Generating Pictorial-Based Representation of Mental Images for Video Monitoring

Chuan-Heng Hsiao[1] Wei-Chia Huang[1] Kuan-Wen Chen[1] Li-Wei Chang[2] Yi-Ping Hung[1]
[1]Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan
E-mail:{chshou, hung}@csie.ntu.edu.tw
[2]Tepper School of Business
Carnegie Mellon University, Pittsburgh, PA, USA
E-mail: liweic@andrew.cmu.edu

ABSTRACT
Multi-camera systems have been widely used in many video surveillance applications. When an event happens and is monitored across multiple cameras, it is easy for an expert to generate the corresponding spatial representation to comprehend the series of event. However, it is not trivial for users new to the environment. With support from psychological evidences, we propose an approach to mimic generating pictorial-based representation of mental images when a target is moving across the views of cameras. First we conduct a ball-rolling experiment to compare this approach with others. The empirical results demonstrate that the performance of users with this approach is significantly better than others. We suggest that it is because this approach is better for users to preserve spatial representation of the environment while transiting views between cameras. Then we propose a framework to realize this approach. The demonstrations in different situations indicate the validity of such framework.

Author Keywords
Mental images, surveillance, user study, intelligent visualization.

ACM Classification Keywords
H5.2 [Information interfaces and presentation]: User Interfaces.

INTRODUCTION
Multi-camera systems have been widely used in many video surveillance applications, such as airports, banks, and other famous buildings and places. When an interesting event happens, usually it is monitored with more than one camera. For example, a suspect may walk through the corridors and across the views of multiple cameras. Observers watch multiple video streams and integrate information to understand the series of events.

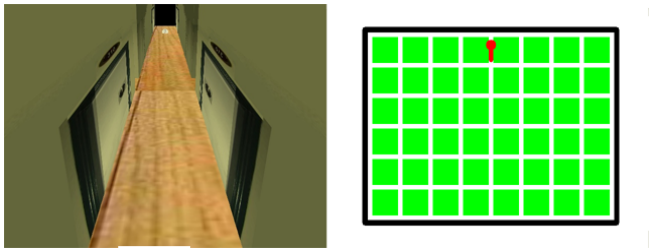
IUI'09, February 8–11, 2009, Sanibel Island, Florida, USA.

It is a common practice to show multiple streams on display simultaneously. However, the characteristics of human vision system allow observes to pay attention to one video at a time [8]. Such approach requires observers to adjust the view point of the spatial representation of the monitored environment internally when switching the view from one camera to another. The adjustment can be very difficult if the environment is sophisticated, or the observer is unable to understand the geometrical relationship among cameras very quickly.

Girgensohn et al. [3] suggested that compared with single main screen, users prefer the arrangement that putting the views of neighboring cameras beside the main screen geometrically correspondingly for better spatial representation. However, their approach requires users to specify the views where the targets appear. For an intelligent video surveillance system which is capable of detecting and tracking interesting events, the corresponding views switch automatically along with the detected events, and users may not be able to follow the changes in time.

Our goal is to generate mental images for the users to monitor the video streams obtained from a multi-camera network in a more intuitive way. Mental image is a classical concept appeared in cognitive science textbooks [7]. According to Stanford Encyclopedia of Philosophy [14], mental imagery resembles perceptual experience, but occurs in the absence of the appropriate stimuli, and mental image is the picture-like representation of visual mental imagery. D'Esposito et al. [2] provided physiological evidence showing that visual association cortex engaged during the generation of mental images. Slotnick et al. [13] further demonstrated that visual mental imagery induced retinotopically organized activation of early visual areas in human brain. With the definition of mental images mentioned above, to understand the series of the event when watching multiple video streams, we assume that observers generate corresponding mental images to interpolate the views among cameras to understand the series of the event when comprehending an event in multi-camera systems. For simplicity, pictorial-based representation of mental images will be abbreviated as "*PRMI*" in the following of this paper.

For an intelligent multi-camera surveillance systems which can automatically detect and track the events, like an expert



(a) (b)

Figure 1. Experiment Settings. More details are illustrated in text.

guiding novice users monitoring the environment, we suggest that it will be helpful for observers to comprehend the series of events if a virtual camera, which is termed "*mental camera*" in this paper, can mimic generating the series of PRMI between cameras along with the series of events and directly show the PRMI on display.

