Biological Visual Attention Guided Automatic Image Segmentation with Application in Satellite Imaging

M. I. Sina, A.-M. Cretu and P. Payeur
School of Electrical Engineering and Computer Science, University of Ottawa,
800 King Edward Avenue, Ottawa, Ontario, Canada, K1N6N5
Email: {msina070, acretu, ppayeur}@uottawa.ca

ABSTRACT

Taking inspiration from the significantly superior performance of humans to extract and interpret visual information, the exploitation of biological visual mechanisms can contribute to the improvement of the performance of computational image processing systems. Computational models of visual attention have already been shown to significantly improve the speed of scene understanding by attending only the regions of interest, while distributing the resources where they are required. However, there are only few attention-based computational systems that have been used in practical applications dealing with real data and up to now, none of the computational attention models was demonstrated to work under a wide range of image content, characteristics and scales such as those encountered in satellite imaging. This paper outlines some of the difficulties that the current generation of visual attention-inspired models encounter when dealing with satellite images. It then proposes a novel algorithm for automatic image segmentation and regions of interest search that combines elements of human visual attention with Legendre moments applied on the probability density function of color histograms. The experimental results demonstrate that the proposed approach obtains better results than one of the most evolved current computational attention model proposed in the literature.

Keywords: Biological Visual Attention, Saliency, Bottom-up attention, Top-down attention, Legendre moments, Segmentation, Region of interest, Satellite imaging.

1. INTRODUCTION

The human visual system processes huge volumes of visual data efficiently and accurately. One of the interesting aspects of human vision is its capability to concentrate the interpretation effort only in selected salient regions in a scene, while discarding all the irrelevant information.This behavior has been reproduced in several computational attention models proposed in the literature. While the research on the topic evolved significantly in time, most of the proposed biological visual attention-inspired systems have been tested for a limited number of images in simplistic scenarios. There are only few attention-based computational systems that have been used in practical applications dealing with real data and up to now, none of the computational attention models was demonstrated to work under a wide range of image contents, characteristics and scales. The ability to deal with such data has become a critical requirement to ensure highly automated processing and full use of the growing volume of images in a timely manner.

A particularly challenging application is found in the geospatial software industry, which is in the middle of a revolution, powered by a concomitant digital revolution including phenomena such as Microsoft Virtual Earth, Google Maps, and Google Earth. The ever increasing availability of high resolution digital satellite and aerial images serve to attest the mainstream adoption of the geospatially enabled web and other spatial information production tools for digital maps and geographical information systems. However, the lack of sufficient performance and the limited automation capabilities of current image processing techniques pose critical challenges to the efficient use of the huge amount of data made available from aerial and space based assets. This justifies a strong research interest in the development of new intelligent algorithms capable to interpret large datasets of complex geospatial and satellite images.

This paper evaluates the capability of a current computational attention model to deal with such complex images as satellite images and puts in evidence some difficulties encountered. A novel algorithm for automatic image segmentation and regions of interest search is then proposed. The latter combines elements of visual attention with image moments. Its performance is evaluated and its suitability is demonstrated for satellite images that often contain large visually similar regions, such as rivers, fields, forests, clouds, etc. The paper does not concentrate on the biological aspects of human

Human Vision and Electronic Imaging XVII, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Huib de Ridder, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8291, 82911N • © 2012 SPIE-IS&T
CCC code: 0277-786X/12/$18 • doi: 10.1117/12.911996

vision system, but rather focuses on the technical issues of the implementation and the description of the proposed novel algorithm.

The organization of the paper is as follows: section 2 provides a brief literature review and section 3 examines the technical implementation issues and the evaluation of a computational attention model that draws inspiration from one of the most evolved current solutions proposed in the literature. In section 4, issues related to image moment based approach for region of interest search are described. Section 5 presents the proposed algorithm for image segmentation and region of interest search and its evaluation. Finally, section 6 concludes the paper.

2. LITERATURE REVIEW

While examining human visual attention, psychological studies have shown that there are two major categories of features that drive the deployment of attention: bottom-up features, derived directly from the visual scene, and top-down features detected by cognitive factors such as knowledge, expectation, or current goals. Most computational implementations are based on bottom-up features that can capture attention during free viewing conditions. A measure that has been shown to be particularly relevant is the local image saliency, which corresponds to the degree of conspicuity between that location and its surround. In other words, the responsible feature needs to be sufficiently discriminative with respect to its surroundings in order to guide the deployment of attention.

Building on the Feature Integration Theory proposed by Treisman and Gelade[1], the model devised by Koch and Ullman[2] is seminal in this field since many other systems are actually derivatives of this model. Milanese's implementation[3] of the model of Koch and Ullman makes use of filter operations for feature map computation. One of the most popular bottom-up visual attention systems is the one proposed by Itti[4] and has been later extended to incorporate top-down influences as well as task relevance information[5-7]. The VOCUS system of Frintrop[8], that brings several improvements to Itti's attention model[4], comprises both bottom-up and top-down influences.

Most of the proposed computational solutions have been tested mainly on indoor scenes or for a limited number of images. It is only in the latest years that attention-based computational systems started to be studied in practical applications dealing with real data. Frintrop and Jensfelt[9] use a sparse set of landmarks based on a biologically attention-based feature-selection strategy and active gaze control to achieve simultaneous localization and mapping of a robot circulating in an office environment and in an atrium area. In a similar manner, Siagian and Itti[5,6] use salient features derived from attention together with context information to build a system for mobile robotic applications that can differentiate outdoor scenes from various sites on a campus [5] and for localization of a robot [6]. In Rasolzadeh *et al.* [10], a stereoscopic vision system framework identifies attention-based features that are then utilized for robotic object grasping. Rotenstein *et al.* [11] propose the use of mechanisms of visual attention to be integrated in a smart wheelchair for disabled children application to help visual search tasks.

