos(c_{r1}/c_r) & \text{if} \quad c_r \neq 0 \wedge c_{r2} \leq 0 \\ 360^o - \arccos(c_{r1}/c_r) & \text{if} \quad c_r \neq 0 \wedge c_{r2} > 0 \end{cases}
\end{aligned}
\qquad (1)
$$

Segmentation gives clue of the moving objects and can be used to limit search process only on foreground region instead of performing exhaustive search. From foreground binary mask, bounding boxes is extracted. This way, our system assume that when a person is entering scene, his bounding box should be well defined and not under merging or occlusion so that initialization process can created good appearance model. In this stage, several problems that might affect subsequent steps are merging and occlusion. This situation is shown in Fig. 2. Our approach to handle these problems will be explained later after we outline appearance model for people in scene.

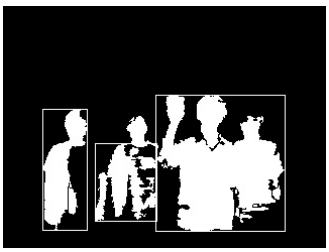

Figure 2. Bounding box from foreground regions

2.3 Integral histogram

The Integral histogram structure is first formalized by [1] as extension of integral image used in [15]. The integral image structure holds at the point *(x,y)* sum of rectangular region defined by top left corner of the image and the point *(x,y)*. This structure allows computation of pixels sum in any rectangular region by using four integral image values at four corners of the region.

Extending this idea into histogram representation is basically building integral image for each bin of the histograms. That is by counting the cumulative numbers of pixels that falling into each bin. Thus complexity and speed up of the integral histogram compared to standard histogram structure depends on how many bins used. Computational requirement for several scenarios of data dimension is given in [1]. The integral histogram structure is illustrated in Fig. 3 below.

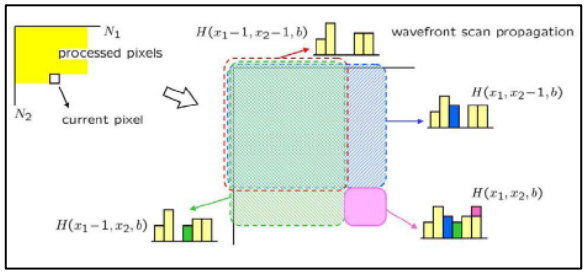

Figure 3. Integral histogram structure

After integral histogram is computed which depends on search region and size of bins, histogram extraction over any rectangular region of any size has similar computational cost. So in evaluating hypotheses of rectangular objects in several positions and scales as used in [2] is equal to cost of histogram comparison. And given disparity information from stereo camera, search over scale can be reduced by estimating scale directly with a linear equation.

2.4 Disparity layer

Disparity can add more information to build robust appearance model. This takes advantages of stereo's system ability to segment objects at different depth level. We build integral histogram structure for disparity image excluding invalid pixels for area with little visual texture. Let $\{v_v\}_{v=1..k}$ and v_M be its normalized histogram representation of a patch and its maximum probability. If we assume that the bounding box enclose tightly the foreground object, then disparity pixels which falls into maximum probability bin M, are the most likely to represent depth of object. This situation can be seen in Fig. 4.

We also segment disparity image into layers to represent scale variation and to find corresponding depth of objects given merging or partial occlusion. One can find scale of objects by using following linear equation [3]

$$s = dK \qquad (2)$$

s is scale, d is disparity value and K is constant which can be estimated or measured in simple calibration step. If an objects template is initialized with scale s = 1 and has disparity value d corresponding to maximum bin M as explained above, then when its observed that disparity value changes to d', objects scale can be assumed has changes to new scale s' by linear equation (2) above.

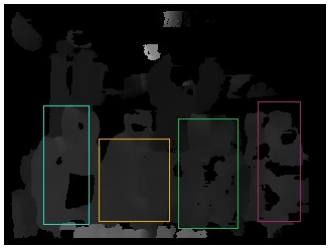

Figure 4. Objects bounding box with maximum probability represent depth of objects

2.5 Appearance model

Color histogram provide good model to represent objects because it's rather invariant to object/camera motion or shape change assuming that objects has constant appearance throughout the scene. Although histogram itself several issues mainly the loss of spatial information and color histogram by itself is not robust to illumination changes, fragment based template as proposed in [2] can overcome this limitation.

