How panoramic visualization can support human supervision of intelligent surveillance

Alexander M. Morison David D. Woods
Human Systems Integration
Ohio State University
Columbus, OH

James W. Davis
Dept. of Computer Science and Engineering
Ohio State University
Columbus, OH

In video-based surveillance people monitor a wide spatial area through video sensors for anomalous events related to safety and security. The size of the area, the number of video sensors, and the camera's narrow field-of-view make this a challenging cognitive task. Computer vision researchers have developed a wide range of algorithms to recognize patterns in the video stream (intelligent cameras). These advances create a challenge for human supervision of these intelligent surveillance camera networks. This paper presents a new visualization that has been developed and implemented to integrate video-based computer vision algorithms with control of pan-tilt-zoom cameras in a manner that supports the human supervisory role.

SUPERVISION OF SMART CAMERAS

In every day activities people move through many areas monitored by security cameras. Surveillance centers use these cameras to meet safety and security goals by looking for unusual human activities and anomalous events in the physical areas under camera surveillance (Haering et al., 2008). Surveillance centers tend to be quite similar. One or more walls of monitors show live video feeds from sensors (e.g., video cameras, news organizations). The centers are staffed 24/7; interesting events occur relatively infrequently, and the centers communicate with security personnel in the physical world and with other security and safety related organizations.

The video camera, which is still the workhorse sensor technology in surveillance, can be thought of as a stand-in for human security personnel. Each video sensor, fixed mounted but with pan, tilt, and zoom (PTZ) capability, can monitor a portion of the total space to be surveilled, which can be spatially large. Security personnel, rather than patrol the physical area, often monitor video feeds and use their experience to look for and identify anomalous patterns of activity. The task of monitoring is challenging given the large area under surveillance, difficulties in recognizing anomalous activity (e.g., high context sensitivity), a low base rate of anomalous events, a large number of video sensors feeding the surveillance center, and the ability to capture only a fraction of the entire area at any given moment.

One goal of computer vision algorithms in video surveillance is to reduce the need for and the burden on human security personnel by creating smart automation to monitor the array of sensor feeds. The result has been the development and deployment of intelligent algorithms to detect human motion (Gavrila, 1999), track people moving through a scene (Aggarwal and Cai, 1999), and analyze the types of motions people carry out (Wang et al., 2003). The design question is how to couple responsible supervisory human security personnel to the results of the algorithmic analysis of the actual video feeds from the scene, or, in other words, what kinds of supervisory displays are needed for smart surveillance systems?

This paper presents a new concept for supervisory visualizations of smart surveillance systems designed for single smart PTZ cameras. The visualization is based on a static panoramic frame of reference that captures the entire space of views for a single smart PTZ camera. This is overlaid with a brightness coded activity map that represents outputs from smart algorithms monitoring for human activity. Scan-path algorithms are overlaid on to this capturing how the PTZ camera will monitor the space based on the output of smart algorithms.

The scan panoramic display serves as a longshot display (Woods, 1984; Woods and Watts, 1997) for human supervisors to understand how diverse sets of smart algorithms capture the flow of human activity through a physical space being monitored. In addition, the display also conveys how the scan-paths of a camera will progress over a scene based on the output of these smart algorithms. A longshot display provides an overview of the system status, orientation, and movement between detailed views. The longshot is always displayed in parallel with detailed views to minimize the attention re-orienting costs associated with moving between isolated detailed views.

The panoramic representation also provides a base for human supervisors to interact with the smart camera (visual programming interface). This provides a means to meet the directability functional requirement for human supervision of automation (Woods and Hollnagel, 2006).

The software to create these visualizations has been built and tested on actual video feeds from PTZ cameras that monitor human activities on the Ohio State University campus (see Figures 2(a-c)) for a smart algorithm (detecting translating motion) and three different scan-path algorithms (Davis et al., 2007a,b). In addition, the base panorama visualization can be built automatically and continuously (Sankaranarayanan and Davis, 2008a,b).