In geo-imaging applications, satellite images contain a complex array of features (e.g. cities, forests, etc). While the recognition of these features by a human operator can be quite efficient provided a given amount of training, the most advanced computational solutions primarily rely on atmospheric and photogrammetric models. Moreover, most of the computational techniques currently used for image feature extraction and classification are generalizations of algorithms which are neither specifically designed for geospatial applications nor fully automated. As a result, the false positive rate of decision is very high and several features of interest remain undetected. According to industry reports, the current algorithms achieve only about 25-30% accuracy when running on large collections of geospatial images, thereby leaving a substantial opportunity for new approaches to improve upon the current state-of-the-art techniques in terms of performance. In spite of the promising results of biologically inspired models, following an extensive research in the literature, a single paper[12], was found that applies the concepts of visual attention to aerial images but only shows extremely limited experimental results on a single image.

3. IMPLEMENTATION OF A BIOLOGICALLY-INSPIRED VISUAL ATTENTION MODEL AND ITS EVALUATION FOR SATELLITE IMAGES

The computational attention model implemented (in C++ and OpenCV software library[13]) and tested in the context of this paper follows the key steps of the VOCUS system of Frintrop. A full description of the algorithm can be found in Ref. 8. The system works with the three features most commonly used by the current generation of visual attention systems, namely intensity, orientation and color. In an attempt to make the algorithm scale-invariant, VOCUS employs both Gaussian and Laplacian pyramids for different features[14]. Particularly, image pyramids with 5 layers, denoted S_i , i

= {0, 1, 2, 3, 4} (with S_0 being the input image) are used. The layers S_0 and $S_1$are discarded, as in VOCUS, in an attempt to reduce noise and also to reduce the consumption of computational resources.

Initially, feature maps are computed at different scales for each feature separately (the computation can be done in parallel) and each of them is weighted by a uniqueness weight function. Then these maps are summed up and normalized into a so-called conspicuity map. There is one conspicuity map for each of the three features used. Finally, the three conspicuity maps are summed up and normalized in the saliency map.

3.1 Computation of intensity map

The intensity map shows the intensity variations in the input image. Two different types of intensity maps are computed: on-center maps and off-center maps. The on-center map highlights the regions of the image where a bright region is surrounded by a relatively darker region. The higher the difference in intensity, the brighter the region looks in the on-center map. Conversely, the off-center map highlights the regions where a dark region is surrounded by a relatively brighter region. The computation of intensity map is performed for each scale of the constructed Gaussian pyramid, after converting the color image to gray scale image, and with two different radii (3 and 7); therefore there is a total of 12 intensity maps (6 for the on-center map and 6 for the off-center map). This computation is achieved by taking a rectangular region around each pixel of the images. After that, an across-scale addition is performed on both on-center maps and off-center maps by scaling up the smaller images of the pyramid to the size of the biggest image of the same pyramid and then performing a pixel-wise addition. After computing both on- and off-center maps, they are fused together by first applying the uniqueness function and then summing up the two maps which is followed by a normalization procedure. The normalization is done so that all the values fall in the range [0, M], where M is the largest value among the two maps. The uniqueness function will be discussed in further detail in the section 3.4. This fused map is called the conspicuity map of the intensity feature.

3.2 Computation of orientation map

According to biological studies, the human visual mechanisms do not respond equally for all orientations of edges in a scene, but rather for a specific set of orientations. For example, the response is stronger for edges situated at 0, 45, 90 and 135 degrees of orientation. This behavior motivates the use of an edge detector that highlights edges of a scene with these particular orientations. Gabor-like filters were identified as the best fit for this purpose [15, 16]. In the computation of orientation maps, VOCUS follows the procedure devised by H. Greenspan et al. [17]. The procedure begins by computing a Filter-Subtract-Decimate (FSD) pyramid [18], a variant of Burt and Adelson's pyramid [14], of the grayscale converted input image. Each layer of the FSD pyramid is then modulated with complex sinusoids of orientations 0, 45, 90 and 135 degrees. Next, the pyramid layers are low-pass-filtered with a 5-sample Gaussian-like separable filter (1/16, 1/4, 3/8, 1/4, 1/16). In this way, four pyramids are obtained, one for each orientation. An across-scale addition is performed on each pyramid to compute the corresponding orientation map, highlighting the edges of corresponding orientation. Finally, the four maps are fused together by pixel-wise addition followed by a normalization process similar to the computation of intensity map. The resulting map is called the conspicuity map for orientation feature and highlights the edges having the four considered orientations.

3.3 Computation of color map

Color plays an important role in human vision. Biological and psychological experiments proved the existence of dedicated photosensitive cells to identify greenness against redness (and vice versa) or blueness against yellowness (and vice versa). To mimic this behavior, the input RGB color image is first converted into CIE LAB color image, where L measures the luminance of the pixel, A the amount of green and red (smaller values represent green color and larger values represent red) and B the amount of blue and yellow (smaller values represent blue and larger values represent yellow). A Gaussian pyramid is then constructed based on the LAB image, but without considering the L component to reduce the effects of illumination change. Four more pyramids are considered for each of the colors and filled up with values that represent the color distance. Color distance is simply the Euclidean distance in the A-B space. In the pyramid for red the distances are computed with respect to maximum red coordinates: (127, 0). But for blue coordinates: (0, -127) is used since that coordinates represent maximum blue. Note that the pyramids for individual colors are of a single channel but the original LAB image pyramid has three channels. Next, the on-center map of each pyramid is computed to highlight sudden spikes of colors in the input image. This leads to four color maps corresponding to the four colors. Finally, the uniqueness function is applied and the four maps are added together followed by a normalization process similar to other feature maps in order to compute the conspicuity map for the color feature.