People body region is described by multiple sub regions (patches) with histogram representation (see Fig. 5 for example). From IHLS color, we compute saturation-weighted hue histogram expressed by following equation:

$$W_\theta = \sum_x S_x \delta_{\theta H_x} \qquad (3)$$

Where H_x and S_x are hue and saturation value at point x and δ_{ij} is Kronecker delta function. In this way, we have color information of objects in one dimensional histogram, thus reducing computational time to evaluate the metric

between patches histograms.

Figure 5. Patches of the appearance model for each people

Each rectangular patch also has its depth histogram. Color appearance of objects is more stable then depth since if the people move forward or backward than depth information will be changes drastically. Color information also will change slowly by illumination changes and new color appearance in the bounding box. Thus histogram representation of color and depth will be updated as following

$$H^{t}(k) = (1-\alpha)H^{t-1}(k) + \alpha H^{new}(k) \qquad (4)$$

Parameter α is update rate and k is bin of the histogram. Depth and color histogram has different α since each will changes at different rate.

3. Fragment-Based Tracking for Multiple People

In this section we will describe fragment based tracking using robust method to evaluate patches vote map and extend initial concept in [2] to use disparity information and tracking multiple people in scene.

3.1 Fragment based tracking

Fragment based tracking utilize fast calculation of integral histogram to compute multiple histogram of sub region in one rectangular template [2]. By comparing multiple location hypotheses we calculating vote map for next best location of current model. In this algorithm, tracking can be categorized as target representation and localization, where target representation is patch template and localization is performed by histogram matching in the patch's neighboring region. As mentioned previously, this method is possible to run at real time frame rate by using integral histogram structure.

Given an object O represented by template image T which contains a patch P_T, in tracking process we wish to find the position and the scale of a region in image I which is similar to patch T. Tracking is performed by exhaustive search in the neighboring region by calculating similarity of histograms. Given image patch $P_{T;(x,y)}$ where (x,y) is the hypothesis of objects position in current frame and patch P_T, if $d(Q,P)$ is some measure of similarity between patch Q and patch P, then

$$V_{PT}(x,y) = d(P_{I;(x,y)}, P_T) \qquad (5)$$

is the vote map corresponding to template patch P_T which gives scalar score of every possible position of the patch in current frame I [2]. In practice, any similarity metric can be used to obtain the vote map.

After all defined patches gives several vote maps, next is to combine this information to constitute new tracked object's position. To combine patches vote map, one can sum the vote maps and look to the position which gives minimal sum (if histogram metric calculate dissimilarity) or maximal sum (if histogram metric calculate similarity). But this direct approach is prone to occlusion since occluded single patch may contribute significant value to the sum. This can lead to wrong tracking estimation.

One intuitive way to combine all vote maps is proposed in [2] by using robust statistics with LMedS-type estimator expressed as

$$C(x,y) = Q'th \text{ value in } \{V_P(x,y) \mid \text{patches P}\} \qquad (6)$$

where

$$\{V_P(x,y) \mid \text{patches P}\} \qquad (7)$$

is the sorted set of obtained vote map and $Q'th$ is Q smallest score (they measure dissimilarity of patches by EMD distance). Q shows the expected inlier measurement which can be interpreted as percentage of template's target is visible. One desirable properties of such robust estimator is that outliers which will be rejected automatically can be assumed as occluded patches or partial pose change [2]. As note in [2], one issue in intensity only tracker is scale space search although the robustness of method shown tracker stability. By using scale and disparity relation in Equation 2, scale can be estimated directly and computational cost for matching process can be reduced.

3.2 Multiple Template Tracking

People in scene have difference of appearance based on color and depth information. To evaluate patches of multiple templates created in initialization step, we create overlapped grid in a rectangle region which contains possible people appearance based on foreground mask and calculate histogram of each grid. The grid size and step in x and y directions are adopted as the smallest patch size and half x and y size of the smallest patch respectively. Then for each template, we evaluate vote map patches and determine the new location of each object in current frame using method as explained in section 3.1.

Multiple labeling from parts based tracking can be performed by calculating observation likelihood and then assigning current parts into single objects using MAP solution [8]. Similar method is used to classify pixel into corresponding color blob as objects appearance [14]. By using robust statistics to combine all vote maps of patches in a template as expressed by Equation 6, tracking of multiple objects is actually similar. Thus the implementation to multiple people tracking is actually straightforward.

