LA-UR- 10-06973

Approved for public release; distribution is unlimited.

Title:	Visual Attention Based Detection of Signs of Anthropogenic Activities in Satellite Imagery
Author(s):	Skurikhin, A.N.
Intended for:	Applied Imagery Pattern Recognition Workshop, Washington DC, USA, 10/13/2010-10/15/2010.

Form 836 (7/06)

Visual Attention Based Detection of Signs of Anthropogenic Activities in Satellite Imagery

Alexei N. Skurikhin
Los Alamos National Laboratory
Los Alamos, NM 87545 USA
alexei@lanl.gov

Abstract — **With increasing deployment of satellite imaging systems, only a small fraction of collected data can be subject to expert scrutiny. We present and evaluate a two-tier approach to broad area search for signs of anthropogenic activities in high-resolution commercial satellite imagery. The method filters image information using semantically oriented interest points by combining Harris corner detection and spatial pyramid matching. The idea is that anthropogenic structures, such as rooftop outlines, fence corners, road junctions, are locally arranged in specific angular relations to each other. They are often oriented at approximately right angles to each other (which is known as rectilinearity relation). Detecting the rectilinearity provides an opportunity to highlight regions most likely to contain anthropogenic activity. This is followed by supervised classification of regions surrounding the detected corner points as man-made vs. natural scenes. We consider, in particular, a search for anthropogenic activities in uncluttered areas.**

Visual attention; cueing; corner detection; spatial pyramid; man-made object detection; satellite imagery

I. INTRODUCTION

High resolution commercial satellite imagery provides a unique opportunity for geographic profiling of activities of interest. Imagery supplied by satellites, such as IKONOS, QuickBird, WorldView and GeoEye, reveals many details previously unobservable in satellite images. It has been shown that the new generation of commercial satellites is capable of making a useful contribution to a wide variety of applications, such as land use mapping [15, 17], topographic mapping [4], natural disaster assessment [18], change analysis [5, 21], detection of man-made objects [14].

Observations from space using high resolution commercial satellite imagery can be used to detect "hot spots", containing combinations of man-made structures, surface disturbances, and contextual factors, which could be of interest. This can trigger a further scrutiny of locations found interesting. Such use of commercial imagery can result in substantial reductions in broad area search efforts, as well as can make relevant information available on a much timelier basis. It should be recognized that although there has been recent significant progress in general purpose image segmentation and object recognition, an automated broad area search for signs of anthropogenic activities in high-resolution satellite imagery remains a difficult problem. The main challenges to address are spatial and temporal transferability of search, non-conformities (such as signature suppression), and huge size of image data that should be processed in a reasonable amount of time for result to be of practical value. While exploitation of lower spatial resolution imagery (e.g. ASTER 15-m imagery) may be acceptable, when objects of interest constitute compact cluster of several hundred meters diameter, it is not always applicable. Therefore, to provide robust identification of "interesting" sites, it is necessary to develop tools that can also perform search in high resolution satellite imagery.

The method put forth in this paper treats broad area search as a goal-driven visual attention problem. The intent is to reduce the amount of incoming image data to task-relevant parts of the image and direct subsequent processing to such image parts. These parts, regions of interest, are the locations of where object/site categorization and recognition are applied. This is in contrast with recognition approaches which look for objects of interest by applying all the computational resources uniformly across the whole image. Instead, visual attention is known for its capability to gate image information to most informative parts of the image [23]. Visual input is usually first decomposed into a set of general feature maps, such as intensity, color, orientation, which feed, in a bottom-up manner, into a master saliency map to identify most salient regions that stand out of the local image background [6, 7]. This is known as spotlight approach and it is based on Treisman's feature integration theory [19]. The type of performance which can be expected from models of this type, e.g. [6], critically depends on one factor: only object features explicitly represented in feature maps lead to pop-out.

