MULTI MOVING PEOPLE DETECTION FROM BINOCULAR SEQUENCES

Yang Ran   *Qinfen Zheng*

Center for Automation Research
Institute of Advanced Computer Studies
University of Maryland,
College Park, MD 20742-3275, USA
{rany, qinfen@cfar.umd.edu}

ABSTRACT

A novel approach for detection of multiple moving objects from binocular video sequences is reported. First an efficient motion estimation method is applied to sequences acquired from each camera. The motion estimation is then used to obtain cross camera correspondence between the stereo pair. Next, background subtraction is achieved by fusion of temporal difference and depth estimation. Finally moving foregrounds are further segmented into moving object according to a distance measure defined in a 2.5D feature space, which is done in a hierarchical strategy. The proposed approach has been tested on several indoor and outdoor sequences. Preliminary experiments have shown that the new approach can robustly detect multiple partially occluded moving persons in a noisy background. Representative human detection results are presented.

1. INTRODUCTION

Automatic human detection is an important issue for many applications [1]. Since the visual information from a single image is quite limited, there is a growing interest in using multiple cameras and/or video sequences for human detection. Stereo and motion are two widely used cues for human detection. Although significant progresses have been achieved [3], using either of the two alone does not provide satisfactory and un-ambiguity detecting results. Establishing dense stereo correspondence is a challenging task involving massive search and comparison. On the other hand, motion correspondence is often complicated by presence of shadow and/or illumination variations. Studies on fusing the two have been reported in [2], practical solution for integration of motion and stereo clues remains to be an open issue.

In this paper, we propose a novel approach for detecting moving human from binocular videos. It uses a fast and accurate sub-pixel accuracy motion estimation technique to extract object motion information, which significantly reduces ambiguity and computation cost in establishing dense stereo correspondence. In our approach, both motion consistency between the two cameras and stereo disparity map are used for background subtraction and moving object segmentation/grouping. Main contribution of this work is integration of correspondence from motion and stereo. The motion correspondence significantly improves the background subtraction process; while stereo correspondence trims down the searching computation. Fig. 1 shows a flow chart of our approach. Detailed description is presented in Section 2. Section 3 gives representative experimental results. Summaries and discussions are presented in Section 4.
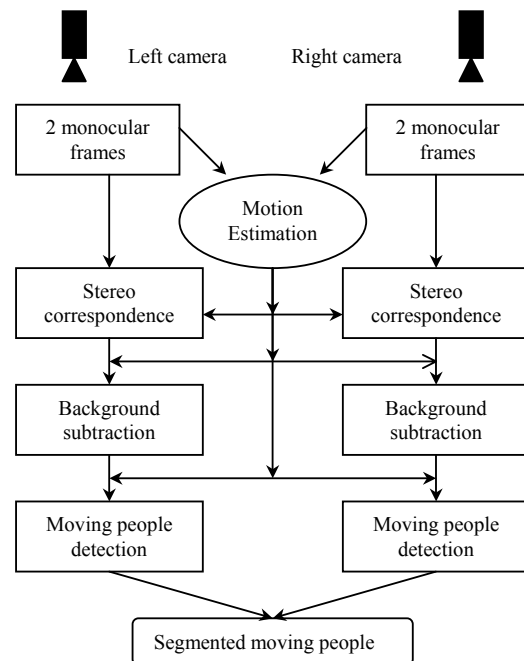

Figure 1. Flow chart of the proposed approach

2. ALGORITHM DESCRIPTION

2.1 Sub-pixel Motion Estimation

Object motion between two consecutive frames often does not confine to the spatial sampling grid (i.e. the object usually does not move in discrete pixel displacements), we need to estimate the object motion vector at sub-pixel accuracy. We use a spatial and temporal derivative analysis to compute motion flow [4] in both cameras. Assume the displacement of pixel (i,j) between f1 and f2 is $(\delta X, \delta Y)$:

$$f_2(i,j) = f_1(i - \delta X, j - \delta Y) \qquad (1)$$

The frame difference could be written as

$$d(i,j) = f_1(i,j) - f_2(i,j) = f_1(i,j) - f_1(i - \delta X, j - \delta Y)$$
$$\approx \frac{\partial f_1(i,j)}{\partial X}\delta X + \frac{\partial f_1(i,j)}{\partial Y}\delta Y \qquad (2)$$

