LLNL-TR-609988

LAWRENCE LIVERMORE NATIONAL LABORATORY

Benchmark Imagery for Assessing Geospatial Semantic Content Extraction Algorithms -- Annual Report for FY2012

W. T. White, P. A. Pope, J. Goforth, L. R. Gaines

January 14, 2013

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Benchmark Imagery for Assessing Geospatial Semantic Content Extraction Algorithms
Annual Report for FY2012

W. Travis White III,[1] Paul A. Pope,[2] John Goforth,[1] Lucinda R. Gaines,[2]
30 September 2012

[1]Lawrence Livermore National Laboratory (LLNL)
[2]Los Alamos National Laboratory (LANL)

Summary of Activities in FY2012

The primary objective of this project is to create a set of benchmark imagery for validating algorithms designed to extract semantic content from geospatial images of industrial facilities. The goal is not to develop algorithms but to provide images that can be used to do so. The secondary objective is to demonstrate how this imagery could be used for algorithm validation. The project is a collaborative effort between Los Alamos National Laboratory (LANL) and the Lawrence Livermore National Laboratory (LLNL).

The two top-level tasks for this fiscal year were to compile a prototype suite of annotated imagery and to demonstrate the use of the images in a prototype methodology for verification and validation (V&V) of geospatial, semantic extraction algorithms [1]. We completed both of those tasks, met all scheduled milestones and delivered all required deliverable items for the year. The latter included the following:

- An initial suite of annotated, overhead photographs of 114 industrial facilities.
- A demonstration of the use of the Benchmark Imagery for algorithm V&V
- A proof-of-principle synthetic image of a fictitious nuclear power plant.
- Documentation [2] – [6].

In addition, we participated in two reviews and one technical conference.

- Independent Program Review by Dr. Steven Schubert *et al.* [7]
- VMRD2012 Joint Program Review presentation [8]
- ASPRS2012 poster [9]

As detailed in the report, the process of compiling and annotating the images produced a suite of imagery useful for testing geospatial, semantic extraction algorithms. Exercising a V&V protocol with the present benchmark imagery then provided insights into useful improvements to the benchmark imagery suite. Two key — but potentially expensive — recommendations were to annotate much more completely and to segment objects of interest. The proof-of-principle synthetic image took us through the major steps of synthetically generating a spatially detailed composite scene. It opened the door to testing a variety of geospatial, semantic extraction algorithms with synthetic images. We anticipate that with further work, we shall learn what improvements should be made both in the imagery suite and in emerging algorithms.

The body of this report has 4 more sections: Background, Compiling a Prototype Suite of Imagery, Demonstrating the use of the Imagery in a Prototype Methodology for Algorithm V&V, Creating a Proof-of-Principle Synthetic Image. The report ends with a Conclusions section.

1. Background

An important portion of nuclear non-proliferation information is derived from overhead imagery. As the quantity and quality of overhead imagery grows, the data stream is eventually going to exceed unaided human capacity to handle. Since image interpretation ultimately is word-based in nature, tools that automatically label and interpret images in words are needed (Figure 1 and Figure 2). In parallel, testing standards must be developed for these emerging tools, including suites of test images and methodologies for using the test images. In short, new tools for the automatic extraction of semantic content from geospatial imagery of industrial facilities are needed to better support NA-22's mission in nuclear non-proliferation [10]. Collecting an initial suite of test images and demonstrating their utility are the underlying *raison d'être* for — and goal of — this project.

Geospatial Semantic Content Extraction

"Inference (interpretation) of the implied significance of objects or activities from images that include information of their layout and location"

Geospatial (adj.): of or relating to the relative position of things on the Earth's surface.

Semantic (adj.): of, pertaining to or arising from the different meaning of words or other symbols.

Figure 1. *Key terminology used in this report.*

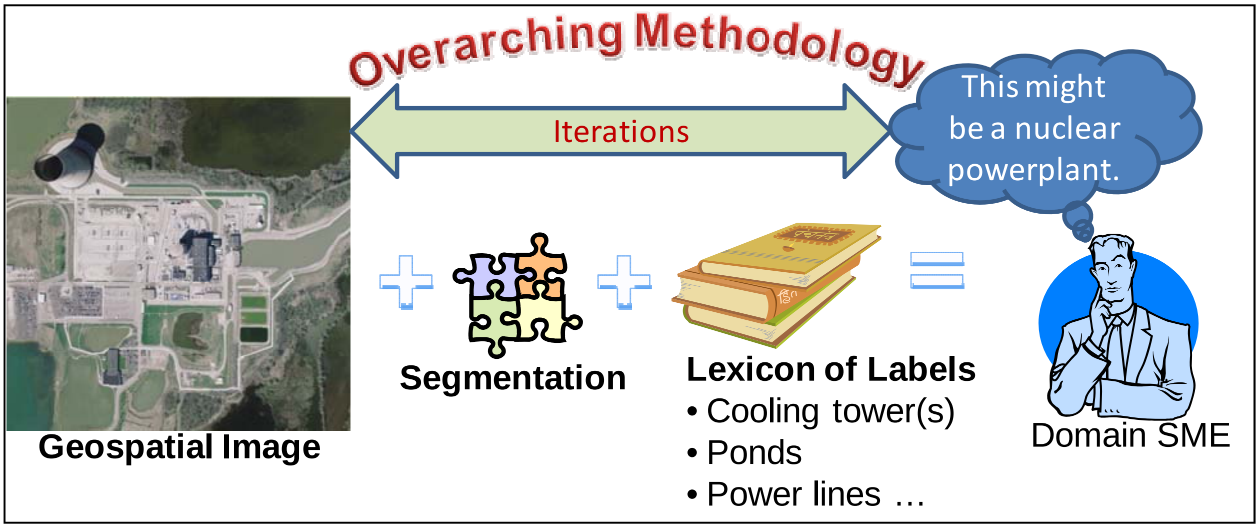

Figure 2. *Simple schematic illustration of an iterative approach to converting images to understanding to assist a domain subject matter expert (SME).*

The actual process of automatic image interpretation, illustrated by the simple cartoon in Figure 2, is non-trivial. The overarching methodology must contain lists of rules, objects, and relationships among those objects that can be mapped onto the input image(s). These may or may not be well specified or complete. For example, what are the rules that enable a human analyst to determine the boundaries of a facility from looking only at an image or set of images? The lexicon of labels that an image analyst would use (*pond, tower, pile, …*) might be different than the labels that an image-processing algorithm would use (*polygon, edge, point, …*). How should one design a segmentation algorithm to align with the lexicon of labels that a human would assign to objects in the scene? How does the vocabulary of labels limit possible interpretations of the image? What range of features must the geospatial image include for an algorithm to function? These are but a few of the issues that arise and that are active areas of computer vision research as applied to automated interpretation of overhead imagery.

From the preceding discussion, one can easily see that algorithms that support automated image interpretation may apply only over a narrow range of conditions. When applied to images that do not meet those conditions, the algorithms break. One desired property of a suite of benchmark imagery is that it should contain enough variety to test algorithms to their breaking points. This would help users understand the range of validity of the algorithms. Furthermore, it seems intuitively correct that the training set and the test set of images should be different when evaluating an algorithm.

Another goal of creating a benchmark imagery suite is to stimulate creation of new algorithms. It is the *Field of Dreams* approach: “Build it, and they will come.” [11] Since much of the creativity is concentrated within the academic community, another desired feature of the Benchmark Imagery is that it be affordable (free) and distributable (unclassified and unlicensed) to academics.