Recent advancements of spotlight approach include models which integrate top-down information, e.g., [12, 16, 22], to bias bottom-up saliency computation in order to enhance saliency of areas of interest. The idea is that competition to gain visual attention occurs not only between individual features but also between object like entities, which are formed as a result of image segmentation. Therefore, one of main challenges of methods of this kind is the development of segmentation of object like entities.

A different approach that employs visual attention to narrow the focus for further processing was presented in [20]. The approach relies on bottom-up saliency-based attention model, as in [6]. Once the bottom-up analysis that is based on the integration of general feature maps identifies a set of most salient pixel patches, learning and recognition of objects are done using only the indentified patches. Both learning and

recognition use attributed interest points that are computed using Scale Invariant Feature Transform (SIFT) [10].

In this paper, we present a goal-driven cueing and recognition, where cueing is driven by task related features and object recognition is applied only to the regions identified by cueing. Instead of analyzing general features such as color, orientation and contrast, as in [20], we focus on features that are task related, such as corners. This has an advantage over the use of general features which can be useful for narrowing the search for man-made objects but do not necessarily carry any task relevant information. Therefore, it can lead to misdetection. We selected corners as indicators of rectilinearity, which provides an opportunity to highlight regions most likely to contain signs of anthropogenic activity. While this feature is general to recognize specific objects, we hypothesize that it contains enough information to pre-select regions most likely to contain a variety of man-made objects in different backgrounds. Then, preselected regions are processed by supervised classifier to categorize preselected regions as man-made vs. natural scenes. Categorization is based on the spatial pyramid matching approach [8] which has demonstrated a very good performance on several computer vision datasets. For evaluation we compare our two-tier approach against recognition without cueing. When recognition based on the spatial pyramid matching is used without cueing, it is subsequently applied to multiple locations across the whole image (which is known as a "sliding window" approach).

The next two sections of the paper describe our approach and experimental results. The conclusions and discussion are stated in the fourth section.

II. Corner-Based Cueing and Recognition

Our approach, CC-SPM, to broad area search consists of two major components: corner based cueing (CC) and recognition based on spatial pyramid matching (SPM). SPM based recognition follows corner detection and is applied to the rectangular image patches centered on the detected corner points. Overall goal is to achieve both high precision and recall by incorporating general domain related knowledge into the broad area search for man-made structures.

Corner detection is based on the Harris corner detector [3]. The detection is done using the second moment matrix, also called the auto-correlation matrix, M. The matrix describes the gradient distribution in the local neighborhood of a point and it is defined by:

$$M = G(\sigma) * \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \quad (1)$$

where I_x, I_y are the local image derivatives and $G(\sigma)$ is the Gaussian kernel of scale σ. We used the cornerness measure of Noble [13], C, which is defined as:

$$C(x,y) = det(M)/(trace(M)+\varepsilon), \quad (2)$$

where $det(M)$ is the determinant of M and $trace$ is the trace of M. Once the cornerness computed for each pixel, they are filtered by thresholding that eliminates all the pixels with cornerness values below the threshold. This threshold is selected in a semi-adaptive manner using pre-specified percentile of a set of pixel cornerness values.

Image block of size 300×300 pixels, which approximately corresponds to the size of 180×180 m, is created around each corner that is preserved by thresholding. These regions are then classified using the spatial pyramid matching approach [8]. This approach models image region (block) using representation similar to the bag of visual words scheme [2]. SPM constructs the image representation by iteratively partitioning the image block into increasingly fine rectangular sub-blocks and then concatenates histograms of SIFT descriptors found inside each sub-block (Fig. 1). SIFT descriptors are computed for pixel patches of size 16×16 pixels with spacing of 8 pixels.

For a compact representation, a visual vocabulary is built by clustering SIFT descriptors using the k-means algorithm. The vocabulary size is determined by the number of clusters pre-specified for the k-means algorithm. Sizes for visual vocabularies can be an order of several hundreds and more and depend on problem at hand.