COGNITIVE COMPLEXITIES

This section highlights some of the issues in human supervision of smart video surveillance by exploring the analogy between experienced human security personnel moving through the scene of interest and intelligent processing of the video stream from multiple cameras placed in and around the scene of interest.

Patrolling

The patrol or in-scene agent is embedded on the ground within the physical environment they are observing. The in-scene patrol officer directly perceives the world they are moving through as a continuous physical topology. The scene of interest is a series of views the patrol officer takes and the path he or she follows. The observing behavior of the patrol officer defines a field of view (FOV) that is not scalable (i.e., cannot externally expand or shrink the FOV). The in-scene patrol officer is sensitive to the temporal evolution of activity only at a human scale (i.e., cannot see patterns of activity defined over different temporal scales, e.g., the last month). Additional constraints, such as physical structures, layout of physical forms, and the environment influence what an in-scene agent can observe and where they can move. For an in-scene agent, all possible view directions are, for our purposes, represented by a full sphere. Also, the maximum distance between consecutive points of observation is defined by the type of environment (e.g., city, suburb, etc.) and the mode of transportation (e.g., by foot, segway, or car). Finally, the embedded agent has the ability to directly interact with the environment by moving objects, speaking with people, and by visible presence.

Surveilling

The out-of-scene agent understands, moves around, and interacts with the world with a different set of constraints. Within the surveillance center, personnel are external to the environment being observed. Consequently, they do not have a single perceptual experience but multiple, narrow "keyhole" views generated by video sensors on the world. These video sensors, depending on their configuration in the world, can create opportunities to view the world at multiple spatial scales (e.g., from the rooftops of buildings at differing heights). In addition, a region of interest is spatially scalable by organizing a set of cameras (i.e., 2 or more) to observe an area larger than the viewable field of any single camera. Surveillance should be sensitive to events defined over multiple temporal scales from extremely short (notice the bag left behind) to extremely long (the organization of a protest gathering). Activities of interest can also play out across multiple temporal and spatial scales, such as when a small organized protest in one place interacts with other events that transforms the situation into a chaotic violent confrontation that spills out over a wider area.

Movement through the environment differs for an an out-of-scene agent as compared to an in-scene agent. Interaction with the world for an out-of-scene agent consists of switching between different cameras. But there are a large number of video feeds mapped onto a set or even a wall of display monitors creating the potential for a form of data overload. Selecting among the feeds creates, in some sense, a virtual patrol even though the sequence of selecting camera feeds can create tortuous paths and jumps. The virtual paths are not constrained by the spatial topology, rather, only by the configuration of the sensor network and the method of monitoring and controlling the network (e.g., through a mapping of camera feeds to monitors with a single control for all cameras). Distance between points of observation is no longer meaningful given the structure of the control room (i.e., wall of monitors). View directions are restricted to a downward pointed hemisphere for the majority of PTZ cameras, as opposed to the full sphere for the in-scene agent. The context for out-of-scene surveillance creates risks for impaired spatial understanding of the actual physical environment, relationships, and activities.

Smart surveilling

Intelligent algorithms create an opportunity to overcome the complexities that arise from trying to understand in-scene human activities from a distant surveillance center (such as the problem of selecting among a very large number of video feeds for display onto a set of monitors). The current trend is to allow the automation to detect, track, and alarm human activities that could be anomalous. This creates a human supervisor–automation system design problem. Commercial surveillance system designers typically assume that alerting human supervisors to potentially anomalous behaviors and popping up the relevant video feed is an acceptable base design for human supervisors even though years of human factors research have demonstrated that this is a very poor joint system design which produces a variety of predictable problems and failures (e.g., (Woods and Sarter, 2000; Woods, 1995)). These include the false alert problem, getting lost effects in navigating over multiple cameras (Guerlain, 2006), and spatial disorientation from view sequences that jump from place to place (Woods, 1984).

