Towards Monitoring Human Activities Using an Omnidirectional Camera

Xilin Chen Jie Yang
Interactive Systems Lab, School of Computer Science, Carnegie Mellon University
{xlchen, yang+}@cs.cmu.edu

Abstract

In this paper we propose an approach for monitoring human activities in an indoor environment using an omnidirectional camera. Robustly tracking people is prerequisite for modeling and recognizing human activities. An omnidirectional camera mounted on the ceiling is less prone to problems of occlusion. We use the Markov Random Field (MRF) to present both background and foreground, and adapt models effectively against environment changes. We employ a deformable model to adapt the foreground models to optimally match objects in different position within a pattern of view of omnidirectional camera. In order to monitor human activity, we represent positions of people as spatial points and analyze moving trajectories within a time-spatial window. The method provides an efficient way to monitoring high-level human activities without exploring identities.

1. Introduction

Monitoring human activity has many applications such as video surveillance and human computer interaction. It requires understanding interaction among people and between people and environments. Automatically tracking people is prerequisite for analyzing and understanding human activity. Human activities can be monitored at different levels of details by tracking different features. For example, some tasks require a system monitor who is doing what, when, and where, in what kind of mood. Such a system needs to track every detail of a person, from facial features to body parts as well as objects in the scene. Many other tasks, on the other hand, do not need such detailed information. For example, we can ignore many details if we only want to know how many people are in a scene, how long they have stayed in the scene, where they are in the scene, and what pattern they are in the scene (meeting or alone). In such a case, we even can represent a person using a spatial point (*x, y, z*). In this research, we are interested in monitoring human activities at this level. More specifically, we are interested in monitoring human activities at a coarse level in an indoor shared working space such as a laboratory. We would like to track moving trajectories of people in the space and use such information to model and identify certain human activities without detailed information. By combining knowledge of the environment, the system will be able to determine where a person is located; how people are moving in the space; and if a person is working alone or having a meeting with other people.

Occlusion is a major challenge for the existing systems in monitoring human activities. A solution is to use multiple cameras to view people from different angles. This increases complexity and expense of the system. Furthermore, a normal video camera has a limited viewing angle and is directional. The human target could be occluded from multiple normal video cameras if he/she is among a crowd. An omnidirectional camera mounted on the ceiling, on the other hand, has 360 degree viewing angle and, by virtue of being above the 'action,' is less prone to problems of occlusion. Although an omnidirectional camera has a limited resolution, it causes little problem for our application where we represent a person as a point.

The work presented in this paper is related to video surveillance. A good review of many state-of-the-art video surveillance systems can be found in the special issue of the IEEE Transactions on PAMI (August, 2000) and October 2001 special issue of the Proceedings of the IEEE as well as IEEE Workshops on Visual Surveillance in the last three years (1998, 1999, 2000). These technologies include feature-based, edge-based, boundary-based, and model-based approaches. Feature-based approaches, however, have problems when targets are small or/and with deformations and occlusions. An effective method is to combine model-based approach and feature-based approach. Many researchers have used various features to help initialize models [6, 9, 14, 18]. Several people tracking systems have been developed for multi-person whole-body tracking. University of Maryland has developed a series of outdoor person-trackers named W^4 (separated people, grayscale camera), W^{4}S (separated people, stereo camera), and Hydra [10]. The Robotics Institute at Carnegie Mellon University has created an elaborate system that classifies and tracks multiple people and vehicles as they move about outdoors, under the DARPA VSAM project [4]. Microsoft Easyliving project has developed multiple people tracking system using multiple stereo cameras [11]. MIT AI lab has derived dense stereo models for object tracking using long-term, extended dynamic-range imagery, and by detecting and interpolating uniform but unoccluded planar regions [5]. Orwell *et al.* has reported a tracking system that uses multiple cameras and tracks multiple people walking in a parking lot [13]. Rosales and Sclaroff described a multi-person tracking system that unifies object tracking, 3D trajectory estimation, and action recognition from a single video camera [17]. Rehg *et al.* presented a multi-person tracking system for an interactive kiosk that uses a pair of widely space color cameras [16]. Boult *et a*l. used an omnidirectional camera to track multiple, camouflaged soldiers in outdoor scene [3]. Rees first introduced the concept of an omnidirectional camera for television [15]. There are two classes of methods to obtain an omnidirectional view. The first one is a single camera-based system, which can be a catadioptric system [12] or a fish-eye lens based system, and the other one is a multi-camera system, such as multi-camera networks [7], FlyCam [8]. There are several omnidirectional cameras commercially available from different companies.