4. Experimental Results

We evaluate our algorithm in challenging situation where color similarity of foreground and background objects still posing problem in foreground segmentation step, and similarity of people appearance model. In this sequence, one person entering the scene one by one giving initial time for initialization of appearance model. Then one by one each person is moving forward causing merging of foreground segment and partial or full occlusion of other person. Results of our algorithm can be seen in following images.

Figure 6. Tracking result

5. Conclusion and Future Works

We have proposed novel method for multiple people tracking utilizing fragment based IHLS histogram representation which computed rapidly using integral histogram. A fragment based region representation is shown to be robust against occlusion and scale issue by using color and disparity information. Multiple people labeling is maintained by creating online appearance representation for each people detected in scene and calculating fragment vote map. The system will be implemented for robot system to maintain consistent labeling and tracking of people.

Acknowledgement

This research has been supported in part by MIC & IITA through IT Leading R&D Support Project and Culture Technology Research Institute through MCT, Chonnam National University, Korea.

References

[1] F. Porikli, "Integral Histogram: A Fast Way to Extract Higtograms in Cartesian Spaces", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, pp. 829-836, June 2005.

[2] A. Adam, E. Rivlin and I. Shimshoni, "Robust Fragments-based Tracking using the Integral Histogram." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2006.

[3] D. Beymer and K. Konolige, "Real-Time Tracking of Multiple People Using Stereo," IEEE Frame Rate Workshop, 1999.

[4] A. Hanbury, "Circular Statistics Applied to Colour Images," 8th Computer Vision Winter Workshop, Valtice, Czech Republic, February 2003.

[5] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.25, no.5 pp. 564- 577, May 2003.

[6] P. Pérez, C. Hue, J. Vermaak, M. Gangnet, "Color-Based Probabilistic Tracking." ECCV 2002: 661-675.

[7] Changjiang Yang, R. Duraiswami, L. Davis, "Fast multiple object tracking via a hierarchical particle filter," Tenth IEEE International Conference on Computer Vision, vol.1, pp. 212- 219 Vol. 1, 17-21 Oct. 2005.

[8] Qi Zhao, Jinman Kang, Hai Tao, Wei Hua, "Part Based Human Tracking In A Multiple Cues Fusion Framework," 18th International Conference on Pattern Recognition, 2006, vol.1, pp. 450- 455, 20-24 Aug. 2006.

[9] S. T. Birchfield, S. Rangarajan, "Spatiograms Versus Histograms for Region-Based Tracking," Proceedings of the IEEE CVPR 2005, California, volume 2, pages 1158-1163, June 2005.

[10] Qi Zhao and Hai Tao, "Object tracking using color correlogram," in IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05) in conjunction with ICCV, pp. 263 - 270, Beijing, China, October 2005.

[11] S. Bahadori, G. Grisetti, L. Iocchi, G.R. Leone, D. Nardi, "Real-time tracking of multiple people through stereo vision," The IEEE International Workshop on Intelligent Environments 2005, pp. 252- 259.

[12] N.A. Setiawan, Seokju Hong, Jangwoon Kim, Chilwoo Lee, " Gaussian Mixture Model in Improved HLS Color Space for Human Silhouette Extraction." ICAT 2006, LNCS 4282 pp. 732-741.

[13] Cheolmin Choi, Jungho Ahn, Seungwon Lee and Hyeran Byun. "Disparity Weighted Histogram-Based Object Tracking for Mobile Robot Systems." ICAT 2006, LNCS 4282 pp. 584-593.

[14] M. Balcells-Capellades, D. DeMenthon and D. Doermann, "An Appearance-based Approach for Consistent Labeling of Humans and Objects in Video," Pattern Analysis and Applications, pp. 373-385, November 2004.

[15] P. Viola and M. Jones, "Robust Real Time Object Detection," In IEEE ICCV Workshop on Statistical and Computational Theories of Vision, 2001.

[16] Ji Tao, Yap-Peng Tan, "Color appearance-based approach to robust tracking and recognition of multiple people," Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. vol.1, pp. 95- 99 Vol.1, 15-18 Dec. 2003