Intelligent Automated Motion Imagery Acquisition

Volodymyr Kindratenko
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign (UIUC)
kindr@ncsa.uiuc.edu

Abstract

Producing high quality information content-rich motion imagery from multiple data sources is both labor-intensive and expensive. Only recently attempts have been made to devise intelligent automated motion imagery acquisition systems. It is clear, however, that a more complex approach involving real-time data processing for guiding a multiple camera system, precise spatial registration, deployment of a distributed, scalable computational infrastructure, and so forth, is required to provide an adequate solution to this problem. A more complex task is to automatically devise and maintain a global model of the scene of interest based on the motion imagery from multiple data sources in real-time. The goal of this white paper, therefore, is to identify relevant application areas and examine the underlying technology, current R&D directions, and future trends in development of intelligent automated motion imagery information acquisition systems.

1. Introduction

Motion imagery (MI) is the process of capturing, storing, indexing, retrieving and analyzing images captured over time. Images typically are from video captured with a digital movie camera, as well as time based sequences, such as those captured from a security camera, satellite, or aerial surveillance vehicle. One aspect of MI is motion imagery acquisition, that is, how MI data is captured and stored. Automated motion imagery acquisition implies that the data is obtained automatically without human intervention, which is the case in many MI applications. Intelligent automated motion imagery acquisition (IAMIA) implies that instead of recording all the available MI data, the system intelligently and automatically decides what data to acquire based on either the MI's content or other external or internal factors. The system may not only be empowered to decide if the next portion of MI data needs to be stored, but also if only a specific subset of all data is of interest. In addition to the usual MI acquisition equipment and software, a typical IAMIA system includes a decision-making component; it may also include additional sensors and computerized camera control.

IAMIA techniques are used in many application areas, e.g., surveillance, target acquisition, intelligent transportation systems, industrial inspection, teleconferencing, etc. Although they vary significantly depending on the application, they share a common ability to make a decision if the present MI data is of relevance for current application needs and if it must be further processed, or discarded. This white paper identifies IAMIA application areas, technology, current state, and future directions.

2. State-of-the-Art in IAMIA Applications

IAMIA is currently used in a number of applications, yet there exists great potential for this technology to be used in many more areas. Unmanned video surveillance (UVS) and a multimedia presentation room (MPR) are two applications in which IAMIA plays a key role. Robotics and industrial inspection (II) applications as well as airborne video surveillance (AVS), satellite imagery (SI), and intelligent transportation systems (ITS) utilize IAMIA to some degree.

2.1. Unmanned Video Surveillance

Civilian video surveillance systems have been developed since the 60s. First-generation analog systems had no information processing capabilities and video surveillance has been mainly performed by human operators who interpreted the visual information on multiple monitors. Second-generation systems have limited visual information processing capabilities, primarily in the form of motion detection, but they still deliver all visual information to a human operator for the final decision, or store it for later examination. Such systems have been developed for different tasks such as human activity monitoring, object detection, traffic control, etc. Third generation distributed digital surveillance systems currently under development are capable of intelligent data acquisition and decision-making. Only information relevant to an event of interest is delivered to the human operator. More on UVS can be found in [Foresti et al. 2000].

Military reconnaissance, surveillance, and target acquisition systems use motion imagery mainly in combination with other sensors, therefore they are not discussed here.

Digital event recording video system (DERViS), described in [Cherevan et al. 2000], is an example of a second-generation UVS system with minimal IAMIA capabilities. The system is designed to work with existing UVS systems to enhance the operator’s ability to review visual data acquired in real-time just before an event of interest occurred and to focus on the event of interest as it unfolds. Motion-activated video capturing, as described in [Rostovtseva et al. 2000] and [Wildes 1998], is another example of second-generation UVS systems. In these systems, an event recording is activated based on simple motion detection. The GOTCHA! [http://www.gotchanow.com] and SCS [Rostovtseva et al. 2000] commercial systems implement motion detection via direct image analysis (more on this can be found in [Haering et al 2001]) whereas systems based on proximity infrared (PIR) cameras, manufactured by a number of companies, e.g., Kalatel [http://www.kalatel.com], detect motion via an external PIR sensor.