3.4 Computation of the bottom-up (BU) saliency map

At this stage, three conspicuity maps are built, one for each feature. The fusion of these maps leads to the saliency map of a scene. Before fusing the maps a uniqueness function is applied on each of the maps which suppresses maps with many peaks but promotes maps with fewer peaks. In order to achieve this, at first the number of local maxima, denoted by *m* is computed in a certain range and a weighting function, W(X), is defined as follows:

$$W(X) = X / \sqrt{m} \tag{1}$$

where X is the map itself. In this function, a smaller value for *m* leads to higher weight and a higher value of m leads to a smaller weight. The range used to compute *m* is [M/2, M] where M is the maximum value present in the entire map. Finally all three maps are summed up pixel-wise and normalized according to the procedure described in section 3.1. The resulted saliency map highlights the regions of the input image that are most likely to attract human attention. The region around the pixel with the highest value within a certain threshold is discovered using a flood fill approach and is named Most Salient Region (MSR). The MSR represents the focus of attention. Following a similar approach, one can compute the next MSR or top k-MSRs in order to identify the order in which objects in the scene are attended.

3.5 Learning and searching mechanism

The bottom-up saliency map shows the level of interest of regions in a scene to an unbiased observer. On the other hand when the observer is looking for something specific, for example, a red car, then to that observer red objects and car shaped objects will be more attracting, even though such objects do not belong to top salient regions of the scene. This type of behavior is modeled using top-down saliency where objects of interests are deliberately given more weight so that they become more salient.

VOCUS's learning mechanism is based upon the maps computed in the course of computing the saliency map. Overall, thirteen maps, namely two intensity maps, four orientation maps, four color maps, and three conspicuity maps, are used in the learning and recognition phases. At first, a training image is provided along with a selected region of interest (ROI). Each ROI is represented with a rectangle identified by the image coordinates that surround an object of interest. The algorithm attempts to learn the object in the given ROI. There are a few restrictions on how to choose the training images and the ROI[8]. In order to learn the object in the ROI, VOCUS first computes all thirteen maps described above and computes thirteen weights w_i , $1 \leq \text{i} \leq 13$ corresponding to the maps X_i , $1 \leq \text{i} \leq 13$ as follows:

$$w_i = \frac{m_{i_{MSR}}}{m_{i(image-MSR)}} \tag{2}$$

where $m_{i_{MSR}}$ and $m_{i_{(image-MSR)}}$ represent the mean intensity of the MSR, and the mean intensity of the rest of the images excluding the MSR respectively, in the map X_i. In the search phase, it computes again all thirteen maps Y_i from the test image and computes the top-down saliency as following:

$$I_{TD} = \sum_{w_i > 1} w_i Y_i - \sum_{w_i < 1} \frac{1}{w_i} Y_i \tag{3}$$

Next, the global saliency is computed by taking a convex combination of both bottom-up and top-down saliency maps. In the exploration mode, more weight is given to the bottom-up component, and in the search mode top-down component takes more weight. The top-down map highlights the target objects in the test image. The important thing to note here is that the system not only considers the object in the ROI, but also the background of the object. This forces the test image to have similar background over the search phase in order to generate a successful search.

3.6 Evaluation of VOCUS performance on satellite images

A set of satellite images obtained from Mapquest[19] was used to evaluate the performance of the VOCUS computational attention model. Fig. 1 shows an example of two input images used for experimentation along with their corresponding bottom-up saliency maps (with brighter points denoting more salient regions) computed by VOCUS.