Another key issue is what should be the balance among actual photographs and synthetic images in the Benchmark Suite? Image analysts work with actual photographs, but algorithm developers can benefit from using real, synthetic and composite images. Each type of image has its own advantages. Real images automatically incorporate all the physics of the scene and all the engineering characteristics of sensors. Composite and synthetic images enable the content in a scene to be entirely configurable and the ground truth to be perfectly known. Creating facilities, objects, and environment in a synthetic image can be dictated by the end user and their needs. This is very important when testing algorithms on situations where imagery is limited or where control is needed over the scene. Facilities can be generated in different stages of construction. Images can be “captured” at different angles and different time of day. Effects such as fog, clouds, and lighting can be simulated. Sensor and mission characteristics such as over-exposure, distortion, blur, and a variety of viewing geometries can be modeled. Also, synthetic images may be more distributable than real images, because they tend to be less sensitive in nature (unclassified). Therefore, the Benchmark Imagery Suite should contain a mix of real, composite and synthetic images.

In the next section, we discuss the initial suite of imagery, which consists entirely of real images and their associated annotations.

2. Compiling a Prototype Suite of Imagery

The work in this task was accomplished in the following sequence of steps:

1. Identify a method for classifying types of industries and their observable characteristics (done in FY2011)
2. Compile a list of candidate facilities and locations
3. Identify types and sources of imagery to be considered
4. Collect imagery
5. Annotate imagery
6. Check the annotated imagery

The first step in compiling the prototype suite of benchmark imagery was to identify a system for classifying types of industry and their features. From our work in the previous year, we chose to use the work of Chisnell and Cole [12], which divides all industries into six categories and provides a vocabulary of names for industrial objects. To the list of Chisnell and Cole's industry types, we added "semiconductor plant," which did not exist at the time that Chisnell and Cole wrote their work.

Next, a list of candidate industrial facilities and their locations was prepared, compiled from open-source information, including two EPA databases, an OSHA database the Energy Justice database [13] and the *North American Industry Classification System* (*NAICS*) by the US Census Bureau. A separate gazetteer was formed for four of the six main industrial types identified by Chisnell and Cole: heavy manufacturing (HM), chemical processing (CP), heat processing (HP) and mechanical processing (MP). A unique alphanumeric code was assigned to each site. The region of interest around each site was chosen to be nominally one square kilometer; minimum and maximum latitude and longitude values were then calculated for each site, based on the center geographic coordinate.

Having selected the types and locations of the candidate industrial sites, we had to identify sources of imagery. To accommodate all overhead remote sensing imaging modalities that currently exist (radar, IR, visual, broad-band, multispectral, hyperspectral, and lidar, for instance), requires a large number of different algorithms and images. Rather than try to collect enough benchmark images to validate all possible algorithms, we chose to simplify the situation, focusing only on visual-band classical imaging this year. Such overhead imagery is the most readily accessible, affordable and distributable of all of the imaging modalities.

A ready source of visual-band overhead imagery at ground sampling distances (GSD) on the order of tens of cm is the USGS National Map Viewer, http://nationalmap.gov/viewer.html (formerly called the Seamless Data Warehouse http://seamless.usgs.gov/). These images, which are GeoTIFF, [14] carry terrestrial coordinates of the scene embedded in their headers. Although restricted to images over the United States, the images are in the public domain and can be obtained free of both cost and licensing restrictions. This permits wide-spread sharing of the benchmark imagery suite among academics and small companies engaged in algorithm development. Figure 3 through Figure 5 show examples of several different industry sites obtained from the National Map server.

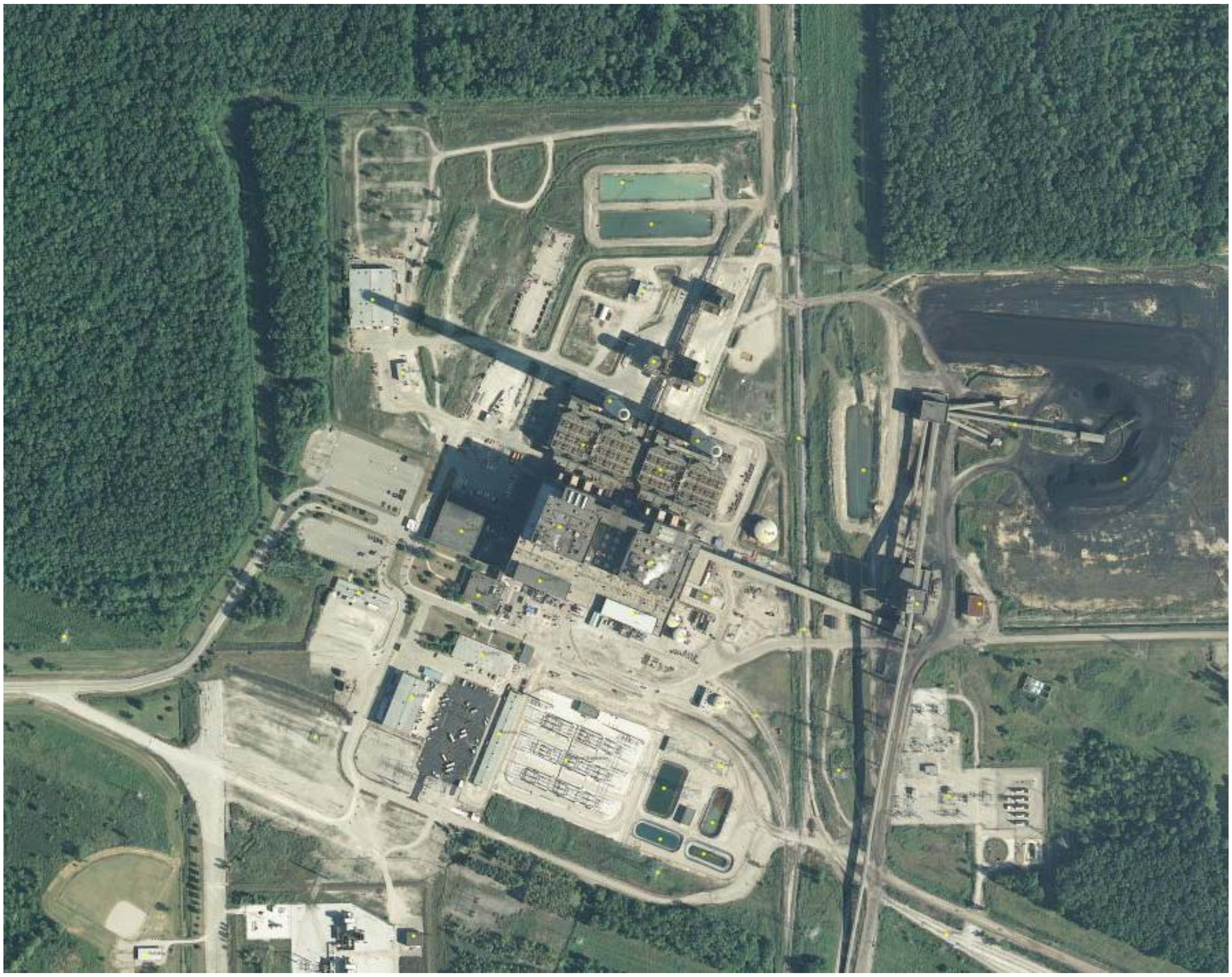

Figure 3. *Illustration of a Heat Processing Plant, HP0058_C8, Belle River power plant, a coal-fired plant located in China Township, Michigan, Latitude 42.774676, Longitude -82.494964, GSD: 0.3m.*