Third generation distributed digital surveillance systems, such as described in [Regazzoni et al. 2000] and [Karuppiah et al. 2001], utilize multiple data sources and perform distributed data analysis and data fusion to derive a high-level scene description which is then delivered to the human operator. The system described in [Regazzoni et al. 2000] consists of a set of modules each of which is responsible for a specific task, such as image acquisition, change detection, focus of attention, etc., and communicate with each other over a local area network. The key element of the system is an intelligent camera – a conventional digital camera with some image processing capabilities and motorized control. Several such cameras are connected to the same hub, thus forming surveillance cells. Hubs themselves are capable of performing advanced data processing and communicating the results to a control center. This architecture provides progressive specialization of the operations performed by each task. Several scenarios of system utilization have been proposed. In an agent-based system, each camera is managed by a camera agent responsible for detecting and tracking moving objects and creating object agents which are responsible for progressively identifying the nature of events and describing underlying activities. Extracted motion trajectories are compared and classified to derive a set of observed activities and the control center is notified when a predefined type of activity is detected. In another scenario, a virtual multiple sensor system, a single mobile camera can be used to provide continuous coverage for a large space that would typically require multiple cameras. New applications of third generation surveillance systems include crowding estimation, people tracking, activity analysis and recognition, ID verification, traffic monitoring [Zhu et al. 2000a], etc.

2.2. Multimedia Presentation Room

Until now, real-time presentation recording required several human operators who directed cameras to capture various events and a production director who decides what video stream has to be recorded or transmitted at any given moment of time. Only recently have attempts been made to automate these tasks, particularly for such applications as classroom lecture recording, meeting capture, and video teleconferencing. In these applications, IAMIA is a key element that enables high information content motion imagery to be extracted from multiple data sources.

To this date, several automated video conferencing systems have been proposed [Wang et al. 1998], [Buxton et al.], [Foote et al. 2000]. The core of the system described in [Wang et al. 1998] is a real-time talking head tracking software that utilizes multiple cameras and microphone arrays. Audio data from microphones is used for coarse position estimates for initial camera aiming. Once the camera is pointed in the direction of the talking person, position refinement, framing, and motion tracking are done via image analysis using source motion as the primary face detection criterion. The camera is continuously adjusted to follow the talking head. In another approach [Cutler et al. 2000], a time-delayed neural network is trained to detect and correlate mouth motions captured via MI with audio data from a microphone. Other systems, e.g., [Foote et al. 2000], rely only on analysis of motion imagery to position the camera for tracking regions of interest.

The meeting capture system described in [Chiu et al. 1999a], [Chiu et al. 1999b], and [Chiu et al. 2000] is capable of automatically acquiring and storing MI from multiple video sources. Data from different sources is sorted in real-time on a switch so that the best obtainable image for each type of activity is acquired together with notes taken by meeting participants. Although, in this system all data is stored even when it does not provide useful content, it is not difficult to implement an IAMIA approach where only high information content MI data is captured and stored.

Several lecture room tracking applications have been proposed [Bianchi 1998], [Cruz et al. 1994], [Mukhopadhyay et al. 1999], [Rui et al. 2001]. The common approach is to have multiple cameras to capture the lecturer, stage, screen, and podium and to record only the most content-rich motion imagery. Thus, the system should not only be capable of closely tracking individual events, but also recognizing in real-time their overall ranking and instructing the recording device what stream should be captured. The core of the system described in [Rui et al. 2001] is a motorized camera equipped with additional sensors to localize the area of interest. Two such sensors have been evaluated: a wide-angle camera and a microphone array. The wide-angle camera setup is used for tracking the lecturer, and the microphone array setup is used to localize the talking audience. In both cases, the output from the sensors is processed in real-time and is used to direct the camera allowing the events of interest to be closely followed. In other systems, e.g., [Bianchi 1998], [Cruz et al. 1994], video from the camera itself is analyzed and used to adjust the position of the camera.

2.3. Airborne Video Surveillance and Satellite Imagery

In AVS and SI systems, the main focus is on real-time motion imagery analysis and scene understanding. Precision video registration (accurate geolocation of moving and stationary targets in video surveillance imagery) [Kumar et al. 1998], multiple target surveillance (simultaneous tracking of multiple vehicles) [Cohen et al. 1998], activity monitoring (monitoring of several areas of the battle space for distinctive motion activities) [Cutler et al. 1999], and 3D scene reconstruction (reconstruction of 3D terrain and building maps based on aerial MI) [Hanson et al. 2001], [Schultz et al. 2000] are just a few examples of AVS/SI applications. In these applications, an IAMIA occurs on a different level than in UVS and MPR applications. Instead of adjusting the camera to closely follow regions of interest, the acquired imagery is processed and adjusted to match the model. The AVS/SI MI output therefore contains processed video data that fits a certain model rather than just a stream of raw video.

