MULTI-SOURCE FEATURE EXTRACTION AND VISUALIZATION IN URBAN ENVIRONMENTS

Edward M. Mikhail
Head, Geomatics Engineering
1284 Civil Engineering Building
Purdue University
West Lafayette, IN 47907-1284, USA
mikhail@ecn.purdue.edu

IC-13

KEY WORDS: Sensor Model, Invariance, Hyperspectral, Building Extraction, Road Grid, Visualization

ABSTRACT

Basic research is being performed by a team composed of specialists in photogrammetry, spatial image analysis, remote sensing, computer vision, and visualization, for the purpose of efficiently extracting urban features from multi-image sources and construction and visualization of the resulting database. The team members work cooperatively such that the effort is an integrated research. Topics discussed include: sensor modeling for data registration, photogrammetric invariance, DEM supported classification of hyperspectral imagery, DEM and thematic data supported building extraction, DEM supported road-grid extraction, and visualization in support of photogrammetry and exploitation research.

1 INTRODUCTION

Extraction of information from imagery has been the domain of photogrammetry, remote sensing, and image understanding/computer vision for many years. To be sure, the types of imagery used and the theories and techniques applied have varied somewhat from one of these three disciplines to another. Nevertheless, the primary objective of all is to obtain correctly labeled features which are geometrically and positionally accurate enough to be useful for a variety of applications. The practice in the past has been for researchers and practitioners in each of these three areas to work essentially independently from others. Of course, each area was aware of the activities of the others, and attempted to adapt and use methodologies developed by the others to the extent possible by their understanding of such methodologies. The increased prevalence of imagery in digital form, and the introduction of new sources of data, brings to focus the inadequacy of such independent pursuit of a similar goal. It has become quite apparent that combined integrated team research by experts in these fields is likely to yield significantly more than what can be expected from the sum of the individual efforts. Nowhere can this be more apparent than in the extraction and visualization of labeled spatial features in urban environments. This task has been, and continues to be, the most demanding in time and effort. In order to meet this challenge, and to put in place a team to address this problem in an integrated fashion, the US Army Research Office, under the Multidisciplinary University Research Initiative, MURI, awarded a 5-year project to Purdue University as the lead institution. The MURI team members and their speciality

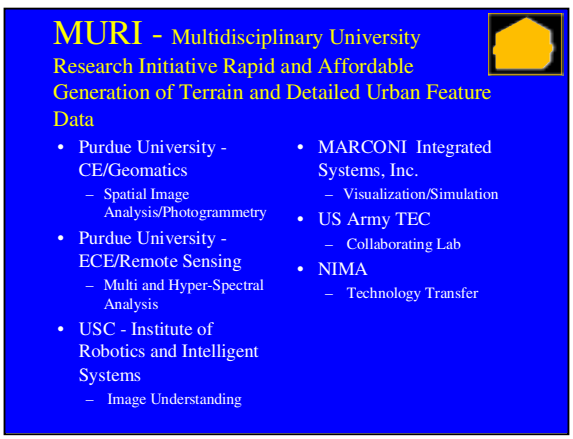

Figure 1. MURI Team Members

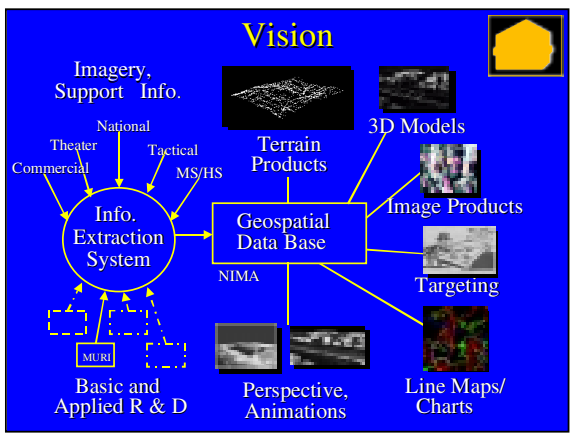

Figure 2. Vision

areas are depicted in Figure 1: Photogrammetry and Remote Sensing within Purdue University, the University of Southern California, and BAE SYSTEMS (formerly GDE then Marconi) as an industrial partner. The title of the project is: Rapid and Affordable Generation of Terrain and Detailed Urban Feature Data.

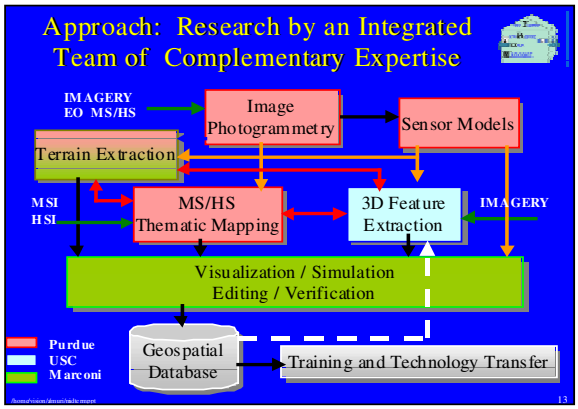

Figure 3. MURI Approach

The overall vision relative to which the MURI Research Center has been established is shown in Figure 2. Figure 3 illustrates the approach taken by the Center and the interaction between the Team members to accomplish the vision. The primary goal of the research is to economically construct an accurate three-dimensional database suitable for use in a visualization environment. Many sources of data are considered: hand-held, aerial, and satellite frame imagery, including video, both panchromatic and color; multi-spectral and hyper-spectral imagery; SAR, IFSAR and LIDAR data; and other information sources such as digital elevation models. Because of the diversity of the sources, rigorous mathematical modeling of the various sensing systems is imperative in order to accomplish accurate registration and fusion. Section 2 is therefore devoted to this aspect of the research, followed by the important task of spatial feature extraction in section 3, and database construction and visualization in section 4. The paper ends with section 5 on conclusions and recommendations for future research directions.

2 SENSOR MODELING AND MULTI-SOURCE REGISTRATION

Since several different data sources are considered as input to the feature extraction module, it is imperative that they are "registered" with respect to each other and relative to the terrain object space. In the case of imagery, registration means that the mathematical model of the sensor acquiring the imagery is rigorously constructed and recovered . Two types of passive sensors will be discussed: Frame and Push-broom, each of which will be discussed in a separate subsection. Accurate sensor models are also important for the generation of digital elevation models which are not only a product in their own right, but are also used in support of other tasks such as hyperspectral image classification and cultural feature extraction as will be discussed later.

2.1 Modeling For Frame Singles and Sequences

Frame imagery has been the most common form and its modeling has therefore been discussed extensively in the photogrammetric literature over the years. Each frame is assigned six exterior orientation (EO) elements, and usually three geometric interior orientation (IO) elements. When considering uncalibrated digital cameras, the IO elements are often augmented by several more parameters that account for some or all of the following: skewness, differential scale, and radial and decentering lens distortion coefficients. These are explicitly carried as parameters in the pair of photogrammetric collinearity equations for each ray. Since the equations are non-linear, it is important to have reasonable approximations for the unknown parameters. Such approximations are sometimes difficult to obtain particularly for unusual image acquisition geometries of oblique aerial photography and hand-held imagery. The linear invariance-based formulation is useful for quickly deriving approximations for camera parameters. One formulation is for a pair of overlapping images in which the image coordinates are related by the *Fundamental Matrix*, *F*, or

$$\begin{bmatrix} x_1 & y_1 & 1 \end{bmatrix} F \begin{bmatrix} x_2 & y_2 & 1 \end{bmatrix}^T = 0 \qquad (1)$$

Although *F* has 9 elements, only 7 are independent. As an example, this technique is applied to a pair of convergent video frames, Figure 4. After *F* is estimated, relative camera transformation matrices for each of the two video frames can be extracted from the fundamental matrix. Then projective model coordinates can be computed for any known ground control point visible on the two frames. Using the known ground control point coordinates in the 3D orthogonal system and their corresponding projective model coordinates, fifteen elements of the 4x4 three-dimensional projective transformation matrix can be estimated. Now the true camera transformation matrices are computed by multiplying the relative camera transformation matrices by the projective transformation matrix. Finally, the real camera parameters are extracted from the camera transformation matrices.