Each cluster is treated as a word in the visual vocabulary. Counting number of occurrences of SIFT descriptors in each cluster results in histogram representation for each image block. This counting is done by comparing the SIFT descriptors to the cluster centers and associating every descriptor to that cluster to which the distance to its center is the minimal one. The resulting histograms are the bag of visual words representation for the image block and its partitions.

Matching of image blocks is based on the weighted intersection of their integrated histograms, which are represented as long concatenated vectors of histograms corresponding to different resolutions and partitions for each of image blocks. The classifier that works with this bag of visual words representation is the support vector machine classifier (SVM).

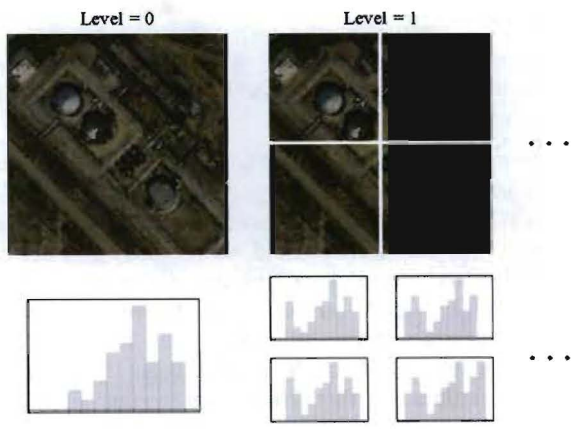


Figure 1. An illustration of the spatial pyramid representation using two levels of resolution. For each level of resolution and each partition feature histograms are constructed.

III. EXPERIMENTAL RESULTS

A. Data Preparation

For training and evaluation, we used high-resolution commercial satellite imagery provided by Digital Globe and Google Earth. Spatial resolution of the images is ~0.6 m.

For training SVM-based recognizer, we created a dataset of 512 images, each image of size 300×300 pixels. The dataset contained images of two categories "natural" and "man-made" scenes. There were 300 images in the "man-made" category and 212 images in the "natural" category. Sample images from the training dataset are shown in Figs. 2 and 3. Accuracy of the SPM based classification on the training dataset was 96%.

For evaluation, we used a set of ten high-resolution images; average image size is 10,000×10,000 pixels. Table 1 summarizes the evaluation dataset description. The images contained man-made structures in different landscapes, such as desert, mountains, forest and farm fields. Man-made structures varied from industrial facilities to residential areas. The dataset contained examples of man-made structures such as nuclear power plants, small villages, towns, mines. Because we focus on a search for anthropogenic activities in uncluttered areas (i.e. non-urban, remote), nine images were images of relatively uncluttered geographical areas. The last image is the image of an area with multiple man-made structures of different types, including residential housing. The motivation was to evaluate corner-based efficiency in the presence of man-made clutter.

Man-made structures that were contained in ten images were manually labeled with polygonal shapes. These labels of man-made areas were used to test CC-SPM and SPM methods. The size of anthropogenic signs was approximately of 200×200 m and larger. Therefore, the size of training images approximately matched the size of sites we looked for. Most of the created labels occupied areas 300×300 pixels and larger. To challenge the recognition and evaluate false negative rate, we also labeled several man-made areas as small as ~20×20 m. Detection of man-made site was interpreted as true positive if area of the intersection of the image block (300×300 pixels) recognized as man-made area and the labeled area was at least 1% of the image block area. Such a low area threshold was used to counterbalance the fact that the recognizer was optimized for the detection of much larger areas, and it was not trained to recognize small man-made areas.

B. Performance Evaluation

The effectiveness of SPM and CC-SPM detection is evaluated by precision, recall and false negative rate. Precision, recall and false negative rate are defined as functions of true positives (correct detections), false positives (incorrect detections) and false negatives (missed targets) as in

$$\text{Precision} = \text{NmTPs}/(\text{NmTPs}+\text{NmFPs}), \quad (3)$$

$$\text{Recall} = \text{NmTPs}/(\text{NmTPs}+\text{NmFNs}), \quad (4)$$

$$\text{FNR} = \text{NmFNs}/\text{NmTargets}, \quad (5)$$

where NmTPs is the number of true positives, NmFPs is the number of false positives, FNR is the false negative rate, NmFNs is the number of false negatives and NmTargets is the number of man-made areas (labeled targets) in the image. Therefore, the precision shows how accurate the prediction of the target is and the recall shows the percentage of correctly detected targets with respect to the total number of targets. FNR, false negative rate, shows the percentage of the targets missed.

