LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-480881

Design of Benchmark Imagery for Validating Facility Annotation Algorithms

R. S. Roberts , P. A. Pope, R. R. Vatsavai, M. Jiang, L. F. Arrowood, T. G. Trucano, S. Gleason, A. Cheriyadat, A. Sorokine, A. K. Katsaggelos, T. N. Pappas, L. R. Gaines, L. Chilton

April 21, 2011

IEEE International Symposium Geoscience and Remote Sensing
Vancouver, Canada
June 24, 2011 through June 29, 2011

Disclaimer

DESIGN OF BENCHMARK IMAGERY FOR VALIDATING FACILITY ANNOTATION ALGORITHMS

Randy S. Roberts [1], *Paul A. Pope* [2], *Raju R. Vatsavai* [3], *Ming Jiang* [1], *Lloyd F. Arrowood* [6], *Timothy G. Trucano* [4], *Shaun Gleason* [3], *Anil Cheriyadat* [3], *Alex Sorokine* [3], *Aggelos K. Katsaggelos* [7], *Thrasyvoulos N. Pappas* [7], *Lucinda R. Gaines,* [2] *and Lawrence K. Chilton* [5]
{[1] Lawrence Livermore, [2] Los Alamos, [3] Oak Ridge, [4] Sandia, [5] Pacific Northwest} National Lab, USA
[6] Y-12 National Security Complex, [7] Northwestern University, USA
Email: roberts38@llnl.gov

ABSTRACT

The design of benchmark imagery for validation of image annotation algorithms is considered. Emphasis is placed on imagery that contains industrial facilities, such as chemical refineries. An application-level facility ontology is used as a means to define salient objects in the benchmark imagery. Instrinsic and extrinsic scene factors important for comprehensive validation are listed, and variability in the benchmarks discussed. Finally, the pros and cons of three forms of benchmark imagery: real, composite and synthetic, are delineated.

Index Terms— Benchmark imagery, Algorithm validation, Ontology, Benchmark variability, Real annotated imagery, Validation using synthetic imagery

1. INTRODUCTION

Searching large image databases for remotely-sensed industrial facilities is a complex and difficult task [1, 2]. Part of the difficulty, as illustrated in Figure 1, lies in the fact that facilities are complex geospatial arrangements of functionally interdependent objects. One approach to this problem is to label and ascertain the relative spatial locations of objects in the imagery, and use these attributes as keys for the search [3]. An important step in the development of such auto-annotation algorithms is a verification and validation (V&V) strategy [4]. A properly designed and implemented V&V strategy establishes and quantifies the conditions under which an annotation algorithm can be applied to imagery with an expectation of success. Furthermore, a key component of the V&V methodology is a large, well-designed set of benchmark imagery [5, 6].

In this paper we propose a methodology to design benchmark imagery for the V&V of facility annotation algorithms. Rather than taking an ad-hoc approach of seeking and annotating available facility imagery, we propose to design the benchmarks by specifying the attributes of the imagery, and then acquire imagery that best meets the specifications. In the context of facility annotation algorithms, benchmark imagery refers to imagery of facilities in which objects relevant to the purpose of the facility have been annotated.

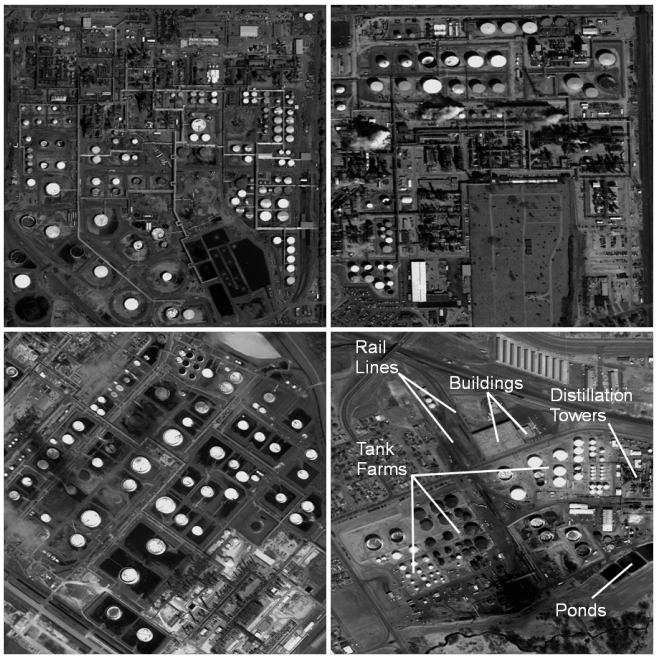

Fig. 1. Four real imagery examples of chemical refineries illustrating the large layout variation within a single industry type, with the last example manually annotated. (Copyright Digital Globe 2008.)