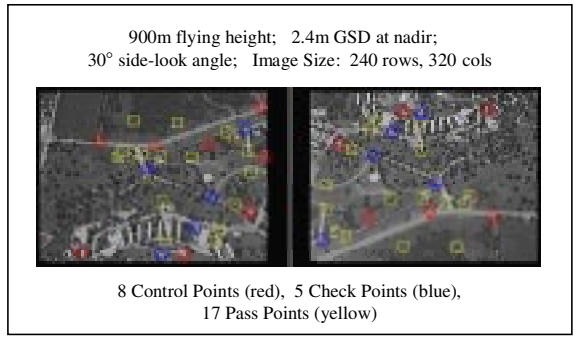

Figure 4. Invariance Applied to a Video Pair

The parameters estimated from invariance may then be used as good approximations in a rigorous non-linear photogrammetric solution. Table 2.1 shows the control point and check point RMS results for both the invariance and the rigorous photogrammetric solutions. In this case, the recovered parameters for each camera included the 6 EO parameters, the three geometric IO parameters, and one radial lens distortion coefficient, K_1. Although the linear invariance technique was helpful in obtaining initial approximations for camera parameters, the use of rigorous photogrammetry with added IO parameters significantly improved the RMS results.

Case	Control Point RMS (m)				Check Point RMS (m)			
	X	Y	Z	R	X	Y	Z	R
Invariance	0.94	3.25	1.58	3.73	6.14	2.68	2.09	7.02
Rigorous Photogrammetry	0.45	0.84	1.69	1.94	1.13	2.87	1.47	3.41

Table 2.1 Two-Frame Video Triangulation Results

Another useful application of invariance is in *image transfer*. Image transfer is an application performed on a triplet of images. Given two pairs of measured image coordinates, the third pair can be calculated using a previously established relationship among pairs of image coordinates on all three images. Six techniques, including two based on the F-matrix (see equation (1)), three based on the so-called trilinearity equations (Theiss, et al, 2000), and one collinearity technique, have been investigated for image transfer. As an example, a data set over Fort Hood consists of two near vertical aerial frame photographs taken at 1650 meters above mean terrain, and one low-oblique aerial frame photograph taken at a flying height of 2340 meters with a 25 degree (from the vertical) side-looking angle; see Figure 5. Nineteen reference points measured on each of the three photographs were used to establish the image coordinate relationships. Then, for 18 check points the image coordinates from two photographs were used to compute the transferred positions on the third, and the transferred positions were compared to their measured values. The results for all of the models are shown in Table 2.2.

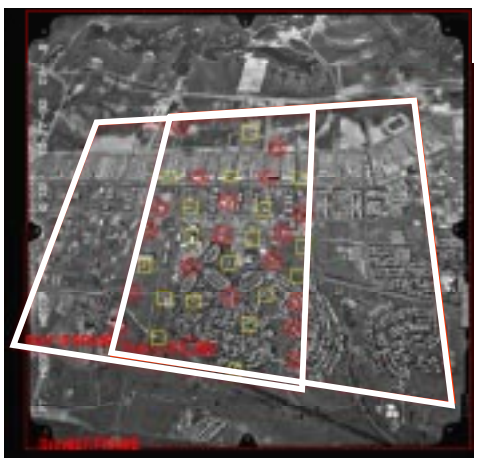

Figure 5. Fort Hood: Footprints of 2 Vertical Aerial Frames Overlaid on 1 Oblique Frame

Model	x RMS (pixels)	y RMS (pixels)
Trilinearity 1*	0.46	0.58
Trilinearity 2**	0.47	0.61
Trilinearity 3**	0.47	0.61
Collinearity	0.47	0.61
F-matrix, 3F's	0.54	0.62
F-matrix, 2F's	0.50	0.62

Table 2.2 Image Transfer Experiments with Fort Hood Data

* After scaling image coordinates to range from -1 to +1.
** After rotating image coordinates by 90 degrees.

As noted by the asterisks below Table 2.2, the raw image coordinate data must be augmented for Models 1-3 in order to obtain those results. Since Model 1 does not rigorously linearize with respect to the observations, the image

coordinates must be scaled in order to prevent the solution from becoming unstable. A degenerate case occurs for Models 2 and 3 for this particular case of aerial photography where the air base direction between the two near vertical photographs is parallel to the image x coordinate direction, and an independent subset of the trilinearity equations is selected.

2.2 Modeling For Non-Frame Imagery

One of the significant accomplishments of the MURI project has been the integration of remote sensing analysis with the task of extraction of urban features. This has been made possible by the availability of high spatial and spectral resolution image data such as generated by sensor systems known as HYDICE and HyMap.

2.2.1 HYDICE Modeling (Push-broom)

One HYDICE image contains 320 columns, and typically consists of four major frames each containing 320 lines, resulting in a 320 column by 1280 line image for each of the 210 bands of the hyperspectral sensor. At constant time intervals associated with each individual line of the pushbroom scan, 320 by 210 pixel arrays called minor frames are exposed; see Figure 6(a). Since the geometric distortions that exist among the 210 bands are negligible, rectification is performed on just one of the bands which depicts features on the ground clearly. The same transformation may then be applied to any of the other bands, or later to the thematic image after each pixel has been classified.

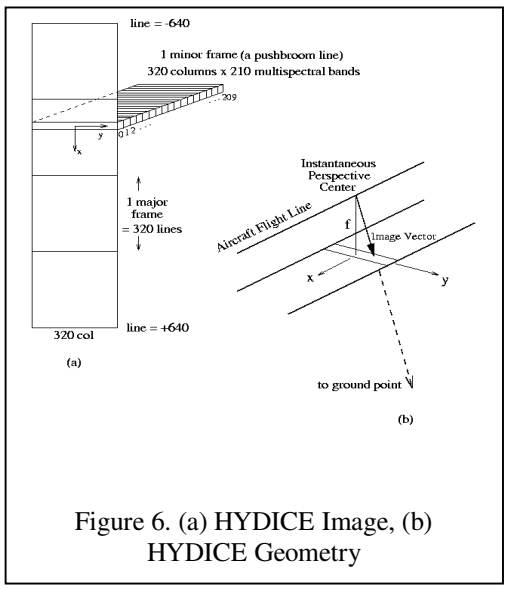

Figure 6. (a) HYDICE Image, (b) HYDICE Geometry

Mathematical modeling includes sensor and platform models. The objective of sensor modeling is to relate pixels on an image to coordinates in an orthogonal 3-dimensional sensor coordinate system (SCS). At any given instant of time, we can imagine the HYDICE sensor positioned along its flight trajectory at the instantaneous perspective center, coinciding with the origin of the SCS; see Figure 6(b). At this time instant 1 minor frame consisting of a line of 320 pixels is exposed.

Platform modeling involves determining the exterior orientation of the instantaneous perspective center, i.e. origin of the SCS, with respect to the ground coordinate system. Three items are considered: the data that is recorded in real time in the header of the HYDICE imagery; piecewise polynomial as platform model; and the concept of a Gauss-Markov process and its application to platform modeling.

There are six time-dependent elements of exterior orientation consisting of three coordinates for position and three angles for orientation. At one second intervals, the easting, northing, and height are recorded from the Global Positioning System (GPS), which is operating in differential mode on board the aircraft. When functioning properly, the standard deviations on the horizontal and vertical components of position are 0.4 and 0.9 meters, respectively. The GPS data are used as a priori values, although they are not fixed, in both of the platform models considered.

Roll, pitch, and yaw angular values and rates are supplied by the inertial navigation system (INS) of the aircraft for every minor frame of the HYDICE image; i.e., for each line. These data express the orientation of the aircraft with respect to an inertial ground system in terms of three non-sequential angles. A flight stabilization platform (FSP) is used aboard the aircraft to keep the orientation of the sensor roughly constant by compensating for changes in the orientation of the aircraft. Three non-sequential angles for the FSP are recorded for each minor frame. Errors in the INS data prevented it from being fully exploited in our experiments.

The piecewise polynomial approach involves the recovery of polynomial coefficients for each of the six elements of exterior orientation. A different set of coefficients may be recovered for each segment of an image that has been divided into sections. Constraints on the parameters, such as continuity, may be imposed at the section boundaries. Although sufficient for the modeling of satellite pushbroom scanners, which are in a stable orbit, this method appears to be too rigid for the modeling of an aircraft flight path, and therefore a more flexible approach was sought.