Our results are reported with the cornerness threshold that is equal to the 99.99-th percentile of the computed cornerness for image pixels and recognition was done using image blocks of size 300×300 pixels. Feature vector was constructed with number of levels, that is equal to 3, and a vocabulary size, that is equal to 200. Training and classification were done with support vector machine classifier. The actual implementation of the SVM was done based on the libSVM software [1].

Table 1 shows the results obtained with SPM and CC-SPM approaches. SPM column corresponds to a "sliding window" approach, while CC-SPM column corresponds to the SPM-based recognition focused on areas which were pre-selected by corner based cueing. In this table, the number of image blocks that were processed in the course of performing object detection is also shown. Recall that in the case of CC-SPM, the number of image blocks equals to the number of detected corners. As can be seen from the table, corner-based cueing increased the precision from 43.5% to 81.02%, while the recall changed from 96.85% to 90%. Therefore, the use of goal-driven knowledge in the form of corners makes possible to achieve both high recall and high precision. SPM complemented corner detection by classifying corner surrounding regions as man-made vs. non man-made. Fig. 4 shows examples of successful elimination of non-man-made areas surrounding the detected corners. It should be noted that high precision and recall were achieved by CC-SPM with the one seventh of the number of image blocks that were evaluated by "sliding window" approach employed by SPM.

Analysis of the results obtained for individual images has revealed that the most difficult cases that decrease the precision are arrangement of man-made structures embedded in the mountain-like landscape. Figs. 5a illustrates false positive detection over the mountain landscape. We hypothesize that the reason is due to the presence of elongated ridge like structures that might seem as man-made structures for SPM-based recognition. Fig. 5b shows another example of false positive detected of desert-like landscape, where dried river beds might look like man-made structures as well. While the use of multiple scales (not only blocks of 300×300 pixels) can potentially improve the precision over the mountain landscape, the dried river bed might require a different approach. Fig. 6 shows a comparison of the detection results obtained with SPM and CC-SPM methods. It is obvious that corner based cueing has drastically reduced the number of false positives in this case.

Figure 2. Sample images of man-made areas used for training man-made region recognition algorithm.

Figure 3. Sample images of nonman-made areas used for training man-made region recognition algorithm.

TABLE I. DETECTION RESULTS

	SPATIAL PYRAMID MATCHING (SPM)	CORNER-BASED CUEING FOLLOWED BY SPATIAL PYRAMID MATCHING(CC-SPM)
NUMBER OF IMAGES IN THE DATASET	10	
AVERAGE IMAGE HEIGHT (PIXELS)	10,313	
AVERAGE IMAGE WIDTH (PIXELS)	10,363	
THE CHANCE PROBABILITY OF DETECTING MAN-MADE AREAS IN THE DATASET (×100%)	6.929	
SIZE OF THE BLOCK (PIXELS)	300 × 300	
NUMBER OF IMAGE BLOCKS PROCESSED (FOR THE WHOLE DATASET)	28,456	3,950
AVERGAGE PRECISION (×100%)	43.508	81.02
AVERAGE RECALL (×100%)	96.855	90.003
AVERAGE FN R (×100%)	11.158	40.113

An important characteristic of broad area search for signs of anthropogenic activities is the false negative rate. As it can be seen from the table 1, the false negative rate increased approximately 3 times while going from SPM to CC-SPM. False negatives can be categorized into two groups: (1) misdetection by SPM (Fig. 5c), and (2) misdetection of corners (Fig. 5d). Misdetection by SPM is most frequently done for areas containing small clusters of man-made structures, such as the ones shown in Fig. 5c. This can be explained by the nature of feature vector that is used by SPM. The feature vector was built using SIFT descriptor, and can be thought as texture descriptor. Therefore, a search for small structures using such the descriptor with fixed scale might not be an optimal strategy.