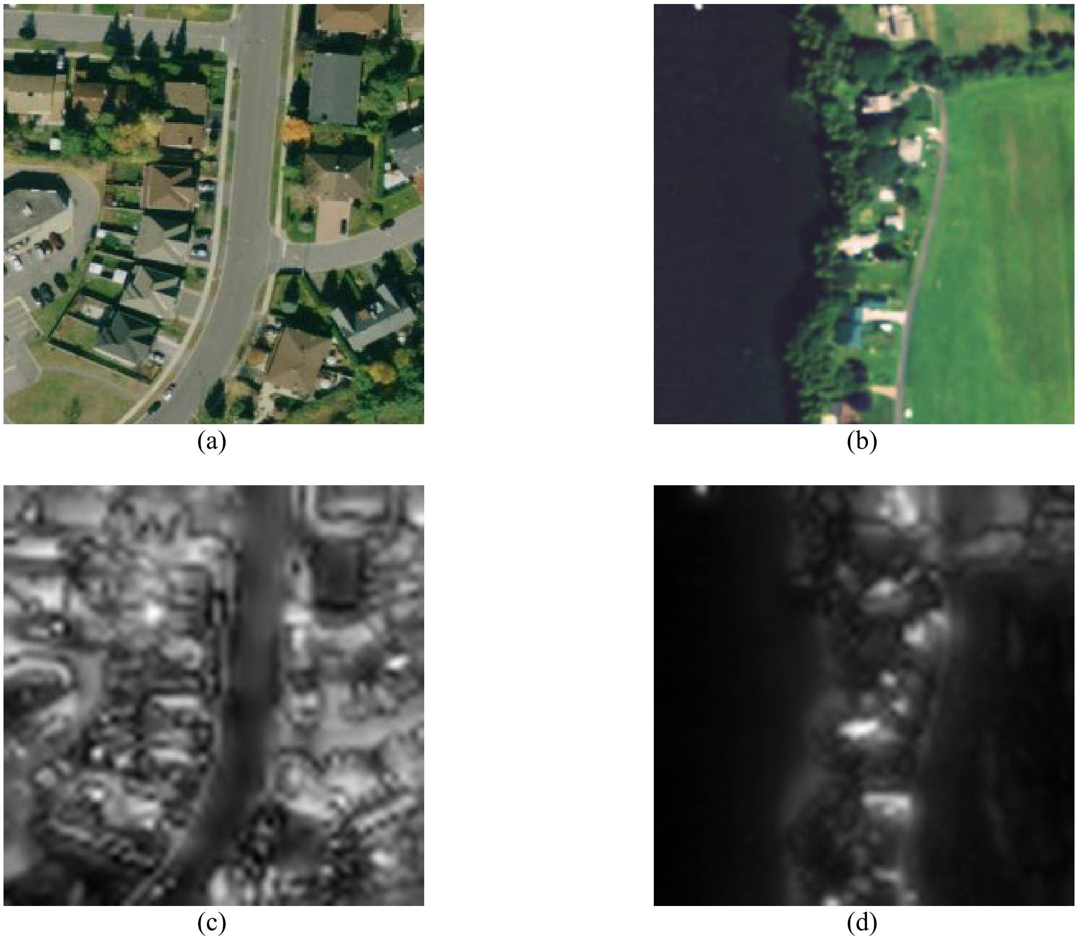

Figure 1. a) and b) input satellite images, c) and d) corresponding saliency maps as computed by VOCUS. Homogeneous regions have either low saliency or uniform saliency.

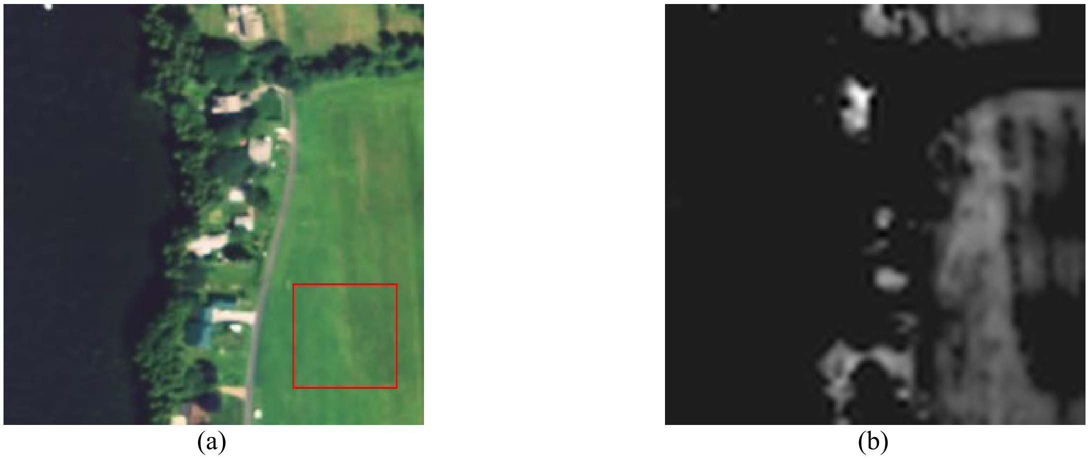

Figure 2. a) Selection of a ROI (depicted by a red rectangle) for learning purpose, and b) self-test result when the same input image is used as the test image for searching. The system could identify most of the green fields and discard the dark blue lake.

The system generates fairly plausible saliency maps. It can be observed that regions different from their surroundings are highlighted and, at the same time, homogeneous regions have a low and uniform saliency. In Fig. 2(a), a ROI of Fig. 1(b) is then selected and used to generate the top-down saliency map. When the same image is used as the test image, the system can identify most of the green fields and discard the dark blue lake, as shown in Fig. 2(b). It can be noticed that some parts of the trees on the bank of the lake are highlighted, whereas other parts remain dark. Next, the image of Fig. 1(a) is used as a test image and the search result is shown in Fig. 3. In this case, since the contents of the training and the test images are significantly different, the system's behavior is unpredictable. One would expect that only the green regions would be highlighted. But since the image in Fig. 3(a) has different background and object contents than the training image, the learning/search mechanism of VOCUS erroneously highlights parts of the road. It is clear that a new approach is required to improve the performance of the computational visual attention models for such satellite images.

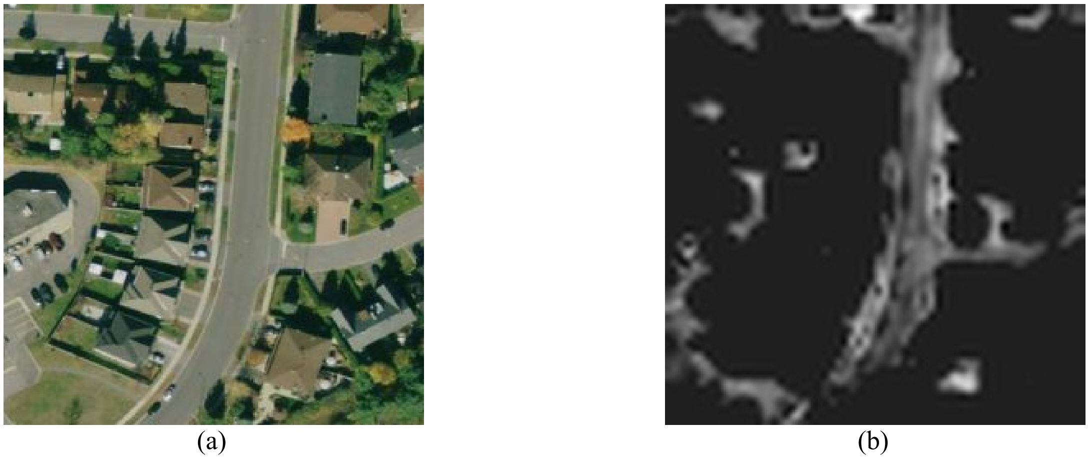

Figure 3. a) Image used as test image for the search, and b) top-down map obtained when combining the training on the image shown in Fig. 2(a) with the new test image.