In the Gauss-Markov approach, six parameters per line are carried to model the instantaneous exterior orientation for each pushbroom line. Parameters for each image line are tied, or constrained, stochastically to those of the previous image line. This model allows for greater flexibility for linear feature constraints to contribute to parameter recovery

thereby improving rectification accuracy. The criterion for a first order Markov process is that the probability distribution function $F[x(t)]$ of the random process $x(t)$ is dependent only on the one most recent point in time.

For each line of imagery in which a point is observed, two collinearity condition equations are written as in the case for the piecewise polynomial model. Assuming that the interior orientation of the sensor is known, there is a total of 6L parameters carried in the least squares adjustment, where L is the total number of lines in the image. Therefore, for each line in the image starting with the second line, six equations are written which reflect the constraints resulting form the Gauss-Markov process.

Although these six equations per line are treated as observation equations in the least squares adjustment algorithm, they are essentially constraint equations effectively reducing the number of unknown parameters from 6L to 6. Therefore, a unique solution may be obtained if three control points are available.

As the number of observed points corresponding to control points or linear features increases, the redundant measurements can contribute significantly to the recovery of exterior orientation elements in the vicinity of the observation. This effect occurs if the weights assigned to the constraint equations are low enough to allow the parameters to vary significantly from one line to the next. When the platform model provides this flexibility in parameter recovery, there are greater than six independent parameters being estimated; therefore the redundancy is less than 2P - 6, where P is the number of control points.

Similarly, the second order Gauss-Markov process, $(x(t))$, can be defined as a Gaussian random process whose conditional probability distribution is dependent only on the two previous points. For each scan line in the image staring with the third line, six equations are written. These 6L-12 equations reduce the number of unknown parameters from 6L to 12.

2.2.2 HyMap Modeling (Whisk-broom)

The HyMap (Hyperspectral Mapping) sensor uses a whiskbroom scanner, unlike the HYDICE sensor that uses a pushbroom scanner. It sweeps from side to side as the platform moves forward. Therefore each image pixel, which is collected at a different time, requires its own set of six exterior orientation elements. To simplify this situation, it is assumed that the
time to complete one scan line is small enough to consider one exposure station for each scan line. Then, each scan line of a whiskbroom image can be modeled as a panoramic image, instead of a framelet as used in the pushbroom model. Modeling from line to line remains the same for both imaging modes.

2.2.3 Control Features

The most common control feature used in the triangulation of multispectral imagery as well as traditional frame photography is the control point. Control point coordinates are commonly obtained from a field survey or a GPS survey. In our data set, however, the control point coordinates were easily and reliably obtained from the triangulation of pass points in frame photography that included the entire HYDICE coverage. The image coordinates (x,y) are line and sample values measured on the HYDICE imagery.

Linear features offer some advantages over control points in their use as control features. Linear features are abundant in urban imagery and are often easier to extract using automated tools. When used in overlapping imagery, point to point correspondence is not required. Furthermore, without knowing their absolute location on the ground, linear features and their inherent constraints can contribute significantly to a solution by reducing the ground control requirements.

Although the term linear feature encompasses any continuous feature with a negligible width and includes such parameterizations as splines, we limit consideration to straight line segments. Although there are several different parameterizations of a straight line in 3D space, we choose to model object space lines using two end points. At a given scan line, an image vector, which is rotated to the ground coordinate system, should be on a plane that is defined by three points: two end points of the line in the ground space and the position of the instantaneous perspective center.

2.2.4 Experimental Results

Two HYDICE images are used. The first data set was collected over the Washington DC mall in August 1995 (Figure 7). Its ground sample distance is about 3.2 meters. Its flight height was about 6320m. From Figure 7, the straight line features like roads and building edges along the flight direction display a modest degree of roll-induced "waviness".

Figure 7. HYDICE Imagery, Washington, D.C.

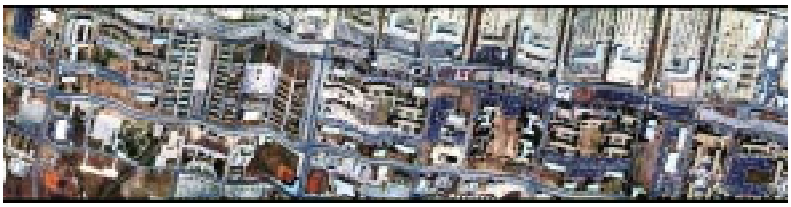

Figure 8. HYDICE Imagery, Fort Hood

The second data set, flown over the urban area of Fort Hood, Texas in October 1995, is shown in (Figure 8). Its ground sample distance is 2.2 meters, respectively. Its flight height was about 4430m. As can be seen from Figure 8, straight roads along the in-track direction are severely wavy.

Using the first Gauss-Morkov model, the orthorectified images corresponding to Figures (7) and (8) are shown in Figures (9) and (10) respectively.

3 URBAN FEATURE EXTRACTION

3.1 Hyperspectral Analysis

Remote sensing techniques have been used for many years for the classification of multispectral imagery. However, until recently, most of this type of imagery did not have sufficiently fine spatial resolution to make it useful in urban environment. Now imagery such as HYDICE and HyMap discussed in section 2 offer excellent urban data. Figures 7 and 9 showed the original and orthorectified 3-color images of an airborne hyperspectral data flightline over the Washington DC Mall. Hyperspectral sensors gather data in a large number of spectral bands (a few 10′s to several hundred). In this case there were 210 bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. This data set contains 1208 scan lines with 307 pixels in each scan line. It totals approximately 150 Megabytes. With data that complex, one might expect a rather complex analysis process, however, it has been possible to find quite simple and inexpensive means to do so. The steps used and the time needed on a personal computer for this analysis are listed in Table 3.1 and described as follows:

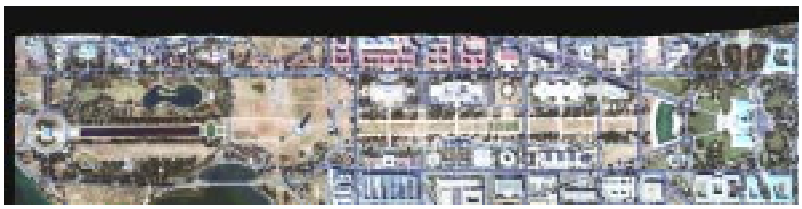

Figure 9. Ortho-rectified Image (Gauss-Markov, Washington, D.C.)

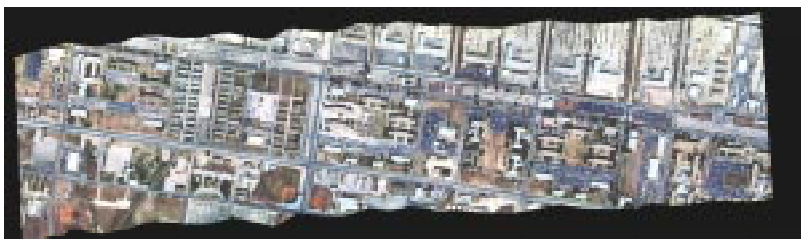

Figure 10. Ortho-rectified Image (Gauss-Markov, Fort Hood)

Define Classes. A software application program called MultiSpec, available to anyone at no cost from http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/, is used. The first step is to present to the analyst a view of the data set in image form so that training samples, examples of each class desired in the final thematic map, can be marked. A simulated color infrared photograph form is convenient for this purpose; to do so, three bands are used in MultiSpec for the red, green, and blue colors, respectively. (See Figures 7 and 9).

Feature Extraction. After designating an initial set of training areas, a feature extraction algorithm is applied to determine a feature subspace that is optimal for discriminating between the specific classes defined. The algorithm used is called Discriminate Analysis Feature Extraction (DAFE). The result is a linear combination of the original 210 bands to form 210 new bands that automatically occur in descending order of their value for producing an effective discrimination. From the MultiSpec output, it is seen that the first nine of these new features will be adequate for successfully discriminating between the classes.

Operation	CPU Time (sec.)	Analyst Time
Display Image	18	
Define Classes		< 20 min.
Feature Extraction	12	
Reformat	67	
Initial Classification	34	
Inspect and Add 2 Training Fields		. 5 min.
Final Classification	33	
Total	164 sec = 2.7 min.	. 25 min.