Shepard and Metzler [12] demonstrated the ability of mental rotation and the linear relationship between the reaction time and the angle of rotation. In addition, Kosslyn et al. [6] demonstrated that the reaction time is proportional to the distance between two locations in an imagined map. These two psychological evidences imply that given the transition route, the extrinsic parameters of mental camera can be interpolated linearly when switching the view from one camera to another.

In the following section, we design a ball-rolling experiment to compare the performance among baseline method (single-main-screen method), Girgensohn et al.'s [3] method, and PRMI method (presenting based on analogy to PRMI).

EXPERIMENT: BALL ROLLING IN A MAZE

In this experiment, a virtual ball was rolling along the corridors in a virtual environment. Participants were required to determine the end point of the ball. The reason to conduct the experiment in a virtual environment was that we could easily manipulate multiple cameras in a complex environment.

For the Girgensohn et al.'s [3] method, although they mentioned the benefits of displaying the main screen along with a 2-D map to present the spatial configurations of the environment, what we focused on here was how the main screen presents to the users. In some extreme situations, the users may concentrate on the behavior of the target tracked on the main screen, and it is difficult to monitor both the main screen and the 2-D map at the same time. Therefore, in our experiment, as shown in Figure 1, only the main screens of the three different methods were presented on the left to the participants and participants responded in the corresponding map on the right.

Apparatus and Participants

A 17" LCD display with IBM-PC compatible system was adopted, and 9 university-level students participated in this experiment.

Stimuli

Figure 1 demonstrates the spatial organization of the virtual space, where the 3D structure is presented in Figure 1(a), and the map of the whole virtual space is presented in Figure 1(b). The virtual space consisted of 8 × 6 square rooms, and each room was 20 × 20 × 10 feet. The textures of the walls were all the same, but with different background colors in different regions. The background colors were not shown on the map. There were 6-foot wide corridors between rooms. Since there were 8 × 6 square rooms, there were 9 × 6 blocks of vertical corridors and 8 × 7 blocks of horizontal corridors with 7 × 5 intersections.

As illustrated in Figure 1(a), a 2-foot-in-diameter ball was placed in the center of the corridors. There were textures on the ball to help participants visualize the ball rolling on the corridors. For each trial, initially the ball was placed on the top-center of the vertical corridor and rolled downward at a speed of 8 feet per second, as indicated as the red spot and red line on the map. When the ball entered into an intersection, the ball may move forward, turn left, or turn right at the intersection. The ball moved along six blocks and then stopped at an intersection.

Cameras were put in every center of the corridors. Therefore, there were total 110 cameras. Each camera was placed at a height of 9 feet and tilted downward 40 degrees. For the cameras in the vertical corridors, they may view upward or downward. For the cameras in the horizontal corridors, they may view leftward or rightward. The views of the cameras were facing along the rolling route of the ball.

While the ball was rolling along the corridors, the view on the main video screen automatically switched to the view of the new corresponding camera as the ball was leaving the field of view of the main camera. There were three methods presenting transitions between cameras. The baseline method was that the video immediately switched to the view of the new camera. Girgensohn et al.'s method in this experiment was that there were views of neighboring cameras showing at geometrically corresponding neighboring regions of the screen beside the view of main camera, and the views switched immediately to the view of the new cameras. PRMI method was that the video smoothly switched to the view of the new camera for 0.25 second with linear interpolation of the parameters between cameras over time.

Procedure

For each trial, first the video of the ball rolling along the corridors was demonstrated on the left of the screen and the map was shown on the right of the screen. When the video finished, the video frame disappeared and a "Go" phrase