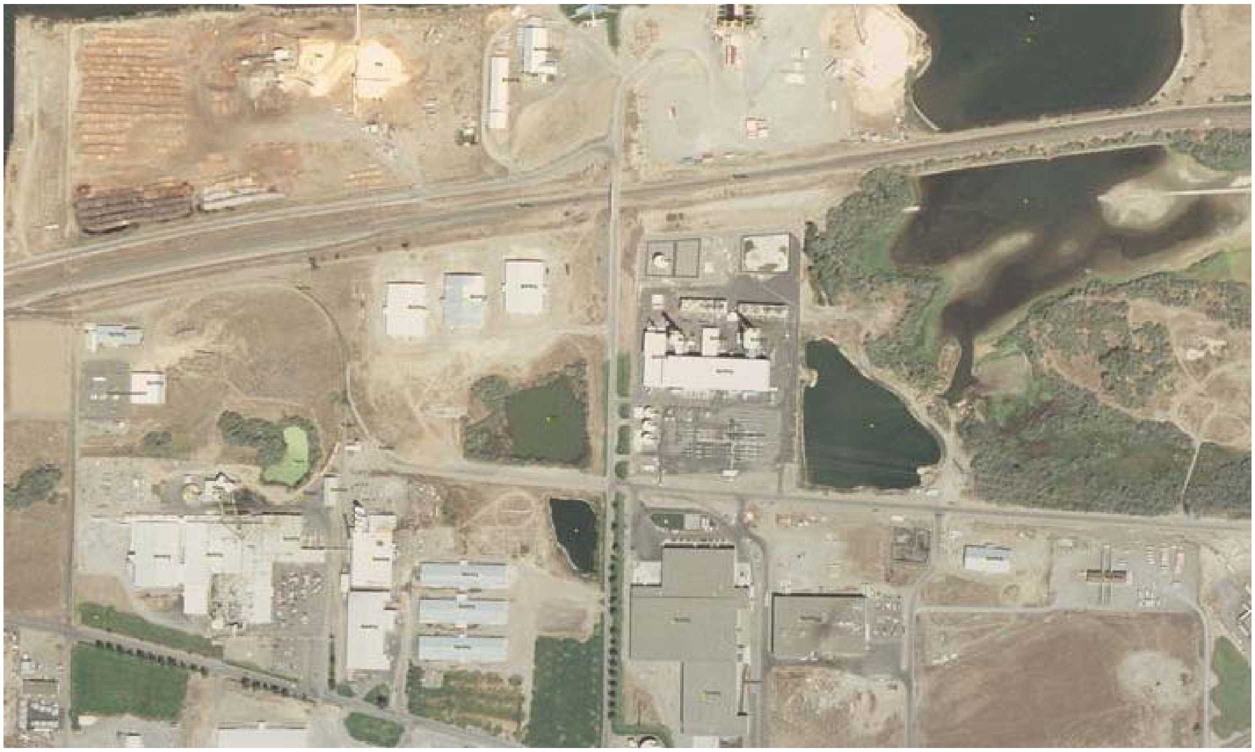

Figure 4. *Illustration of a chemical processing plant, CP28_VDE, Pacific Ethanol Columbia, LLC, located in Boardman, Oregon. Latitude 45.84797, Longitude -119.673825, GSD 0.5m.*

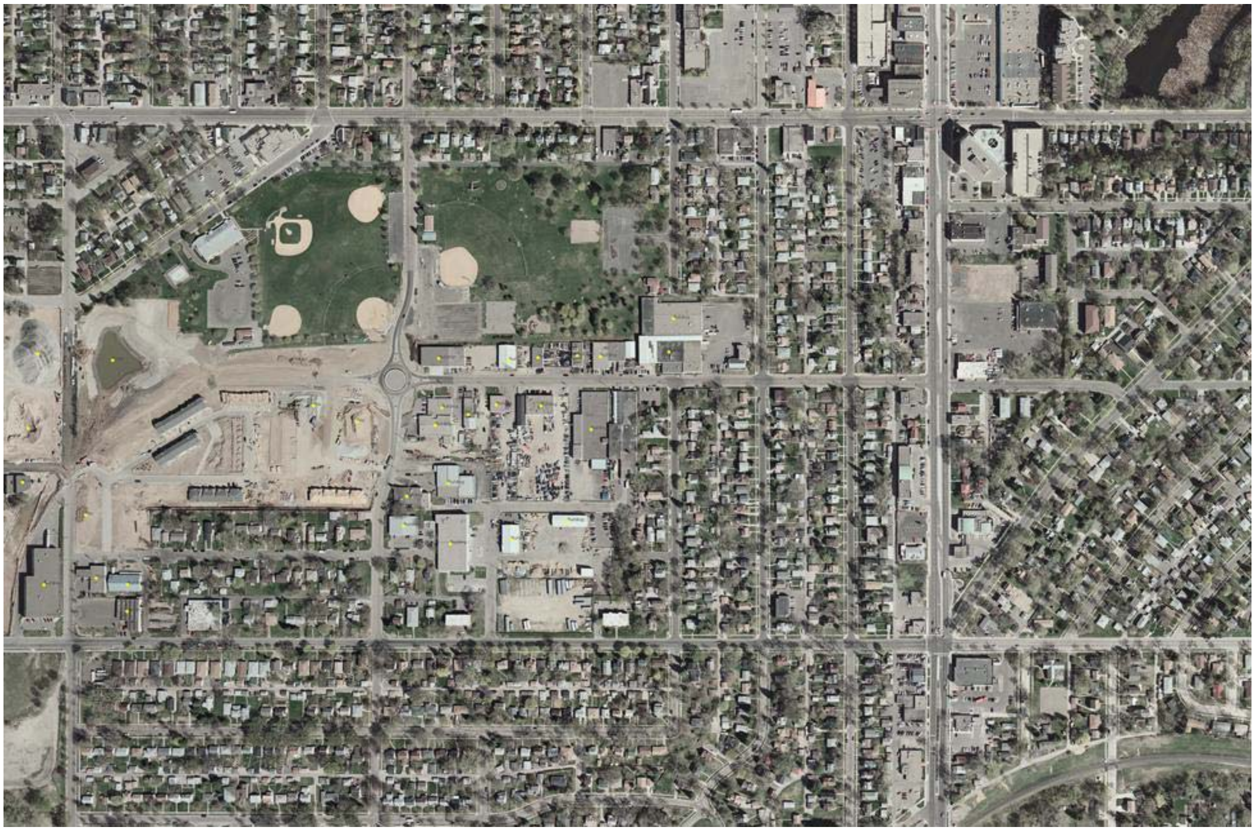

Figure 5. *Illustration of a heavy manufacturing plant, HM45_V3F, Invest Cast, Inc, an iron and steel foundry, located in Northeast Minneapolis, Minnesota, Latitude 45.037613, Longitude -93.251851, GSD: 0.3 meters.*

Next we wrote a software tool to download image chips from the USGS server using the geographic information of each site as provided in the gazetteers. We downloaded over three hundred images and organized them by industry type in separate folders. Occasionally an image would have to be discarded. Sometimes the image was blank or grossly misshapen; other times it did not contain the intended industrial plant. We manually inspected each downloaded image and rejected obviously incorrect pictures. Also, one square km was sometimes not large enough to contain all of a desired facility. When our images were undersized, we removed the images from the benchmark suite and added the site to a working list of images to be re-downloaded.

Then we annotated the images using a controlled vocabulary derived primarily from Ref. [12]. The labels appearing in Chisnell and Cole's work often included detailed functional descriptions such as "gravel storage" instead of the simpler "pile" or "concentration building" instead of the more generic "building." For purposes of this work, we deliberately made the initial controlled vocabulary somewhat generic, as detailed ground truth was not available to confirm the functionality of the buildings, towers, tanks, and ponds in the images (Table 1).

Term	***Definition***
Building	"A structure consisting at a minimum of a foundation a roof and supports."
Conveyor	"A structure used to move material (e.g. a conveyor belt)."
Cooling Unit	"A piece of equipment used to dissipate or remove heat to the surrounding environment from other equipment or a building through the use of an intermediate fluid (e.g., water refrigerant); includes cooling towers and chillers.
Crane	"Equipment used to lift and possibly move heavy objects."
Electrical Substation	"A collection of structures used for the transmission transformation distribution or switching of electric power."
Kiln	"Equipment (e.g. furnace or oven) for drying baking or burning material."
Pile	"A heap of material."
Pond	"A body of liquid smaller in size than a lake sometimes artificially formed (e.g. by damming a stream)."
Railline	"A linear structure consisting of a foundation (e.g. compacted material) upon which rests horizontally oriented bars of metal to facilitate the movement of locomotives, railcars, etc."
Stack	"A vertical structure used to convey exhuast; usually relatively tall in height compared to surrounding structures."
Tank	"A structure of a wide variety in shape (e.g. cylindrical spheriodal, etc.) for holding liquid or gas."
Tower	"A vertical piece of equipment (e.g. watch tower fractionating tower, cracking unit, etc.) — usually relatively tall in height compared to surrounding structures."