For a local window n*n, all the equations contribute to:

$$\overline{I} = G\overline{D} \qquad (3)$$

where $\overline{I}$ is the column vector containing frame difference in local window; $\overline{D}$ is the motion vector matrix $(\delta X, \delta Y)^T$; G is a two-column matrix representing derivatives. Motion vector D can be computed as

$$\overline{D} = \begin{pmatrix} \delta X \\ \delta Y \end{pmatrix} = (G^t G)^{-1} G^t I \qquad (4)$$

Using of sub-pixel accuracy motion flow significantly cleans foreground extraction and furthermore simplifies stereo matching procedure.

2.2 Stereo Matching Using Motion

Binocular stereo matching is the process of obtaining depth information from a pair of calibrated cameras. Stereo provides a mechanism to group motion pixels according to their 3D location (range). Matching is typically performed on either pixel-wise dense map or sparse features follow by interpolation between the feature points. Pixel-wise matching usually involves intensive computations. Here we only interest in detection of moving objects. Sub-pixel motion estimation helps localizing stereo matching operations to "changing" regions. In addition, motion vectors provide additional clues to disambiguate objects with similar color but different motion directions and/or velocities. Simple statistical analysis reveals that focus on change regions reduces the total computation to about 20% of the original cost. Furthermore, with motion clues we can reduce the size of matching window to achieve faster processing speed and better localization.

In our implementation, we first transform the input video to YUV format and apply the Laplacian of Gaussian operator to the color-intensity video. Next "sum of absolute difference" is used to find the best match. This stereo algorithm checks 16 disparity levels and performs post-filtering, cross camera consistency verification and 4x range interpolation. The resolution correspondence map is obtained densely in changing regions and sparsely in background regions.

The major advantage of this framework is that the motion constraint is inherently enforced in the computation. In addition, depth representation based on motion estimation is exploited to guarantee depth smoothness.

2.3. Background Subtraction

Background subtraction is a critical step for moving person detection and tracking. The basic idea is to subtract the current stereo pair from corresponding reference (background) images to get the foreground. It typically requires having good static background models of the intensity or chromaticity of the scene [5, 6]. This requirement may not always be satisfied in applications involving operations over extended time period.

In our system, we use a three-level scheme that integrates uniqueness, homogeneity, and continuity assumptions about the foreground object motion and color distribution. The three assumptions lead to following three constraints:

1. *Uniqueness*: Each region/pixel from one image should match to at most a region/pixel at the other image of the stereo pair.
2. *Homogeneity*: Motion field for the same moving object should be spatially smooth within an image.
3. *Continuity*: Motion field for the same object should be similar over consecutive frames.

From the first constraint, if the corresponding region/pixel cannot be found, the moving region/pixel is reset to be background. This rule helps in reducing false motion detection caused by sensor noise. From the second constraint, if the motion estimations within a neighborhood various significantly, the corresponding motion detection is also reset to be background. And this rule helps in reducing false detections caused by clutter motions such as waving of tree branch and/or leaves. The third constraint checks the motion smoothness of the moving objects.

We classify a pixel (i,j) as belonging to foreground if and only if one of the following two rules is satisfied.

Weak Rule: $\max\{ D_{i,j}^L, D_{i,j}^R \} \geq \min\{ T_C, T_O \}$ (5)

or, *Strong Rule*: $\min\{ D_{i,j}^L, D_{i,j}^R \} \geq \max\{ T_C, T_O \}$ (6)

where

$$D_{i,j}^\alpha = \min\{ \overline{d}_{i,j}^\alpha, d_{i,j}^\alpha \}$$

$$\overline{d}_{i,j}^\alpha = \frac{1}{w \times w} \sum_{i,j \in w} d_{i,j}^\alpha$$

$$d_{i,j}^\alpha = abs[(m_n^\alpha(i,j) - m_{n-1}^\alpha(i,j)) \times (f_n^\alpha(i,j) - f_{n-1}^\alpha(i,j)]$$

$m_n^\alpha(i,j)$: Normalized optical flow in horizontal direction with index (i,j) from the n-th frame of camera α. It takes value in [-1, 1].