The rest of this paper is organized as follows: We briefly discuss the advantage of omnidirectional tracking and its model in section II, and then we give the models for background and object in section III. In section IV, we analyze some human activity patterns from the trajectories using a time-spatial window based method. Some experimental results are given in section V, and the conclusion is in section VI.

2. The omnidirectional camera model

The catadioptric omnidirectional camera Cyclo-Vision's ParaCamera is applied in our tracking task. The advantage of the catadioptric omnidirectional camera is that it can provide a wide scope of view to improve the objects' visibility from the top in the indoor environment, such as office or shopping mall, because the objects are easily occluded if they are monitored from a corner mounted normal camera. Figure 1 illustrates coverage of an omnidirectional camera in a typical indoor environment.

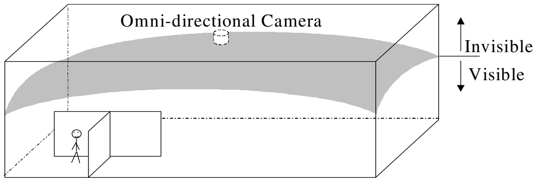


Figure 1. An illustration of indoor tracking with an omnidirectional camera

A general catadioptric camera is an optical system combined with a mirror and lens. For wide view purpose, the mirror is convex, usually a half ball, paraboloid or a cone. The imaging model can be as figure 2(a) or figure 2 (b), which depends on the adjusting of aperture. [1] gives some geometry properties of omnidirectional camera. The main difference between ordinary perspective camera and catadioptric camera is that the resolution of the image captured by the former is homogeneous, while that provided by the later is heterogeneous which has a high resolution in the center and low resolution at the surround, which is similar with creatural vision.

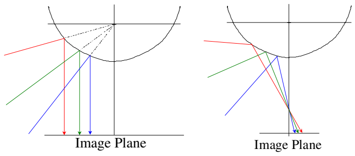


Figure 2. Two types of omnidirectional camera models

3. Tracking model

The purpose of the tracking is to find one or multiple objects that we are interested in and to keep marking them in a 2-D image and further to try to analyze their actions in 3-D space. First, we will model the tracking as an optimal problem based on the background and foreground transformation model.

3.1. The background model

The initialization of the background and its update along with time are two key points in background estimation. Usually, the background can be stable for a period, but it can also be changed in several ways:

1. Gradual lighting change, such as sunset or sunrise;
2. Fade in / fade out lighting change, such as the cloud passing;
3. Suddenly lighting change, such as the lamp being turned on or turned off;
4. Partly background change, such as some objects being added into or removed from the view scope for permanent or a period.

All these cases should be considered in modeling the background. The background can first be regarded as a 2-D field with limited support area, and evolves with time *t*. It is illustrated as Figure 3. The image we get can be regarded as the background image covered with some object's images.

Suppose the support set of the image is $\Lambda = \{(1,1), (1,2), \cdots, (1,n), (2,1), \cdots, (m,n)\}$ and m, n are the height and width of the image respectively. The support set of object i at time t is $\Lambda_t^{O_i} \subset \Lambda$. The background support set at time t is $\Lambda_t = \Lambda \backslash \{\bigcup_i \Lambda_t^{O_t}\}$.