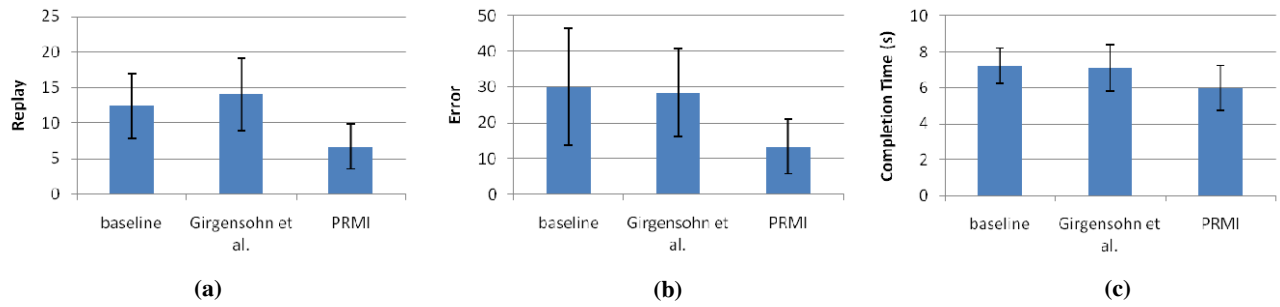


Figure 2. Results of the Experiment. More details are illustrated in text.

was shown on the left screen, and participants clicked on the intersections of the map.

The trial was completed as the participants clicked on the correct intersection of the end point of the rolling route. Otherwise, the clicked wrong intersections were marked on the map. Participants could press right button and the video replayed again.

Before the experiment, participants were instructed and were required to practice on two routes for each method to familiarize the procedure. One practicing route was a straight route, and the other was with one turn. A welcome screen was shown on display in the beginning of the experiment, and participants pressed the button and the first trial began. After finishing each trial, participants could take a break and then pressed the button again, and the next trials continued. A "Thank you" screen appeared in the end of the experiment.

There were 12 routes for each method. Therefore, there were total 36 trials during the whole experiment. Each route consisted of at least 3 turns. The order of the trials was randomized. Participants were specifically notified that the completion time was measured.

To avoid complicated evaluation, for each trial, the completion time was measured as from the end of the last time of the video played to the corrected intersection clicked. In addition, only the participants who answered more than 8 trials correctly (no replays and no errors) in each method would be considered.

Hypothesis

Three hypotheses are proposed as follows:

(H1) Among the three methods, the number of replays is the least in PRMI method.

(H2) Among the three methods, the number of errors is the least in PRMI method.

(H3) Among the three methods, the completion time is the least in PRMI method.

The reasons for the above three hypothesis is based on the assumption that PRMI method provides more direct spatial organization of the whole environment to help users represent the space in mind more easily.

Results and Discussion

The result of number of replays is demonstrated in Figure 2(a). The result of number of errors is demonstrated in Figure 2(b). The result of the completion time is demonstrated in Figure 2(c). 5 participants meet the criterion for evaluating completion time.

One-way ANOVA is evaluated in this experiment. From the results, we can observe that the performance with PRMI method is significantly better than the other two methods in number of replays ($F = 9.835$, $p < 0.002$) and number of errors ($F = 4.175$, $p < 0.05$). We suggest that with PRMI, users can obtain spatial representation of transition trajectory more directly. Therefore, users can directly follow the trajectory and need no additional effort to generate the spatial representation internally.

On the other hand, the difference of performance between baseline method and Girgensohn et al.'s method is not statistically significant. We suggest that it is because the system automatically switched the views. Therefore, users may not follow which view of the neighboring camera will be the next on the main screen. The other reason is that because the spatial organization is very similar for every camera, the views of the neighboring cameras may not provide additional spatial information for the users.

The result of the completion time is not significantly different among the three methods ($F = 1.788$, $p > 0.2$). We suggest that one of the reasons is because we only measure the time from the end of the last replayed video to clicking the correct intersection. Therefore, the result may be confounded with that some participants may replay the video for several times until they are certain of the answer, and click on the correct intersection quickly.

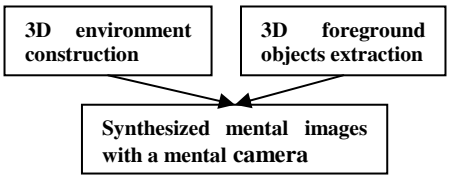


Figure 3. Framework of realizing PRMI for video monitoring.