Past research on coordinating human-agent activities in such joint systems has specified basic functional requirements for effective designs: observability, directability, directed attention, and shifting perspectives (Woods and Hollnagel, 2006). The task for human factors of smart surveillance systems is to develop specific visualizations that meet these functional requirements

Extended perception and smart surveilling

The design direction we have been exploring for a wide range of new sensor capabilities and systems is called extended perception. In this paradigm new technology extends a remote human observer's ability to perceive and explore the world as if they were present in the scene (Murphy and Burke, 2008). For the case of a smart PTZ camera in a surveillance task, we conceptualize the visualization opportunity created by computer vision algorithms as: (a) support the out-of-scene agents ability to take virtual patrols as if they were exploring a continuous space, and (b) integrate the structure of activity and events in the monitored physical scene extracted by computer vision algorithms with a direct view of that physical scene.

The visualization design, for the case of a single PTZ camera, first, requires a visible spatial frame of reference that surveillance personnel can modify, i.e., is directable (Woods, 1995). Figure 2(a) shows a panoramic frame of reference that captures all of the views possible from a fixed PTZ camera in the monitored scene (panels (a) through (c) show the base panorama from three different cameras that are part of the research surveillance network on the OSU campus).

Second, the visualization requires an overlay that captures the results of the intelligent processing of activity in the scene. We chose a history of translating motion through the monitored scene as a baseline and representative exemplar of smart algorithms (Davis et al., 2007a). Detecting translating motion is an interesting problem in computer vision, and it is often a base for more sophisticated algorithms such as tracking a person moving through a scene. Translating motion or activity paths can also serve as a backdrop for displaying the output of algorithms that detect specific patterns of activity such as walking versus running. The visualization uses brightness coding to provide an overlay that indicates those areas where the activity algorithms have seen translating motion, cumulated over a past temporal window. The brighter the area the more motion the algorithm has seen in that position over the time interval. Figures 3 show the brightness coding overlay for the actual motion histories for the cameras/scenes of OSU campus in Figures 2. Note the darkest areas correspond to the rooftops and structures where the cameras are mounted (generally high on buildings) where human activity occurs very rarely and bright areas correspond to roads and pathways.

Given the base frame of reference and the activity history overlay, one can now consider the scan path of the camera or the spatial-based virtual patrol–where should or will the camera point next? Scan-path algorithms use the motion history data to tailor the PTZ camera movement to the activity in the physical scene. Figure 4 illustrates scan paths for three different scanning algorithms for the motion history data of the scene in Figure 2(a) and the brightness coding overlay in Figure 3(a) (see (Davis et al., 2007a,b) for details of the scanning algorithms). The scan-path in Figure 4(a) moves probabilistically from location to location to sample areas with high activity (probabilistic jump), while the scan-paths in Figure 4(b-c) create smooth continuous pathways. There are many different criteria (e.g., activity value, staleness of data, operator comprehensibility) that should be considered and balanced in designing any automatic scan-path algorithm. The algorithms presented balance these criteria differently resulting in distinct scan-path behaviors.

EXTENDED PERCEPTION DISPLAY

The panoramic frame of reference is constructed from individual images taken by the PTZ camera and combined through an image-based stitching process. The panoramic construction process uses a mapping that converts the cameras pan and tilt orientation to an x, y pixel position. The inverse mapping converts pixel position to camera orientation and is the foundation for communication between supervisor and smart algorithms.

The smart algorithm implemented and demonstrated separates patterns of translating motion from background noise. The algorithm accumulates individual pixel differences between consecutive images in to a single motion history image (Bradski and Davis, 2002). Over 6 seconds (72 images) the salience and robustness of translating pedestrians, cyclists, and vehicles in the motion history image emerges against background noise sources such as camera noise, changes in illumination, and random motion (e.g., moving tree leaves and branches).