2.4. Intelligent Transportation Systems

ITS research includes several unrelated topics such as autonomous ground vehicles, autonomous underwater vehicles, autonomous aerial vehicles, etc. MI plays a significant role in these systems, but it is not the only information channel – other sensory data plays an even more important role. Most of the systems do not require an IAMIA as they are based on an approach in which MI is acquired through a fixed wide-angle camera [Baten et al. 1998], [Mandelbaum et al. 1998a], [Mandelbaum et al. 1998b], [Hansen et al. 1997]. However, one can point out similarities between the ITS and AVS/IS applications with respect to how the acquired imagery is used to fit a certain model and provide the feedback in response to that model.

2.5. Robotics

Flexible, re-configurable vision systems can provide an extremely rich sensing modality for sophisticated robot platforms. In robotics, MI is used for two functions: navigation and task-specific operations. The main purpose for the navigation system is to allow the robot to move in the environment. Task-specific operations can be anything ranging from surveillance to inspection and maintenance. IAMIA is particularly important for robotics applications because of the requirement for many robots to operate autonomously for long periods of time with minimal communication capabilities and limited computing resources. Real-time processing is essential for dynamic and unpredictable environments, and it is important for visual sensing to rapidly focus attention on important activity in the environment.

RABART III [Everett et al. 2001] is an example of a robotic system that uses MI both for navigation and surveillance tasks. For example, when the robot is instructed to enter the next door on the right, the camera immediately turns 45 degrees right of center to acknowledge the behavior request and provide a better view of the doorway detection process. As soon as the door is detected and the penetration behavior is invoked, the camera pans to compensate for the platform's rate of turn in order to keep the door opening in the center of its field-of-view.

Intruder detection and assessment algorithms operate upon the output from the video motion detection system and a 360-degree array of PIR sensors. The PIR data is used to pan the surveillance camera to the center of any zone with suspected intruder activity. The video output is then used to track and keep the intruder in the center of the visual field using a combination of robot head and body movements.

The core of the cooperative multiple robot system described in [Zhu et al. 2000b] is a panoramic annular lens camera system that can capture its surroundings with a 360-degree view. Multiple robots are used, each equipped with such a camera system. In the two-robot scenario for human searching, one of the robots is assigned the role of "monitor" and the other the role of "explorer". The role of the "monitor" is to monitor movements in the environment, including the motion of the "explorer". The role of the "explorer" is to follow a moving object of interest and/or find a "better" viewpoint for constructing virtual stereo geometry with the camera on the "monitor". Mutual awareness of the two robots is important for their dynamic calibration of relative orientations and the distance between the two panoramic cameras. Thus, the system allows for an IAMIA by continuously changing its configuration, e.g., moving from one location to another, one robot at a time. Similar robotic systems have been described elsewhere [Thayer et al. 2000].

2.6. Industrial Inspection

Most of the visual II applications are based on 2D computer vision techniques and do not require any IAMIA. This approach only works in a well-defined environment with all the setup parameters known in advance. However, in dynamic and unpredictable environments, IAMIA is mandatory. Only recently have attempts been made to introduce such capabilities in II systems. One such example is Authentication Server from Iridian Technologies, Inc. [http://www.iridiantech.com] – a vision system that uses face tracking and rapid pan-tilt-zoom-focus to acquire high-resolution image of a person's eye and performs positive identification based on iris pattern. The person does not have to stay very close to the camera in a pre-defined position, instead the camera is dynamically adjusted to locate and follow the person’s eye as he/she walks by. Another example is IMAGO 100 system by IMAGO Machine Vision, Inc. [http://www.imagotrackers.com] – a video-based ground to air target tracker. The system identifies and tracks moving objects in real time with a pan/tilt camera. Its applications include military target tracking and golf ball trajectory analysis. Another area of interest is 3D geometry reconstruction from multiple images [Criminisi 2001].

3. State-of-the-Art in IAMIA Technology

A so-called “smart camera”, functional diagram as shown in Fig. 1, is the main building block of a generic IAMIA system. It consists of an MI acquisition camera, guiding sensor, motorized camera control, and a decision-making subsystem. Data from the guiding sensor is processed by the decision-making subsystem and instructions are transmitted to the motorized camera control to adjust the camera’s parameters so that only relevant MI is acquired. Guiding sensor data as well as the intermediate processing results obtained by the decision making subsystem can also be output together with the MI data.