2. DESIGN OF BENCHMARK IMAGERY

The design of benchmark imagery involves several considerations including intrinsic scene attributes (objects in the scene and their geospatial relationships), extrinsic scene attributes (e.g., illumination, sensor characteristic, etc. that describe how a 3D scene is mapped into a 2D image), and annotation labels. Central to our approach is an application-level ontology that provides a principled means to determine ob-

To appear in *Proc. IGARSS 2011, July 24-29, Vancouver, Canada*

jects that compose a facility, objects that constitute clutter, and standardized annotations. Realizing that *comprehensive* validation requires benchmarks that span a wide range of intrinsic and extrinsic factors, we consider both the variability and uncertainty associated with benchmark imagery. In passing we note that the design of benchmark imagery is closely related to the metrics and processes used to validate the algorithms. While we are aware of this relationship, and implicitly account for it in our design methodology, we defer discussion of the validation process and associated issues to future publications.

2.1. Ontology for Scene Objects and their Annotations

Following the outline presented in [4], we consider the use of an application-level ontology as a means to specify physical objects that populate the scene in benchmark imagery. Ontology is a useful tool to represent subject-matter expert (SME) knowledge about industrial facilities [1, 4]. Typically ontologies utilize formal languages like the Web Ontology Language (OWL) and the Knowledge Interchange Format (KIF) to represent domain-specific or more general knowledge. They are capable of depicting various levels of generalization and aggregation as conceptualized by the domain experts. For example, an ontology of industrial facilities would contain a classification of industrial facilities (manufacturing, power generation, etc.), classifications of the elements of industrial facilities (building, pipe, access roads, storage tanks, etc.), and relations that tie facility elements into explicit geographical features with observable qualities.

By integrating domain knowledge from a large number of experts and recognized knowledge sources such as textbooks and standards, ontologies are capable of describing variability of the concepts (e.g., typical ranges in planimetric area for building footprints) and the relationships between them (e.g., next-to and far-away). Ontology also provides a means of defining the objects, and standardized annotations, that are salient, pertinent, or unrelated to the purpose of the facility.

2.2. Intrinsic and Extrinsic Factors

Automation of facility annotation requires some degree of image segmentation. Hence, a comprehensive validation of image annotation algorithms must consider the underlying image segmentation algorithms. Both intrinsic and extrinsic factors play a role in image segmentation, as data driven segmentation algorithms tend to rely on the image attributes such as color (or, in general, pixel intensity), edges, and texture (or combinations of these attributes). Table 1 lists factors that influence segmentation (and therefore annotation) of imagery containing facilities. The list is by no means comprehensive, but rather serves to illustrate factors to consider when designing benchmark imagery. This list of factors is based on image interpretation studies, such as [7], and the experience of the authors when designing collection campaigns and analyzing facility imagery.

Each factor in Table 1 contains three levels describing variations of the factor. Facility location and compactness contribute to the scene information, and based on the level of information required, the possible segmentation solutions can vary. For example, in a dense neighborhood a possible segmentation solution would identify spatially co-located similar objects (group of buildings or group of cars) as a single segment. In a suburban neighborhood the adjacency of different classes of objects, like trees and buildings, can impact segmentation performance. Roof type is another parameter that can impact the segmentation performance — for example, multi-faceted roofs tend to be over-segmented. More than the building size and shape dimensions, the adjacency of buildings is often a confounding factor for segmentation algorithms. Sensor characteristics, ambient light conditions, phenological variations, and climatic conditions also influence the segmentation solution space. These parameters directly impact low-level visual elements such as color, edges and local textures.

2.3. Variability and Uncertainty of the Benchmarks

Variability in benchmark imagery is a key to robust algorithm validation [4, 5]. Within the context of validation testing, there are two facets of benchmark variability. First there is the variation in image content that is representative of the range of images that may occur for the intended application of a given algorithm. We assume that this variation can be characterized, or "parameterized," by factors intrinsic to the scene, such as different arrangements of buildings and landscaping, and by factors extrinsic to the scene such as viewing geometry, illumination and sensor artifacts. When selecting a set of benchmark images for validation testing, representing this variation is understood as designing a set of tests that cover, to some degree, the anticipated variety of images that will confront the algorithm in its intended application. Because even a small amount of intrinsic and extrinsic variability in this sense can give rise to an enormous amount of benchmark imagery, due to the explosion of possible variable combinations, this variability poses significant problems for defining and organizing a benchmark test set to achieve appropriate levels of coverage.