Using the same mapping that generates the panoramic frame of reference, the output of the smart algorithms is transformed into a panoramic representation. Instead of a set of images as the input to the mapping function, however, the input is the result or output of the smart algorithms sampled across the pan-tilt viewspace. The full process for generating an activity map requires moving the camera to a pan/tilt position, capturing a sequence of images, performing the motion analysis, storing the results, and then moving the camera to a new pan/tilt position. One complete pass of the entire scene is sufficient to generate a single activity map over a short temporal window ($\sim$ 20 min). Collecting and merging multiple passes (single activity maps) of the scene results in a global activity map such as shown in Figure 3.

EXPLORING VIRTUAL PATHWAYS

We introduced a representation to subsidize the raw camera view from surveillance cameras. Integrating spatial structure, activity data, and algorithm generated scan-paths over time supports observability and directability for human supervisors. This smart surveilling or extended perception redefines the unit of analysis for surveillance from a sequence of single camera feeds to virtual pathways or patrols through the viewable space. The scan-path panoramic display and temporal displays support these virtual pathways through spatial-, temporal-, and activity-based frames of reference. These frames of reference are inherently coupled and a virtual pathway necessarily defines each of these dimensions, however, for clarity, we define virtual pathways and the forms of exploration for each dimension individually.

Spatial

The spatial pathway of a PTZ camera differs from that of an in-scene agent, which was defined as a moving point of observation with all possible view directions represented by a full sphere. The spatial-based virtual patrols for a PTZ camera are, instead, a sequence of pan and tilt positions within a downward pointed hemisphere, from a fixed location in space. The scan-path panoramic display in Figure 4 supports exploration of the viewable scene for a PTZ camera by making observable for the human supervisor the virtual pathway or sequence of pan and tilt positions. An intrinsic quality of this longshot display is that not only can a human supervisor apprehend what the camera will see in the future, but also what the camera will not see. This display provides a mechanism for the human supervisor to act on this information to re-direct the scan-path algorithms, through the activity overlay, to explore the viewable scene through a different virtual pathway. The exploration through spatial-based virtual pathways are thus a collaboration between human supervisor, smart algorithms, and scan-path algorithms with the scan-path panoramic display as the medium of communication and interaction.

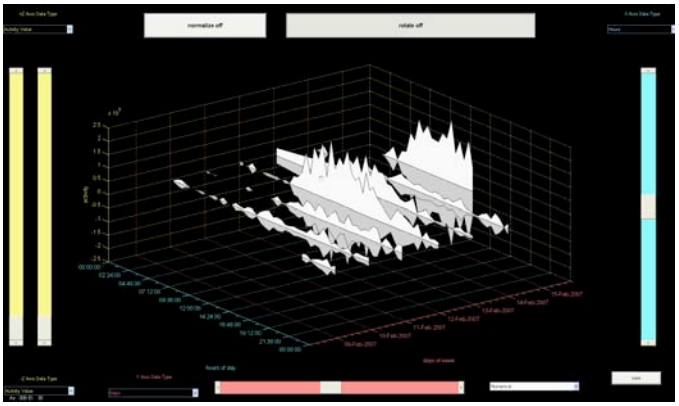


Figure 1: The temporal display allows users to explore different temporal rhythms through two independently scalable temporal scales versus smart algorithm activity output. In this case, the display plots hours versus days to capture the rhythms of different days over a one week period.

Temporal

The out-of-scene supervisory agent must monitor and explore across multiple temporal rhythms. Temporal-based virtual pathways are a new construct for analysis of the temporal dimension. The relevant temporal rhythms may occur over different temporal scales (minutes vs. hours), temporal intervals (last month vs. last week), or in different temporal patterns (Mondays and Wednesdays vs. Tuesdays and Thursdays). A temporal-based virtual pathway is defined by a temporal window size, scale, location, and orientation for a PTZ camera and exploration of the temporal dimension along a virtual pathway consists of adjusting these different dimensions. The display in Figure 1 captures these temporal dimensions and allows a supervisor to create a temporal-based virtual pathway. The current scan-path algorithms do not incorporate temporal information, however, this is a natural extension for smart algorithms that will likely inform new designs for temporal-based displays for video surveillance.