Table 1. *Controlled vocabulary used in the Benchmark Imagery annotation as of 8/31/2012.*

We divided the images to be annotated among six annotators, with two people being responsible for each of the three industry types. To speed up the annotation process, we prepared a point-and-click software tool, Geospatial Image ANalysis Tool (GIANT). The tool, which was mouse driven, allowed users to point to an object in the image and apply a label from a drop-down menu that contained the controlled vocabulary. Annotators were instructed to label 30 to 50 objects in each image. Objects were labeled by selecting a single pixel within the boundary of the object. Thus, most pixels in the image had no label attached to them. (The pixel classification was not exhaustive across an image.) After all of the images were annotated, there were two sets of annotations for each image. The annotations were then merged pair-wise and the final annotated images were checked for correctness by a trained image analyst. No attempt was made to segment the annotated objects in the images by drawing borders around them. In the process of annotating the images, annotators identified additional new user features and recommended a list of potential improvements for GIANT for future implementation.

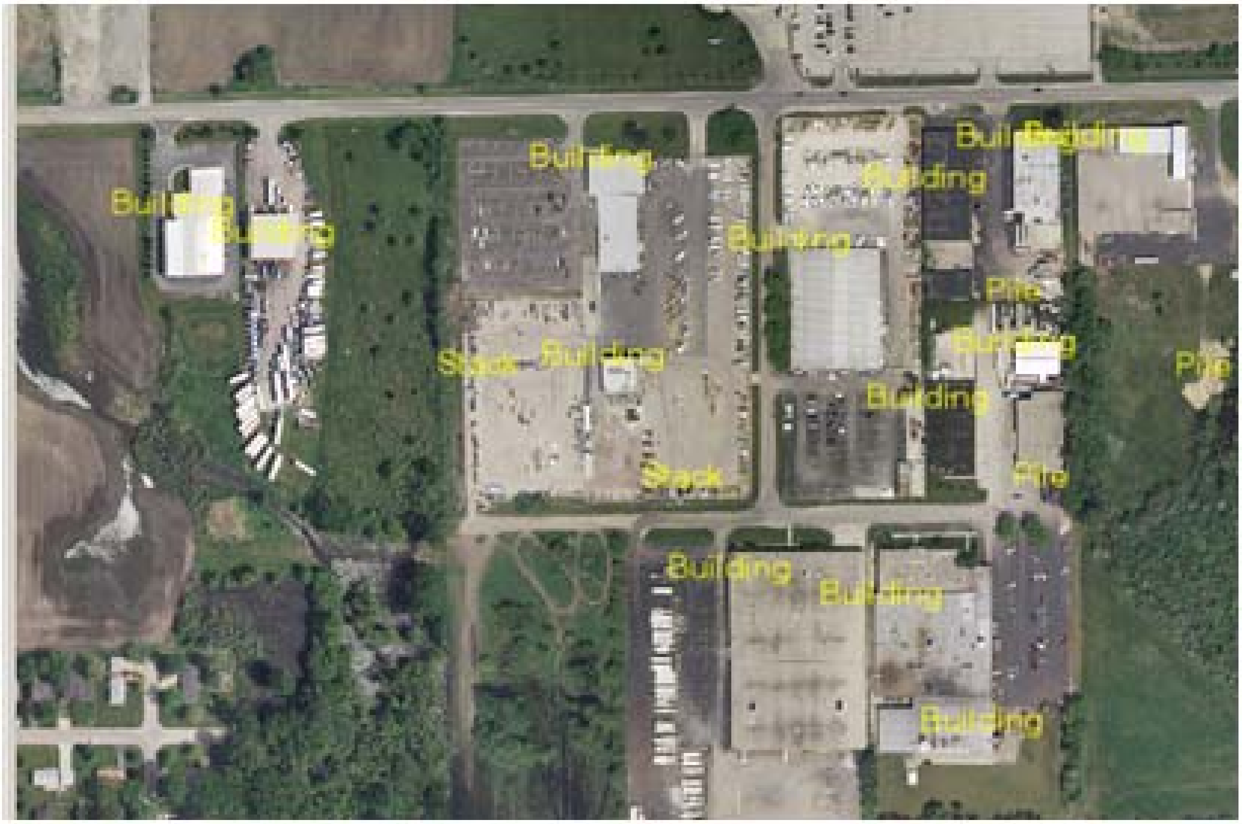

Figure 6. *Portion of an annotated image. Heavy Manufacturing Plant, HM5, Aurora Metals Division, L.L.C, Montgomery, IL, Lat.: 41.72736, Lon.: -88.365065.*

The prototype benchmark imagery suite was then complete. It contained 114 GeoTIFF images of three different types of industrial facilities, as categorized by reference [12] plus some semiconductor plants:

- Heat Processing (38 images) — *e.g.*, electrical power plants.
- Heavy Manufacturing (34 images) — *e.g.*, steel foundries, metal casting plants.
- Chemical Processing (30 images) — *e.g.*, oil refineries.
- Semiconductor Plants (12 images) — *e.g.*, computer chip companies.

3. Demonstrating the use of the Imagery in a Prototype Methodology for Algorithm V&V

In this task, the work proceeded in four major steps:

1. Define V&V methodology
2. Obtain algorithm(s) to test

3. Test algorithm(s) using the Benchmark Imagery Suite
4. Interpret results from the V&V tests

3.1 Define V&V Methodology.

In previous years, NA-22 sponsored a GeoSpatial Validation (GSV) working group that recommended a protocol for validating geospatial algorithms. [15] We have chosen to use the V&V methodology from that group (Figure 7).

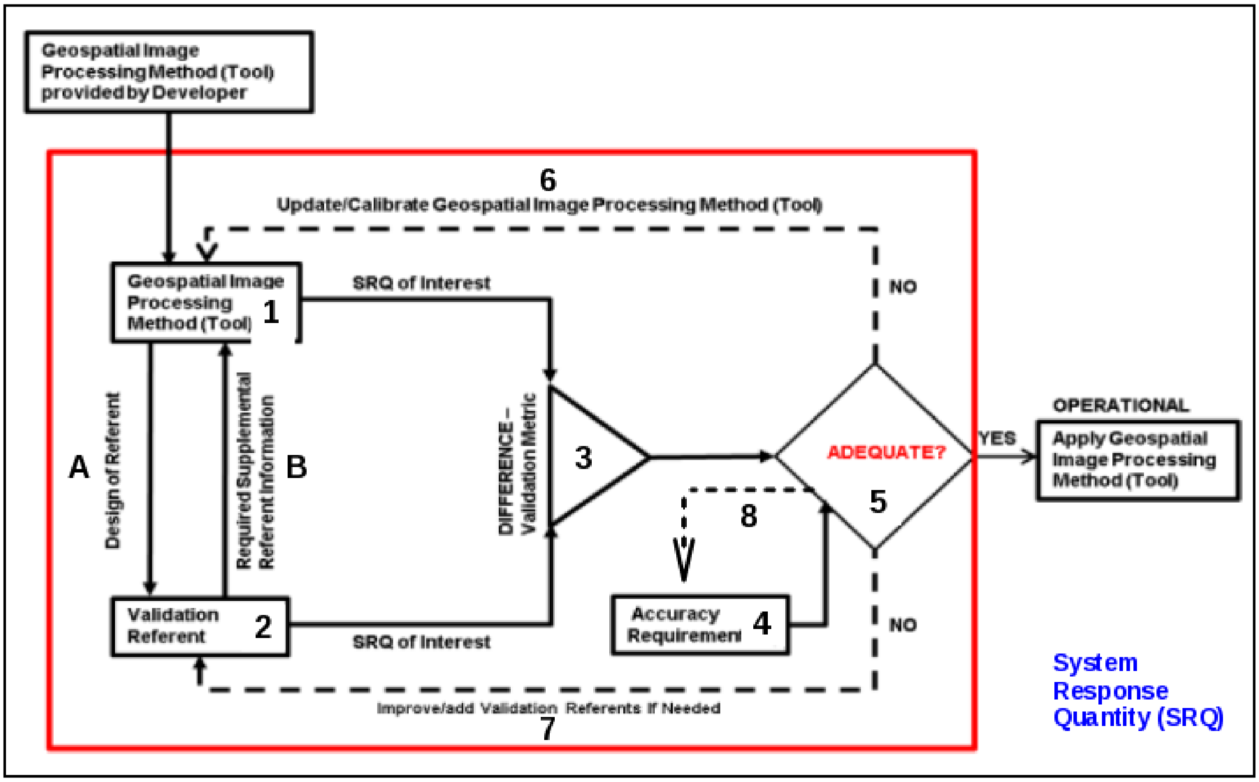
Geospatial Image Processing Method (Tool) provided by Developer
6
Update/Calibrate Geospatial Image Processing Method (Tool)
Geospatial Image Processing Method (Tool)
1
SRQ of Interest
NO
A
Design of Referent
Required Supplemental Referent Information
B
DIFFERENCE – Validation Metric
3
ADEQUATE?
5
YES
OPERATIONAL
Apply Geospatial Image Processing Method (Tool)
8
Validation Referent
2
SRQ of Interest
Accuracy Requirement
4
NO
Improve/add Validation Referents If Needed
7
System Response Quantity (SRQ)