4. EXPERIMENTATION WITH MOMENT INVARIANTS-BASED SALIENCY SEARCH

Because the original VOCUS system faces challenges in the search mode when the background is different than the one in the training mode, an attempt is made to make the search independent from the background with the use of moment invariants. Image moments of various types and their invariants have been researched and used in practice for a long time. The field is vast and hence this article focuses more on applications of moments rather than on theoretical background. Moment invariants are a set of mathematical formulae which remain constant for a particular pattern even if the pattern undergoes some transformations to which it is invariant. The transformation can be any basic transformations (translation, rotation and uniform scaling) or a combination of basic transformations, uniform contrast change or affine transformation depending on the invariants being used. They therefore appear as a promising research direction to better describe the properties of a ROI. Experimental tests were run with a set of classical moment invariants, namely, Complex moments[20], Hu invariants[21] and Legendre moments[20, 22]. Hu invariants are invariant to rotation, translation and uniform scaling. The Legendre moment invariants used are uniform contrast invariant but they are not rotation invariant by nature; they can be made affine invariant[21]. On the other hand, the complex moment invariants used are rotation/translation/scaling invariant but are not invariant to uniform contrast change.

These different moment invariants are computed from the saliency map of the given ROI of the training image. Then similar moment invariants are searched for in the saliency map of the test image in an attempt to make the search independent to background. Fig. 4 shows the results of the application of various moment invariants for the same image and same ROI as in Fig. 2(a). Unfortunately none of them are even close to what the original ROI depicts. A close look into the ROI reveals a small area where the grass is a bit yellowish. This part of the ROI is the MSR of the ROI.

However, when the entire image is considered, the yellowish grass that generated saliency almost vanishes due to the normalizing functions used by VOCUS. Therefore further refinement is required to reliably identify ROIs in such complex images.

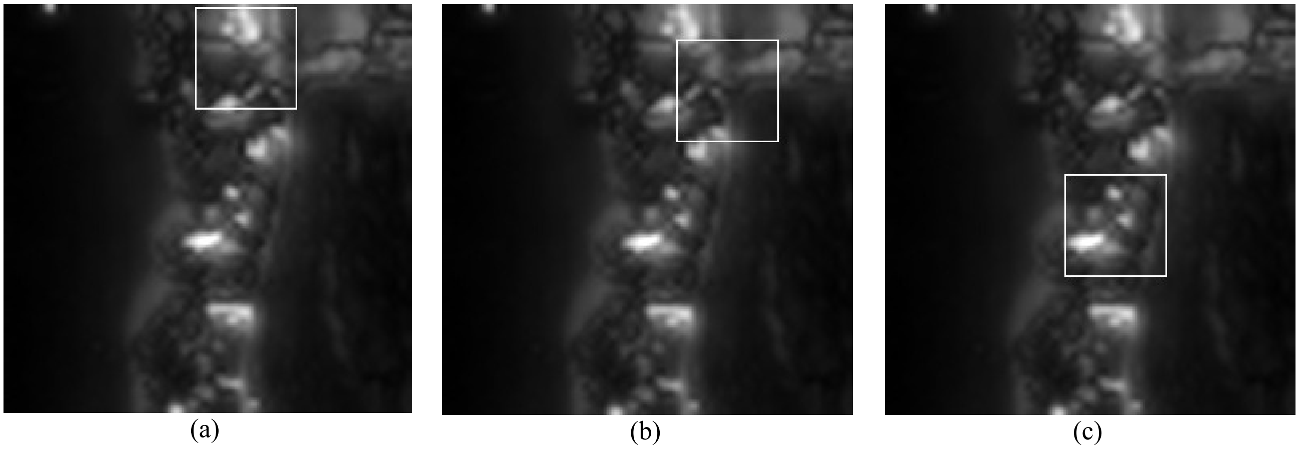

Figure 4. Using the same ROI from the training image of Fig. 2(a) and the same test image (self-test), white rectangles show the best match found by (a) seven Hu invariants, (b) five Complex moment invariants up to the 3rd order, and (c) three Legendre moment invariants up to the 3rd order. None of the rectangles correspond to what really is in the original ROI.

5. PROPOSED ALGORITHM FOR SALIENCY AND MOMENTS BASED SEGMENTATION AND ROI SEARCH

In the proposed approach, the saliency map is viewed from a different perspective. While the MSRs show the regions that attract attention, the least salient regions (LSR) direct the observation towards regions that are homogeneous. Their homogeneity makes them less salient and therefore they are ignored in VOCUS[8]. The LSRs revealed to be good locations to be learned particularly due to their homogeneity. If a model is learned from an LSR and used later on to search for similar model parameters in the image, similar regions would be grouped together. This leads to the segmentation of an image based on visual similarity.

5.1 The proposed algorithm

The novel algorithm proposed in this paper is based on the bottom-up saliency map built by VOCUS and Legendre moments applied on the probability density function of histograms. Any saliency detection system could be used instead of VOCUS. After the computation of the bottom-up saliency map, the proposed solution finds the LSR of the map. If the size of the LSR is not sufficiently large, it is discarded and the next LSR is found. A sufficiently large region is required for the learning model to work. The proposed learning model is based on the RGB color histograms of the region in the original image corresponding to LSR of the saliency map. The histograms are subsequently converted into a probability density function (PDF). From the PDF, *f(x)*, Legendre moments are computed up to a certain order *n* using the formula:

$$L(n) = \int_{-1}^{1} P_n(x) * f(x) dx \tag{4}$$

where $P_n(x)$ is the *n*th-order Legendre Polynomial defined as:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left(x^2 - 1\right)^n \tag{5}$$

