

Surveillance and Monitoring Using Video Images from a UAV

G rard Medioni and Ram Nevatia

Institute for Robotics and Intelligent Systems
University of Southern California
Powell Hall Room 204, MC-0273
Los Angeles, California 90089-0273
<http://iris.usc.edu/Outlines/vsam-project.html>

Abstract

We present a methodology to perform the analysis of a video stream, as generated by an Unmanned Air Vehicle observing a theater of operation. The goal of this analysis is to provide an alert mechanism to a human operator. We propose to first detect independently moving objects, then to track and classify them, and to infer behaviors using context. We propose to use, as context, the information contained in a site model, which we will register with the image. We present a technical approach, together with a demonstration plan and an evaluation procedure.

1 Introduction

Continuous surveillance and monitoring in battlefield and urban environments is becoming feasible due to the easy availability and lowered costs of video sensors. Such sensors may be deployed on stationary platforms, be mounted on mobile ground vehicle, or be airborne on board Unmanned Air Vehicles (UAVs). While the multiplicity of such sensors would permit close surveillance and monitoring, it is difficult to do so by relying on purely manual, human resources. Not only would the cost of humans observing sequences from these multiple sensors be prohibitive, but unaided humans may have difficulty remaining focused on the tasks. Typically, long periods may pass before any event of interest takes place; it is easy for human attention to wander in such a situation, and significant events may be missed. Even partial automation of the pro-

cess to indicate possibly significant events to a human will considerably improve the efficiency of the process. Note that the automatic analysis need not completely define the threat, but enough evaluation must be done so that false alarms can be kept within acceptable limits.

The key task of video surveillance and monitoring (VSAM) is to observe moving vehicles and humans, and to infer whether their actions pose a threat that should be signalled to the human monitor. This is a complex task and, in general, requires integration of information from multiple sensors. Further, the deployment of, and control of the sensors, may depend on the perceived events. We plan to focus on data from a single UAV (though information from multiple UAVs could be integrated).

The UAV video introduces several constraints. The UAVs fly at fairly high altitudes, so the resolution and the field of view are limited. This limits the kinds of judgements that can be made, nonetheless, we believe that several significant events of interest can be detected and used to cue a human monitor, vastly reducing the amount of data that the human needs to observe.

The most important information that can be extracted from a video sequence is that of moving vehicles in the scene. We should be able to detect these moving vehicles, estimate their speeds and trajectories, and observe their behavior (within the constraints of available resolution and time). Motion detection is made difficult as both the observer and some elements of the scene may be moving. It may be hard to estimate 3-D trajectories due to lack of resolution.

* This research is supported in part by the Advanced Research Projects Agency of the Department of Defense and is monitored by U. S. Army.

Motion by itself, however, is not a sufficient indication of a threatening or otherwise interesting activity. In most natural scenes, there is significant amount of normal vehicle motion. It is *unusual* motion patterns that are of interest. We believe that the use of a site model, and context, can help us separate the mundane from the unusual. For example, normal traffic flow on a highway should not be signalled, but abnormal speeds or a certain aggregation of vehicles may represent a significant event. Even more complex behaviors may consist of a number of vehicles or humans acting cooperatively, and of the pattern of these activities. The actual behaviours of interest will be decided upon by consultation with the user community.

Our approach requires us to have at least crude models of the site being observed; for the monitoring system to also have an ability to recognize features such as roads and buildings generically is beyond the scope of this effort. In addition, we will need models of what is normal and unusual behavior, and how it depends on context.

Our approach consists of three major steps. The first is the detection and tracking of moving objects, the second serves to relate these vehicles to features known in a map or site model, and the last is used to infer the behaviors.

As we must deal with moving objects and moving observer, we plan to first detect egomotion, which is manifested globally. Egomotion is then used to register frames to detect independently moving objects, and to track them. Image to model correspondence will help in relating the observed motion to relevant features on the ground. We expect techniques using low level features such as lines and curves to suffice for matching in this task. As video images come continuously, tracking from frame to frame is a much easier task than looking at isolated frames. Finally, behavior analysis will be based on the interaction between vehicle trajectories and the features around them. Certain speeds and trajectories in certain contexts indicate a threatening behavior. More complex behaviors will be analyzed by observing actions of groups of vehicles, rather than just single vehicles.