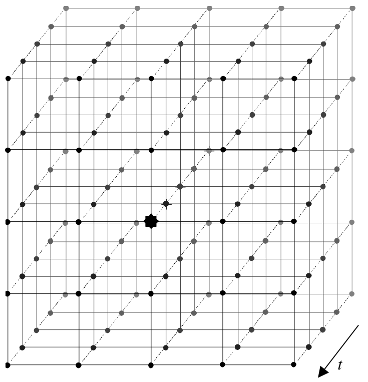


Figure 3. 2-D background grid evolving along with time

The backgrounds at time t and t-1 can be regarded as related, so the estimation of background is given as equation (1).

$$\arg\max_{\Lambda_t, B_t} \{p(B_t, \Lambda_t \mid I_t, B_{t-1})\}, \quad \textbf{(1)}$$

where B_t is the ideal background image at time t, I_t is the observed image at time t. For those position occupied by objects, we have $B_t(i,j) = B_{t-1}(i,j)$ iff. $(i,j) \in \bigcup_k \Lambda_k^{O_k}$. From Bayesian rule, we have

$$P(B_t, \Lambda_t \mid I_t, B_{t-1}) = \frac{P(I_t \mid \Lambda_t, B_t, B_{t-1}) p(\Lambda_t, B_t, B_{t-1})}{P(I_t, B_{t-1})}. \quad \textbf{(2)}$$

Because the observed image I_t is directly from the background B_t at the time t, so we have

$$P(I_t \mid \Lambda_t, B_t, B_{t-1}) = P(I_t \mid \Lambda_t, B_t), \quad \textbf{(3)}$$

which can be considered as the process of an image produced from an ideal background. Usually, we can model the imaging procedure as an independent noise addition.

$$I_t(\mathbf{X}) = B_t(\mathbf{X}) + n(\mathbf{X}) \quad \mathbf{X} \in \Lambda_t,$$

and we assume that the noise is the white Gaussian noise $n(\mathbf{X}) \sim N(0, \sigma^2)$, we have

$$P(I_t \mid \Lambda_t, B_t) = \prod_{\mathbf{X} \in \Lambda_t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{[B_t(\mathbf{X}) - I_t(\mathbf{X})]^2}{\sigma^2} \right\}. \quad \textbf{(4)}$$

$p(\Lambda_t, B_t, B_{t-1})$ describes the transformation from B_{t-1} to B_t at the support set Λ_t. From Hammersley-Clifford Theorem [2], we can have

$$p(\Lambda_t, B_t, B_{t-1}) = \frac{1}{Z_p} \exp[-U(\Lambda_t, B_t, B_{t-1})/T], \quad \textbf{(5)}$$

where $Z_p = \sum_{\Lambda_t, B_t, B_{t-1}} \exp[-U(\Lambda_t, B_t, B_{t-1})/T]$ is the partition function, T is a temperature related parameter, which controls the speed of background evolution, and $U(\Lambda_t, B_t, B_{t-1}) = \sum_c V_c(\Lambda_t, B_t, B_{t-1})$ is the energy function base on the clique c, where we use a causal first order clique as shown in figure 4, and the clique energy at position (i, j) is defined as,

$$U(\Lambda_t, B_t, B_{t-1}) = \sum_{\mathbf{X} \in \Lambda_t} \left[\frac{\partial B_t(\mathbf{X})}{\partial t}\right]^2 + \left[\frac{\partial^2 B_t(\mathbf{X})}{\partial t \partial x}\right]^2 + \left[\frac{\partial^2 B_t(\mathbf{X})}{\partial t \partial y}\right]^2. \quad \textbf{(6)}$$

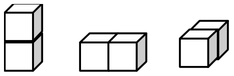


Figure 4. First order clique

From equations (1)-(6), the background estimation is gotten as following,

$$\arg\min_{\Lambda_t, B_t} \left\{ \sum_{\mathbf{X} \in \Lambda_t} E_N(\mathbf{X}) + E_P(\mathbf{X})/T \right\} \quad \textbf{(7)}$$

where,

$$E_N(\mathbf{X}) = \frac{[B_t(\mathbf{X}) - I_t(\mathbf{X})]^2}{\sigma^2},$$

$$E_P(\mathbf{X}) = \left[\frac{\partial B_t(\mathbf{X})}{\partial t}\right]^2 + \left[\frac{\partial^2 B_t(\mathbf{X})}{\partial t \partial x}\right]^2 + \left[\frac{\partial^2 B_t(\mathbf{X})}{\partial t \partial y}\right]^2.$$

It is easy to get a trivial solution $\Lambda_t = \varnothing$ if we only have this background model (7). This is almost impossible for a tracking system unless the object is too close to the camera. We will give the object model as another constraint.