Figure 7. *Algorithm verification and validation cycle proposed by the NA-22 GSV Working Group.*

Shown in the figure, an algorithm (Box 1) is supplied by a developer and placed into validation testing (red box). It is then given input from a source of ground truth, or “Validation Referent” (Box 2). The algorithm calculates a result that can be measured, a “System Response Quantity (SRQ).” A geospatial SRQ could be, for example, a list of all cooling towers in a set of images and the corresponding geocoordinates and areal footprints. The SRQ computed by the algorithm is then compared to the pre-tabulated, true value (Box 3). In this report, we refer to the differences between the algorithmic outputs and the ground truth values as “validation metrics.” In the example just mentioned, validation metrics could include the difference in the number of cooling towers identified vis-à-vis the actual number or the position and size errors of the detected towers. The validation metrics from Box 3 are then compared with a set of Accuracy Requirements (Box 4). If the validation metrics for the algorithm are within acceptable limits, then the algorithm is considered to be validated for the SRQs being tested (Box 5). Otherwise, the algorithm is rejected.

As drawn, the V&V process flow applies to an R&D environment, where a failure of validation might not indicate a problem only with the algorithm. Rather, an unsuccessful validation test could be a result of errors either in the ground truth, the accuracy requirements, or the algorithm, requiring improvements in any or all of them. The three dashed lines numbered 6, 7 and 8 in Figure 7 accommodate feedback to support the R&D validation cycle. This situation applies to the Benchmark Imagery Project, which is a research and development effort.

(Note that in situations in which a set of validation benchmarks and acceptance requirements have been adopted as standards, a failure to validate is attributed entirely to having a bad algorithm, and the algorithm is flatly rejected.)

3.2 Obtain algorithm(s) to test.

At the beginning of this project, we planned to use existing commercial software such as ERDAS Imagine, Arc GIS or eCognition for demonstrating the use of the benchmark imagery for V&V [1]. Upon investigation, we concluded that such software was not of interest, due to the high purchase price, steep learning curve and lack of automation. We sought to find more-automated software if possible. As the project developed, it became apparent that automated or semiautomated algorithms for identifying and classifying industrial facilities from overhead images were not available. Therefore, to demonstrate the process of using our current suite of images for V&V, we created our own algorithm, a simplistic building finder aimed at detecting large rectangular buildings with flat, bright white roofs. Since the purpose of the demonstration was not for us to develop an algorithm but to illustrate the use of our benchmark images in the V&V process, it was not required that the algorithm be robust.

The algorithm followed a rules-based design such as is commonly described in elementary remote-sensing textbooks [16]. It used a combination of spectral clustering, morphological operations (erode and dilate), and spatial pattern matching. The spectral clustering employed a look-up table derived at the outset of the exercise. That look-up table derived from a single, subjectively chosen, representative image and was calculated only once during the exercise. When running the algorithm, it produced groups (clusters) of all pixels in an image that were “roof-colored” [17]. The morphological operations eliminated isolated pixels and smoothed the remaining pixel clusters. The spatial pattern matching compared the shape of the final pixel clusters to rectangles, using three parameters to do so: area, rectangularity and aspect ratio [6]. Our building-detection rule was that each of those three spatial pattern parameters had to be within one standard deviation of its expected value. Both the standard deviation and the expected value were determined during initial algorithm training, prior to the beginning of testing (Sec. 3.3). A wrapper written around the algorithm permitted batch operation of the code for efficiently testing large numbers of images.

The semantic aspect of the algorithm was that it attempted to match a word from the constrained annotation vocabulary (Table 1, Sec. 2) with objects in images. “Building” was chosen as the particular word to match, because buildings were a key industrial feature of all images in our data. The geospatial aspect of the algorithm was that it used the spatial resolution (i.e., the ground sample distance (GSD)) of each image to calculate the area of the pixel clusters.

3.3 Test algorithm(s) using the Benchmark Imagery Suite.

The V&V demonstration consisted of two main parts: training and testing. We used the Benchmark Imagery Suite to provide all images for both training and testing. For training, a subset of 16 Heavy Manufacturing imagery and annotations were used. For testing, 76 other images were used: Heavy Manufacturing (16 images), Heat Processing (38 images) and

Chemical Processing (22 images). Training data were used to collect statistics on — and to set detection thresholds for — the spatial pattern-matching parameters. The representative image for color classification was chosen empirically by visual inspection of all sixteen images contained in the Heavy Manufacturing training set. Testing was conducted to assess the accuracy of the algorithm when operated with the values of the detection thresholds.

Four tests were run. For the first test, the sixteen Heavy Manufacturing images used for training were used for testing. For the second test, the set of sixteen test images for the Heavy Manufacturing category were used. The third test used the twenty-two test images from the Chemical Processing category. The fourth test used the thirty-eight test images from the Heat Processing category.

3.4 Interpret results from the V&V tests.

Results are summarized in Table 2. In this section, we use those results to draw inferences about the benchmark imagery and about the algorithm.

Test	N_imgs	N_bldg (truth)	N_bldg (alg.)	N_true_positive	tp%	Minutes
1 HM_train	16	350	2578	63	18.0%	2.2
2 HM_test	16	252	2388	56	22.2%	2.0
3 CP_test	22	551	4715	133	24.1%	17.1
4 HP_test	38	882	6309	221	25.1%	7.7

Table 2. *Summary of Results of V&V Tests 1 through 4. N_imgs denotes the number of images in a test. N_bldg(truth) denotes the number of buildings that had been labeled by the annotators (ground truth) in each test set. N_bldg(alg.) denotes the number of buildings detected by the algorithm in each test set. N_true_positive is the number of algorithmically detected buildings that had also been labeled by the annotators. tp% is the ratio of N_true_positive to N-bldg(truth) expressed as a percentage. Minutes denotes the time required to process each test set and store the output via LAN.*

The first observation from Table 2 is that the true positive (building-detection) rate was about 20% in all four tests. The low true positive rate indicates that the algorithm did not find most of the buildings that had been labeled by the image annotators, even though the algorithm itself classified every pixel in the image. Because the buildings were, by and large, very easy to discern to the human eye, the low true-positive rate implies that the quality of the algorithm is low. Such a result is not surprising, since accurate segmentation of objects of interest from imagery is an on-going, difficult problem in computer-vision [18], [19]. Because they enabled us to measure the performance of the building-finder algorithm, *the images of the Benchmark suite, with their current point annotations, proved useful for the purpose of making a decision about the validity of a simple algorithm*.

