A COMPUTATIONAL METHOD TO EMULATE BOTTOM-UP ATTENTION TO REMOTE SENSING IMAGES

X. Chen [a*], H. Huo[a], F. Tao[a], D. Li[b], Z. Li[a]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, No.800 Dongchuan Road, Shanghai, China - weedcx@gmail.com, (huohong, tfang, lizhiqiangmy)@ sjtu.edu.cn

[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Hubei Wuhan, China - dli@wtusm.edu.cn

Youth Forum

KEY WORDS: Image interpretation, Modelling, Image understanding, Texture Analysis, Process modelling

ABSTRACT:

In this paper, we propose a computational model which is capable of emulating the expert's bottom-up attention to remote sensing images. The bottom-up visual attention is a relatively primary step in neuroscience, and it can perfectly perform recognition if combined with context. Thus, efficient and fast bottom-up model is in need to give convenience to process context in following step. Our computational model well conforms to these conditions. The model cut down uncertain complication of visual attention by introduction of textons based on neurobiology and information entropy. First, our model processes images extremely rapidly while achieves relatively high hit rates. Second, our model provides rarity hierarchy by converting unique or rare visual attributes to number rare attribute for future processing. Third, our results provide size, shape and location information for the future context attention computation.

1. INTRODUCTION

Nowadays, vast amounts of remote sensing data achieved by a great deal of sensory receptors, satellites and other instruments are ready to be processed in time, but the processing ability still lies behind. A novel and promising solution is to analyze remote sensing data automatically by emulating the experienced interpreters' psychological procedures(Lloyd 2002). To meet such challenge, the first and the key step is to get the size, shape and the location of latent attentive regions efficiently during simulating, which relates to "visual attention", a term in psychology.

Visual attention has not been thoroughly understood in neurobiology and psychology so far, but it is still get a lot of attention for its grand potential (Fabrikant 2005). Visual attention enables people to select most relevant information to ongoing action (Chun & Wolfe, 2001), and it will be helpful to interpret remote sensing images by focusing on the most informative places. Though there are not only bottom-up model like that of Itti (Itti 1998) but also models based on task-specific attentional bias like Tsotsos's (Tsotsos 1995), it is not clear how factors concerning with tasks can be formally predicted or be incorporated into the mathematical model(Davies 2006). According to recent work on visual attention, the perceptual saliency critically depends on the surrounding context (Itti 2001), and the bottom-up model provides the crude data for future processing. Thus, the bottom-up implementation is the practical way and an important step to emulate interpretation. Itti and Koch depend on neurobiology to construct a framework for understanding of visual attention, such as "saliency map" to code stimulus conspicuity, "inhibition of return" to prevent from being attended again. Some psychologists have already attempted to employ Itti and Koch's model to update aerial photogrammetry (Davies 2006). They prove that the distribution of visual attention is measured by the use of a relatively simple, low-cost method, instead of full eye tracking. Through the contrast among the prediction of model, experts' and participants' performance, the model's results are more similar to experts than novices. The reason is that the low-level visual patterns guide experts' attention, which means it is the important or special data that attract experts instead of the content. Some researchers employ Itti's model to empirically assess the effectiveness of dynamic displays for learning, knowledge discovery, and knowledge construction based on the relationship of perceptual salience and thematic relevance in static and animated displays (Fabrikant 2005). However, the model does not optimally predict people's visual attention (Davies 2006) and can not be considered a typical design solution (Fabrikant 2005). Even to relatively simple natural scene images, the model's hit rate is not larger than 0.5076 and false alarm rate is between 0.1433 and 0.2931(Hou 2007). Furthermore, Itti's model also falls to depict interesting areas accurately.

In light of foregoing analysis, this study is intended for detecting the size, shape and location of latently salient texture regions comprising a specific texton according to rarity based on neurobiology and information theory. We extract textons based on the grey level information at lowest cost; use the entropy measure to categorize the rarity hierarchy, which is one of key factors in visual attention.

[*] Corresponding author: Tao Fang. E-mail: tfang@sjtu.edu.cn; phone:+86-021-34204758

The paper is organized as follows. The study site and data is described in Section 2. The former methods are presented briefly and compared to the new one in Section 3. Result and discussion are shown in Section 4. Final conclusions are given in Section 5.