3.2 The object model

The object model is divided into two parts: the appearance model and motion model. The appearance of the object is changed while the object itself or some other objects move (occlusion), lighting condition change, etc. The appearance changes not only its silhouette but also the intensity distribution, which can be modeled as a MRF. Usually, the object's motion keeps continuously in both speed and direction, and the change of speed and direction is smooth, which can be formed with Kalman filter.

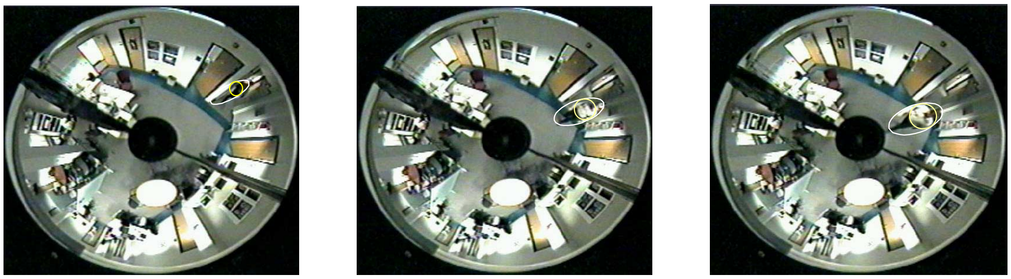


Figure 5. An example of object size changing as the position changes

Suppose the position of the object i in a 2-D image is $\mathbf{X}_t^i$ at time t. From the Kalman filter, we have

$$\begin{aligned}\mathbf{X}_{t+1}^i &= \Phi_{t+1}\mathbf{X}_t^i + \eta_{t+1}^i \\ \mathbf{D}_{t+1}^i &= \mathbf{H}_{t+1}\mathbf{X}_{t+1}^i + \xi_{t+1}^i\end{aligned}, \quad \textbf{(8)}$$

where, $\mathbf{H}_{t+1}^i$ is the identity matrix. $\eta_{t+1}^i \sim N(0,\mathbf{Q}_{k+1}^i)$, $\xi_{t+1}^i \sim N(0,\mathbf{R}_{k+1}^i)$, and $E[\eta_{t+1}^i \xi_{t+1}^i] = 0$.

The object appearance model is similar to the background model after motion compensation. The appearance model can be expressed as

$$\arg\min_{\Lambda_t^i, O_t^i} \left\{ \sum_{\mathbf{X}\in\Lambda_t^i} E_N(\mathbf{X},\dot{\mathbf{X}}_t^i) + E_P(\mathbf{X},\dot{\mathbf{X}}_t^i)/T \right\}, \quad \textbf{(9)}$$

where

$$E_N(\mathbf{X},\dot{\mathbf{X}}_t^i) = \begin{cases} \dfrac{\left[B_t(\mathbf{X}) - B_{t-1}(\mathbf{X}-\dot{\mathbf{X}})\right]^2}{\sigma^2} & \text{iff.} \mathbf{X}\in\Lambda_t^i \text{ and } \mathbf{X}-\dot{\mathbf{X}}\in\Lambda_t^i \\ \left[\dfrac{\max\left(B_t(\mathbf{X}), B_{t-1}(\mathbf{X}-\dot{\mathbf{X}})\right)}{\sigma}\right]^2 & \text{otherwise} \end{cases},$$

and

$$E_P(\mathbf{X},\dot{\mathbf{X}}_t^i) = \left[B_t(\mathbf{X}) - B_{t-1}(\mathbf{X}-\dot{\mathbf{X}})\right]^2 + \left[\frac{\partial B_t(\mathbf{X})}{\partial x} - \frac{\partial B_{t-1}(\mathbf{X}-\dot{\mathbf{X}})}{\partial x}\right]^2 + \left[\frac{\partial B_t(\mathbf{X})}{\partial y} - \frac{\partial B_{t-1}(\mathbf{X}-\dot{\mathbf{X}})}{\partial y}\right]^2.$$

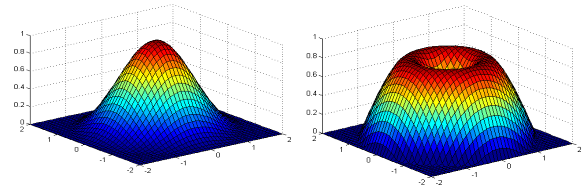
Therefore, the tracking problem can be formulated as minimizing both background model (7) and object model (9), we can model the tracking as equation (10).

