Cooperative Moving Object Segmentation using Two Cameras based on Background Subtraction and Image Registration

Zhigao Cui*, Aihua Li, and Ke Jiang
502 Faculty, Xi’an Institute of High-Tech, Xi’an 710025, China
*Corresponding author, Email: cuizg10@126.com

Abstract—Moving camera, such as PTZ (pan-tilt-zoom) camera, has been widely applied in visual surveillance system. However, it’s difficult to extract moving objects because of the dynamic background caused by the camera motion. In this paper, a novel framework for moving object segmentation exploiting two cameras collaboration is presented by combining background subtraction and image registration method. The proposed method uses one static camera to capture large-view images at low resolution, and one moving camera (i.e., PTZ camera) to capture local-view images at high resolution. Different with methods using a single moving camera, the moving objects can be effectively segmented in the static camera image by background subtraction method. Then image registration method can be applied to extract moving region in the moving camera image. To deal with the resolution and intensity discrepancy between two synchronized images, we design a practical three-step image registration method, which has higher registration accuracy than traditional feature based method. Experimental results on outdoor scene demonstrate the effectiveness and robustness of proposed approach.

Index Terms—Motion Segmentation; Moving Camera; Background Subtraction; Image Registration; Resolution And Intensity Discrepancy

I. INTRODUCTION

Intelligent visual surveillance [1-2] has become one of the most active research areas in computer vision community due to the rapid demands of security and safety in variety of applications such as public, military, and commercial field. The aim of intelligent visual surveillance is to replace traditional passive surveillance system and accomplish the entire surveillance task as automatically as possible. There are several key technologies in visual surveillance [3]: segmenting of moving objects, tracking of such objects at each frame, and understanding the behavior or activity of the tracked objects. Among these, moving object segmentation is the base step of the total intelligent visual surveillance system, and usually has great influence on the other two high-level steps. Thus it is important to improve the moving object segmentation step in terms of accuracy and efficiency.

The aim of moving object segmentation is to detect regions corresponding to moving objects from surveillance video sequences [4]. According to the usage of different cameras, the moving object segmentation problem can be divided into two classes: segmentation under stationary background and segmentation under nonstationary background. In the case of stationary background, the static camera and the surveillance scene are relative fixed. Background subtraction is a widely used approach due to its accuracy for motion segmentation. It usually chooses a model to describe the background, and then a learning rate is given to dynamically update the background model. According to a certain criteria, such as the absolute difference between current image and background model, every pixel of the current image will be segmented as moving objects or background. The common used background models include median filtering model [5], mixture of gaussian model [6], codebook model [7], nonparametric kernel density estimation model [8], and so on. Recently, Barnich and Droogenbroeck in [9] propose to model the background color distribution at a stochastic select and update strategy. Numerous results have demonstrated that their methods have better performance than other methods in terms of both accuracy rate and computation speed.

With the rapid improvement of camera manufacturing, moving cameras (i.e. PTZ cameras) have been widely used in visual surveillance environments [10-11]. Thus it is important to segment moving object from images taken with such types of viewing sensors. However, it is more difficult than stationary background due to the dynamic background caused by camera motion. In such cases of nonstationary background, the traditional background subtraction method cannot be applied directly. Extensive research has been carried out to solve this problem. These methods can be categorized into two types: background compensation based method and background mosaic based method. Both of the two methods can be reduced to image registration problem. Background compensation based method [12-14] first estimates the global transformation caused by the camera motion between consecutive images using image registration method. Then the background regions are compensated and registered ideally, and the moving objects can be segmented pixel by pixel. Global transformation parameters estimation is the key technology in the background compensation method, and it usually consists
of three major procedures: 1) choosing a transformation model to describe the camera motion; 2) extracting corresponding features between consecutive images; 3) estimating the model parameters based on matching features. The performance of background compensation based method is not sensitive to the scene environment and its computational cost is low. However, it faces the problem that it can’t segment entire pixels of moving objects. Background mosaic based method [15-16] usually creates a panorama image of the whole scene background beforehand. Then at each captured image, we can obtain the corresponding background region within the mosaic background using image registration technique. Finally, traditional background subtraction method can be used to obtain the moving object regions. This method can extract the whole moving regions compared with the background compensation method. However, due to the dynamic change of illumination and scene structure, it’s difficult to update the mosaic background model [16].