2. STUDY SITE AND DATA

One of our study sites are rural places located in the Xianghua Town and Chenjia Town of Chongming County in Shanghai, China mainland. Chongming County is adjacent to the East China Sea and situates at the mouth of the Changjiang (Yangtze) River. The area is flat, fertile land and characterized by complicated spatial patterns with architecture regions, rivers, pools, field, beach, sea and so on. The aerial photos were taken in December 2006; their resolution is 0.25m/pixel.

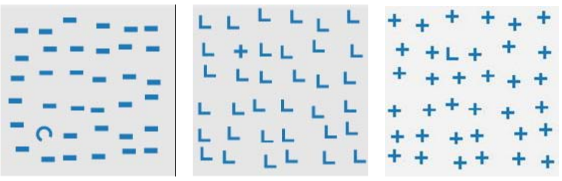
The other imagery tested in this work is taken on 22 February 2003 over Mount Wellington near Hobart Tasmania by IKONOS. It is 4-meter 8-bit multispectral (red, blue, green, near infrared) imagery. The imagery is separated into sub-images with 1280*1024 pixels to analyze individually in order to observe for observers in the most convenient way.

3. METHOD

Davies's experiments involve a semi-focused visual search or change detection task, where participants are looking for anything unusual or different from anticipations (Davies 2006). We follow their key principle, unusual things are salient, and detect the unusual or different regions made up by the same texton. This section is divided in three parts: First, we introduce the psychological knowledge to derive our model's fundamental. Second we describe the metrics to estimate the rarity. Third, we review the classic Itti and Koch's model and our former model based on it. Finally, we introduce the new model.

3.1 The psychological base of our model

Our model excludes colour and direction information from consideration. On the one hand, the Superior Colliculus (SC) straight correlates to the eye's motion and one of Superior Colliculus's tasks is to direct eyes to the arresting areas. The afferent pathway to spatial and motion have no ability of distinguishing colour and direction (M. MANCAS 2007). On the other hand, only conspicuous stimuli pop out and form the salient regions in human's pre-attention stage. Itti refers that the salient stimulus should have unique or rare visual attributes over the entire visual scene (Laurent Itti 2005).In intuition, it is also the salient parts that attract the attention of human observers in a given image and some researchers believe that the important components are unusual or rare(MANCAS 2007; Escalera 2007). Under such assumption, the results have little to do with extracted features, and only concern with the minority. The characteristic nicely fit with the performance of Human Vision System (HVS). A recognized example is shown in Figure 1 (Hou 2007) which comes from psychological patterns. We can observe easily the rare stimuli attract human attention without any psychological background. We develop our work based on the above two neurobiology rules. First, we turn the images into gray level, and then take histogram information into consideration and disregard the orientation information at the same time.



(a) (b) (c)

Figure 1. Simple and clear saliency discrimination. The most distinct atom is also the scarce. The curve is the rarest object in (a); intersection in (b); inverse intersection in (c). In these psychological experiments, the corresponding salient objects are as the same as the rare. A lot of psychological experiments demonstrate that there is a strong link between rarity and saliency.

3.2 Quantification of approximate saliency by rarity based on textons

The distribution of visual attention to aerial photogrammetry varied greatly among experts and among novices (Davies 2006). The individual differences remain obstacles of the traditional psychological measure. An alternative way to detect the salient regions is to count approximate saliency estimation defined by Shannon entropy (Sergio 2007). The self-information contains an amount of information known to describe the amount of surprise of a message inside its message set.

Let T be the texton dictionary and $\mathrm{T}=\{t_i, i\in I\}$, t_i be $i-th$ texton, n_i be the number of t_i, N be the total number of textons, and p_i be the proportion of t_i in all the textons. Then $p_i = n_i / N$, and the distribution of textons is