$$\arg\min_{\Lambda_t, B_t, \Lambda_t^i, O_t^i} \left\{ \begin{aligned} &\sum_{\mathbf{X}\in\Lambda_i} E_N(\mathbf{X}) + E_P(\mathbf{X})/T \\ &+\lambda \sum_i \sum_{\mathbf{X}\in\Lambda_t^i} E_N(\mathbf{X},\dot{\mathbf{X}}_t^i) + E_P(\mathbf{X},\dot{\mathbf{X}}_t^i)/T_i \end{aligned} \right\} \quad \textbf{(10)}$$

λ is a factor.

3.3 Adapting object model to an omnidirectional camera

As we have mentioned in section II, the catadioptric omnidirectional camera provides a deformative view. Under the view of the omnidirectional camera, the shape and size of the object will be changed with the view angle's changes. Figure 5 is such an example. The above object model is based on perpendicular projection, and this model can also be used for a perspective projection when an object is far enough from the camera. In order to use the object model for omnidirectional tracking, view angle-based compensation must be done before object match. Based on the characteristic of the camera we used, we calculate the factor of compensation as figure 6. Figure 6(a) is the factor in X and Y direction and Figure 6(b) is the factor in Z direction which is the optical axis direction. We assume that the object movement is only in X and Y direction in figure 6. The captured scene in the image is located within a circle, whose radius is 2f, and f is the focus length of the paraboloid. The object's dimension in X and Y direction will be maximum when the object is located on the optical axis, and will disappear at the circle. The object's dimension in Z direction will be invisible at the center and the circle, and will reach maximum at the radius 0.8629f. We can estimate the size changes for each object using such a match factor map.



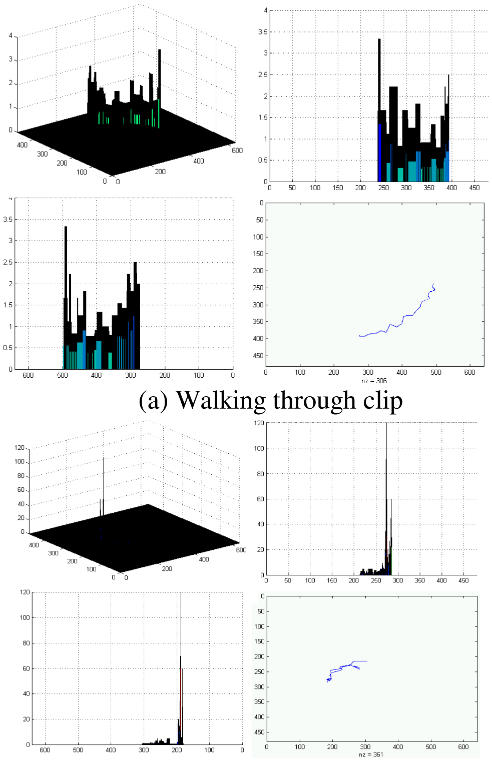
(a) factor in X-Y direction (b) factor in Z direction

Figure 6. Horizontal and vertical match factor map for the omnidirectional camera

4. Activity and scene modeling from trajectories

When a person is monitoring a spot, he can easily judge an object's purpose even if he can't see the object in detail, which implies that we can monitor object's behavior from trajectory, therefore we define a hierarchical behavior model. At the lowest level of the model, it contains essential information such as moving or stopping and sitting or standing, which can be observed from the tracking sequence. At a higher level we can distinguish some different activities, such as working alone or having a meeting, etc., which can't be observed directly. These activities can be observed via tracking moving trajectories of people in a scene. For example, we can define a meeting as two or more trajectories coming from same or different directions and staying in the scene for a period of time.

We use a time-spatial window to analyze individual trajectory. 5 seconds' duration is used as the time window. The trajectory within this time window forms the spatial window. The time overlap window is used for each clip, and the overlap time is 2.5s. If the object stays at a spot for a period, we just accumulate the histogram. From the histogram we can infer some human activities. Figure 7 shows some typical patterns for different activities. The top-left in each group is the spatial histogram. The top-right and bottom-left are horizontal and vertical view of the histogram, and the bottom-right is the top view of the trajectory.