Each of these steps poses significant image understanding (IU) challenges. While there is much knowledge in the field, that is relevant to solving

such problems, the techniques have not been put together to yield complete systems.

As a means of background, we start by describing the elements of such a complete system, the EPSIS Billboard Replacement System, which shares some common characteristics with our scenario, then proceed with the details of our technical approach.

2 The EPSIS Billboard Replacing System

2.1 Background

It is a common practice to place billboards advertising various products and services during sports events. These billboards target not only the spectators at the stadium, but also (and mostly) the viewers of the TV broadcast of the event. This fixed advertising is therefore limited, as the billboards might be advertising products out of context for the TV audience, especially for international events.

We are presenting a system to automatically substitute, in real-time, one billboard by another, synthetically created, billboard. It aims at replacing the billboards in the scene in such a way that it should be transparent to the viewer. It allows a local TV station to plant its own advertisement billboards regardless of the original billboard, thus increasing the overall effectiveness of the advertising.

The process by which we accomplish this goal is therefore the composition of a video stream and a still image, to create a new, smoothly blended and photo-realistic video stream.

Editing of images or image streams is fast becoming a normal part of the production process[1]. Many recent movies, such as Terminator 2, Forrest Gump, Casper, ID4 seamlessly blend live images with Computer Generated Imagery. The mixing of multiple elements is performed primarily by *screen matting*, in which the background is of almost constant color, generally blue or green. This approach requires a very controlled studio environment and operator intervention for optimal results.

Our system must instead function without active cooperation, in real-time (therefore automatically, without operator intervention), and in a non controlled environment. Furthermore, it must also adapt the model to fit the observed billboard. It involves the “intelligent,” automatic manipulation of images and image streams, *based on their contents*.

The system receives as input a TV broadcast signal, must identify a given billboard in the image flow, track it precisely, and replace it with another pattern

(fixed or animated), broadcasting the replaced signal, in real-time, with only a short, constant delay. Figure 1 presents an example frame of billboard



Figure 1 Two examples of billboard replacement in a Video Sequence

replacement.

2.2 Requirements and challenges

The fundamental requirement that the system perform *on-line* in *real-time*, imposes major constraints on the design and implementation:

- No human intervention is possible.
- No on-screen errors are permitted. The system has to include self quality control mechanisms to detect problems and revert to the original signal when they occur.
- Complex high level algorithms are limited due to the need for implementation in real time.
- No cooperation from the field is expected, in order to allow the system to operate independently from the imaging process (e.g. at the down link).

The contribution of such a system, for which a patent was issued[2], thus resides both in the design and implementation of the individual modules (finder, tracker, replacer), and in the management of failure and uncertainty for each of these modules, at the system level, resulting in reliable replacement.

2.3 Implemented Solution

2.3.1 Overall System Design

The task of the system is to locate a planar, rectangular target billboard in the scene, detect camera switches, track the billboard throughout the sequence (between camera switches), and replace it with a new billboard. The direct naive approach would be to inspect the incoming frames, search for the billboard and replace it. Unfortunately, this approach is not sufficient, as it may be impossible

to locate the billboard in the current frame: This may be due to large focus or motion blur, or to the billboard being occluded, or to the fact that only a small part of it may be in the field of view. The billboard may therefore be found only in a later frame of the sequence, and it is not advisable to start replacing then, as this would be offensive to the viewer. Instead, replacement should be performed on the whole sequence to avoid billboard switches on screen.

Our system relies on modular design, and on a pipeline architecture, in which the search and track modules propagate their symbolic, low-bandwidth results throughout the pipe, and the replacement is performed at the exit of the pipe only, therefore relying on accumulated information. This allows the system to make replacement decisions based on complete sequences, thus avoiding mid-sequence on-screen billboard changes.

The *Finder* module searches for the target billboard in the entering frames and passes its results to the Updater, which propagates them throughout the buffer. It first extracts “interesting” points (corners, or other “busy” formations) in the image, then selects the interest points which are most likely to come from the target billboard based on color information. It then finds a set of corresponding points between model points and image points, using an affine-invariant matching technique proposed by Lamdan and Wolfson[5], and uses these correspon-

dences to find the precise (to a sub-pixel resolution) location of the billboard.