As it was expected, majority of false negatives are due to misdetection of corners either due to non-optimal setting of the cornerness thresholding or lack of strong corners in man-made areas (Fig. 5d), such as road junctions in rural areas. Non-optimal thresholding on cornerness was particularly crucial in the case of the image containing multiple spatially separated man-made areas. In our dataset such image contained approximately 16.2% area occupied by 437 spatially separated man-made areas. Due to the thresholding, 617 corners were preserved for further analysis for SPM-based recognition. These corners did not hit all the labeled man-made areas. Performance achieved on that specific image was: Precision = 96.11, Recall = 79.7 and FNR = 34.55. A simple lowering of the threshold does not work as it would result in more number of false positives over the mountain landscape, which is the most difficult case for SPM based recognition. A possible avenue for future research to address this problem is the integration of cornerness directly into the feature vector.

Overall, for the used dataset, the corner based cueing had problems for areas, which either do not have clearly expressed corner structures (e.g. electric poles), have weak corners, or for the images cluttered with spatially separated man-made areas. Recognition using spatial pyramid matching had most problems for the areas that contained man-made like looking structures, such as mountain landscapes or desert landscape with dried river beds.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed and evaluated a two-tier approach to broad area search for signs of anthropogenic activities. Results from experiments on high-resolution (~0.6m) commercial satellite image data showed the potential applicability of this approach and its ability of achieving both high precision and recall rates. The main advantage of combining corner-based cueing with general object recognition is that the incorporation of domain specific knowledge even in its more general form, such as presence of corners, provides a useful cue to narrow the focus of search for signs of anthropogenic activities. Combination of corner based cueing with spatial pyramid matching addressed the issue of corner categorization. An important practical issue for further research is optimizing the balance between false positive and false negative rates.

While the results presented in the paper are encouraging, the problem of an automated broad area search for signs of anthropogenic activities remains challenging. Logical extension of this work is to perform more experiments on a larger set of satellite imagery with manual labels tuned to specific application scenarios. Further research is necessary to optimize the balance of false negatives and false positives. The optimization of false alarms and false negative rate might be achieved via a number of steps. Goal-driven cueing should include fusing the outputs of several interest point detectors, such as different types of corner and rectilinearity detectors, as well as detectors of other cues associated with anthropogenic activities, such as circularity, collinearity and curvilinearity. Such detectors can be built upon linear scale-space theory [9], e.g. Harris-Laplace corner detector [11]. Further research to improve supervised SPM-based classification might include extension of the feature vector by including attributes of goal-driven interest points and/or attributes of pixel patch based on image pre-segmentation.