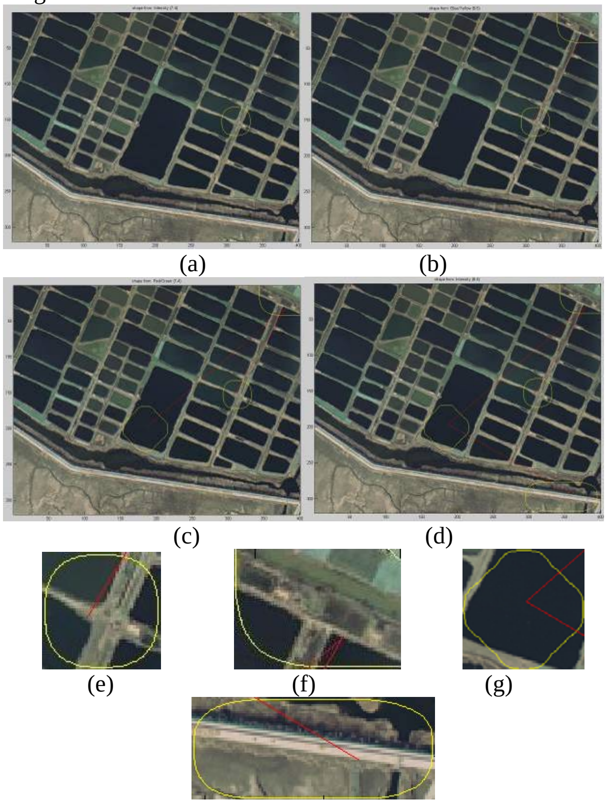
$$[\mathrm{T}, p_i] = \begin{pmatrix} t_1, t_2, \cdots, t_n \\ p_1, p_2, \cdots, p_n \end{pmatrix} \quad (1).$$

The self-information of texton is defined as $I(t_i) = -\log(p_i)$, and it describes the amount of surprise of a message inside the message set. In Itti's surprise research results, the rare messages are surprising, so they draw people's attention. The rarest areas comprise the same class of texton which correspond to the smallest amount. By information theory, the least amount often corresponds to the maxima of the self-information. We consider regions comprising textons with maximum function value is the most salient.

3.3 Classical Itti model and its modulation based on texture

Computational models of visual attention take some feature representation as input, and return a location on which attention should be focus. In Itti and Koch's model, the image is filtered in low-level visual feature channels at multiple spatial scales, for colour, intensity, and orientation. The colour and orientation channels have several sub-channels, each of which generates nine-scale pyramidal representation of filter outputs with Gaussian smoothing between scales. Then the model performs centre-surround operations between filter output maps at different scales within each sub-channel. The results are feature maps, and they are combined into three "conspicuity maps" with normalization operator through across-scale addition. The three “conspicuity maps” are summed into the final input to the saliency map and the maximum of the saliency map denotes the most salient image location. At last, the model deploys “inhibition of return” to subsequently bias the attention to salient places (Siagian 2007).

We generate the attentive locations based on Itti’s theory. The MATLAB implementation of this method can be downloaded from http://www.saliencytoolbox.net. All the images are down-sampled to the maximum area for Itti’s method because the larger images will generate better results in our experiences. As figure 2[†] shows, the vertex of the red line is the centre of attentive location, and the yellow occluded curve point out the range of attention.


(a) (b)

(c) (d)

(e) (f) (g)

(h)

Figure 2. Demonstration of the Itti and Koch’s model detecting the most four salient regions with an input image from Chongming aerial data set. The yellow close curves and the red lines are drawn by their tool box automatically. The yellow curves describe the attention range. From (a) to (d), the number of yellow curves add one every time. The following four images (e), (f), (g) and (h) are the attended original regions. Nevertheless, the attended locations seem scarce with distinct geographical meaning.

Although locations in the Itti’s saliency maps are more likely to be fixed upon by human subjects than random, the computation results seem of less meaning. A natural and well founded measure to detect salient texture blobs is to use the blobs as a substitute for the points; extract texture feature every 3*3 window; construct nine-scale Gaussian pyramids; implement Center-Surround, across-scale combination and normalization; combine these conspicuity maps and grade the ground objects; neglect the “Inhibition of return” mechanism (Chen 2007). An example is shown in figure 3. One of the shortcomings of this method is great computation consumption, another is it only detects several intermittent interesting regions.

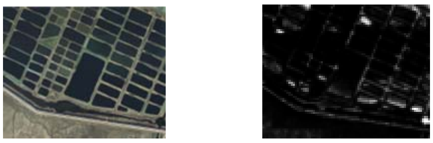

(a) original image (b) saliency map

Figure 3. Demonstration of the modulated model. The attentive areas are light regions. Although the results are of some meaning, the saliency ranks of most regions are still unknown.