The **Global Motion Tracker (GMT)** module estimates the motion between the previous and current frames, *regardless of whether the billboard was found or not*. This is used as a mechanism for predicting the billboard location in the frames in which it was not found. The prediction is necessary to ensure continuity of replacement, since we do not want the billboards to switch back and forth between the original and the new one in front of the viewer. The GMT also performs the task of camera switch detector.

Since we are interested in the motion of the camera and not in a per pixel motion, we take a global approach, and use an iterative least squares technique on all pixels of the image[3]. The images are first smoothed and the spatial and temporal derivatives computed. Using this information, estimates of the motion parameters are computed. Using these estimates, Frame $t+l$ is warped towards Frame t , and the process is repeated. Since Frame $t+l$ gets closer to Frame t at every iteration, the motion parameters should converge. The accumulated parameters are then reported to the Updater. We have implemented the algorithm at multiple levels of resolution. A Gaussian pyramid is created from each frame[4]. At the beginning of a sequence, the algorithm is applied to the lowest resolution level. The results from this level are propagated as initial estimates for the next level up, up to the highest level. This allows for recovery of large motions.

An improvement to the global motion algorithm allows for accurate and stable results, even in the presence of independently moving obstacles in the scene. This is achieved by scaling the coefficients of the motion equations inversely proportional to the temporal derivatives. Moving obstacles do not match when the images are warped according to the camera motion. Therefore, pixels corresponding to obstacles produce high temporal derivatives, and consequently contribute less. The improved results allow for long propagation of estimates along the sequence.

The **Replacer** performs the graphic insertion of the new billboard, taking into account variations from the model due to lighting, blur and motion.

Given the coordinates of the billboard corners in the current image, the Replacer module replaces the image contents within these corners (the billboard)

with the new desired contents (usually a new billboard). Because the human eye is quite sensitive to sharp changes in colors, we correct the gain and offset of the replaced billboard to make it appear close to the average intensity of the image. Note that we currently assume that the original billboard is unoccluded. Mechanisms which allow for detection of obstacles in front of the billboard are currently under development with promising results.

The **Updater** handles communication within the buffer and also manages the *Measure Of Belief (MOB)* associated with the information passed along, due to the MOB of each of the modules, and a decay related to the length of the propagation. The information about scene changes is also used so that the *Updater* does not propagate the predictions beyond the scene change markers.

It collects data from all the other modules, and corrects missing or inaccurate information within a processed sequence. We can visually think of the system as a circular buffer, holding a frame and a frame attribute in each of its cell. The Updater manipulates these attribute records only, which are composed of a small number of parameters, and processes *all* attribute records in the buffer in one frame time.

Figure 2 presents the overall system architecture. As the frame at time t comes in from the video source on the right, the Finder searches for the billboard. At the same time, the Global Motion Tracker (GMT) computes the camera motion between the previous and current frames, and stores it in an attribute record. If the billboard is found, its four corners are recorded in the attributes record, and the Updater unit predicts the location of the billboard in all the (previous) frames from the first frame of the sequence to frame $t-1$, based on the computed motion, and updates the attribute records accordingly. As the frame is about to be displayed, the Replacer performs the insertion.

Let us consider the difficult case where the billboard is slowly entering into view, as a result of a pan or zoom. In this case, the billboard cannot be found initially by the Finder. As the frames continue to come in, the Global Motion Tracker computes the camera motion between frames, regardless of whether the billboard was found or not. The camera motion parameters found are stored in the frame attribute record to be accessed by the Updater. When the billboard is reliably found in some frame,