In this paper, we propose a novel algorithm framework for motion segmentation under nonstationary background using two cameras collaboration. In our method, we use one static camera to capture large-view images with low resolution, which can correspond to motion segmentation problem under stationary background. And the other one is a moving camera (i.e. PTZ camera) to capture local-view images with high-resolution, which corresponds to motion segmentation problem under nonstationary background. Because that the moving object can be easily segmented in static camera view. Thus using the static view as a bridge, the moving region in the moving camera can also be accurately segmented by image registration method. Since the resolution and intensity discrepancy between two camera images increase the difficulty of image registration. We propose a practical three-step method to deal with this problem. Overall, compared with methods using a single moving camera, the proposed algorithm framework has the following advantages:

1) Both panoramic and detailed information of the moving object can be obtained at the same time, which is very useful for behavior analysis and evidence collection.

2) The static camera image can serve as the assistance to guarantee the entirely segmentation of the moving region in moving camera image, which is difficult or impossible for the background compensation based method. Besides, the problems of background mosaic based method (background model construction and adaptation) can also be avoided.

The paper is organized as follows: Section II gives an overview of the proposed approach. Section III describes the proposed three-step image registration method. Section IV presents the cooperative moving object segmentation method based on the registration model estimated in Section III. The experimental results are provided in Section V. In Section VI, we summarize the paper with some conclusions.

II. OVERVIEW OF THE PROPOSED APPROACH

In this paper, we denote $I'_S$ and $I'_D$ as the static camera and moving camera image at $t$-th frame, $A'_{SD}$ as the registration model between $I'_S$ and $I'_D$. The moving regions in $I'_S$ and $I'_D$ are denoted, respectively, as $F'_S$ and $F'_D$. Fig. 1 gives an overview of the proposed algorithm, which contains two parts: three-step image registration (presented in Section III) and cooperative motion segmentation (presented in Section IV). In the three-step image registration part, we estimate the registration model $A'_{SD}$ between $I'_S$ and $I'_D$. After $A'_{SD}$ is known, the moving regions in $I'_D$ can be extracted using $A'_{SD}$ and moving regions in $I'_S$. The major steps of the two parts are described as follows.

Three-step image registration algorithm
- Input: $I'_S$ and $I'_D$, $t = 0, ..., \text{Lastframe}$.
- Output: $A'_{SD}$.
- Major steps:
  - Step 1. Use mean-shift tracking algorithm to get the rough registration region in $I'_S$.
  - Step 2. Use matching SURF feature points between rough registration region of $I'_S$ and $I'_D$ to estimate the initial registration model.
  - Step 3. Use histogram equalization method to make the intensity of $I'_S$ and $I'_D$ comparable, then refine the initial registration model.

Cooperative motion segmentation algorithm
- Input: $I'_S$, $I'_D$, and $A'_{SD}$, $t = 0, ..., \text{Lastframe}$.
- Output: $F'_S$ and $F'_D$.
- Major steps:
III. THREE-STEP IMAGE REGISTRATION

The key problem of our method is registration of the moving camera image $I'_p$ to the static camera image $I'_s$. However, there are two difficulties in image registration: 1) because of different FOVs (field of view) of two cameras, the images are captured at different scales or resolution; 2) because of the self-adjustment of different cameras to the illumination condition, there exist large intensity inconsistence in two cameras images. An example to illustrate the two problems is shown in Fig. 2. During last decades, lots of image registration algorithms have been proposed. These methods can generally fall into two categories: area-based methods and feature-based methods [17]. Area-based methods use part of the whole image to estimate the registration model parameters. The main advantage of such type of method is that it can provide relatively high registration accuracy. However, it cannot handle the resolution and intensity discrepancy between images. Recently, the local invariant feature has been widely used in image registration for its invariant to image scale, rotation, and illumination. Thus, the feature-based method can allow large discrepancy between images, although it is less accurate due to the unpredictable of distribution of matching feature points. Here, we propose a three-step image registration method. The proposed method has combined the merits and simultaneously avoided the shortcomings of both methods.

![Figure 2. An example to show the difficulties in image registration: scale discrepancy and intensity inconsistence. (a) Image $I'_p$ with intensity($I'_p$) = 193; (b) Image $I'_s$, with intensity($I'_s$) = 151.](image)

Before present the major steps of proposed registration method, we first select an appropriate model to describe the geometric transformation between two synchronized images. Let $x_i = (u_i, v_i)$ and $x'_i = (u'_i, v'_i)$ be a pair of correspondence points in two views. Then the geometric relation between $x_i$ and $x'_i$ can be expressed as

$$x'_i = f(x_i)$$

where $f$ is the geometric transformation between two images, which is determined by the prior knowledge of the scene as well as the sensor geometries [18]. The commonly used geometric transformation models include similarity, affine, and projective transformations. In our study, we assume that the optical centers of two cameras intersect at one point. Then the group of geometric transformations between two synchronized images may undergo a group of projective transformations. To further simplification, we choose the six-parameter affine transformation model (see Eq. (2)). Note that the affine transformation is an acceptable approximation if the cameras are far enough from the scene or the baseline is relatively small. When the two situations are not satisfied, we can use more complicated transformations to improve the accuracy.