The second facet of variability is acknowledged by the fact that given any specific benchmark image, there is variability in the information in the image (e.g., additive white Gaussian noise, cropping, compression) simply thought of as "noise." The presence of noise, or aleatory uncertainty (see [4]) in the benchmark influences the mechanisms of comparing the algorithm performance with the benchmark and in interpreting what these comparisons mean for purposes of validation. The role of this facet of variability is not a component of validation test problem design, rather it is a component of

Factor	Levels (3)
Facility location	Urban, Suburban, Rural
Facility size	Small, Medium, Large
Compactness	Sparse, Moderate, Dense
Roof type	Flat, Sloped, Multi-faceted
Building size	Small, Medium, Large
Time of day*	Mid-Morning, Noon, Dusk
Sensor view angle*	Nadir, Low-oblique, High-oblique
Spatial scale*	Small, Medium, Large
Visibility*	5, 10, 20 (km)
Cloud cover*	Clear, Broken, Thin Cirrus
Season*	Summer, Fall, Winter
Climate zone*	Tropical, Temperate, Arid

Table 1. Short list of intrinsic and extrinsic factors and associated levels. Extrinsic factors are marked with an asterisk.

interpreting what the results of the test mean.

While we previously observed in [4] that validation must typically engage epistemic uncertainty, here we assume that each benchmark image has complete knowledge associated with it, and therefore no epistemic uncertainty. In principle, the performance of the algorithm could have aleatory uncertainty (for example, stochastic algorithms are used) and epistemic uncertainty (specific numerical errors are unknown), but we also ignore these uncertainties in the present discussion.

Even with the small number of factors and levels listed in Table 1, it is evident that a very large quantity of imagery is needed for comprehensive validation. In general, the total number of benchmark images is $T = N \times L^F$ where F is the number of factors, L is the number of levels and N is the number of images per factor and level combination. In our example, $L = 3$, $F = 12$, and say that we desire $N = 3$, which yields an estimate of $T \approx 1.6$ million images in the validation ensemble. This amount of imagery could be reduced by careful subselection of factors and levels. Still, the volume of imagery involved in validation indicates that non-traditional benchmarks, such as synthetic imagery, should be considered.

3. SOURCES OF BENCHMARK IMAGERY

Design and acquisition of remotely-sensed facility imagery that are suitable for benchmarks is non-trivial even with the essential scene elements specified as above. One approach is to acquire imagery using freely available sources such as the internet. A potential drawback to this approach is that wide distribution of the benchmark imagery might be prohibited by copyright and legal issues [5]. Another potential issue with internet imagery is the lack of image metadata, which is often useful when processing remotely sensed imagery. Given these drawbacks, our preference is to obtain facility imagery directly from image providers. This approach also allows us the opportunity to specify collection attributes and potentially achieve the benchmark imagery specifications.

Real imagery is available from several sources in a range of costs, quality and spatial resolutions. The US Geological Survey Earth Resources Observation and Science (EROS) Center provides aerial photography and satellite imagery at minimal cost. High resolution imagery from commercial vendors can currently be purchased at prices ranging from approximately US$500 to $2.5k per full-frame image. Manually annotating facility imagery is a difficult, tedious, and potentially boring task that is prone to error. The human capital required for manual annotation is difficult to ascertain, and greatly depends on the detail and precision of annotation. For example, annotating obvious objects such as buildings and large storage vessels can be produced by a novice in a short amount of time. However, relatively large errors in omission (completeness of the annotation task) and commission (misinterpretation of objects) are likely to be incurred. Annotating to a greater level of detail, e.g. annotating specific processing units such as fractional distillation towers, can require several hours of effort by an experienced image analyst. We prefer the latter level of annotation detail, as that level provides clues to the purpose of the facility.

Given the expenses involved, it would be costly to develop a set of benchmark imagery with appropriate variability using only real imagery. This fact compels us to consider the use of composite and synthetic imagery to augment our set of real imagery as part of our benchmarks. Real imagery can be used to guide the development of models for synthetic imagery, and can serve as the foundation for composite imagery. Features of interest for the synthetic and composite benchmarks are determined from the ontology, and would include such objects as new tanks, piping and specialized vehicles.