The use of Legendre moments is advantageous for several reasons. First, they are computed using Legendre polynomials which can be performed efficiently using a recurrence relation by pre-computing the polynomial values in a table. Second, Legendre polynomials are orthogonal on [-1, +1] which is deemed to have strong recognition power. Moreover, Legendre polynomials are numerically stable. The minimum size of the LSR and the number of orders up to which the Legendre moments of the PDF are computed are parameters to be tuned in the proposed approach. All the parameters and their effects are discussed in the next section. At this stage, the entire input image is searched for similar Legendre moment values up to that order. For this task a sliding window of a particular size is considered and for each position of the window the Legendre moments of the PDF of color histograms of that window are computed. If the values are similar to those of the LSR then the center pixel of the window is marked to be similar to the LSR. The measure of similarity is based on Euclidean distance considering that the moment values form a vector in *n*-dimensional space. If the distance is less than a threshold, it is considered to be similar. The whole procedure is repeated for the rest of the unmarked region of the image. Gradually, the marked regions grow and unmarked regions shrink. In those cases where there is no region similar to the current LSR, the latter is considered a segment by itself. It is also possible to have small regions scattered around the image that may or may not be similar to other parts of the image. They remain unmarked and are called orphan regions. The algorithm does not need any seed point or any user intervention, since it selects the best region to learn a model from by itself. After setting the parameters, it operates in a fully automated manner and reports on the segments of the image. Morphological operations can be applied as a post-processing step to fill up small holes. The pseudo code of the proposed algorithm is given below:

```
BIO-SEGMENT (I)
   1. Compute saliency map M of image I
   2. Find the LSR of M and compute the area S of the LSR
   3. If S >= minimum required LSR area:
   4. Then Ref_Moments = compute histograms of the region corresponding to LSR
      from I and subsequently Legendre moments up to order n for all color
      channels
   5. Wnd_Moments = slide a window of a given size W over I and compute Legendre
      moments as in step 4
   6. D = compute Euclidean distance between Ref_Moments and Wnd_Moments
   7. If D < threshold (T):
   8. Then add the center pixel of the current window to the current segment and
      discard the pixel from I
   Repeat steps 2 – 8 until no more LSR can be found.
```

5.2 The choice of parameters

There are five parameters to be set in the proposed method, namely the minimum size of the LSR (S), the number of bins for histograms (B), the maximum order used for Legendre moments (n), the similarity threshold (T), and the size of the sliding window (W). The minimum size of the LSR is crucial. If it is too large, the system may not find any suitable region, resulting in an increased number of orphan regions. On the other hand, too small regions may not be good enough for learning. The number of bins used in histogram can be used to tune robustness; a higher value can capture too much detail about the region, leading to a higher number of segments and an increased computation time. A smaller value may increase robustness, but a too small value may falsely match with visually different regions and lead to a smaller number of segments. The maximum order of moments used has similar effects to the choice of the number of bins. A similarity threshold is also used to tune robustness/rigidness. Too large values may match with visually different regions and too small values may force the system to match only regions looking exactly the same. The most important parameter is the size of the sliding window. It determines the fineness of the segmentation. If it is too large then near region boundaries include a larger portion of other regions and the detection fails. If too small, it does not have enough pixels to compute a representative histogram. Among all these restrictions, it is still possible to find a proper set of parameters that work over a given class of images. Parameters values used during the experiments, estimated by trial-and-error are: S = 441, B = 128, N = 5, T = 0.4, W = 11x11.

5.3 Results

Fig. 5 shows three input images (two of them from Fig. 1) and their corresponding segmented image. Different shades of red and green show visually similar regions. Black regions are the orphan regions. Fig. 5(d) shows that the empty field as well as the lake entirely share the same red shades, respectively. The bushes are also properly segmented although different shades of red are visible in some places. This is due to the effect of the size of sliding window. In Fig. 5(f) green field and bushes are well segmented. The trees are also segmented well with few holes and noise which may be removed by application of morphological operations. The algorithm provides fairly good results for the segmentation of very complex images as the one illustrated in Fig. 5(b).