Activity

The out-of-scene supervisory agent through new smart algorithms monitors activity patterns over multiple spatial and temporal scales. New algorithms are constantly created to detect new types of activity and as the data extracted from these algorithms increases, the potential for data overload also increases. Escaping from data overload requires new forms of organization (Woods et al., 2002) and the spatial longshot provided by the scan-path panoramic display is precisely tuned to this requirement. Independent of the type of smart algorithm or resulting data, if the point of extraction is the video feed, then the pan-tilt positions necessarily provide a spatial frame of reference and is therefore transformable into a spherical overlay representation, as illustrated in Figure 3. While integration of this data into the current visualization emphasizes the usefulness of the panoramic longshot, these data also create a new activity-based dimension for exploration, which can be supported by new forms of activity-based virtual pathways. An area for future research is to understand how activity-based virtual pathways inform the organization of algorithm overlays, what manipulation of the activity dimension are necessary (scaling, translating, etc.), and how does the activity-based virtual pathway construct inform the design of new smart algorithms.

Summary

This paper presented a panoramic display for supervisory visualization of smart algorithms for a single PTZ camera within a video-based security surveillance context. This display integrates the capability of video sensors, computer vision algorithms, and cogitive systems principles to overcome the cognitive challenges inherent in understanding a distant environment through a video feed with a narrow field-of-view.

References

Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440.

Bradski, G. R. and Davis, J. W. (2002). Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184.

Davis, J. W., Morison, A. M., and Woods, D. D. (2007a). An adaptive focus-of-attention model for video surveillance and monitoring. *Mach. Vision Appl.*, 18(1):41–64.

Davis, J. W., Morison, A. M., and Woods, D. D. (2007b). Building adaptive camera models for video surveillance. In *Applications of Computer Vision, 2007. WACV '07. IEEE Workshop on*.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Guerlain, S. (2006). Software navigation design. *Applied Spatial Cognition: From Research to Cognitive Technology*.

Haering, N., Venetianer, P. L., and Lipton, A. (2008). The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19(5-6):279 – 290.

Murphy, R. R. and Burke, J. L. (2008). From remote tool to shared roles. *Robotics & Automation Magazine, IEEE*, 15(4):39–49.

Sankaranarayanan, K. and Davis, J. W. (2008a). An efficient active camera model for video surveillance. In *Applications of Computer Vision, 2008. WACV 2008. IEEE Workshop on*, pages 1–7.

Sankaranarayanan, K. and Davis, J. W. (2008b). A fast linear registration framework for Multi-Camera GIS coordination. In *Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on*, pages 245–251.

Wang, L., Hu, W., and Tan, T. (2003). Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601.

Woods, D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics*, 38(11):2371–2393.

Woods, D. D. (1984). Visual momentum: a concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies*, 21(3):229–244.

Woods, D. D. and Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. CRC Press.

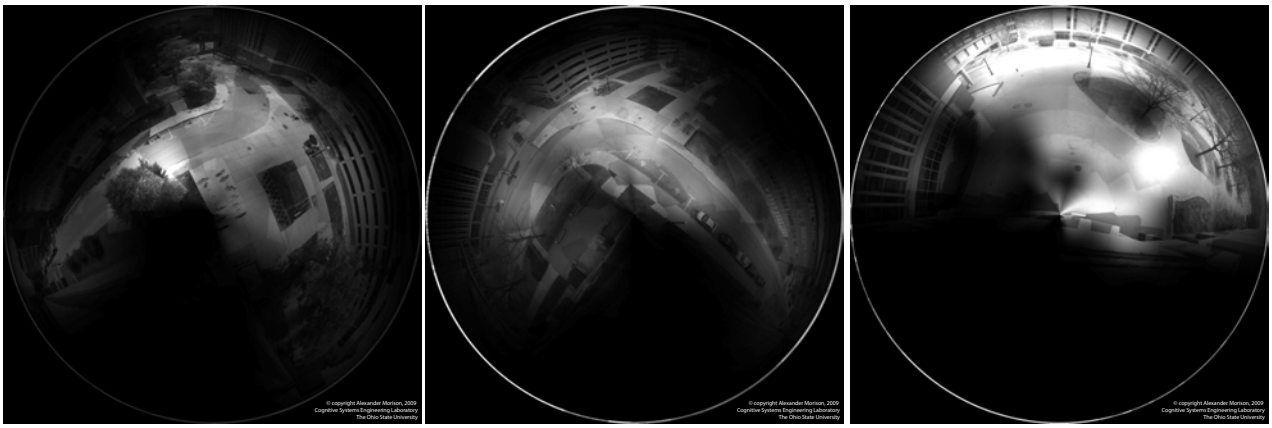
Woods, D. D., Patterson, E. S., and Roth, E. M. (2002). Can we ever escape from data overload? a cognitive systems diagnosis. *Cognition, Technology & Work*, 4:22—36.