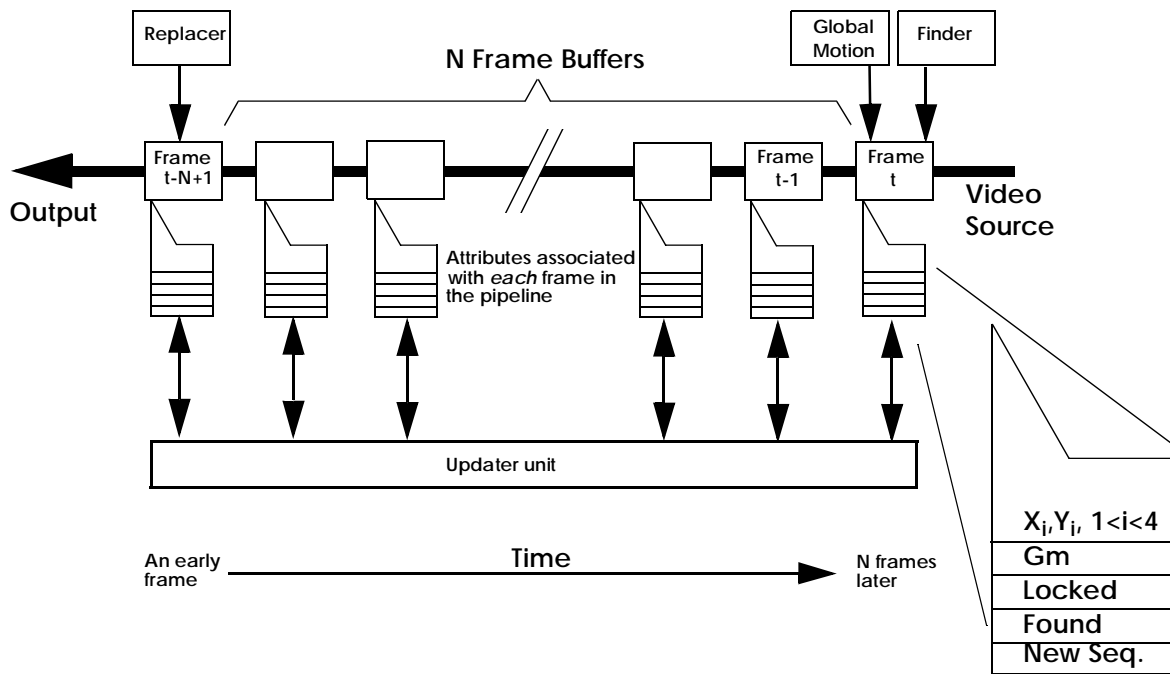


Figure 2A block diagram of the system.

t , of the sequence, the Updater module uses the motion parameters computed earlier, to predict the location of the billboard in all the frames from the first frame of the current sequence up to frame $t-1$. Since this is a very simple computation (not image based), involving low bandwidth communication, it can be performed for the whole buffer in one frame time. As the images reach the end of the buffer, we know the location of the billboard, either directly from the Finder, if it was found in this frame initially, or via a prediction from the Updater, using the motion information.

The combined use of the Global Motion Tracker, the delay buffer and the Updater mechanism, allow the system to, in essence, go back in time without having to process the image again, and to use information from the current frame to locate the billboard in earlier frames. This enables the system to perform well under varied conditions, such as occlusion and entering billboards. The system is also very robust to failure of specific modules, as it can overcome failure in some frames by using information from the other frames of the sequence. It is important to note that each image is processed once only, and that each module works at frame rate, thus the system works in real-time, introducing only a constant delay.

This design can guarantee that no offensive substitution will take place, as long as a whole sequence

fits in the buffer. Otherwise, in case of a problem occurring after replacement is started, a smooth fade back to the original billboard is used. In practice, a buffer of the order of 3 seconds (180 fields in NTSC), covers a large percentage of sequences in which the billboard is present.

2.3.2 The Machine

A design somewhat simpler than the one described here has been made operational by Matra CAP Systèmes using off-the-shelf components, and used by Symah Vision for live broadcasts.

This successful aggregation of computer vision and computer graphics techniques should open up a wide avenue for other applications, which are either performed manually currently, or simply abandoned as too difficult.

On a different note, it is interesting to note that such a system also casts some doubts as to the authenticity of video documents, as predicted in fiction such as *Rising Sun*. It shows that digital video documents can be edited, just like audio and photo documents.

3 Research Issues and Approach

We plan to develop a system for analysis of video image sequences from a single Unmanned Air Vehicle (UAV) with the objective of detecting impor-

tant events that present a threat or are significant in other ways and alert a human monitor to them. Our goal is not complete automation but reliable operation while minimizing false alarms for the human, resulting in a great reduction on the time that the human must devote to monitoring such video streams.