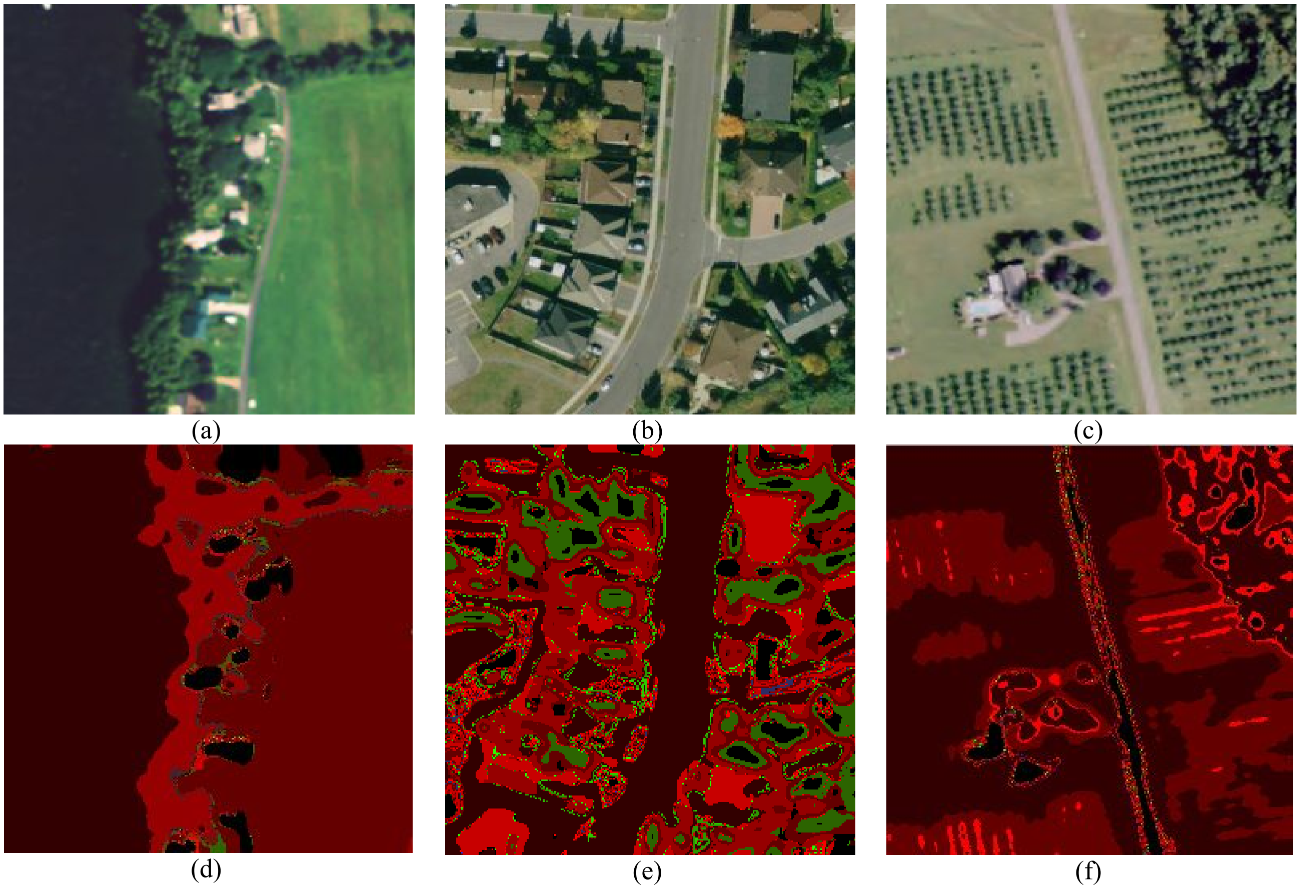

Figure 5. a) - c) input images, and d) - f) corresponding segmentation where shades of red and green depict visually similar regions.

The proposed algorithm also allows searching for a given region from a training image in another test image without computing the saliency map of the training image. In this case, the given ROI is learned the same way by computing the Legendre moments of the color histogram's PDF from a training image instead of automatically selecting it from LSR and the entire test image is search for similar Legendre moments as described in section 5.1. It therefore allows for the identification of similar regions in a test image. Some experimental results are presented in Fig. 6. Fig. 6(a) is the training image with the ROI depicted by a red rectangle, that is the same image as in Fig. 2(a) to enable comparison with VOCUS. The images in Fig. 6(b) and Fig. 6(c) are used as test images. Binary images on the bottom row of Fig. 6 show the results of a search for the learned ROI, with white regions indicating visual similarity. It can be noticed that the proposed segmentation is accurate and obtains significantly better results when compared to the ones obtained by the VOCUS system. One can notice a better identification of the green fields in Fig. 6(d) and Fig. 6(e) in spite of the higher complexity when compared the ones obtained in Fig. 2(b) and in Fig. 3(b) respectively, where the VOCUS learning mechanisms are used.