A second observation from Table 2 is that the true-positive detection rate was lowest for Test 1, which is the opposite of what one might expect. For that particular test, one would expect the performance of the algorithm to be the best, since the training set and the test data were identical. A more detailed examination of the individual segmented images within each test set showed that the accuracy of the algorithm did not segment buildings consistently [6]. Detection percentages per image ranged from a low of 0% to a high of 75%, with a mean value around 25%. One interpretation of this variability in the output of the algorithm is that *the current set of images appeared to fully exercise the algorithm*.

A third observation from the table is that the number of buildings detected by the algorithm in each of the four tests was about an order of magnitude greater than the number of buildings that actually were labeled by the annotators. Upon human inspection of the images, one could see that quite often the "building" corresponded to an unlabeled region of the image. Hits or misses in unlabeled regions could not be tested for accuracy by the computer, whose only knowledge of ground truth came from the 30 to 50 annotated pixels per image. Had the annotations been more exhaustive, the detection statistics (true positives, false positives, true negatives and false negatives) would have been more complete. The conclusion is, *annotate the benchmark images exhaustively to improve the V&V detection statistics of the algorithm.* By "exhaustively," we mean that there are no unlabeled regions in an image. In other words, all objects in an image must belong to some class, one of which could be "none of the above."

A fourth recommendation comes from knowledge of the design of the algorithm, not from the measured results in Table 2. It is based on the fact that the particular algorithm under test used spatial measures that were properties of a polygon (area and perimeter, for example). Polygons are extended but finite objects with closed boundaries. They are not point or line objects. Precisely for this reason, evaluating the effectiveness of such measures requires information about the extent (boundary) of the corresponding regions in a test image. The annotation process used this year did not include defining the boundaries of the labeled objects. The recommendation is, *annotation of the benchmark imagery should include demarcation of the boundaries of the labeled, extended objects.*

Exhaustive annotation and segmentation are, of course, exceedingly expensive and may be humanly impossible when performed manually on hundreds of images containing several millions of pixels each. To what degree should the annotation and segmentation be exhaustive and what methods should be used to obtain exhaustive annotation are unresolved issues at this time. This problem is not unique to this project; it is a current research issue for the computer vision community [20]. For example, should the Benchmark Imagery suite have fewer images but segment and annotate them all in great detail; should it have more images but with a small number of selected classes of object manually identified in as complete a manner as possible; should we increase the number and variety of images in the set but exhaustively annotate only a few?

A fifth recommendation about the Benchmark Imagery annotation is, *annotated objects should be segregated into those that make up the "figure" (i.e., are part of the industrial facility) and those that make up the "ground" (i.e., are part of the surroundings)*. For example, an industrial park in the center of an urban area will be surrounded by buildings that are not part of a facility: houses. The annotation scheme must somehow distinguish between industrial buildings and non-industrial buildings. A motivation for this recommendation is that the process of training a learning algorithm (*e.g.*, a neural network [21]) requires well defined examples of objects of interest and objects not of interest. For the Benchmark Imagery suite a goal is to support development of algorithms that can identify and classify industrial facilities. At this time, the problem of how to identify the boundaries of a specific facility from imagery alone is an open issue. The solution may lie, in part, in incorporating geospatial information into algorithms.

In the V&V demonstration performed with the Benchmark Imagery suite in FY2012, we created and tested a simplistic building-finder algorithm. Our validation metrics were also simple, and our statistics incomplete, due to the partial (sparse) point-like annotation. Furthermore, we had no user-supplied accuracy requirements on the algorithm. In the future, other characteristics such as annotation confidence, image quality, amount of clutter, etc. might also prove useful for conducting V&V testing [22].

So far, we have emphasized how the V&V demo led to recommendations about the Benchmark Imagery suite. Before we close this section, though, we comment briefly on the algorithm that we used. We attribute much of the inaccuracy of the building finder algorithm first to the limited spectral resolution (3-bands) of the training data and second to the fact that the algorithm only made use of the visual portion of the electromagnetic spectrum. Buildings are not spectrally unique in the visual wavelength region of the electromagnetic spectrum. This leads to inaccurate segmentation. Next, the threshold values on the spatial pattern parameters were only plus-or-minus one standard deviation away from the mean. Often in defining a detection algorithm one varies the threshold to determine the value that optimizes the ratio of the probability of detection to the probability of a false alarm. That was not done in this study, as our objective was not to optimize the performance of the algorithm but to use it as a vehicle for stepping through the process of V&V.

Remarks and suggestions about the Benchmark Imagery Suite:

- The images of the Benchmark suite, with their current point annotations, proved useful for purposes of making a decision about the validity of a simple algorithm.
- The current set of images appeared to fully exercise the algorithm.
- Annotate the benchmark images exhaustively to improve detection statistics of the algorithm.
- Annotation of the benchmark imagery should include demarcation of the boundaries of the labeled, extended objects.
- Objects should be segregated into those that make up the “figure” (i.e., are part of the industrial facility) and those that make up the “ground” (i.e., are part of the surroundings)

Comments on the test algorithm:

- Much of the inaccuracy of the building finder algorithm is most likely attributable to the limited spectral resolution of these data, and the fact that the algorithm only used the visual bands.
- Our objective was not to optimize the performance of demonstration algorithm but to use it as a vehicle for stepping through the process of V&V.

4. Creating a Proof-of-Principle Synthetic Image

When complete, the Benchmark Imagery suite will contain a combination of actual overhead photographs, composite imagery and synthetic imagery. Although most of the work in FY2012 centered on actual photographic images, in the final third of the year we began work in composite and synthetic images. We created a proof-of-principle synthetic GeoTIFF [14] image that combined a synthetic image of a fictitious nuclear power generation facility with actual overhead imagery. In this section, we describe the composite image and the process for creating it.