$$\begin{bmatrix} u'_i \\ v'_i \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}$$

(2)

A. Step1: Obtain the Rough Registration Region

To segment moving objects from the moving camera image, previous methods either build a panoramic background model or use several consecutive images to make the background relative stationary. Here we use one static camera as a bridge, and then convert nonstationary background problem to the stationary background problem. However, the major challenge we faced is the registration of two synchronized images due to the huge scale difference. Although feature-based method such as SIFT [19] or SURF [20] are widely used for scale invariance. The performance of these features decrease when the moving camera image only occupies a small part of the static camera image. So we need to identify a local region in the static camera image $I'_s$, which is denoted by rough registration region in this paper.

We propose to apply mean-shift tracking [21] method to determine the rough registration region in $I'_s$. Firstly, we use mean-shift method to get trajectory of motion object at each frame. Then the trajectory center is smoothed within forty consecutive frames in order to decrease the computational errors. Finally, taken the smoothed center as the region center, the rough registration region can be identified. For our application, the image size of both cameras is $320 \times 240$. And the rough registration region is set to be $150 \times 150$ in our study. An example is provided in Fig. 3.

![Figure 3. Rough registration region estimation for the image in Fig. 2](image)

B. Step2: Estimate the Initial Registration Model

After getting the rough registration region in $I'_s$, we use SURF features for image registration due to its...
efficient computation by the use of integral images. Furthermore, it can achieve similar performance to the SIFT keypoints algorithm. Firstly, we detect and describe SURF keypoints in the rough registration region of $I'_s$ and the whole image $I'_o$. Then, the keypoints are matched under the criterion of vectorial angle cosine between descriptors vectors. Let $x'_i (i = 1, ..., n_i)$ and $x'_o (j = 1, ..., n_o)$ are extracted keypoints in two images. Then the vectorial angle between descriptors of $x'_i$ and $x'_o$ can be expressed as

$$\theta(x'_i, x'_o) = \arccos(\text{des}^i \cdot \text{des}^o) = \arccos \left( \frac{\text{des}^i \cdot \text{des}^o}{\|\text{des}^i\| \cdot \|\text{des}^o\|} \right)$$

(3)

where $\text{des}^i$ and $\text{des}^o$ are the normalized descriptor vectors of keypoints $x'_i$ and $x'_o$. The best matching candidate is found if the ratio between minimum angle and second minimum angle is smaller than a certain threshold. Once SURF feature correspondences are found, the initial affine transformation model should be estimated. However, the correspondences may include a large number of outliers, which are not compatible with the affine model, so a robust algorithm should be used. In our study, Random Sample Consensus (RANSAC) method [22] is used to remove the outliers and collect the inliers. This approach is proven to be effective even when the correspondences are severely corrupted by a large percentage of outliers. After getting a set of inliers, the Least Squares Algorithm (LSA) is employed to estimate the initial affine registration model $A_{02d}$. Fig. 4 shows the SURF feature matching result for the image in Fig. 2.

![Figure 4. SURF feature matching result for the image in Fig. 2](image)

C. Step3: Refine the Registration Model

Using the initial affine model derived from previous step, we can warp the moving camera image $I'_o$ towards the static camera image $I'_s$. Since the features may not be well distributed over the image, the two images will not align exactly. Although this, we can obtain a rough overlap region, and then the area-based registration method can be applied to improve the registration accuracy. The general principal is to estimate an affine transformation model that minimizes the sum of squared differences (SSD) between pixels in image $I'_s$ and $J'_o$, where $J'_o$ is the warped image using initial affine model estimated in step2. Ordinarily, the calculation of SSD is only implemented over the pixels of overlap regions between two images.

As shown in Fig. 2, the two images have large intensity inconsistence due to self-adjustment of different cameras to the illumination condition. So histogram equalization method should be performed beforehand to make the intensity of two images comparable. In our study, we choose a piecewise linear intensity mapping model with three linear segments to deal this problem. The designed mapping model can be written as

$$J'_o(x,y) = \begin{cases} k_1I'_s(x,y) & 0 \leq I'_s(x,y) < a \\ k_2I'_s(x,y) + b_1 & a \leq I'_s(x,y) < b \\ k_3I'_s(x,y) + b_2 & b \leq I'_s(x,y) \leq 255 \end{cases} \quad (4)$$

where $k_1, k_2, k_3, b_1, b_2$ are the parameters of piecewise linear function, $a$ and $b$ are the knots. The parameters of $k_1, k_2, k_3, b_1, b_2$ can be calculated by compare the cumulative intensity histograms of overlap regions between images $I'_s$ and $J'_o$. Then the other parameters can also be determined to ensure the consistency of intensity mapping model.