(a) Walking through clip

(b) Looking for something clip

Figure 7. Examples of activity patterns

5. Experiments and discussion

The sampling rate is 12 fps in our experiments. The first experiment is to build up a background model in a dynamic environment. Building up a background is the most important in tracking using the background subtraction method. The background's evolution speed is controlled by the temperature *T* in the MRF model. We use a high temperature at the beginning, and then gradually replace it with a lower temperature. The algorithm of building up the background is adaptive to various scene conditions including background only sequence, lighting condition change sequence, and objects on spot sequence. Figure 8 is an example of background building with an object on the spot. The first row of images is the two image correspondence to the begin and the end of background setup. The other two rows are the under building background. At the beginning of the background setup, some black areas within the circle are under construction, and we can dynamically update the background while tracking the object and obtain a complete background when the object is out of the spot.

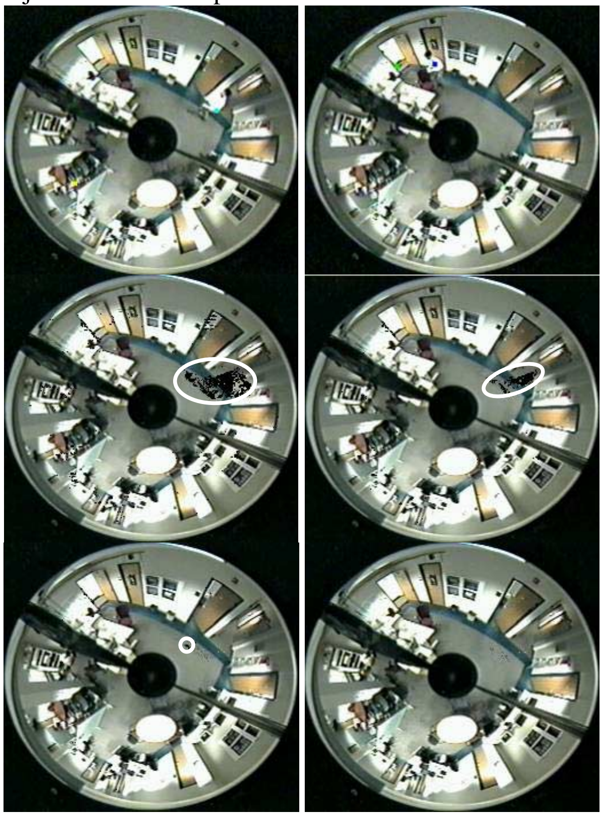
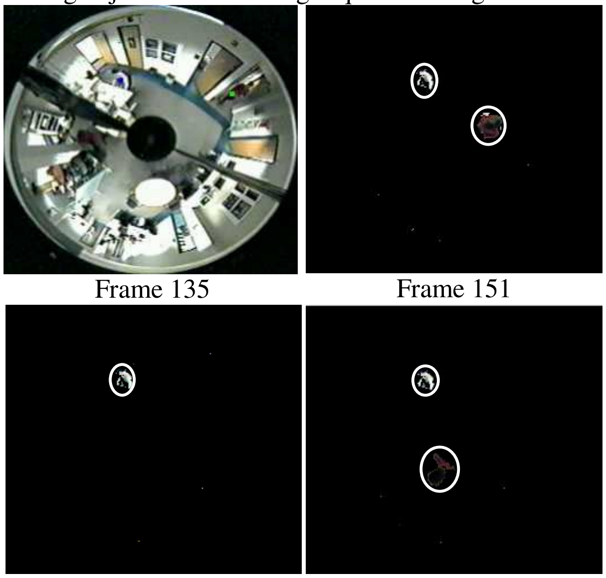


Figure 8. Background evolutions with time changing

Another experiment is to test the ability of tracking and monitoring an object with changes in size and lighting, and the ability to monitor multiple objects. Figure 9 is an example of multi-object tracking. Only the top left one with the background, the others are only moving objects. Although one of the objects passes through the blind zone as the frame 155, it is continuously tracked by the system. This shows the beneficial properties of the Kalman filter. In this example, we track the stable and moving object at the same time. It can be seen that the size and lighting of the moving object in the tracking sequence change.