$f_n^\alpha(i,j)$: Normalized color/intensity value with index (i, j) from the n-th frame of camera α. It takes value within [0, 1].

T_C : Median of the normalized color/intensity value within the $w \times w$ clique window

T_O : On-site absolute threshold

α: Camera L or R.

In criterion (5) and (6), we take into account the difference from the two cameras. The strong and weak rules correspond to different criteria in change detection. The weak rule aims at detecting strong single pixel motion, while the strong rule focuses on collectively motion among connected pixels. $d_{i,j}^{\alpha}$ and $\bar{d}_{i,j}^{\alpha}$ carry on-site and average local changes respectively. The change is quantified by the production of normalized motion and intensity components. Hence $D_{i,j}^{\alpha}$ balances the target cost function by taking the minimum of the two values. When both motion and color vectors are available, the two items in computing $d_{i,j}^{\alpha}$ are redefined by the sum of the absolute values from all color components.

Generally speaking, foreground extracted using intensity-based methods has more accurate silhouette but too sensitive to illumination variations. On the other hand, the stereo-based detection algorithm is not affected by illumination changes over a short period. Combining both intensity and range clues in this algorithm provides better foreground detection.

Another problem we address here is shadow, which makes intensity based detection and tracking more challenging. However, it does not cause problems in our approach. We remove false detections caused by shadows through chromacity analysis. Note that, when a shadow is cast on the background, the intensity will be darker while the chromacity will keep unchanged. By verifying motion estimation from both intensity and color components, we can remove false detections caused by shadowing.

A representative background subtraction result is shown in Fig 2. It is taken from an outdoor sequence with tree branch weaving in the background. Shown on the left is an input image acquired by the left camera. Shown in the center is the result of background subtraction obtained by a traditional approach based on color. Shown on the right is the background subtraction result obtained by our approach. As shown, the tree moving and shadows cause false detections in the middle image. But in the right image, the noise caused by the two factors has been removed.

2.4 Human Detection

Most existing algorithms for human detection make use of intensity or color information in the grouping process [5]. Both probabilistic and deterministic approaches have been investigated but most of them use the assumption that different people wear clothes with different color distributions. The background model has to be built without occlusion.

Fig 2. Examples of background subtraction. Left: One input frame; Center: Result obtained using traditional approach; Right: Foreground extraction using the proposed approach.

The basic idea of human detection under occlusion is rather straightforward: grouping foreground pixels according to their homogeneity by the *motion* and *range* estimations. The main novelty here is to integrate range data and motion information in a hierarchical strategy. We use the depth information available from stereo disparity and motion information available from the subpixel motion estimation to build a 2.5D representation of a human. The human detection is achieved through three sequential operations: The first operation is to group foreground pixels into sets according to their range data. This would segment the foreground into different moving planes. The second operation is to group moving planes into blobs with consistent moving directions. The last operation is to segment blobs into moving objects with similar moving velocity. The flow chart of human detection proposed is shown in Fig 3.
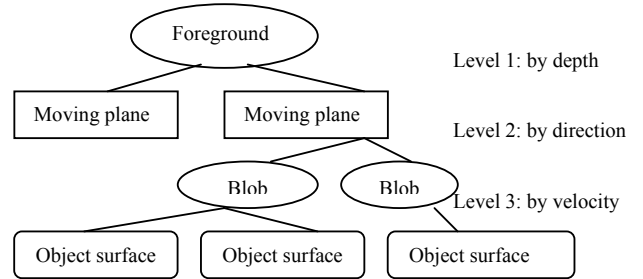

Fig 3 Flow chart for human segmentation under occlusion

Generally, there are four cases in multiple moving person detection:

i. No occlusion
ii. Occluded but at different ranges
iii. Occluded at the same range but move along different directions
iv. Occluded at the same depth, with the same motion direction, but move at different velocities.

The first 3 cases can be addressed using the above hierarchical method. For the last case, we build a motion vector distribution model for each blob. The model we choose is a mixture of Guassians in motion histogram. If the histogram contains multiple peaks, it is likely contains multiple moving objects and need to be split. Otherwise the corresponding blob will be treated as single object.