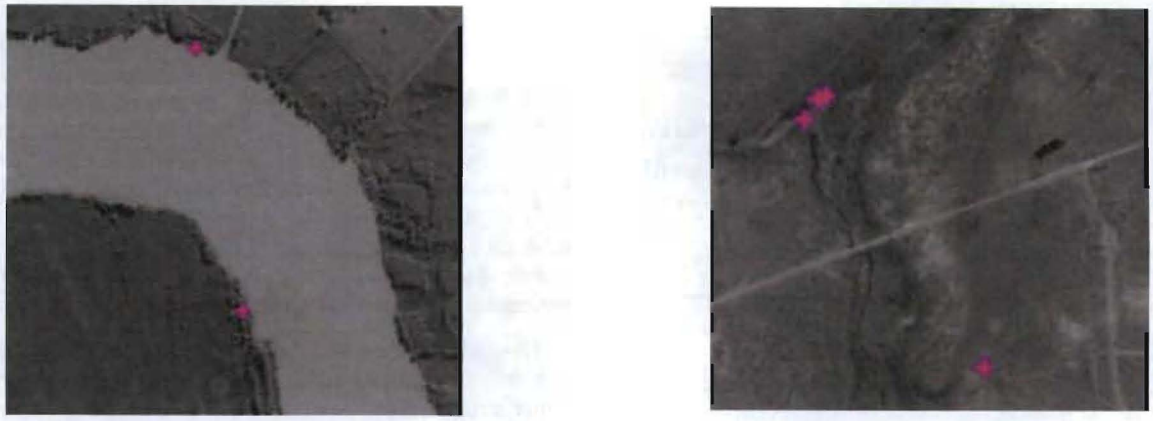
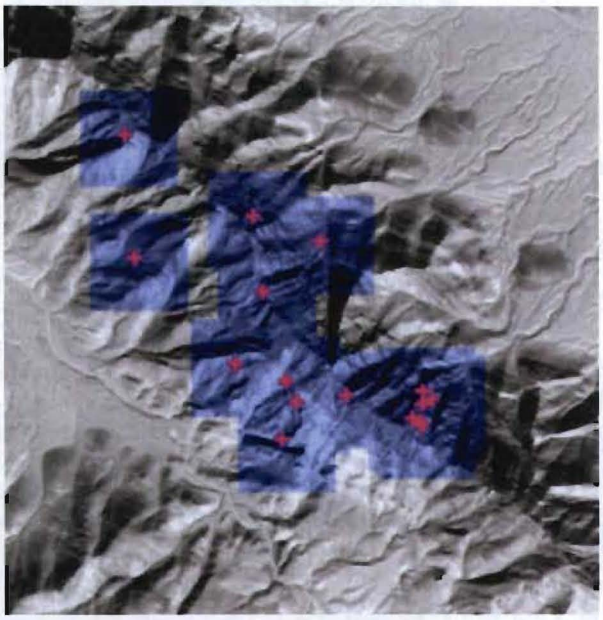
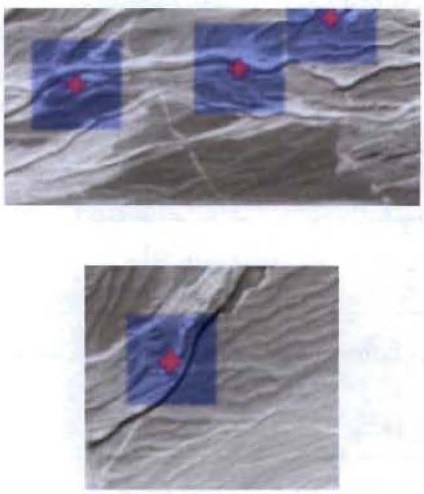


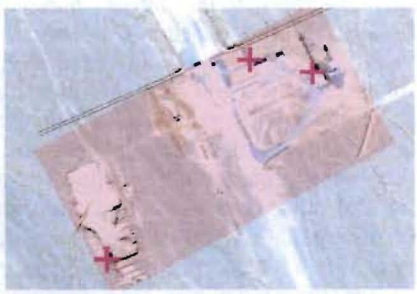
Figure 4. Examples of successful classification of the areas surrounding the detected cotrners as nonman-made areas. Detected corners are shown with magenta "+". No areas are hightligthted because man-made areas were not detected. Image credit: ©Google Earth.



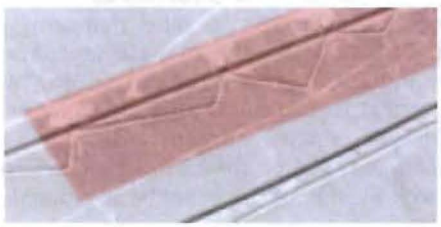
(a) Example of false positive over the mountain landscape.



(b) Example of false positive over the desert-like landscape.



(c) Example of false negative due to misclassification of areas surrounding the detected corners.



(d) Example of false negative due to misdetection of corners.