Several such systems can be merged to form a distributed multiple source MI acquisition system. Alternatively, a single decision making subsystem can be used to control multiple cameras and/or multiple sensors can feed data to multiple decision making subsystems.

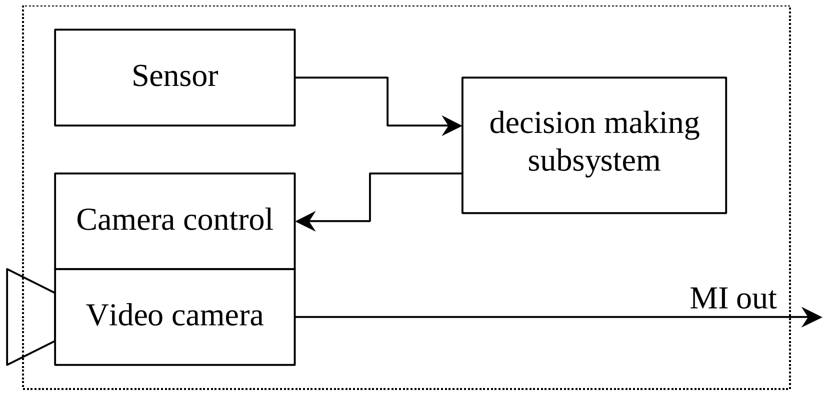

Fig. 1. Basic IAMIA system.

3.1. Guiding Sensor

A guiding sensor is used to obtain information about current conditions in the operating environment. For example, UVS systems frequently utilize motion detection sensors that trigger the MI acquisition based on the motion spotted via a PIR detector [Rostovtseva et al. 2000]. A secondary wide field of view video camera can also be used [Rui et al. 2001]. Alternatively, the main video camera itself can serve as the sensor device [Regazzoni et al. 2000], [Foote et al. 2000], in fact, some cameras have built-in in-image motion detection capabilities [http://www.axis.com]. In this case, the acquired MI is processed by the decision-making subsystem and the camera is guided based on the content of the imagery. In another example, microphone arrays can serve as the sensing device [Wang et al. 1998], [Rui et al. 2001]. In ITS and robotics applications, near-distance radar or infrared imagery can be used for the guiding sensor. The choice of the guiding sensor device is highly application-dependent and, as shown above, may take different forms.

3.2. Motorized Camera Control

Motorized camera control includes pan, tilt, zoom, and focus. Camera control elements can be integrated with the camera, e.g., as in AXIS EVI-D30/D31 model [http://www.axis.com], or the camera can be installed on a motorized platform, e.g., as in Pan and Tilt Head Units developed by Telemetrics Inc. [http://www.telemetricsinc.com].

As with the sensors, characteristics of a particular motorized camera control are highly application-dependent. For example, in some cases the system should be able to provide a 360-degree view, in other cases its zooming capabilities may be of a primary importance. However, in any case, the camera control system should be software-controllable; otherwise it can be a challenge to integrate it with the rest of the system.

3.3. Decision Making Subsystem

The decision making subsystem in a smart camera can be implemented in a number of ways ranging from using a general-purpose PC to a dedicated DSP board integrated with the camera's electronic circuits. Its function is to process sensor readings in real-time and provide guidance instructions to the motorized camera control. When external sensors that require additional data processing are used, such as a microphone array, it is common to use a PC as the main data processing and decision-making device, e.g., as in [Rui et al. 2001]. When the MI from the camera itself is used for decision-making, or no computationally expensive data processing is required, it is possible to embed extra DSP electronics in the camera's circuits, e.g., as in PIR camera [http://www.kalatel.com]. In another approach, external specialized hardware can be used, e.g., as in AXIS 2401 Video Server [http://www.axis.com]. Many ITS systems use Sarnoff Corporation video processors [http://www.sarnoff.com] for real-time MI analysis. AVS & SI applications use proprietary application-specific real-time computing platforms.

The characteristics and capabilities of a particular decision-making subsystem are application-dependent, however one requirement remains the same – the data have to be processed and the decision have to be made in real-time.