Once the parameters have been estimated, we adjust the intensity of $I'_s$ according to the intensity mapping model, and we denote it by $J'_s$. Then the registration model between $J'_s$ and $J'_o$ is close to $3 \times 3$ identity matrix. So we can solve the following optimization problem to refine the affine transformation model and further improve the registration accuracy.

$$A'_{02d} = \arg \min_A \sum_v \left\| J'_s(x_v) - J'_o(g(x_v,A)) \right\|^2$$

(5)

where $A$ is an $3 \times 3$ affine transformation matrix, $v$ is pixels number of overlap regions between images $J'_s$ and $J'_o$. $g(x_v,A)$ is the coordinate transformation function, and is defined as

$$g(x_v,A) = Ax_v$$

$$g(x_v,A) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1(1) \\ x_1(2) \\ x_1(3) \end{bmatrix}$$

We use the gradient based incremental parameters estimation method [23] to iteratively solve above optimization problem. In each iteration, we estimate the least square solution of incremental model parameter by calculating the image gradient Hessian matrix. Then the incremental model parameter is used to update the current estimated affine model. In our study, the initial value of $A$ is set to be $3 \times 3$ identity matrix. The iteration number is set to be twice. The final refined affine model $A'_{2d}$ can be computed by

$$A'_{2d} = A'_{02d}A'_{2d}$$

(7)

To test the effectiveness of our method, we enlarge the rough registration region of image $I'_s$ with a magnification factor 3, and get an $450 \times 450$ image. Then
the registration model between the enlarged image and image $I^*_p$ can be expressed as

$$A^t_{SD} = A^t_{sp} \times \begin{bmatrix} 1/3 & 0 & l - 1/3 \\ 0 & 1/3 & t - 1/3 \\ 0 & 0 & 1 \end{bmatrix}$$

(8)

where $l$ and $t$ are left and top border of rough registration region in image $I^*_s$. We warp the image $I^*_p$ towards the enlarged image using the initial and refined registration model, respectively. The results are shown in Fig. 5. We can see that the registration accuracy is improved by the model refinement step.

![Figure 5. Image registration comparison for the images in Fig. 2. (a) Result using the initial registration model. (b) Result using the refined registration model.](image)

IV. COOPERATIVE MOTION SEGMENTATION

Once the registration model between two synchronized images has been successfully computed, the cooperative motion segmentation can be realized. The detailed implementation procedures are described as follows.

A. Step1: Motion Segmentation in Static Camera Image

For the static camera image, the motion segmentation is quite reliable even using conventional approaches. In this paper, we use a simple median filtering algorithm [5] to extract moving regions for image captured form the static camera. Although using more complex background subtraction algorithms can generate better results. Here, we just select a simple method to show the effectiveness of proposed algorithm framework.

We first store $r$ frames in the video buffer. Then at each pixel position $(x, y)$, the medium of $r$ training frames can be seen as the background intensity value:

$$B^t_s(x, y) = \text{median}(I^1_s(x, y),..., I^{r-t}_s(x, y))$$

(9)

where $B^t_s(x, y)$ is the background intensity value at position $(x, y)$ in the $t$-th frame. Finally, motion segmentation that distinguishes moving objects from the background can be carried out using a threshold $TH$ :

$$F^t_s(x, y) = \begin{cases} \text{Moving objects} & \text{if } |I^t_s(x, y) - B^t_s(x, y)| > TH \\ \text{Background} & \text{otherwise} \end{cases}$$

(10)

In most situations, a learning rate $\lambda$ is introduced to update the background image:

$$B^{t+1}_s(x, y) = \begin{cases} B^t_s(x, y), & \text{if } I^t_s(x, y) \in \text{moving objects} \\ \lambda I^t_s(x, y) + (1 - \lambda)B^t_s(x, y), & \text{otherwise} \end{cases}$$

(11)

To reduce the influence of noise and illuminations, we use morphological filter technique and connected component analysis method to get the final moving regions in image $I^*_s$. Fig. 6 shows the segmentation result for image in Fig. 2.