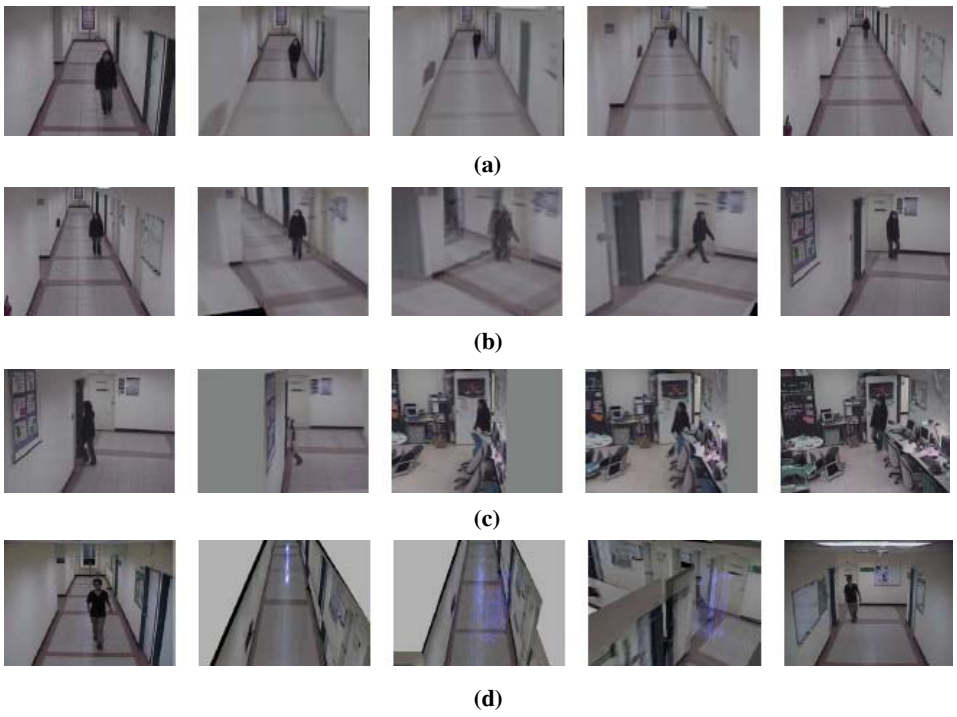


Figure 4. Some generated PRMI in different situations. More details are illustrated in text.

REALIZING PRMI FOR VIDEO MONITORING

Based on the observation of the confirmed hypotheses, we propose a framework capable of automatically detecting targets and generating PRMI along the series of events.

Figure 3 illustrates the general framework of generating PRMI between cameras. Given two cameras and a view transiting from one camera to the other, first the intrinsic and extrinsic parameters of cameras are obtained [15], and the 3D geometry model of the environment is offline pre-constructed [5]. Then the system is put online, and computer vision techniques are adopted to detect targets automatically [4, 9]. With the assumption that the targets are standing on the floor, the 2D images of the targets can be mapped to the 3D geometry model of the environment. Finally, based on the support from psychological evidences that the linear relation between the rotating angle and reaction time [12] and the reaction time is proportional to the distance [6], the pre-constructed 3D geometry model of the environment and the 3D information of the targets are synthesized with a mental camera, whose parameters are the linear interpolation of the parameters of the real cameras over time [1, 11]. The framework can be improved when the components in the framework, such as detection and tracking techniques in computer vision, are improved.

Figure 4 demonstrates some of our implementation of the framework in different situations, where the leftmost and the rightmost images are the images from real cameras, and the middle three images are some generated PRMI between the cameras. It is strongly recommended to watch video for the impression of our implementation (http://www.csie.ntu.edu.tw/~chshou/vt.wmv). In our implementation, the processing speed can reach 22 fps with an Intel Pentium IV 3.4 G IBM-PC compatible system.

The scenario in Figure 4(a) is that one camera is behind the other and both are viewing in same direction, where the size of the target showing on the images is quite different. The scenario in Figure 4(b) is that the two cameras are viewing in orthogonal directions. The scenario in Figure 4(c) is that the views of the two cameras are blocked by a wall. The scenario in Figure 4(d) is that there are some areas which are not viewed from either camera, and possible locations of the target are then modeled with blue lines [9].