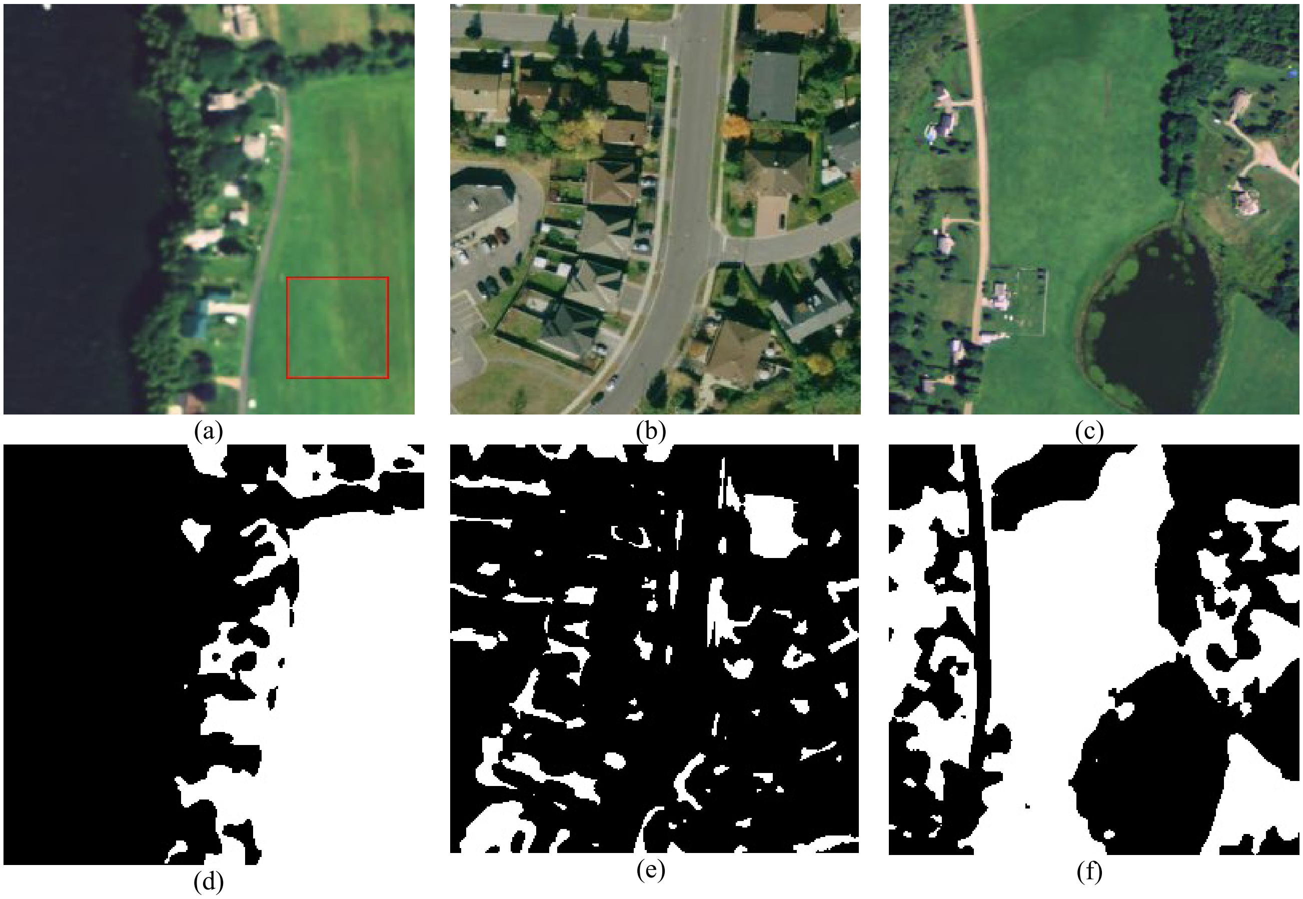

Figure 6. (a) training image with ROI in red rectangle, (d) result of self test where test image is the training image itself. White regions indicate a successful match; (b) and (c) different test images, and (e) and (f) results of the search for similar ROIs.

6. CONCLUSION

This paper described the implementation and evaluation of a computational vision system inspired by the VOCUS system of Frintrop[8] in the context of satellite images. The particular weakness identified is related to the learning mechanism that simulates the top-down component of the visual attention which results in unsatisfactory performance when the training and test images are significantly different. It was also shown that a search for regions of interest strictly relying on moment invariants is not satisfactory either, since the objects' saliency map depends on the other objects present in the scene. A novel method which originally combines the VOCUS bottom-up attention component with Legendre moments of histograms of probability density functions is therefore proposed for automatic segmentation and search over satellite images. Experimental evaluation demonstrated that the proposed approach result in better segmentation and higher ROI search performance when applied on satellite images.

REFERENCES

[1] Treisman A. M. and Gelade G., "A feature integration theory of attention," Cognitive Psychology, 12, 97 – 136 (1980).

[2] Koch C. and Ullman S., “Shifts in selective visual attention: towards the underlying neural circuitry,” Human Neurobiology, 4(4), 219 – 227 (1985).

[3] Milanese R., “Detecting salient regions in an image: from biological evidence to computer implementation,” PhD Thesis, University of Geneva, Switzerland (1993).

[4] Itti L., Koch C., and Niebur E., “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. PAMI, 20(11), 1254 – 1259 (1998).

[5] Siagian C., and Itti L., “Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention”, IEEE Trans. Pattern Analysis and Machine Intelligence, 29 (2), 300- 312, (2007).

[6] Siagian C., and Itti L., “Biologically Inspired Mobile Robot Vision Localization”, IEEE Trans. Robotics, 25(4), 861-873 (2009).

[7] Navalpakkam V., and Itti L., “Modeling the influence of task on attention”, Vision Research, 45, 205–231 (2005).

[8] Frintrop S., “VOCUS: A visual attention for object detection and goal-directed search,” PhD Thesis, University of Bonn, Germany (2006).

[9] Frintrop S., and Jensfelt P., “Attentional Landmarks and Active Gaze Control for Visual SLAM”, IEEE Trans. Robotics, 24 (5), 1054-1065 (2008).

[10] Rasolzadeh B., Björkman M., Huebner K. and Kragic D., “An Active Vision System for Detecting, Fixating and Manipulating Objects in the Real World”, Int. Journal of Robotics Research, 29(2-3), 133-154 (2010).

[11] Rotenstein A. M., Andreopoulos A., Fazl E., Jacob D., Robinson M., Shubina K., Zhu Y., and Tsotsos J. K., “Towards the Dream of an Intelligent, Visually-Guided Wheelchair”, Proc. 2nd Intl. Conf. Technology and Aging, Toronto, Canada (2007).

[12] Xu S., Fang T., Huo H., and Li D., “A novel method of aerial image classification based on attention-based local descriptors”, Int. Conf. Mining Science & Technology, Procedia Earth and Planetary Science, 1, 1133–1139, (2009).

[13] Bradski G., “The OpenCV Library”, Dr. Dobb’s Journal of Software Tools (2000).

[14] Burt P. J., and Adelson E. A., “The Laplacian pyramid as a compact image code,” IEEE Trans. Communications, 31, 532 – 540 (1983).

[15] Daugman J. G., “Uncertainty relations for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters,” Journal of Optical Society of America, 2, 1160 – 1169 (1985).

[16] Kruizinga P., and Petkov N., “Non-linear operator for oriented texture,” IEEE Transactions on Image Processing, 8(10), 1395 – 1407 (1999).

[17] Greenspan H., Belongie S., Goodman R., Perona P. Rakshit S., and Anderson C. H., “Overcomplete steerable pyramid filters and rotation invariance,” Proc. IEEE CVPR, 222 – 228 (1994).

[18] Anderson C. H., “A filter-subtract-decimate hierarchical pyramid signal analyzing and synthesizing technique,” United States Patent 4,718,104 (1987).

[19] Available online. www.mapquest.com

[20] Flusser J., Suk T., and Zitova B., “Moments and moment invariants in pattern recognition”, John Weily & Sons Ltd, West Sussex, United Kingdom PO19 8SQ (2009).

[21] Hu M. –K., “Visual pattern recognition by moment invariants”, IRE Transactions on Information Theory, 8(2), 179 – 187 (1962).

[22] Zhang H., Shu H., Coatrieux G., Zhu J., Wu Q. M. J, Zhang Y., Zhu H., and Luo L., “Affine Legendre moment invariants for image watermarking robust to geometric distortion,” IEEE Trans. Image Processing, 20(8), 2189 – 2199 (2011).