Table 3.1

Reformatting. The new features defined above are used to create a 9 band data set consisting of the first nine of the new features, thus reducing the dimensionality of the data set from 210 to 9.

Initial Classification. Having defined the classes and the features, next an initial classification is carried out. An algorithm in MultiSpec called ECHO (Extraction and Classification of Homogeneous Objects) is used. This algorithm is a maximum likelihood classifier that first segments the scene into spectrally homogeneous objects. It then classifies the objects.

Finalize Training. An inspection of the initial classification result indicates that some improvement in the set of classes is called for. To do so, two additional training fields were selected and added to the training set.

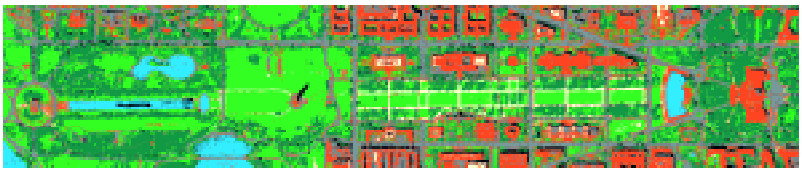

Figure 11. Orthorectified Maximum Likelihood Classification

Final Classification. The data were again classified using the new training set. The result is shown in Figure 11. Note that the procedure used here does not require complex preprocessing such as correction for atmospheric effects, absolute calibration, transformation from radiance to reflectance, etc. The analysis required approximately 2.5 minutes of CPU time on a Power Macintosh G3 machine and 27 minutes of analyst time.

3.2 DEM Supported Hyperspectral Analysis

This is an experimental study where using a fusion of two essentially different types of data proves significantly superior to the individual use of either one or the other. The task is to identify and accurately delineate building roof-tops in the flightline of hyperspectral data of the Washington D.C. Mall, supplemented with digital elevation model (DEM) data for each pixel of the scene, Figure 12.

Figure 12. Digital Elevation Map (DEM)

Experiments using gradient-based algorithms on the DEM data show that its use alone is not sufficient to sharply delineate building boundaries. A spectral classifier does not have region boundary problems. However, building roof-tops in this urban scene are constructed of different materials and are in various states of condition and illumination. This and the fact that, in some cases, the material used in roof-tops is spectrally similar to that used in streets and parking areas make this a challenging classification problem, even for hyperspectral data.

Figure 13. Gradient Operation on Section of DEM

It is shown here that combining hyperspectral and DEM data can substantially sharpen the identification of building boundaries, reduce classification error, and lessen dependence on the analyst for classifier construction.

The information content in the DEM is in the rise in elevation of a given area-element in relation to its neighbor. The use of a gradient operator in identifying building pixels (presumably at higher elevation than ground level) is thus appropriate. A Sobel gradient operator was used on the DEM. A gradient-threshold was applied to the result to obtain Figure 13.

The analyses can now be compared: Spectral analysis focuses on pixel-wise identification of the class rooftop. Note that the task desires the identification of a specific usage (rooftop) in the scene, rather than the material classification provided by spectral analysis. Thus, there is the possibility that spectrally similar materials will be identified with the roof class, regardless of the manner of their usage in the scene. Gradient operator based analysis identifies the building boundaries. In essence, the latter is a scheme to delineate building boundaries, while the other is a pixel classification scheme. Figure 14 shows extracted rooftops.

The output in Figure 13 outlines buildings as objects with thick boundaries. It is possible to thin the delineated scene objects by setting a high threshold on the output of the gradient operator. However this requires operand manipulation on the part of the analyst, and is inefficient.

In general, spectral analysis is more robust over an extended scene. For instance, should the analyst note a different 'type' of building rooftop in isolation, the set of scene-classes can be enlarged and training data included appropriately. On the other hand, analysis of the DEM can be complicated by hilly terrain. In Figure 12, note the rise to the Capitol Hill at the far right end of the DEM. It is evident that this particular section has to be processed in isolation.

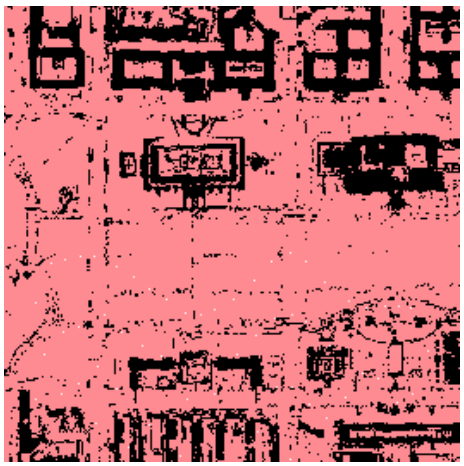

Figure 14. Extracted Rooftops from Spectral Data

In Figure. 14 we can observe considerable speckle misclassifications in the output. In general there is some confusion in separating rooftop - class data from spectrally similar classes asphalt and gravel path.

In highlighting the shortcomings of the respective analyses it has been implicit that the problems associated with one technique can be alleviated through the use of the other. For instance, the last point in the discussion above leads to a significant conclusion. The emergence of inter-class confusion in classification is not a result of "wrong" data. The material used in construction of building rooftops is, quite often, identical to that used in constructing roads, or laying paths. However, the scene-classes are functionally distinct, and this distinction is strikingly apparent in the DEM. This conclusion is key to the solution presented in the next section.

Procedure: Given the disparity in the two types of the data, concurrent analysis is infeasible. Our analysis comprised maximum likelihood classification, as discussed earlier, followed by a thresholding operation on the elevation of all data elements identified as asphalt, gravel path or rooftop. The latter is designed as a Boolean-type operation in which all data (identified as one of the three classes listed above) below a certain elevation are said to be ground-level; the remaining filtered data are thus identified as building-rooftop.

Since there is a large amount of variation in scene elevation, the elevation threshold, discussed above, must be locally determined. The following procedure was adopted towards this task.

Centroid Identification: The DEM was visually examined to identify zones or regions of relatively unchanging terrain. Pixels representative of these zones were identified as zone centroids.

Zoning: The pixel grid was then segmented into zones identified by their respective centroid. The process involved going through the grid and labeling each pixel according to the zone centroid closes to it. The metropolis distance metric was used. The partitioned image is shown in Figure 15. Zone centroids have been highlighted as yellow dots in the figure. Note that only pixels identified as rooftop, asphalt or gravel path are identified in the zoned output. The remaining scene classes have been absorbed into the black background.

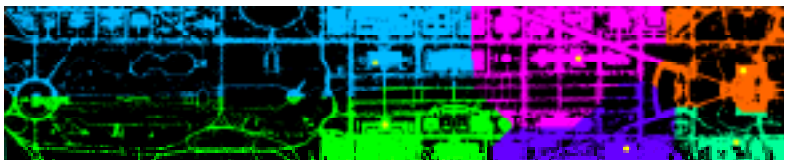

Figure 15. Partitioned Scene with Centroid Identification

Threshold computation: For each zone, the median elevation for the pixels classified as rooftop, asphalt or gravel path is computed. In zones with an insufficient count of rooftop pixels, it is clear that threshold will be biased towards data at ground-elevations. The threshold for a given zone is thus chosen as the average of the median as calculated above, and the elevation of the zone-centroid.

Thresholding operation: The thresholds, thus computed, were used to get the result shown in Figure 16. Note that the rooftops have been color-coded by the identifying zone.

Figure 16. Rooftops Identified via Data-Fusion

Discussion: In the above analysis, we identified the key attributes of the respective datasets available to us. Spectral data is best used in the identification of elemental composition, while the DEM identifies the data element in the functional sense. Data fusion is thus justifiable, with the analysis utilizing the respective attributes of the HYDICE data and the DEM towards the target application.
It is critical to point out that the quality of the fusion of the DEM data and the hyperspectral data depends on the rigor of modeling the hyperspectral sensor. As was shown in section 2, the Gauss-Markov and use of linear features yielded excellent rectification of that imagery. It will also be shown in subsequent section that using computer vision techniques on frame imagery fused with other data depends upon rigorous photogrammetric modeling.

3.3 Building Extraction

Computer vision based three-dimensional urban feature extraction from frame imagery encounters many sources of difficulty including those of segmentation, 3-D inference, and shape description.