The most relevant information that can be extracted from a video sequence is that of moving objects in the scene. We therefore propose to process the video stream to:

- estimate image motion due to the observer (egomotion), and compensate for it,
- detect regions in the image whose motion differ from the above,
- track these tokens over time,
- infer behaviors from this analysis.

The overall approach to the problem is depicted in schematic form on Figure 3. The modules correspond to the major tasks mentioned above.

Note that these tasks require different time frames: while it is possible to estimate egomotion from only 2 frames, reliable tracking of independent objects requires several frames, and behavior inference demands an even longer aggregation of frames.

We now describe our technical approach in some detail..

3.1 Motion estimation and segmentation

We need to detect motion that is independent from the flow induced by the sensor (egomotion) in the image stream. To accomplish this, we propose to first estimate the egomotion, use it to register frames and detect independently moving objects and then to track them to compute their trajectories. We describe these steps below.

3.1.1 Egomotion estimation

Since we are interested in the motion induced by the camera, and not in a per pixel motion, we take a global approach, and use an iterative least squares technique on all pixels of the image [1, 7, 14]. The method, therefore performs the task of *image stabilization*, assuming the sensor motion is limited to pan, zoom and tilt.

However, the motion model may not be able to take reflect the variations of displacements of the object features due to its 3-D geometry. In our scenario,

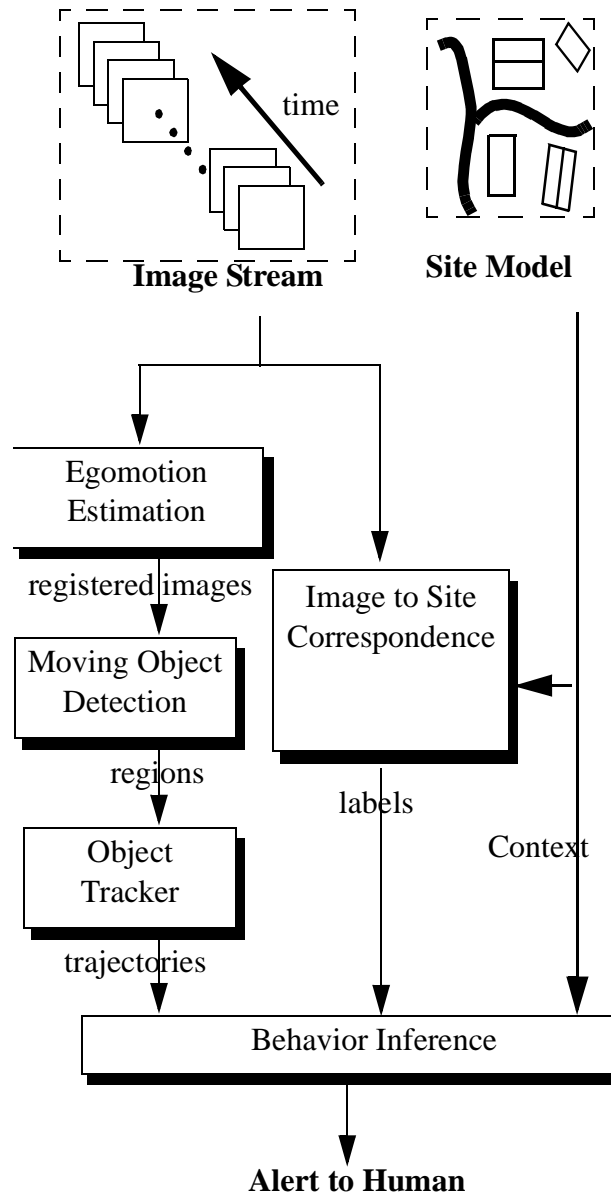


Figure 3 Overview of the Complete System

the sensor is relatively far from the object. Therefore the deviation of the point-of-view, which produces these displacements, is small. Furthermore, even these small deviations can be corrected with coarse knowledge of the terrain, as obtained from the site model, if necessary.

We have also adapted the basic algorithm to function in the presence of independently moving objects in the scene. This is obtained by scaling the coefficients of the motion equations inversely proportional to the temporal derivatives. Moving objects are not registered when the images are warped ac-

ording to the camera motion. Therefore, the pixels corresponding to objects have high temporal derivatives, and consequently less weight in the coefficients. A by-product of the algorithm is the identification of independently moving pixels.