Frame 135 Frame 151

Frame 155 Frame 159

Figure 9. An example of multi-object tracking

6. Conclusion

In this paper, we have proposed to employ MRF models to represent the background and foreground in monitoring human activities using an omnidirectional camera. We use a MRF background model to effectively update environmental changes. We combine the MRF based object model with a motion model to robustly track human motion in a view of an omnidirectional camera. We have presented a method of monitoring human activities by analyzing the trajectory of motion using a time-spatial window. The further work includes: developing more complex behavior models from moving trajectories; automatically estimating geometric parameters of people and scene from an omnidirectional camera; and combining tracking results of an omnidirectional camera with ordinary cameras for robust video surveillance.

Acknowledgement

This research is partially supported by the National Science Foundation under grant No. IIS-0121560.

References

[1] S. Baker, and S. K. Nayar, A theory of catadioptric image formation, Proc. of the Int. Conf. on Computer Vision, pp.35-42, 1998.

[2] J. Besag, Spatial interaction and the statistical analysis of lattice system, Journal of the Royal Statistical Society, Series B, Vol. 36, No. 2, pp. 192-236, 1974.

[3] T. E. Boult, R. Michaels, X. Gao, P. Lewis, C. Power, W. Yin, and A. Erkan, Frame-rate ominidirectional surveillance and tracking of camouflaged and occluded targets, Proc. of 2nd IEEE Workshop on Visual Surveillance, pp. 48-55, 1999.

[4] R. T. Collins, A. J. Lipton, H. Fujiyoshi and T. Kanade, Algorithms for cooperative multisensor surveillance, Proc. of the IEEE, Vol. 89, No. 10, pp. 1456-1477, 2001.

[5] T. Darrell, D. Demirdjian, N. Checka, P. Felzenswalb, Plan-view trajectory estimation with dense stereo background models**,** Proc. of the Int. Conf. on Computer Vision*,* Vol 2, pp. 628-635, 2001.

[6] J. Davis and A. Bobick, The representation and recognition of human movements using temporal templates, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 928–934, 1997.

[7] C. Fermüller, Y. Aloimonos, P. Baker, R. Pless, J. Neumann and B. Stuart, Multi-camera networks: eyes from eyes, Proc. of the IEEE Workshop on Omnidirectional Vision, pp. 11-18, 2000.

[8] J. Foote and D. Kimber, FlyCam: practical panoramic video, Proc. of the Int. Conf. on Multimedia and Expo, pp. 1419-1422, August 2000

[9] I. Haritaoglu, D. Harwood, and L. Davis, W4: real-time surveillance of people and their activities, IEEE Trans. on PAMI, Vol. 22, No. 8, pp.809-830, 2000.

[10] I. Haritaoglu, D. Harwood, and L. S. Davis, Hydra: multiple people detection and tracking using silhouettes, Proc. 10th Int. Conf. on Image Analysis and Processing, pp. 280-285, 1999.

[11] R. Krumm, S. Harris, and B. Meyers, B. Brumitt, M. Hale, and S. Shafer, Multi-camera multi-person tracking for easyliving. Third IEEE International Workshop on Visual Surveillance, pp. 3-10, 2000.

[12] S. K. Nayar, Catadioptric omnidirectional camera, Proc. of 1997 IEEE Conference on Computer Vision and Pattern Recognition, pp. 482-488, 1997.

[13] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. A. Jones, A multi-agent framework for visual surveillance, Proc. of Int. Conf. on Image Analysis and Processing, pp. 1104-1107, 1999.

[14] R. Polana and R. Nelson, Low level recognition of human motion, Proc. of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pp. 77-82, 1994.

[15] D. W. Rees, Panoramic television viewing system, United state patent, No. 3,505,465, April, 1970.

[16] J. M. Rehg, M. Loughlin, and K. Waters, Vision for a smart kiosk, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 690-696, 1997.

[17] R. Rosales and S. Sclaroff, 3D trajectory recovery for tracking multiple objects and trajectory guided recognition of actions, Proc. of 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 117-123, 1999.

[18] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, Pfinder: real-time tracking of the human body, IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 780–785, 1997.