Segmentation is difficult due to the presence of large numbers of objects that are not intended to be modeled such as sidewalks, landscaping, trees and shadows near the objects to be modeled. The objects to be modeled may be partially occluded and contain significant surface texture. 3-D information is not explicit in an intensity image; its inference from multiple images requires finding correct corresponding points or features in two or more images. Direct ranging techniques such as those using LIDAR or IFSAR can provide highly useful 3-D data though the data typically have areas of missing elements and may contain some points with grossly erroneous values.

Once the objects have been segmented and 3-D shape recovered, the task of shape description still remains. This consists of forming complex shapes from simpler shapes that may be detected at earlier stages. For example, a building may have several wings, possibly of different heights, that may be detected as separate parts rather than one structure initially. The approach used in this effort is to use a combination of tools: reconstruction and reasoning in 3-D, use of multiple sources of data and perceptual grouping. Context and domain knowledge guide the applications of these tools. Context comes from knowledge of camera parameters, geometry of objects to be detected and illumination conditions (primarily the sun position). Some knowledge of the approximate terrain is also utilized. The information from sensors of different modalities is fused not at pixel level but at higher feature levels.

Our building detection system is based on a "hypothesize and verify" paradigm. This system can function with just a pair of panchromatic (PAN) images, but can also utilize more images and information from other modalities. This system also incorporates abilities for Bayesian reasoning and machine learning.

3.3.1 Multi-View System, or MVS

A block diagram of the extraction system is shown in Figure 17. The approach is basically one of hypothesize and verify. Hypotheses for potential roofs are made from fragmented lower level image features. The system is hierarchical and uses evidence from all the views in a non-preferential, order-independent way. Promising hypotheses are selected among these by using relatively inexpensive evidence from the rooftops only. The selected hypotheses are then verified by using more reliable global evidence such as from walls and shadows. The verified hypotheses are then examined for overlap which may result in either elimination or in merging of them. This system is designed for rectilinear buildings; complex buildings are decomposed into rectangular parts. Rooftops thus project to parallelograms in the images (the projection is nearly orthographic over the scale of a building). Lines, junctions and parallel lines are the basic features used to form roof hypotheses. Consider the images shown in Figure 18. The images are from the Ft. Hood, Texas, site that has been in common use by many researchers. The low level features composed of lines, junctions between lines and sets of parallel lines are matched among the available views. Two views were used in this example. The set of lines extracted from the image (using a Canny edge detector) to start the process is shown in Figure 19. Roof hypotheses are formed by a pair of matched parallel lines and *U* structures (*U*s represent three sides of a parallelogram). A pair of parallel lines may be matched to parallels in more than one view (when more than two views are used) and each matching pair is considered. Closed hypotheses are formed from these features by using the best available image lines if any, else closures are synthesized from the ends of the parallel lines.

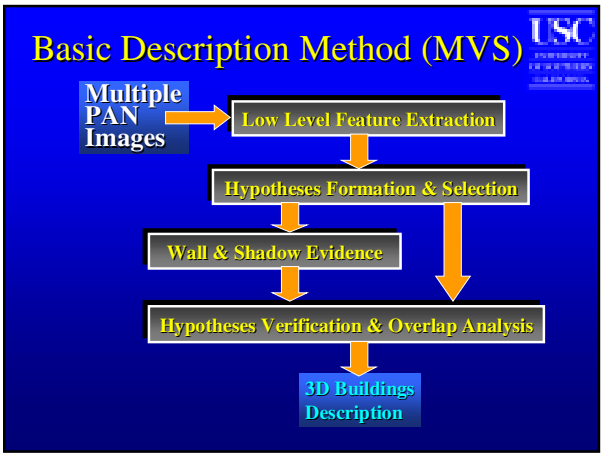

Figure 17. MVS Extraction System

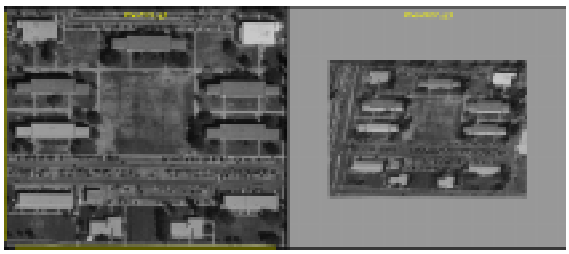

Figure 18. Near vertical (left) and oblique (right) frame photographs of Fort Hood, Texas

The next step is to verify whether the selected hypotheses have additional evidence in support of being buildings. This evidence is collected from the roof, the walls and the shadows that should be cast by the building. Since the hypotheses are represented in 3-D, deriving the projections of the walls and shadows cast, and determining which of these elements are visible from the particular view point is possible. These in turn guide the search procedures that look in the various images for evidence of these elements among the features extracted from the image. A score is computed for each evidence element. Each of the collected evidence parameters is composed of smaller pieces of evidence. A critical question is how to combine these small pieces of evidence to decide whether a building is present or not and how much confidence should be put in it. Results shown in this paper use a Bayesian reasoning approach.

After verification, several overlapping verified hypotheses may remain. Only one of the significantly overlapping hypotheses is selected. The overlap analysis procedure examines not only the evidence available for alternatives but also separately the evidence for components that are not common. Figure 20 shows the wireframes of the detected buildings from the pair of images. Note that while most of the buildings are detected correctly, some are missing.

The system presented above relies on image intensities from multiple overlapping images. The performance of the building detection and description system can be greatly improved if information from other sensors become available. As described above, our system can take advantage of multiple panchromatic (PAN) images even if they are not acquired at the same time. We consider two other sources of a different modality.

The first source of additional information is digital elevation models (DEMs). DEMs may be derived from stereo PAN images or acquired directly by active sensors such as LIDAR or IFSAR. The second source of information is from multi- or hyper-spectral imagery, such as from the HYDICE or HyMap sensors, which is becoming increasingly more available.

DEMs make the task of building detection much easier as the buildings are significantly higher than the surround and accompanied by sharp depth discontinuities. However, DEM data may not be accurate near the building boundaries, and the active sensors may contain significant artifacts. The spectral information makes it easier to decide if two pixels

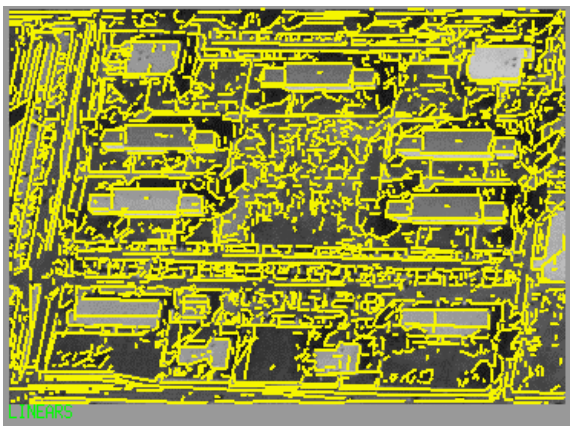

Figure 19. Linear segments from image in Figure 18

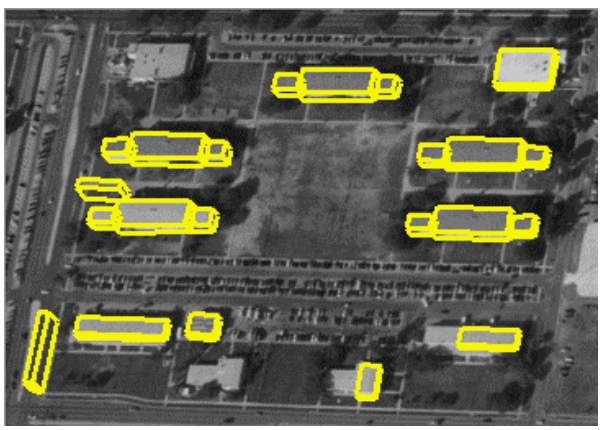

Figure 20. 3D wireframe model of detected buildings

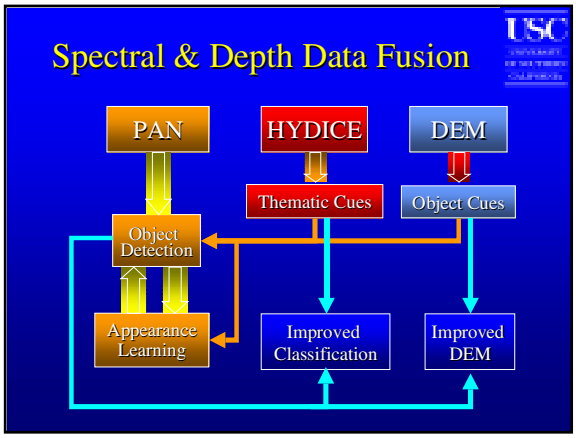

Figure 21. Multi-Sensor Information Integration

belong to the same class, and hence to the same object, or not. However, objects, such as building rooftops, are not always homogeneous in material and the hyper-spectral data is usually of a significantly lower resolution than that of PAN images. For these reasons, we have decided, at least at present, to use DEM and spectral sensors to provide cues for the presence of buildings but to use PAN images for accurate delineation. Figure 21 shows a block diagram of our approach. The left most column denotes the multi-view system described above. If DEM data is available, object cues are extracted from it and supplied to MVS where this information can be used to aid in the process of hypothesis formation and selection. Similarly, HYDICE data is analyzed to produce thematic maps which again aid in the process of hypothesis formation and selection for MVS. These processes are described in some detail next.