3.4. System Integration

An IAMIA system should be able to react in real-time to rapid changes in the environment. To achieve this, each system component should be well tuned for its task and the overall system architecture should be optimized to minimize the overall response time. Application-specific, as opposite for a generic, approach manifests itself in a better overall system. For example, it is possible to build a generic motion-activated surveillance system that uses video as the main sensory data. However, in some applications it may be more beneficial to use PIR sensors instead. Likewise, a generic PC can be used as the decision-making subsystem, however for some applications it may not be fast enough and therefore dedicated computer vision hardware is required. An important aspect is the communication speed between various components in the system and the reliability of the communication channel. All this points out the importance of the right choice of various components in the system, its architecture, and the overall system integration.

4. Conclusions

To date, IAMIA technology has been successfully used in a number of applications ranging from video surveillance to robotics. However, most of these applications are relatively simple, provide small volume coverage, and require a fairly simple computational infrastructure. The challenge ahead is how to integrate multiple distributed smart camera systems, guided by various sensors and installed on various stationary and movable platforms, into a large distributed self-reliant system that would allow monitoring an event of interest on a large scale. The capabilities of such a system should go beyond simple object tracking, as most of today's IAMIA applications do. It should be integrated with geospatial databases, data storage and data-mining facilities, and should have real-time recognition and analysis capabilities available on demand. High-performance computing facilities, such as available at the National Center for

Supercomputing Applications, could play a vital role in both R&D and deployment of such systems by providing computational and network infrastructure necessary for modeling, testing, and running large-scale distributed IAMIA applications.

Bibliography

S. Baten, R. Mandelbaum, M. Luetzeler, P. Burt, and E. Dickmanns, *Techniques for Autonomous Off-Road Navigation*, IEEE Intelligent Systems magazine, special issue on Vision-Based Driving Assistance, Nov-Dec. 1998, pp57-65.

M. Bianchi, *AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations*, In Proc. Joint DARPA/NIST Smart Spaces Technology Workshop, 1998.

W. Buxton, A. Sellen, and M. Sheasby, *Interfaces for multiparty videoconferences*, Video-mediated communication (eds. K, Finn, A. Sellen, & S. Wilbur), Lawrence Erlbaum Publishers.

A. Cherevan, I. Zhukov, V. Murin, G. Rostovtseva, G. Skripka, and S. Sorokin, *Digital Event Recording Video System (DERVIS)*, In Proc. RFNC VNIIEF, Sarov, Russia, 2000.

P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, *Meeting Capture in a Media Enriched Conference Room*, In Proc. 2nd International Workshop on Cooperative Buildings (CoBuild'99), Lecture Notes in Computer Science, Vol. 1670 Springer-Verlag, pp. 79-88, 1999.

P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, *NoteLook: Taking Notes in Meetings with Digital Video and Ink*, In Proc. ACM Multimedia 1999, Orlando, Florida, November 1999.

P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, *Room with a Rear View: Meeting Capture in a Multimedia Conference Room*, In Proc. IEEE Multimedia Magazine 2000.

I. Cohen, and G. Medioni, *Detection and Tracking of Objects in Airborne Video Imagery*, In Proc. Workshop on Interpretation of Visual Motion, 1998.

A. Criminisi, *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*, In Distinguished Dissertations series, Springer-Verlag, London, 2001.

G. Cruz, and R. Hill, *Capturing and playing multimedia events with STREAMS*, In Proc. ACM Multimedia 1994, pp. 193-200.

R. Cutler, C. Shekhar, B. Burns, R. Chellappa, R. Bolles, and L. Davis, *Monitoring Human and Vehicle Activities Using Airborne Video*, In Proc. 28th Applied Imagery Pattern Recognition Workshop (AIPR), Washington, D.C., 1999.

R. Cutler, and L. Davis, *Look who's talking: Speaker detection using video and audio correlation*, In Proc. IEEE International Conference on Multimedia and Expo (ICME), Manhattan, New York, 2000.

H. Everett, G. Gilbreath, D. Ciccimaro, *An Advanced Telereflexive Tactical Response Robot*, Autonomous Robots, Vol. 11, No. 1, 2001.

J. Foote, and D. Kimber, *FlyCam: Practical panoramic video*, In Proc. IEEE International Conference on Multimedia and Expo, vol. III, 2000, pp. 1419-1422.

G. Foresti, P. Mahonen, and C. Regazzoni, eds., *Multimedia Video-Based Surveillance Systems: Requirements, Issues and Solutions*, Kluwer Academic Publishers, Boston, 2000.