3.1.2 Tracking

Pixels identified by the egomotion module as not coherent with the induced displacement are marked as belonging to mobile objects. Given the resolution we can expect from the sensor, we anticipate these mobile objects to consist of only a few pixels in the image. It is therefore unreasonable to expect to try and perform 3-D structure estimation for these objects[16]. We propose instead to represent them only as 2-D regions, and track these regions in time [8, 15]. For vehicles moving on the ground, the image displacement between two frames will be small, but temporally coherent. We will therefore perform a multiframe analysis of the motion, for robustness and accuracy, using a 2-D translational motion model for the image features.

When airborne vehicles, such as helicopters and fixed wing airplanes are observed, the image induced displacement is much larger. The regions corresponding to these objects will also be larger, since they are closer to the sensor. In such a case, we will use the structure of the region contour to disambiguate the tracking process.

3.2 Image to Site Correspondence

To interpret the motion of the observed vehicles, it is useful to geolocate them in reference to the known features of the site. It is important to know if the vehicles are on a road or in the vicinity of certain important buildings. We consider the problem of actually detecting features such as roads from the image sequence itself without prior models to be beyond the scope of this effort. Instead, we propose to make correspondences with prior maps or site models. Approximate locations may be determined just from the knowledge of the observer vehicle parameters. We can expect to obtain highly accurate estimates of the observer location from the GPS navigation system; orientation parameters may have somewhat less precision. However, we cannot expect the navigational parameters to be accurate enough to predict exactly where a feature of interest, such as a road, might be, but we believe that simple image to map (model) matching techniques [6, 11] can suffice to bring them into accurate cor-

respondence.

We will need to continually update the correspondences between the observed features and the map/model features. Since the data is available to us in a continuous stream, the updating process can be much simpler than one of initial correspondence. At each step, we can predict the amount of displacement and can correct by using only a small number of features.

3-D site models, even if they are not very accurate or complete, would help in the process of correspondence. However, it may be possible to make do with 2-D maps if the terrain is relatively flat and the flight of the vehicle is level.

3.3 Behavior Inference

After various vehicle and human motions have been detected and tracked, and some correspondence established between the images and the site features, we still need to interpret the motion to decide if a significant event has taken place. A first step in this process is that of motion interpretation itself. We should be able to tell whether the moving object is on the ground or is airborne as ground objects have some constraints on their motion. We can also estimate the vehicle speed which may provide some constraints on the class it belongs to (*e.g.* tanks don't travel at 100 km/h).

The next step is to try to determine if a significant event is taking place. We will study the kinds of events human monitors detect and the cues they use to detect them. Some examples are: abnormal speeds or trajectories, activity in forbidden areas, and certain kinds of group activities.

We believe that these kinds of activities can be detected by representing the expected behaviors in a symbolic template representation and verifying if the template criteria are satisfied.

4 Evaluation Plan

We describe some proposed metrics and an evaluation methodology below.

4.1 Metrics

We propose the following metrics:

- 1) **Detection rate:** What is the percentage of correctly recognized events? Obviously, only events that are visible, have sufficient resolution and duration can be detected.
- 2) **False Alarm rate:** This measures the frequency of mistaken detection.

4.2 Testing

We intend to develop and test our algorithms directly on data from operational UAVs. We expect that such data will become available to the VSAM research community. To make evaluations, we will need some *ground-truth* to compare with. For UAV data, we may not be able to get the actual ground truth, instead, we may need to rely on the judgement of human observers to see what events they are able to perceive and what their interpretation of the behaviors is.

4.3 Demonstration Plan

Our proposed research is to analyze image sequences observed from a UAV. As we are not likely to have access to UAVs in the field, we will need to demonstrate on stored images in our laboratory. We expect that sample imagery will be available from on-going UAV projects and from the IFD platform. For the early parts of our development, we should be able to use video images available from helicopter flights and other sources that are commonly used in motion analysis experiments.

Given suitable image (and other data), we expect to show the capabilities of detecting vehicles, tracking them through the sequence and indicating when the system believes the behavior to be abnormal. The output of the system can be displayed graphically and also tabulated for comparison with human assessments.