3.4 Clustering textons

Textons, called as texture primitive, texture element and so on, describe the perceptually discriminating micro-structures in textured objects. The textons not only show excellent value in spatial arrangement characters, but also in vision characters representation. Textons are considered as the fundamental units of pre-attentive human visual perception even in natural images, as well as a discrete set which is referred to as the vocabulary of local characteristic features of objects (Zhu 2005). Furthermore, the pre-attentive vision is sensitive to some basic image features and is conjectured a pre-attentive stage to detect some atom--textons (Julesz 1981). These clues inspire us to use textons to emulate attention in our model.

The early texton studies were limited by focus on artificial texture patterns instead of nature images. Later, the dictionary of textons extends to include natural images and some textureless images, and the textons show promising potential to most images. There are two models to extract textons(Zhu 2005). One is the sparse coding with over-complete dictionary based on generative modeling; the other is the K-mean clustering based on discriminative modeling. The traditional K-mean clustering involves a set of filters, and at least the majority of the filters correlate orientation (Leung 1999). For invariant rotating and scaling, the traditional model also needs great computation consumption. The traditional method also set a lot of filters to achieve invariant rotating and scaling. The number of filters reaches 119; sometimes the image lattices need subsample by 4-8 folds (Zhu 2005). As for images without repeated distorted micro-structures, it is also unnecessary to follow the classic methods since the textons are distinct.

The texton extraction is the first and the most consumptive step, it should be implemented effective and efficiently. Aiming at retrieving simple and low-cost bottom-up attention simulation, this model abandons the orientation in clustering textons considering the principle of visual attention in neurobiology and the high computation complexity of the orientation.

Texton extraction in our model follows 3 steps:

1. The image is divided into a set of lattices. For each of the lattices, we calculate its texture features, normalize them and then form feature vectors.
2. The K-mean clustering algorithm is initialized by selecting samples randomly from the data set. Then the

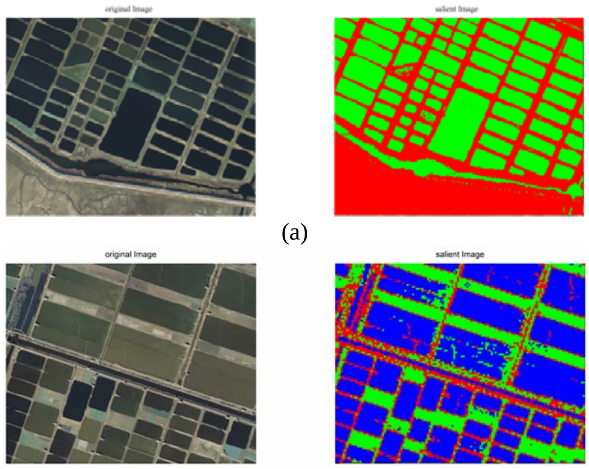
[†] Note that all the images in the paper are chosen randomly to denote fundamental clearly. We do not confine the model to a class of specific problems.

algorithm is applied to these feature vectors to form k' ($k' > \mathrm{k}$) centres. The k' centres should be pruned down by merging centres too close together or those with few data assigned.

3. The K-mean algorithm is applied again to these k' centres to form k centres to achieve a local minimum. These k centers finally constitute the texton vocabulary corresponding to the image.
4. Though results are highly relevant to the choice of k certainly, the method shows robustness and rapidness to our experiments.

3.5 Simulating attention in imagery by extraction of texton

Blobs constructed with a specific texton are considered as a homogeneous class; the connected blobs belonging to the same class are considered as an object. Combined with rarity quantification of saliency, we conclude that less area of regions with a specific texton, are more possible salient regions over the image. In our experiments, the rarity hierarchy is represented consequently by red, green, blue, dark red. From figure 4, one can observe that the size, shape and location of regions have been detected relative accurately and even the trivial ground objects are presented suitable. Though the pure rarity rank does not always accord to the saliency rank, the rarity will still help the future context attention.



(a)

(b)

Figure 4. The attentive regions detected from aerial images. The rarity hierarchy is represented consequently by red, green, blue, dark red, dark green, dark blue and so on. The left image is the original image, and the right one is the result. The square pools with different colour are categorized together rightly; the pools and bank or fields are discriminated properly.

4. RESULT AND DISCUSSION

As foregoing statement, the saliency in remote sensing is close to experts' verdict. Accordingly, we refer to three of our colleagues to take part in our experiments. Each is told to "select interesting regions where objects are presented". If the two of them reported the same place in a certain image, that place would be the attended target place and be labelled by hand.