N. Haering and N. Lobo, *Visual Event Detection*, In Kluwer International Series in Video Computing, Kluwer Academic Publishers, Boston, 2001.

M. Hansen, and L. Wixson, *Vision Technologies for Manned and Unmanned Surveillance*, In Proc. Office of National Drug Control Policy International Technology Symposium, 1997.

A. Hanson, M. Marengoni, H. Schultz, F. Stolle, E. Riseman, and C. Jaynes, *Ascender II: a framework for reconstruction of scenes from aerial images*, In Proc. Automatic Extraction of Man-Made Objects from Aerial and Space Images Workshop (III), Ascona, Switzerland, 2001.

D. Karuppiah, Z. Zhu, P. Shenoy, and E. Riseman, *A Fault-Tolerant Distributed Vision System Architecture for Object Tracking in a Smart Room*, In Proc. ICCV International Workshop on Computer Vision Systems, Vancouver, Canada, 2001.

R. Kumar, H. Sawhney, J. Asmuth, A. Pope, and S. Hsu, *Registration of video to geo-referenced imagery,* In Proc. ICPR 98, Brisbane, Australia, 1998.

R. Mandelbaum, L. McDowell, L. Bogoni, B. Beich, and M. Hansen, *Real-time stereo processing, obstacle detection, and terrain estimation from vehicle-mounted stereo cameras*, In Proc. 4th IEEE Workshop on Applications of Computer Vision, Princeton NJ, 1998, pp. 288-289.

R. Mandelbaum, M. Hansen, P. Burt, and S. Baten, *Vision for Autonomous Mobility: Image Processing on the VFE-200*, In Proc. IEEE International Symposium on Intelligent Control (ISIC), International Symposium on Computational Intelligence in Robotics and Automation (CIRA), and Intelligent Systems and Semiotics (ISAS), Gaithersburg, Maryland, 1998.

S. Mukhopadhyay, and B. Smith, *Passive Capture and Structuring of Lectures*, In. Proc. ACM Multimedia 1999, Orlando, 1999.

C. Regazzoni, and L. Marcenaro, *Object detection and tracking in distributed surveillance systems using multiple cameras*, NATO Advanced Studies Institute on Multisensor Data Fusion, Pitlochry, Perthshire, Scotland, 2000.

G. Rostovtseva, I. Zhukov, A. Cherevan', G. Skripka, S. Sorokin, and V. Murin, *Storage Reliability Control System*, In Proc. RFNC VNIIEF, Sarov, Russia, 2000.

R. Yong, H. Liwei, G. Anoop, and L. Qiong, *Building an Intelligent Camera Management System*, in Proc. ACM Multimedia 2001, Ottawa, Ontario, Canada, 2001.

H. Schultz, A. Hanson, E. Riseman, F. Stolle, Z. Zhu, C. Hayward, and D. Slaymaker, *A System for Real-time Generation of Geo-referenced Terrain Models*, In Proc. SPIE Symposium on Enabling Technolgies for Law Enforcement, Boston, MA, 2000.

S. Thayer, B. Digney, M.B. Dias, A. Stentz, B. Nabbe, and M. Hebert, *Distributed Robotic Mapping of Extreme Environments*, In Proc. SPIE: Mobile Robots XV and Telemanipulator and Telepresence Technologies VII, Vol. 4195, November, 2000.

R. Wildes, *A measure of motion salience for surveillance applications*, In Proc. IEEE ICIP, 1998, pp. 183-187.

C. Wang, and M. Brandstein, *A hybrid real-time face tracking system*, In Proc. ICASSP98, Seattle, 1998, pp. 3737-3740.

R. Wildes, and L. Wixson, *Detecting Salient Motion Using Spatiotemporal filters and Optical Flow*, In Proc. DARPA Image Understanding Workshop, 1998, pp. 349-356.

Z. Zhu, G. Xu, B. Yang, D. Shi, and X. Lin, *VISATRAM: A real-time vision system for automatic traffic monitoring*, Image and Vision Computing, 18 (10), 2000, pp. 781-794.

Z. Zhu, K. Rajasekar, E. Riseman, and A. Hanson, *Panoramic Virtual Stereo Vision of Cooperative Mobile Robots for Localizing 3D Moving Objects*, In Proc. IEEE Workshop on Omnidirectional Vision, Hilton Head Island, SC, 2000, pp. 29-36.