Woods, D. D. and Sarter, N. B. (2000). Learning from automation surprises and going sour accidents. *Cognitive Engineering in the Aviation Domain*, pages 327–353.

Woods, D. D. and Watts, J. C. (1997). How not to have to navigate through too many displays. *Handbook of Human-Computer Interaction*, 2:617–650.

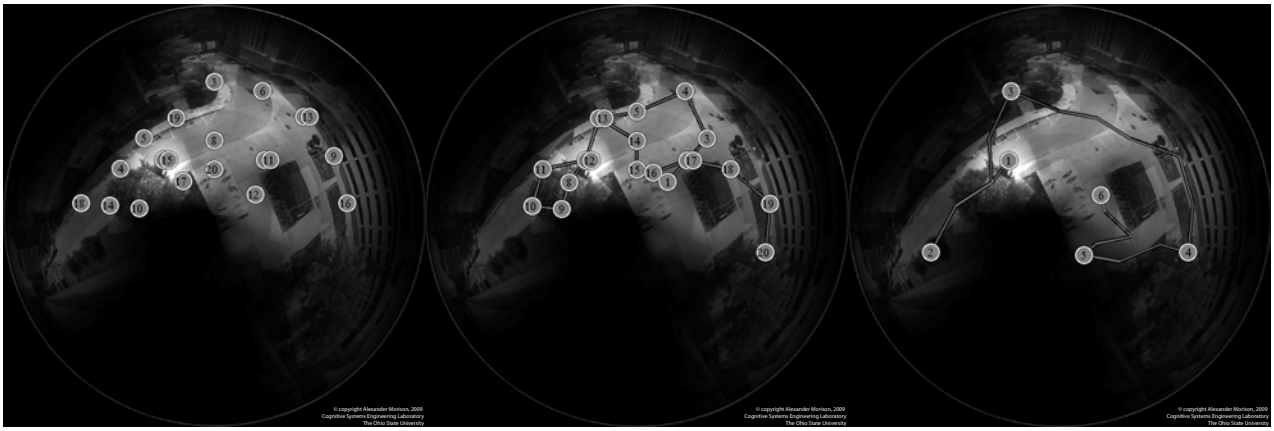
(a) (b) (c)

Figure 2: The panoramic frame of reference for three (panels a-c) separate PTZ cameras on OSU campus that captures all of the views possible from a fixed point relative to the monitored scene.



(a) (b) (c)

Figure 3: The motion history brightness coded overlays for the cameras/scenes in Figures 2(a-c). Note the darkest areas correspond to the rooftops and structures where human activity occurs very rarely and bright regions correspond to locations of expected human activity such as walkways and roads.



(a) (b) (c)

Figure 4: The brightness coded activity map in Figure 3(a), which represents outputs from smart algorithms monitoring for human activity patterns, is overlaid with a scan path that represents how the camera will move to monitor the space. The three different scanpaths are created using three different algorithms which are (a) a probabilistic jump, (b) inhibited probabilistic walk, and (c) reinforcement learning paths.