3. EXPERIMENTAL RESULTS

We have applied our algorithm to a number of stereo sequences acquired by a stationary stereo camera. Two representative results are presented here. The videos are captured at 320x240 resolution, 25 frames per second.

Figure 4 shows an example of detection results in an indoor scene. The background for indoor scene is constant during capturing. Shown in the left column are two input frames (#16 and #25) taken from the left camera. Shown in the central column are motion (foreground) detection results. Shown in the right column are the person segmentation/grouping results, where different individuals are assigned with different gray levels. The first row is the case where no occlusion occurs and the people are at different distances. The second row is the case where occlusion happens and the persons are at different distances. The proposed method successfully detects the moving persons and identifies them as two separated individuals, as demonstrated by the two masks with different gray levels.

Fig 4. Multiple human detection from indoor sequences. Left column: two input frames (#16 and #25) taken by the left image; Central column: foreground subtraction; Right column: human detection.
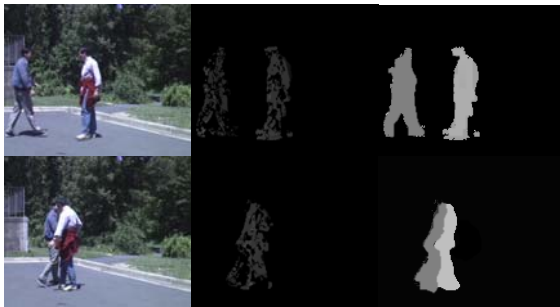

Fig 5. Multiple human detection from outdoor sequences. Left column: two input frames (#597 and #615) taken by the left image; Central column: depth maps obtained by our algorithm; Right column: human detection.

Figure 5 shows an example of detection two people in an outdoor scene. The test demonstrates that even under cluttered background (due to background vegetation motion) and shadows, the proposed method maintains good performance. Instead of showing the detected

foreground in the middle column, we present the raw data of motion detection result. The noisy background and shadows are significantly removed. Using a simple growing and linking method we get clean foreground and subsequently, correct moving persons. Shown in the left column are two input frames (#597 and #615) taken from the left camera. Shown in the central column are depth maps obtained by our algorithm. Shown in the right column are the person segmentation/grouping results.

4. SUMMARY

A novel approach for detection of multiple occluded moving persons from binocular video sequences is presented. By integrating the motion estimation result into every step in the whole detecting process, monocular and binocular correspondences are fused to generate robust detections, which is our main contribution. First an efficient motion estimation method is applied to sequences from each camera. The motion estimation is then used to obtain cross camera correspondence between the stereo pair. Next, background subtraction is achieved by fusion of temporal difference and depth estimation. Finally foregrounds are further segmented into moving objects according to a distance measure defined in a 2.5D feature space. The proposed approach has been tested on several indoor and outdoor sequences. Preliminary experiments have shown that the new approach can reliable detect multiple occluded moving persons from a noisy background. Representative human detection results are presented.

Future work includes using the approach to model scene structure. To associate detections over consequent frames by tracking and/or predicting is another interesting extension. Furthermore, this approach, currently operating on video taken from a stationary camera pair, can be extended to moving stereo platform.

5. REFERENCES

[1] I. Haritaoglu; D. Harwood; L.S. Davis, "W4: Real-Time Surveillance of People and Their Activities", PAMI Vol. 22(8), pp. 809-830. August 2000

[2] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", IJCV, pp. 7-42,Vol 47, No. 1, 2002.

[3] T. Darrell, G. Gordon *et al*, "Integrated person tracking using stereo, color, and pattern detection", CVPR pp. 601-609, 1998.

[4] Q. Zheng, R. Chellappa, "Motion detection in image sequences acquired from a moving platform", ICASSP pp. 201-204, Vol.5. 1998.

[5] A. M. Elgammal, L.S. Davis, "Probabilistic Framework for Segmenting People under Occlusion", ICCV, pp 145-152, 2001.

[6] I.Haritaoglu, D.Harwood, and L.Davis. W4S: A Real Time System for Detecting and Tracking People in 2.5 D, ECCV pp. 877—892. 1998.