3.3.2 DEM Supported MVS

The DEM for the Ft. Hood site, corresponding to the area shown earlier in Figure 18, is shown in Figure 22 (displayed intensity is proportional to elevation.) Note that while the building areas are clearly visible in the DEM, their boundaries are not smooth and not highly accurate. These characteristics prevent direct extraction of buildings from DEM images but clearly can help cue the presence of 3-D objects. The building regions in a DEM are characterized as being higher

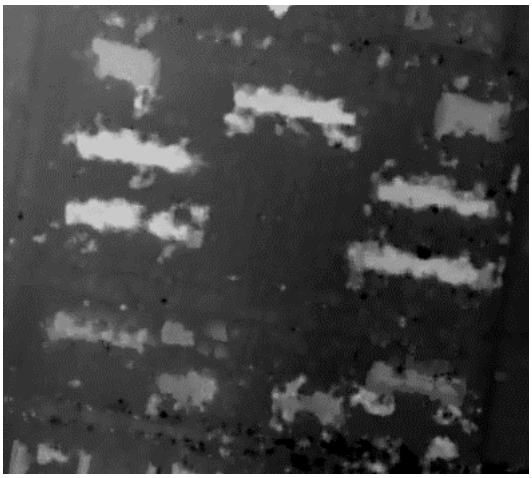

Figure 22. DEM corresponding to image in Figure 18

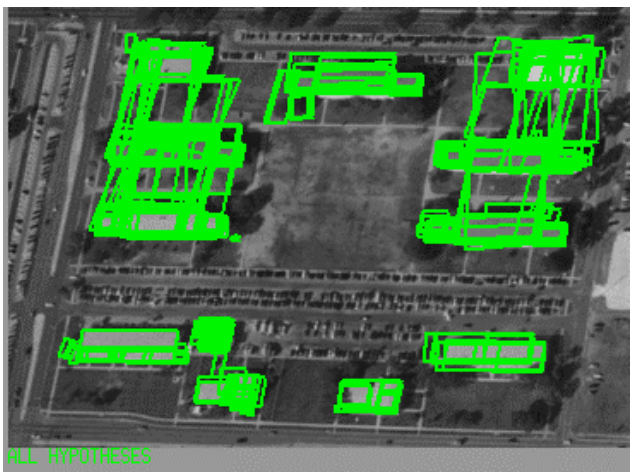

Figure 23. Lines near DEM cues

than the surround. However, simple thresholding of the DEM is not sufficient, as height variations of the magnitude of a single story building can occur even in very flat terrain sites. Our approach is to convolve the image with a Laplacian-of-Gaussian filter that smoothes the image and locates the object boundaries by the positive-valued regions bounded by the zero-crossings in the convolution output. Object cues are used in several ways and at different stages of the hypothesis formation and validation processes; they can be used to significantly reduce the number of hypotheses that are formed by only considering line segments that are within or near the cue regions. The 3-D location of a line segment in the 2-D PAN images is not known. To determine whether a line segment is near a DEM cue region we project the line onto the cue image at a range of heights, and determine if the projected line intersects a cue region. Figure 19 earlier showed the line segments detected in the image of Figure 18; Figure 23 shows the lines that lie near the DEM cues. As can be seen, the number of lines is reduced drastically (81.5%) by filtering without losing any of the lines needed for forming building hypotheses. This not only results in a significant reduction in computational complexity but many false hypotheses are eliminated allowing us to be more liberal in the hypotheses formation and thus including hypotheses that may have been missed otherwise. We also use these cues to help select and verify promising hypotheses, or conversely, to help disregard hypotheses that may not correspond to objects. Just as poor hypotheses can be discarded because they lack DEM support, the ones that have a large support see their confidence increase during the verification stage. In this stage, the selected hypotheses are analyzed to verify the presence of shadow evidence and wall

evidence. When no evidence of walls or shadows is found, we require that the DEM evidence (overlap) be higher, to validate a hypothesis. The 3-D Models constructed with DEM support from the validated hypotheses are shown in Figure 24. Comparing it to Figure 20 shows that false detections are eliminated with DEM cueing. Also, the building components on the top left and on the lower part are not found without DEM support but found with it. Once the buildings have been detected, the DEM can also be improved by replacing parts of the DEM with building models.

3.3.3 MVS Supported by Thematic Data

The HYDICE image strip shown in Figure 10 was classified as described in section 3.1 and rectified, thus producing a useful thematic map. To extract cues we first extract the roof pixels from the thematic map. Many pixels in small regions are misclassified or correspond to objects made of similar materials as the roofs. The building cues extracted from this image are the connected components of certain minimum size.

HYDICE cues are used, in ways similar to those for the DEM cues described above, at different stages of the hypothesis formation and validation processes. The linear segments near HYDICE cues, are very similar to those shown earlier in Figure 23 with an increased reduction in the number of lines (84%). As with the DEMs the HYDYCE evidence helps simplify the hypothesis selection process. The evidence consists of support of a roof hypothesis in terms of the overlap between the roof hypotheses and the HYDICE cue regions. The hypotheses are constructed from matching features in multiple (two in our Ft. Hood example) images and are represented by 3-D rectilinear components in 3-D world coordinates. We can therefore project them directly onto the HYDICE cues image to compute roof overlap. The system requires that the overlap be at least 50% of the projected roof area.

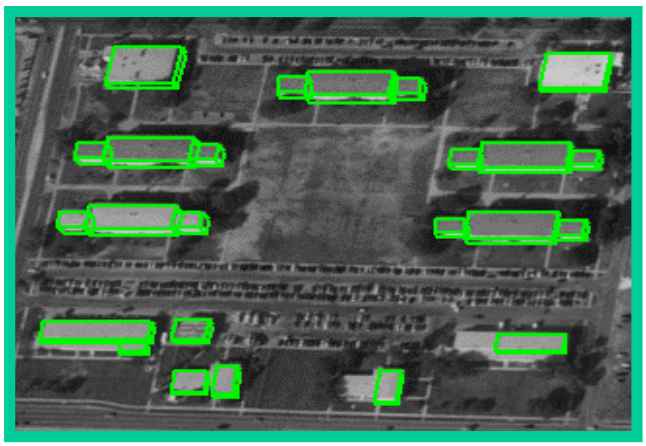

Figure 24. Building components extracted using DEM cues

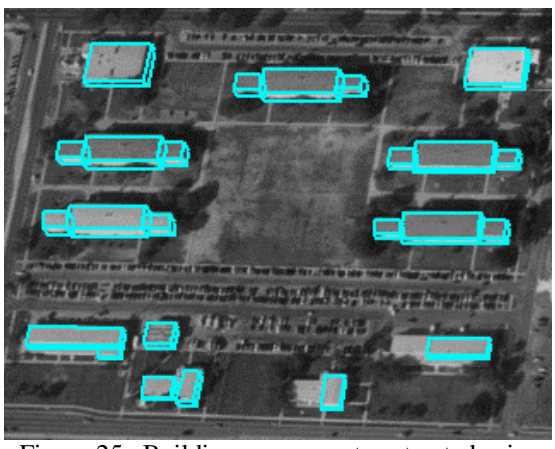

Figure 25. Building components extracted using HYDICE cues

Figure 25 shows the detected buildings using the HYDICE cues. This result shows no false alarms. Once the buildings have been detected, the roof class can also be updated. The performance of the MVS system is very similar using DEM or HYDICE cues. There will be many cases where the quality of the cues from one sensor may be higher. It is appropriate to characterize this quality and combine the support from various sensor modalities. This is the subject of our current work.