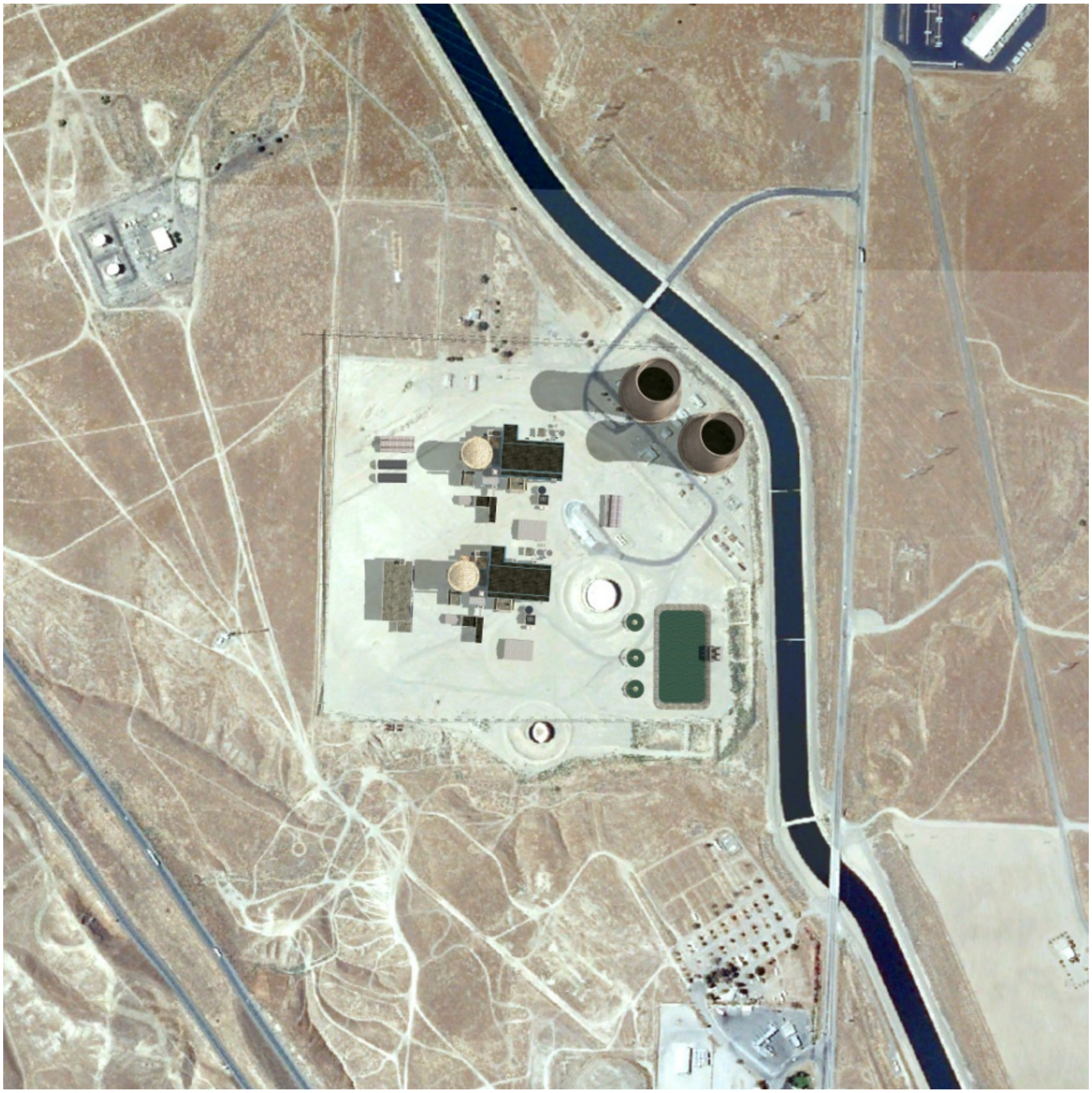

Figure 8. *Proof-of-principle synthetic image of a fictitious nuclear power plant overlaid on an actual scene of the California aqueduct near Kettlemen, CA.*

Figure 8 displays the composite image. The layout of the facility at the center is intended to be illustrative, not a rigorously vetted model. It is based on images of actual nuclear power plants collected in the Benchmark Imagery suite and on discussions with subject-matter experts. It incorporates real terrain elevation data, synthetic lighting, shadows, and post-processing color correction. There is no cloud cover or fog. The camera view is directly overhead. There are 2048 by 2048 pixels with a ground sampling distance (GSD) of 0.79 meters/pixel. The “canvas” or base terrain is made up of real overhead imagery that has been “draped over” elevation data. The elevation data and satellite imagery are for Kettleman City, CA, 35.996N/119.96W. The site was selected in part because there was a large empty area that would provide a suitable “canvas” for placing components. There is also a water source in the area which is required by Nuclear Power Plants (although sometimes out of view and utilized through underground pipes). The lighting/shadows are based on a date and time of August 8th, 2011, 0800 PST. The image has been generated in the visual (EO) band in order to be consistent with the images in the initial Benchmark Imagery set.

The components of the facility Table 3) are synthetic textured 3-D models that are commercially available from TurboSquid 3-D [23]. Each model has textures that define material properties that control how lighting interacts with the textures. To increase realism, these material properties were tuned visually so that the color and textures of the synthetic 3-D components matched the base imagery. Some components that are currently missing will be added to the proof-of-principle image in future iterations. These include parking lots, vehicles, containers, logistical equipment, electrical substation, and other objects.

Component	***Quantity***
Reactor with connected buildings	2
Cooling Tower	2
Large Pool	1
Small Pool	3
Fence	3
Office Building	1
Building A	2
Building B	2

Table 3. *3-D model components used in the proof-of-principal synthetic image of a nuclear power plant of* Figure 8.

To create the synthetic image we used SceneWorks, a real-time 3-D visualization framework developed by LLNL that is used by the Army, Navy, Air Force, NGA, and DHS. SceneWorks leverages the video-game industry and related advances in off-the-shelf video cards. Images take approximately 1/30th of a second to generate, permitting real-time 3-D interaction between the user and the scene. Real-time interaction is a tremendous benefit when setting up the scene, placing components, and tweaking properties such as lighting and materials.

Using the SceneWorks software we imported the terrain based data (overhead imagery and elevation data). Then we placed the 3-D components into the scene and configured the lighting, shadows, and other environmental parameters. Next we positioned the virtual camera (altitude and view angle) and selected the image size (in pixels). Then the pre-processed, synthetic image was captured in SceneWorks. Using Gimp, an open-source Photoshop-like tool, we manually post-processed the image to enhance the brightness, contrast, saturation and lightness. Post-processing improved the realism of the image by visually matching the colors and exposure of the composite image to other similar images in the benchmark imagery set. We stored the output as a TIFF image. After post-processing, we converted the output image from TIFF to GeoTIFF. To do so, we acquired the UTM coordinates of the four corners of the original overhead image using SceneWorks. Then we computed the GSD of each pixel and, using open-source “listgeo” and “geotifcp” tools [14], added the geospatial data (GSD, UTM extents) to the TIFF file as GeoTIFF metadata.

We considered alternate approaches, including using physics models such as RIT’s DIRSIG, a ray tracing application. [24] However, because we are still in an exploratory phase of generating synthetic images for use in the development and validation of semantic extraction algorithms, it is not clear whether having detailed physics-based models is more important than being able to create a wide variety of scenes quickly and easily. Physics-based models are orders of magnitude slower than SceneWorks. For example, with our current SceneWorks hardware (an off-the-shelf laptop), a single frame can take over 20 hours of CPU time to generate with DIRSIG compared to less than 33 milliseconds for SceneWorks. As an initial approach, therefore, we have chosen to use SceneWorks and purchase off-the shelf 3D models of scene components as needed.

5. Conclusions

We have annotated a collection of overhead GeoTIFF images of 114 industrial facilities so that they may be used to evaluate the effectiveness of algorithms for extracting semantic content from overhead images. Each image is a 3- or 4-band VIS/NIR overhead photograph that covers an area approximately 1 km^2 to 1 mi^2 with ground sampling distances ranging from 0.3m to about 1m. Annotations are point annotations (isolated, individual pixels). The number of images and simplicity of annotation, while limited, represent a useful first step, as demonstrated by this effort.

To demonstrate applying the images to algorithm validation, we created a simplistic building-finder algorithm and applied the methodology recommended by the NA-22 geospatial validation (GSV) team to it. (We searched for commercially available software to use in the validation exercise but concluded that such software was far from automated, expensive to buy and expensive to learn.) A noteworthy result of our tests was, the images of the current Benchmark Image suite (even with only point annotations) proved useful for purposes of making a decision about the validity of an algorithm. Furthermore, the variety of the images in the current benchmark suite appeared to fully exercise the algorithm that was tested, with the true-positive detection rate ranging from 0% for some images to 75% for others. Not surprisingly, the validation demonstration pointed out several areas of potential improvement for the benchmark images.

- Annotate the benchmark images exhaustively to improve detection statistics of the algorithm.
- Include demarcation of the boundaries of the labeled, extended objects.
- Segregate (classify) ground-truth objects in each scene into those that make up the “figure” (i.e., are part of the industrial facility) and those that make up the “ground” (i.e., are part of the surroundings)

Increasing the amount of annotation and adding segmentation to the ground truth are expensive, labor-intensive tasks. Segregating ground-truth objects into those that definitely are part of a facility and those that are not requires a source of information about the facility that, at least sometimes, is outside of the existing imagery and metadata. So, a question for next year is, to what extent should the above recommendations be implemented in the Benchmark Imagery suite? That is a question that we plan to address early in the new fiscal year at our annual planning meeting.