CONCLUSION AND FUTURE WORK

Multi-camera systems have been widely used in many video surveillance applications. With support from psychological evidences, we propose an approach to mimic generating pictorial-based representation of mental images when an event happens across the views of cameras. First we conduct the ball-rolling experiment to compare among baseline method, Girgensohn et al.'s method, and PRMI method. The results demonstrate that the performance of users with PRMI method is statistically significantly the best among the three methods. We suggest that it is because users can obtain spatial information more directly with PRMI method. Then based on the observation from the empirical results and support from psychological evidences, we propose a framework to realize this approach. The demonstrations in four different situations indicate the validity of such framework.

There are several issues which we can explore further in the future. First of all, it is common that there are multiple targets moving around the whole environment, which is not discussed in this paper. One possible solution is that the system can track the targets simultaneously and present them side by side. Unlike camera-based arrangement of sub-windows, this setting will be target-based arrangement of sub-windows. In addition, we only used linear interpolation between cameras when transiting one camera to another. This interpolation implies the straight transition route between cameras. We can explore further how to

design transition routes which can really represent the transition routes of the mental images of an expert in different environment settings. Besides, with the improvement of related computer vision techniques, such as detection and tracking, the proposed framework can be improved.

ACKNOWLEDGMENTS

This work was supported in part by the Excellent Research Projects of National Taiwan University, under grant 97R0062-04, and by the Ministry of Economic Affairs, Taiwan, under Grant 96-EC-17-A-02 -S1-032.

REFERENCES

1. Debevec, P. and Malik, J., Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image- Based Approach. In *Proc. of SIGGRAPH* (1996), 11-20.
2. D'Esposito, M., Detre, J. A., Aguirre, G. K., Stallcup, M., Alsop, D. C., Tippet, L. J., and Farah, M. J., A Functional MRI Study of Mental Image Generation. *Neuropsychologia*, 35 (1997), 725-730.
3. Girgensohn, A., Shipman, F., Turner, T., and Wilcox, L., Effects of Presenting Geographic Context on Tracking Activity between Cameras. In *Proc. of CHI* (2007), 1167-1176.
4. Hall, D., Nascimento, J., Ribeiro, P., Andrade, E., Moreno, P., Pesnel, S., List, T., Emonet, R., Fisher, R., Santos Victor, J., and Crowley, J. L., Comparison of Target Detection Algorithms Using Adaptive Background Models. In *International Workshop on Performance Evaluation of Tracking and Surveillance* (2005).
5. Hartley, R. I. and Zisserman, A., *Multiple View Geometry*, 2nd ed, Cambridge University Press (2004).
6. Kosslyn, S. M., Ball, T. M., and Reiser, B. J., Visual Images Preserve Metric Spatial Information: Evidence from Studies of Image Scanning. *Trends in Neurosciences*, 4 (1978), 47-60.
7. Medin, D., Ross, B., Markman, A., *Cognitive Psychology*, 3rd ed. (2000), Harcourt College Publishers.
8. Palmer, S. *Vision Science: Photons to Phenomenology*, MIT Press (1999).
9. Prati, A., Mikic, I., Trivedi, M. M., and Cucchiara, R., Detecting Moving Shadows: Algorithms and Evaluation. *IEEE Transactions on PAMI,* 25 (2003), 918-923.
10. Reeves, W. T., Particle systems – A Technique for Modeling a Class of Fuzzy Objects. *ACM Transactions on Graphics*, 2 (1983), 91-108.
11. Seitz, S. M. and Dyer, C. R., View Morphing. In *Proc. of SIGGRAPH* (1996), 21-30.
12. Shepard, R. N. and Metzler, J., Mental Rotation of Three-Dimensional Objects. *Science,* 171 (1971), 701-703.
13. Slotnick, S. D., Thompson, W. L., and Kosslyn, S. M., Visual Mental Imagery induces Retinotopically Organized Activation of Early Visual Areas. *Cerebral Cortex*, 15 (2005), 1570-1583.
14. Thomas, N. J., Mental Imagery. In Zalta, E. N., ed. *The Stanford Encyclopedia of Philosophy* (Fall 2007). http://plato.stanford.edu/archives/fall2007/entries/mental-imagery
15. Zhang, Z., A Flexible New Technique for Camera Calibration. *IEEE Transactions on PAMI*, 22 (2000), 1330-1334.