3.4 Road Grid Extraction

The system uses a simple three-dimensional road segment and intersection model and known camera parameters, which allows the use of either nadir or oblique views. Roads are assumed to have visible edges without significant occlusions. Since we use the geometric structure of the road and the intersection, vehicles and markings on the road are not a serious drawback. Indeed, for verification, they may be an important feature. We also assume a regular street grid but that the program must detect where the regular grid ends. Some variations in the grid are detected by using a grid model that is smaller (e.g. 1/2 or 1/3) than the actual street grid.

The system requires only a few interactive steps, which could be performed by imperfect automated techniques that have been suggested in the literature. By delaying total automation, we are able to focus on the important issues of using context and grouping for street grid extraction. The only inputs from the user are three points (i.e. the center of three intersections) that give the location, direction and spacing of the street grid. This step can be replaced by automatic methods to find dominant direction and spacings, but these are less reliable and not a focus of this current

work. These three points define an initial model containing four road segments and the intersection that they define. An example of this initial model is shown in Figure 26. Each intersection has four road segments and each road segment connects two intersections. The verification problem is to determine whether each of the road segments exists. Figure 27 shows a small portion of such a grid model which must be verified. Since actual road widths (in meters) vary from scene to scene, we allow the user to adjust the default width to fit the particular scene. A width refinement step, later in the extraction procedure, reduces the need for exact initial widths.

Figure 26. Initial Grid Model

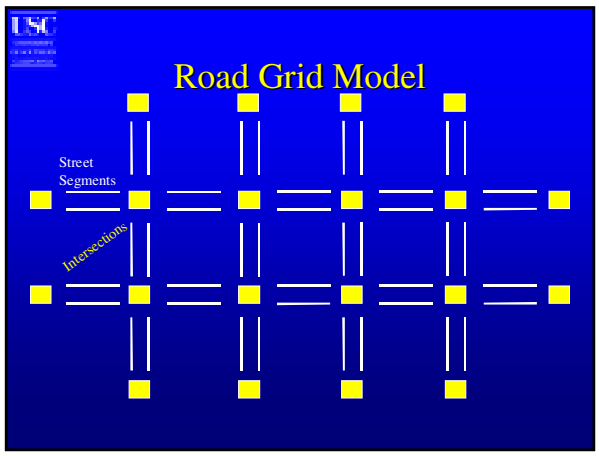

Figure 27. Small portion of road grid model

The grid extraction procedure is composed of two phases, the first tests each intersection (using the four road segments for the model) to find which ones are supported by the image data. This hypothesize- and-verify phase propagates the grid across the entire scene and provides an initial geometric match for the scene. Figure 28 outlines this phase of the procedure. The second phase uses this initial match and tests triples of road segments (three consecutive road segments) to find the best location and width for each triple. Figure 29 illustrates this second phase. These results provide the input for further use of context and refinement using other data sources.

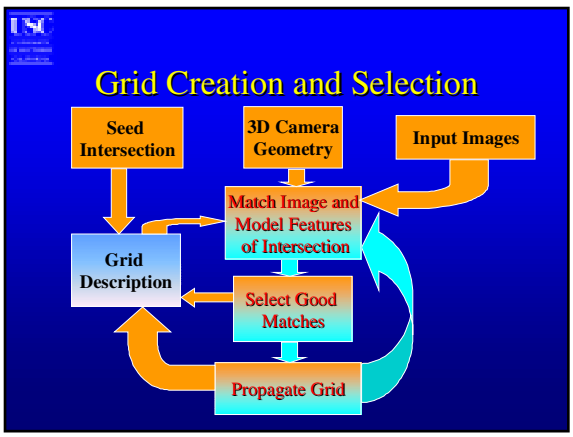

Figure 28. Road Grid Initial Verification

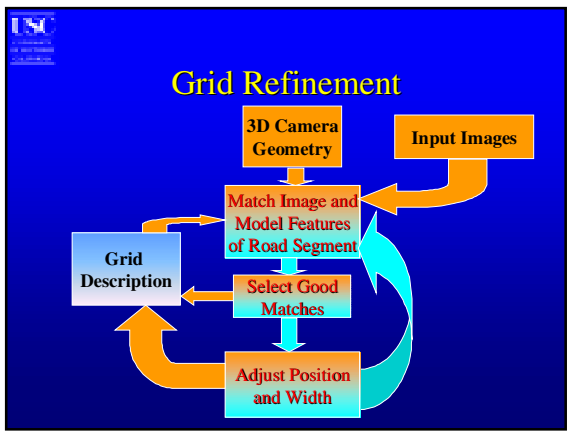

Figure 29. Road Grid Refinement

Use of digital elevation models, or DEM, helps in the refinement of the extracted road segments. While a DEM has many problems and may not be exact, it provides a good approximation of elevation to determine when a road segment is higher or lower than the others in its extended street. Figure 30 shows a small portion of one image with the matched road segments color-coded: consistent segments shown in grays, inconsistent segments in white. In this case, consistent means that the average elevation of the road segment as given by the DEM is similar to the average for the extended street, inconsistent means the elevation is much higher than the average for the extended street.

The DEM is used in two ways, first the road segment is shifted (perpendicular to its primary direction) to a minimum elevation location. Rather than allowing arbitrary shifts, the distance is limited according to the quality of the geometric match (with a perfect match the segment will not be shifted). The results of this refinement step are shown in Figure 31 where the consistency measure has been recomputed using the new locations of the road segments. Even with the shifts

(e.g. in the lower third of image), the worst segments are still bad. We then eliminate these inconsistent road segments from the set of good matches and recompute the extended streets.

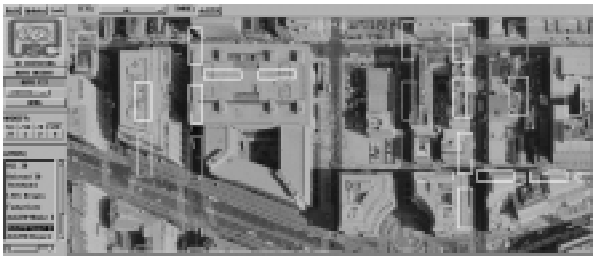

Figure 30. Initial DEM Consistency Measure

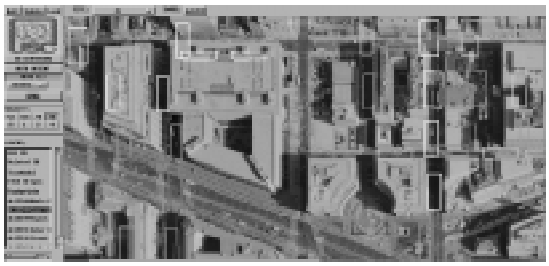

Figure 31. After using DEM

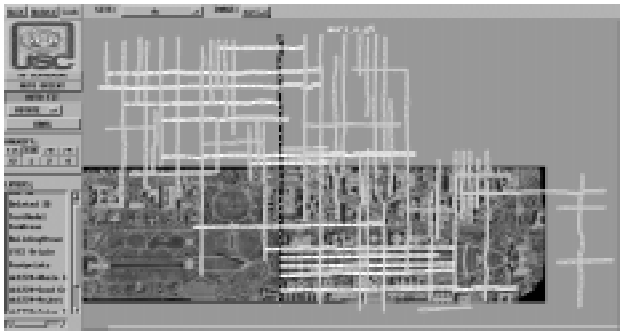

Figure 32. Extended streets combined on one oblique view

This system has been run on several sites one of which is over Washington, DC. The final results of the extracted street grid are shown in Figure 32 projected onto an orthophoto corresponding to the extracted DEM. This orthophoto only covers most of the area of 3 of the images used for the extraction so some extracted roads are displayed off the image. The colors are used to indicate one (arbitrarily) selected street in black, with its intersecting streets in white and all the others in gray. The time for the initial verification (approximately 2000 intersections) was roughly 90 minutes (covering five 2000X2000 images). The refinement using the same 5 images and testing about 3200 road segment triples (some are tested in multiple images) was about 500 minutes. After all the refinement steps, approximately 63km of streets are extracted.