In addition to compiling a prototype suite of annotated images and demonstrating their use in an algorithm-validation protocol, we created a synthetic RGB image of a fictitious nuclear power plant. We embedded it in an actual overhead background scene, producing a composite GeoTIFF picture. In the process, we showed that one can create a realistic looking scene with great spatial detail by leveraging the computer gaming industry, purchasing commercially available 3D models. Implemented on SceneWorks imaging framework, these pre-built 3-D scene components resulted in fast (30 frames/second) image generation. Synthetic and composite images have some key advantages over actual images in that they do not require access to a sensor and they can be tailored systematically to test specific features of algorithms. They can range in complexity from photorealistic pictures to simple site maps. The former may be useful in testing segmentation, classification and detection algorithms; the latter, in providing input to geospatially enhanced semantic graph processing. An open challenge for next year is to explore the usefulness of synthetic images in testing geospatial, semantic extraction algorithms. Our current approach emphasizes realistic geometry and visually appealing appearances. However, an open issue is whether the way we are generating the scenes is as useful for testing algorithms as physically more rigorous but much slower physics-based scene simulation schemes.

We anticipate that as we and others use the images to test algorithms, we shall learn what improvements should be made both in the imagery suite and in emerging algorithms. For example, how detailed must the underlying ontologies be [25] for the benchmarks to be useful? How many images should include drawing the boundaries of the objects (i.e., segmentation)? Should the constrained vocabulary be larger? Should we include other modalities in the image suite? These and other issues can now be addressed.

Acknowledgments

The authors gratefully acknowledge the outstanding support provided by Dr. Randy Roberts, Monica Garcia, Jessie A. Jackson, Arthur G. Jolly, Jo Ellen Neuman, Hiroshi H. Saito, and Mark A. Warmerdam of LLNL and Douglas M. Ranken and Meilen Yen of LANL.

References

[1] Proposal #LL11-FY11-191-PD06 Project Lifecycle Plan, “Benchmark Imagery for Assessing Geospatial Semantic Content Extraction Algorithms (Updated),” Prepared for the U.S. Department of Energy (DOE) National Nuclear Security Administration, Office of Nonproliferation and Verification R&D (NA-22), January 2012.

[2] Paul Pope, Randy Roberts, Luci Gaines, Travis White, John Goforth, “Imagery Analyst Annotation Exercise: Final Report,” Project report to NA-22 Simulations, Algorithms and Modeling Office, March 13, 2012.

[3] Paul Pope, Travis White, Randy Roberts, John Goforth, and Luci Gaines, “Sources, Criteria, and Methods for Creating Benchmark Imagery,” March 30, 2012.

[4] W. Travis White III, Paul A. Pope, John Goforth, Lucinda R. Gaines, “Description of Initial Annotated Benchmark Imagery for Assessing Geospatial Semantic Content Extraction Algorithms,” Technical Report prepared for NA-22, 31 August 2012.

[5] John W. Goforth, W. Travis White III, Paul A. Pope, Randy S. Roberts, Ian G. Burns, Lucinda R. Gaines, “Benchmark Imagery Project Report on Generation of Synthetic Images,” Technical Report prepared for NA-22, September 27, 2012

[6] Paul A. Pope, W. Travis White, John Goforth, and Lucinda Gaines, “Illustration of the Use of the Benchmark Imagery Suite for Verification and Validation of Semantic Extraction Algorithms,” Technical Report prepared for NA-22, September 30, 2012.

[7] Steven Schubert, Panel Coordinator, "Independent Project Review Summary Report," (Official Use Only), Report to NA-22, results of review held at LLNL on 1 February 2012.

[8] Travis White and Paul Pope, ""Benchmark Imagery for Assessing Geospatial Semantic Content Extraction Algorithms," presented at the NA-22 Verification, Monitoring, & Remote Detection (VMRD2012) Joint Program Review Meeting, Held May 15 –17, 2012, Las Vegas, NV.

[9] Paul Pope, Randy Roberts, Luci Gaines, Travis White, and John Goforth, "Results of an Industrial Facility Photo-Interpretation Exercise," poster presented at the 2012 Annual Meeting of the American Society of Photogrammetry and Remote Sensing, March 19 – 23, 2012, Sacramento, CA, LA-UR-12-20239.

[10] NNSA Funding Opportunity Announcement, Funding Opportunity Number DE-FOA-0000568, CFDA Number 81.113, December 29, 2011.

[11] *Field of Dreams*, Phil Alden Robinson (screenplay), Universal Pictures, April 1989.

[12] Thomas C. Chisnell and Gordon E. Cole, "'Industrial Components' — a Photointerpretation Key on Industry," *Photogrammetric Engineering* 24, 590 – 602, March 1958.

[13] The html references that were used to form the industrial facility gazetteer are:

- http://www.epa.gov/ttn/atw/area/facilities.html
- http://www.epa.gov/enviro/html/fii/ez.html
- http://www.energyjustice.net/map/
- http://www.osha.gov/pls/imis/establishment.inspection

[14] See, for example, Niles Ritter and Mike Ruth, "GeoTIFF Format Specification, GeoTIFF Revision 1.0," http://www.remotesensing.org/geotiff/spec/geotiffhome.html.

[15] R. S. Roberts, P. A. Pope, M. Jiang, T. Trucano, C. Aragon, K. Ni, T. Wei, L. Chilton , A. Bakel, "GSV Annotated Bibliography," LLNL-TR-487191, June 14, 2011.

[16] J. R. Jensen, "Chapter 8: Thematic Information Extraction: Image Classification," in *Introductory Digital Image Processing: A Remote Sensing Perspective*, Second Edition, Prentice Hall, New Jersey, 1996.

[17] P. Heckbert, "Color Image Quantization for Frame Buffer Display," *Computer Graphics* 16(3), 1982, pp. 297-307.

[18] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi,"Yet Another Survey on Image Segmentation: Region and Boundary Information Integration," *Proc. of the Seventh European Conference on Computer Vision (ECCV 2002)*, Part III (Lecture Notes in Computer Science Vol.2352), May 28-31, 2002, Copenhagen, Denmark, pp.408-22.

[19] E. Meijering, "Cell Segmentation: 50 Years Down the Road," *IEEE Signal Processing Magazine*, September 2012, pp. 140-145.

[20] A. Hanbury, "A Survey of Methods for Image Annotation," Journal of Visual Languages and Computing, 19, pp. 617-627, 2008.

[21] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing*, ISBN 0-201-50803-6, Addison-Wesley Publishing Co., Menlo Park, CA, September 1993, Section 9.3.3.

[22] F. Tanner, B. Colder, C. Pullen, D. Heagy, M. Eppolito, V. Carlan, C. Oertel, and P. Saillee, “Overhead Imagery Research Data Set (OIRDS): An Annotated Data Library and Tools to Aid in the Development of Computer Vision Algorithms,” *Proc. Of the 2009 IEEE Applied Imagery Pattern Recognition Workshop (AIPR 2009)*, October 14-16, 2009, Washington, DC, USA, pp. 1-8.

[23] TurboSquid, 935 Gravier Street, Suite 1600, New Orleans, LA, 70112, http://www.turbosquid.com/.

[24] J. R Schott., J. E. Mason, C. Salvaggio, J. D. Sirianni, R. A. Rose, E. O. Kuip, D. K. Rankin, "DIRSIG — Digital imaging and remote sensing image generation model: description, enhancements, and validation," RIT/DIRS Report 92/93-51-146, July 1993.

[25] Randy S. Roberts, Timothy G. Trucano, Paul A. Pope, Cecilia R. Aragon, Ming Jiang, Thomas Wei, Lawrence K. Chilton and Alan Bakel, “On the Verification and Validation of Geospatial Image Analysis Algorithms,” 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010), July 25–30 Honolulu, HI, paper 2741.