Figure 5. Examples of false positives and false negatives. Detected corners are shown with magenta "+". False positives are shighlighted with blue color. Areas surrounding the detected corners are misclassified by the spatial pyramid based recognizer. False negative (missed detections) regions are highlighted with red color. False negatives are due to: (1) misclassification of the areas surrounding the detected corners as shon in (c), or (2) misdetection of corners as shown in (d). Image credit: ©Digital Globe.

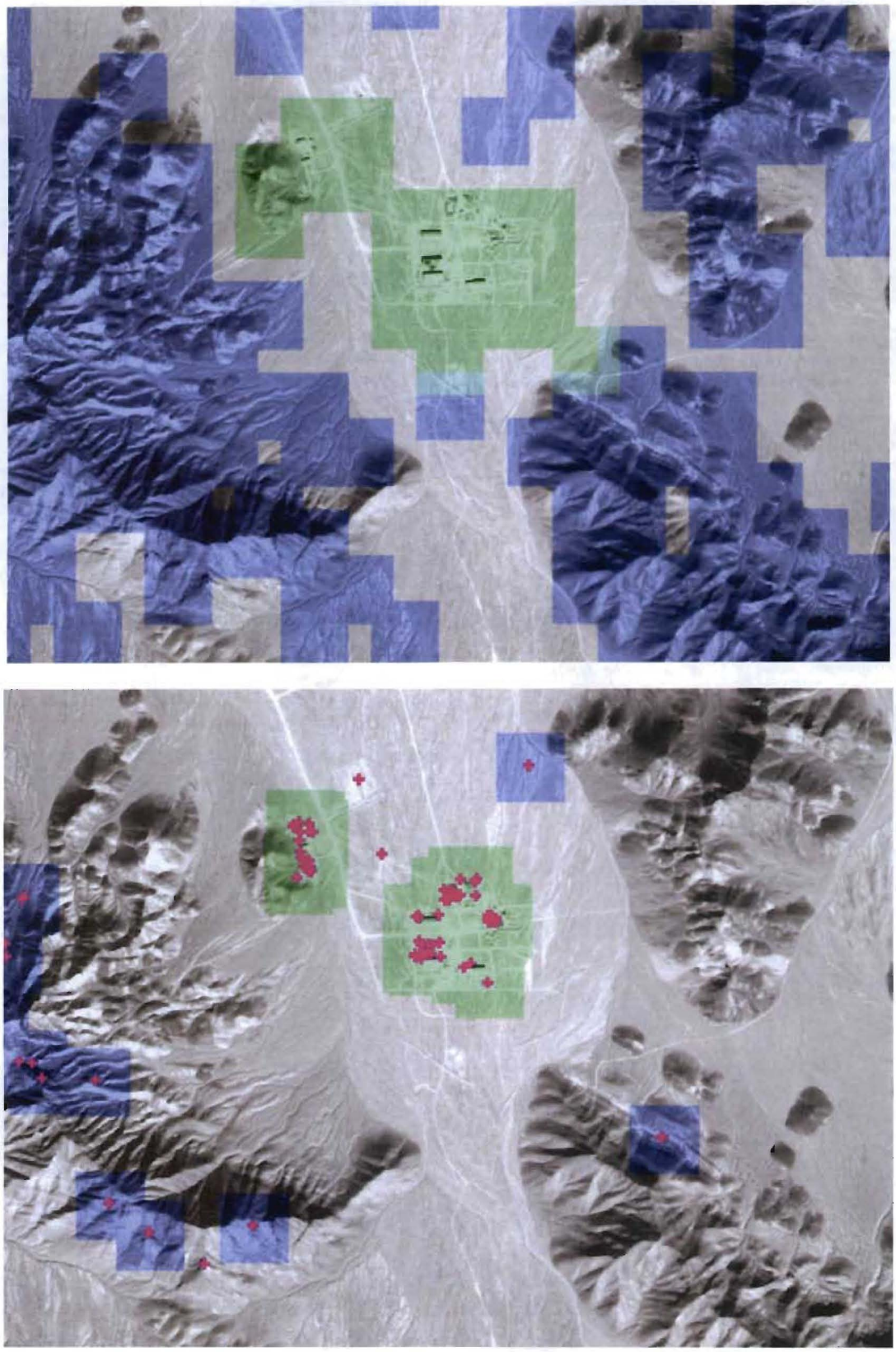


Figure 6. Illustration of how corner-based cueing reduces the number of false positives. 1st row: detection results using SPM. 2nd row: detection result using CC-SPM. Detected corners are shown with magenta "+". True positives are highlighted with green color, false positives are highlighted with blue color, cyan color is used to show the overlap between blocks corresponding to true positives and false positives. Image credit: ©Digital Globe.

REFERENCES

[1] C.-C., Chang, C.-J. Lin, "LIBSVM: a library for support vector machines," Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2001.

[2] G. Csurka, C. Dance, J. Willamowski, L. Fan, C. Bray, "Visual categorization with bags of keypoints," in Proc. Workshop on Statistical Learning in Computer Vision, pp. 1-22, 2004.

[3] C. Harris, M. Stephens, "A combined corner and edge detector," in Proc. 4th Alvey Vision Conference, Manchester, UK, pp. 147-151, 1988.

[4] D. A., Holland, D. S., Boyd, P. Marshall, "Updating topographic mapping in Great Britain using imagery from high-resolution satellite sensors," *ISPRS Journal of Photogrammetry and Remote Sensing*, 60, pp. 212-223, 2006.

[5] R. A. Geerken, "An algorithm to classify and monitor seasonal variations in vegetation phenologies and their inter-annual change," *ISPRS Journal of Photogrammetry and Remote Sensing*, 64, pp. 422-431, 2006.

[6] L. Itti, C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, 40, pp. 1489-1506, 2000.

[7] L. Itti, C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, 2(3), pp. 194-203, 2001.