Synthetic imagery has obvious advantages in terms of mass production of large amounts of imagery with controlled extrinsic variability such as illumination conditions, viewing geometry and sensor artifacts. It can be generated through an image rendering process using a geometrically modeled scene with desired lighting and shading information. Modeling synthetic scenes involving industry facilities requires generating geometric models at the neighborhood or city scale that include both man-made objects, such as roads and buildings, and natural objects, such as grassy fields and trees. These geometric models can be generated interactively using CAD tools, which can be time consuming, or automatically using procedural modeling techniques [8], which have limited fidelity. Proper illumination is crucial for accurately capturing shadows, glints and atmospheric effects in a scene. Developing tools that can produce the appropriate global illumination in a semi-automated fashion is essential for streamlining the rendering process. One of the limitations with synthetic imagery is the amount of man-hours required to model an entire scene. Another limitation is the computational cost of producing photorealistic renderings of a synthetic scene. As is

often the case, there is a tradeoff between rendering quality and computational efficiency.

Composite imagery can be generated by inserting and blending smaller foreground images into a larger background image. Although both foreground and background images can be either real or synthetic, the most common approach is to synthesize features of interest into foreground images and to composite them into a real background image that does not contain such features at the desired locations. There exist a variety of techniques developed in the computer graphics community and the entertainment industry for generating composite images with varying degrees of photorealism and computational efficiency (e.g. Poisson image editing [9], billboarding and z-buffer compositing). Likewise, tools from the remote sensing community, such as DIRSIG [10], can also be utilized due to their high fidelity and wealth of capabilities. One of the limitations with composite imagery is that blending between foreground and background images can create artifacts, such as seams, due to differences in illumination, scale and noise. The problem with these compositing artifacts is that they can cause image analysis algorithms to perform differently than on a non-composite imagery with the exact same content. Potential benefits aside, the use of synthetic and composite imagery for algorithm V&V is not well understood and this topic is an active area of research [11].

4. SUMMARY

In this paper we propose an approach to the design of benchmark imagery. Although our focus is on validation of algorithms that auto-annotate imagery containing industrial facilities, the approach is general and applicable to developing benchmark imagery for other validation problems. We start by creating an application-level ontology, which provides a structured means of specifying the objects, and their spatial relationships, in the benchmarks. Next, important intrinsic scene factors, derived from the ontology, and extrinsic scene factors are listed along with levels for each factor. These two steps define the contents of the benchmark imagery.

The final step is to create the ensemble of benchmark imagery. The number of benchmarks required for comprehensive validation, even for a small number of factors and levels, can be quite large and expensive to acquire. Given these issues, we propose that the ensemble of benchmark be composed of real, composite and synthetic imagery. Synthetic imagery is proposed as a mitigation step against the large number of anticipated benchmarks. Finally, we note that performing comprehensive validation is a formidable task, in part due to the magnitude of the benchmark ensemble. Approaches to the validation process are considered in a future paper.

5. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Alexander Slepoy, Program Manager, Simulations, Algorithms, and Modeling; Office of Nonproliferation Research & Development, National Nuclear Security Administration, for his support of this research.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-480881. v1.02

6. REFERENCES

[1] R. R. Vatsavai et al., "Geospatial image mining for nuclear proliferation detection: Challenges and new opportunities," in *Proc. 2010 IEEE Int'l Geoscience and Remote Sensing Symposium*, July 2010, pp. 48–51.

[2] S. Gleason et al., "Semantic information extraction from multispectral geospatial imagery via a flexible framework," in *Proc. 2010 IEEE Int'l Geoscience and Remote Sensing Symposium*, July 2010, pp. 166–169.

[3] B. Yao et al., "I2T: Image parsing to text description," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, August, 2010.

[4] R. Roberts et al., "On the verification and validation of geospatial image analysis algorithms," in *Proc. 2010 IEEE Int'l Geoscience and Remote Sensing Symposium*, July 2010, pp. 174–177.

[5] T.L. Berg et al., "It's all about the data," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1434–1452, August 2010.

[6] J.A. Shufelt, "Performance evaluation and analysis of monocular building extraction from aerial imagery," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 311–326, April 1999.

[7] T. Chisnell and G. Cole, "Industrial Components—A Photo Interpretation Key on Industry," *Photogrammetric Engineering*, vol. 24, pp. 590–602, March 1958.

[8] Y.I.H. Parish and P. Müller, "Procedural modeling of cities," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 301–308.

[9] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM Trans. Graphics*. ACM, 2003, vol. 22, pp. 313–318.

[10] J. Mason et al., "Validation of contrast and phenomenology in the Digital Imaging and Remote Sensing (DIRS) lab's image generation (DIRSIG) model," 1994, vol. 2269, pp. 622–633, SPIE.

[11] J.P. Bellucci, T.E. Smetek, and K.W. Bauer, "Improved hyperspectral image processing algorithm testing using synthetic imagery and factorial designed experiments," *IEEE Trans. Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1211–1223, March 2010.