Our processing may not be necessarily at real-time speeds though computational efficiency will be a major concern. In later phases of the project, we can use the IFD platforms, if available, for real-time demonstrations.

References

- [1] R. Fielding, *The Technique of Special Effects Cinematography*, Focal/Hastings House, London, 3rd edition, 1972, pp. 220-243
- [2] G. Medioni, G. Guy, and H. Rom, *Video processing system for modifying a zone in successive images*, U.S. Patent # 5,436,672, July 1995.
- [3] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg, *A Three-Frame Algorithm for Estimating Two-Component Image Motion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No. 9, pp. 886-896, Sep. 1992.
- [4] P.J. Burt, *Fast Filter Transforms for Image Processing*, Computer Graphics and Image Processing, Vol. 16, pp. 20-51, 1981.
- [5] Y. Lamdan, J. Schwartz, and H. Wolfson, *Affine Invariant Model-Based Object Recognition*, Robotics and Automation(6), 1990, pp. 578-589.
- [6] P Anandan, P. Burt, K. Dana, M. Hansen, and G. van der Wal, "Real-time Scene Stabilization and Mosaic Construction," *Proceedings of the ARPA Image Understanding Workshop*, 1994, pp. 457-465.
- [7] J. R. Bergen, P. J. Burt, R. Hingorani, and S. Peleg, "A Three-Frame Algorithm for Estimating Two-Component Image Motion," *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, Sep 1992, pp. 886-896.
- [8] W. Franzen, "Structure and Motion from Uniform 3D Acceleration," in *Proceedings of the Workshop on Visual Motion*, IEEE Computer Society, 1991, pp. 14-20.
- [9] W. Franzen, "Structure from Chronogeneous Motion: A Summary," in *Proceedings of the DARPA Image Understanding Workshop*, San Diego, CA, January 1992.
- [10] S. L. Gazit and G. G. Medioni, "Multi-Scale Contour Matching in a Motion Sequence," *Proceedings of the DARPA Image Understanding Workshop*, 1989, pp. 934-943.
- [11] A. Huertas, M. Bejanin and R. Nevatia, "Model Registration and Validation," in *Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, April 1995, Ascona, Switzerland, pp. 33-42.
- [12] M. Irani and P. Anandan, "A Unified Approach to Moving Object Detection in 2D and 3D Scenes," *Proceedings of the DARPA Image Understanding Workshop*, 1996, pp. 707-718.

- [13] Y. C. Kim and K. Price, "Improved Correspondence in Multiple Frame Motion Analysis," in *Proceedings of the DARPA Image Understanding Workshop*, San Diego, CA, January 1992.
- [14] R. MacGregor, T. Russ and K. Price, "Knowledge Representation for Computer Vision: The VEIL Project," *Proceedings of the ARPA Image Understanding Workshop*, 1994, pp. 919-927.
- [15] G. Medioni, G. Guy, and H. Rom, U.S. Patent #5,436,672 *Video Processing System for Modifying a Zone in Successive Images*," awarded July 1995
- [16] G. Medioni and R. Nevatia, "Matching Images Using Linear Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):675-685, November 1984.
- [17] N. Milhau and G. Medioni, "Learning, Recognition and Navigation from a Sequence of Infrared Images," *Proc. of International Conference on Pattern Recognition*, Jerusalem, Israel, October 1994, pp. 822-825.
- [18] T. A. Russ, R. M. MacGregor, B. Salemi, K. Price, and R. Nevatia, "VEIL: Combining Semantic Knowledge with Image Understanding," *Proceedings of the DARPA Image Understanding Workshop*, 1996, pp. 373-380.
- [19] H. Sawhney, S. Ayer, M. Gorkani, "Model-Based 2D and 3D Dominant Motion Estimation for Mosaicing and Video Representation," in *Proceedings of International Conference on Computer Vision*, June 1995, pp. 583-590.
- [20] H. Shariat and K. Price. Motion estimation with more than two frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):417-434, May 1990.
- [21] C. Tomasi and T. Kanade, "The Factorization Method for the Recovery of Shape and Motion from Image Streams," in *Proceedings of the DARPA Image Understanding Workshop*, San Diego, California, January 1992, pp. 459-472.