[8] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in Proc. CVPR, 2006.

[9] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60(2), pp. 91-110, 2004.

[11] K. Mikolajczyk, C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, 60(1), pp. 63-86, 2004.

[12] V. Navalpakkam, L. Itti, L., "Modeling the influence of task on attention," *Vision Research*, 45, pp. 205-231, 2005.

[13] J., Noble, "Finding corners," *Image and Vision Computing*, 6(2), pp. 121-128, 1988.

[14] M. K. Ridd, J. D. Hipple (Eds.), *Remote Sensing of Human Settlements*, American Society for Photogrammetry and Remote Sensing, 2006.

[15] G., Sarp, A. Erener, "Land use detection comparisonfrom satellite images with different classification procedures," in Proc. of the ISPRS Congress, 2008.

[16] Y. Sun, R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, 146, pp. 77-123, 2003.

[17] X. Tong, S. Liu, Q. Weng, "Geometric processing of QuickBird stereo imageries for urban land use mapping: A case study in Shanghai, China," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* , 2(2), pp. 61-66, 2009.

[18] D. M. Tralli, R. G. Blom, V. Zlotnicki, A. Donnellan, D. L. Evans, "Satellite remote sensing of earthquake, volcano, flood, landslide, and coastal inundation hazards," *ISPRS Journal of Photogrammetry and Remote Sensing*, 59, pp. 185-198, 2005.

[19] A. Treisman, G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, 12, pp. 97-136, 1980.

[20] D. Walther, U. Rutishauser, C. Koch, P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, 100, pp. 41-63, 2005.

[21] V. Walter, "Object-based classification of remote sensing data for change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, 58, pp. 225-238, 2004.

[22] J. M. Wolfe, "Guided Search 4.0: Current Progress with a model of visual search," in: W. Gray (Ed.), *Integrated Models of Cognitive Systems*, New York: Oxford, pp. 99-119, 2007.

[23] A. L. Yarbus, "Eye movements during perception of complex objects," in: Riggs, L.A. (Ed.) *Eye movements and Vision*, Plenum Press, New York, chapter VII, pp. 171-196, 1967.