![Figure 6. Motion segmentation result of the static camera image](image)

B. Step2: Motion Segmentation in Moving Camera Image

With two-camera configuration using in our algorithm framework, the motion segmentation problem in the moving camera image can be greatly simplified. Indeed, the extracted moving regions in the static image can be seen as a bridge to help this process. Therefore, the detailed procedure is based on: (1) warping the moving camera image towards the static camera image based on the estimated registration model; (2) comparing the two images pixel by pixel and finding the corresponding moving regions according to the segment result in the static camera image. Fig. 7 illustrates this procedure.

![Figure 7. Illustration for the procedure of motion segmentation in the moving camera image](image)

V. EXPERIMENTAL RESULTS

We now present the experimental results for applying our algorithm framework on video sequences acquired by two PTZ cameras. One of the two cameras is acted as a static camera, which monitors the whole scene with low resolution. The other is acted as a moving camera, which dynamically follow the moving object at high resolution. Images from both cameras are taken at 320×240 pixels of resolution. A total of 933 pairs of images are used test
First, we present the experimental results of our registration method. The proposed method consists of three steps, which enhances the registration accuracy through a model refinement procedure. The method without the refinement step is the traditional feature-based image registration method, as illustrated in Section III. We call this method as the SURF-based method in this paper. In this section, we will compare the registration performance of the SURF-based method and the proposed method. Some frames of the registration result are shown in Fig. 8. The implementation of experiments is the same as the Fig. 5. The first column in Fig. 8 is the SURF matching result of two synchronized images. The second and third columns are the registration results using SURF-based method (unrefined model) and proposed method (refined model). The comparison regions are highlighted by red color and magnified by two times. From the proposed method. The experiment is implemented in C++ utilizing the OpenCV Library, and runs on one computer with Intel 3.0G CPU and 2G memory.

Figure 8. Example of image registration results

Figure 9. The intensity difference mean comparison result of proposed method versus the SURF-based method
enlarged images, we can observe that the registration is improved by the proposed registration method.

For quantitative demonstrations, we use the criteria of intensity difference mean [12] for accuracy evaluation and comparison between SURF based method and proposed method. The intensity difference mean can be computed as

\[
\text{Intensity difference mean}(I_S, I_D) = \frac{1}{WH} \sum_{x=0}^{W} \sum_{y=0}^{H} |I_S(x, y) - I_D(x, y)| \tag{12}
\]

\( J_D \) is the transformed image of \( I_D \) using the registration model. \( W \) and \( H \) are the width and height, respectively, of the overlap regions between \( I_S \) and \( J_D \).

We select 46 pairs of frames from the video sequence at a fixed sampling interval. Fig. 9 shows the comparison result of these frames. As can be seen from this figure, the two methods have similar trend of intensity difference mean, while the proposed method outperforming SURF based method. The average intensity difference mean of proposed method is 7.30, which is apparently smaller than 8.47 of the SURF based method.

We then test the accuracy of our motion segmentation method once the registration model has been successfully computed. Fig. 10 shows the segmentation results of corresponding frames in Fig. 8. The first column of Fig. 10 is the segmentation result of static camera image. The second column is the registration result by warping moving camera image towards the static camera image. The third column is the segmentation result of moving camera image.

VI. CONCLUSION

This paper presents a new framework for motion segmentation under nonstationary background based on two-camera collaboration. Compared with existing methods using single moving camera, the proposed scheme can not only extract entire pixels of moving region but also avoid the panoramic background model construction and adaptation. In addition, we design a practical three-step method to ensure accurate registration of two synchronized images. Experimental results have verified the effectiveness and robustness of our approach.

In the future, we will attempt to extend this work in two aspects: (1) The affine transformation model is used in current experiments with the assumption of a short baseline configuration of two cameras. Future work will attempt to involve more complicated transformation models. (2) Motion segmentation using pixel to pixel comparison is not robust due to the existence of registration error. We will consider this problem in future study.

REFERENCES


Zhigao Cui is currently working toward the Ph.D. degree in the Xi’an Institute of High-Tech, Xi’an, China. He received the B.E. and M.E. degrees in 2007 and 2010, respectively. His research interests mainly include multi-camera visual surveillance, pattern recognition and image processing.

Aihua Li received the Ph.D. degree from Xi’an Jiaotong University, Xi’an, China, in 1998. He is currently a Professor with Xi’an Institute of High-Tech, Xi’an, China. His research interests mainly include fault diagnosis, signal processing, image processing and pattern recognition.

Ke Jiang received the B.E. and M.E. degrees from Xi’an Institute of High-Tech, Xi’an, China, in 2006 and 2009, respectively. His research interests mainly include computer vision, pattern recognition and image processing.