A detailed analysis of these results shows one common error: road segments that are misplaced by the width of the road (i.e. the left side of the model matches the right side of the actual road and the right side of the model matches some other structure parallel to the road). These errors are caused by weak boundaries for the road itself and stronger edges in features parallel to the road. Exact measures of quality are not available, but false negatives are approximately 20% of the total with placement errors in about 30% of the individual road segments.

4 VISUALIZATION

All the activities in the MURI project culminate into a database to be used for a variety of applications. The primary objective here is to develop a 3D Visualization Environment that is suitable for rapidly creating and displaying 3D virtual worlds from the database in order to promote data understanding and rapid model editing. Some of the expected benefits include: (1) an improved model for multi sensor data visualization; (2) enhancement of identification and correction of errors in 3D models and terrain data; (3) model verification; (4) change detection; (5) battle damage assessment; (6) allowance of high fidelity extraction of 3D models in urban areas; (7) support for data understanding through multi-source data fusion; (8) projected textures improve automated extraction algorithms; (9) effective handling of occlusion and foreshortening problems; and (10) generation of ortho-rectified imagery or any camera view in real-time.

The system being developed is called Visualization Environment Supporting Photogrammetry and Exploitation Research (VESPER). The basic elements of photogrammetry are integrated with 3D visualization technology. These include precise camera calibration, position and orientation, overlapping images, and image to ground transformation. Methods presented enable the understanding of multiple overlapping images. Image to ground transformation is accomplished through careful application of projective textures. A Digital Projector with accurate camera information allows imagery to be projected onto terrain and feature surfaces, Figure 33. This method encourages the use of multiple overlapping images in VR. We present results that demonstrate the ability of this process to efficiently produce photospecific VR. Current photospecific visualization tools lack native support for precise camera models.

The fusion of multiple image sources in an interactive visualization environment demonstrates the benefits of bringing the rigors of photogrammetry to computer graphics. The methods presented show promise in allowing the development

of faster tools for model verification, change detection, damage assessment, and photospecific modeling of feature data. The use of digital projectors can simplify the data preparation process for virtual reality applications while adding greater realism.

An example of multi-source data fusion is shown in Figure 34. An ortho-rectified thematic map and frame image are projected and blended on the terrain and feature models. What is new here is that the source images are not being blended in image space or a common ortho projection. They are simply being re-projected through their individual sensor models onto the terrain and feature surfaces. This eliminates all the data preparation steps required for multi-source data fusion and results in a streamlined and rapid process.

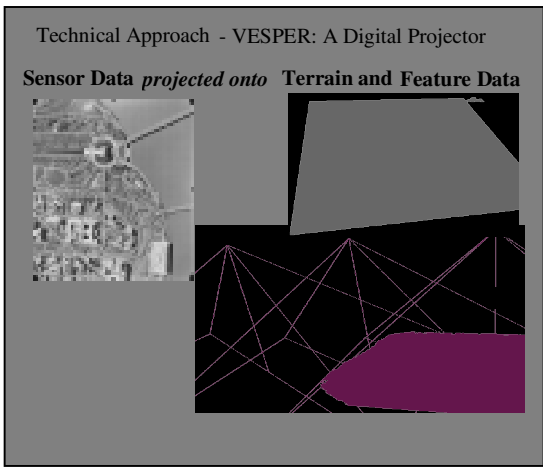

Figure 33

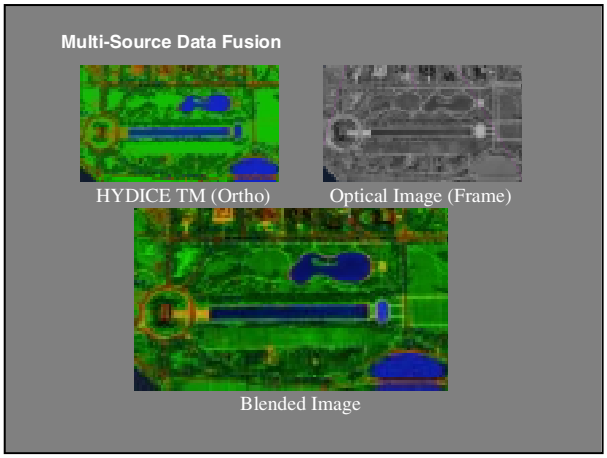

Figure 34

Another example is in Figure 35. Here the fusion of imagery taken with a hand held camera together with aerial imagery enables exciting new capabilities for immersive VR. Notice the depth shadows in the lower left image. This sensor occlusion information can be utilized in Line-of-sight analysis, sensor coverage, etc.

Finally, Figure 36 depicts the on-the-fly generation of an ortho mosaic. True orthographic mosaics can be generated on the fly in real time by simply taking an orthographic overhead view of a scene that uses digital projectors. This will readily support real-time video projected onto topographic map data and retain an orthographic viewpoint if desired (usually the case if you need to read text on the map).

Figure 35

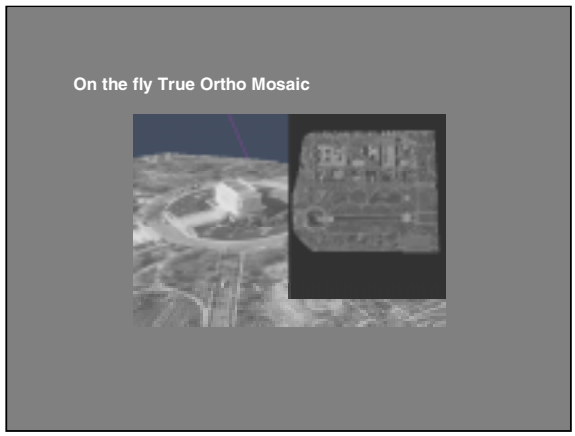

Figure 36

5 CONCLUSIONS

Significant progress has been made by an integrated multidisciplinary team of researchers on urban feature extraction, and construction and visualization of the resultant database. Fusion of multiple data sources based on rigorous sensor modeling has yielded a substantial reduction in effort and improvement of the results. Effort continues toward tighter integration of capabilities from photogrammetry, remote sensing, computer vision, and visualization research.

6 ACKNOWLEDGEMENTS

The research presented is sponsored by the U.S. Army Research Office under Grant No. DAAH04-96-1-0444. The views expressed in this paper do not necessarily reflect the views of the U.S. Government or the U.S. Army Research Office. The author is indebted to all the MURI project principal researchers; from Purdue: J. Bethel, D. Landgrebe, L. Biehl; from USC: R. Nevatia, K. Price, A. Huertas; from BAE Systems: M. Vriesenga, J. Spann, K. Kaufman, L. Oddo. Heartfelt thanks are also due the large number of other researchers and graduate students whose work makes this project so successful.

7 BIBLIOGRAPHY

Bethel, J., Lee, C., Landgrebe, D., "Geometric Registration and Classification of Hyperspectral Airborne Pushbroom Data", Proceedings of ISPRS 19th Congress, Amsterdam, The Netherlands, July 2000.

Madhok, V., Landgrebe, D., "Supplementing Hyperspectral Data with Digital Elevation", Proceedings of the International Geoscience and Remote Sensing Symposium, Hamburg, Germany, June 28 – July 2, 1999.

Nevatia, R., Huertas, A., Kim, Z., "The MURI Project for Rapid Feature Extraction in Urban Areas", ISPRS: Automatic Extraction of GIS Objects from Digital Imagery, Munich, 1999, pp. 3-14, Invited paper.

Price, K., "Road Grid Extraction and Verification", ISPRS: Automatic Extraction of GIS Objects from Digital Imagery, Munich, 1999, pp. 101-106.

Spann, J., "Photogrammetry Using 3D Graphics and Projective Textures", Proceedings of ISPRS 19th Congress, Amsterdam, The Netherlands, July 2000.

Theiss, H., Mikhail, E., Aly, I., Bethel, J., Lee, C., "Photogrammetric Invariance", Proceedings of ISPRS 19th Congress, Amsterdam, The Netherlands, July 2000.