It is hard to evaluate the performance of our model quantitively. If we record eye movements as usual psychological experiments, the position, instead of the sizes and shape, of attended regions can be recoded. Even though we only care about the position, the individual differences still impose great influence on the performance for the complexity of remote sensing images (Davies 2006). Hence, we draw a conclusion that it is impossible to define interesting regions' saliency hierarchy precisely at least in the short term. We use the baseline of "region of interest" and saliency map for reference. If several regions connected, these regions are set the same index and considered as a whole region. If the region's area is larger than a threshold t, the region is considered as an attentive object. If the region's area is smaller than t, its neighbours in a circle with a fixed radius are considered as parts of a latent attentive object. The homogeneous regions merged together are not considered twice, but the heterogeneous regions merged together have right to merge their neighbours to form a new object again.

We take the labelled regions as the standard ones. If a region fall in the right corresponding range, the region is thought as "hit". If a region is categorized in a false class, it is though as "false alarm". The *Hit Rate* (HR) and *False Alarm Rate* (FAR) measure how well the regions correspond to human perception. They can be obtained as follows:

$$HR_j = \frac{\sum_i n_{i,j}}{\sum_i N_{i,j}} \tag{2}$$

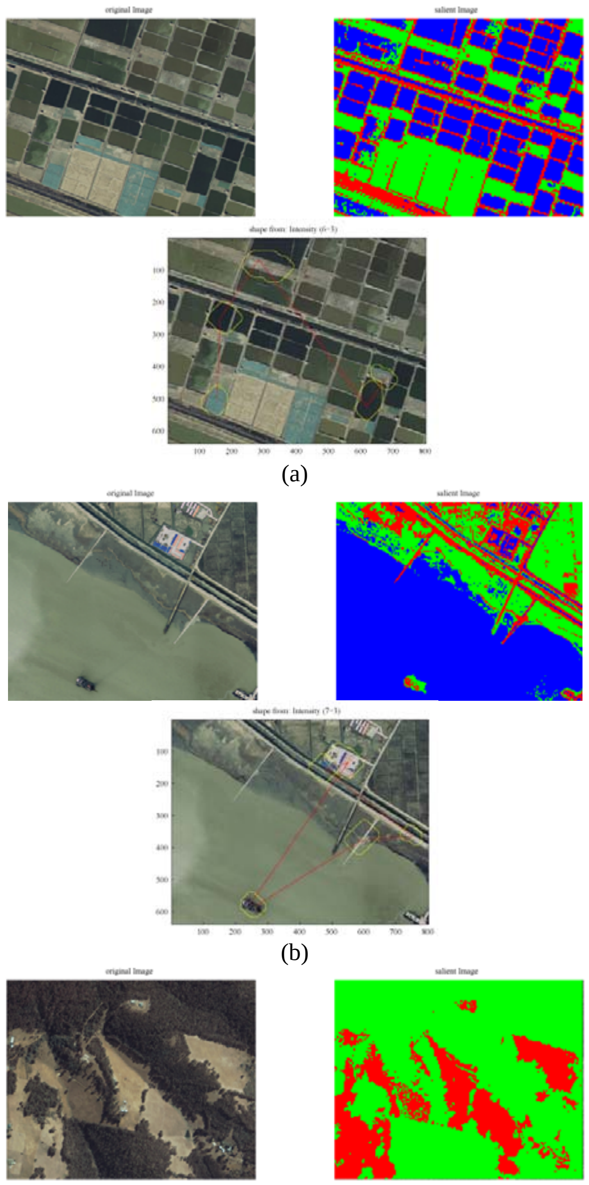
$$FAR_j = \frac{\sum_i fn_{i,j}}{\sum_i n_{i,j} + \sum_i fn_{i,j}} \tag{3}$$

where $n_{i,j}$ denote the number of objects from our model at $i-th$ rarity rank of $j-th$ image; $N_{i,j}$ the number of objects at $i-th$ rarity rank of $j-th$ image by human perception; $fn_{i,j}$ the number of objects which are detected at $i-th$ rarity rank and need to be excluded from in $j-th$ image.

From the experiments, the HR and FAR vary from this photo to that photo. The HR is often larger than 85%; the FAR can reach 50% in the worst situation when set a bad threshold t. Larger t, smaller FAR; more complicated an image, higher FAR possibly.

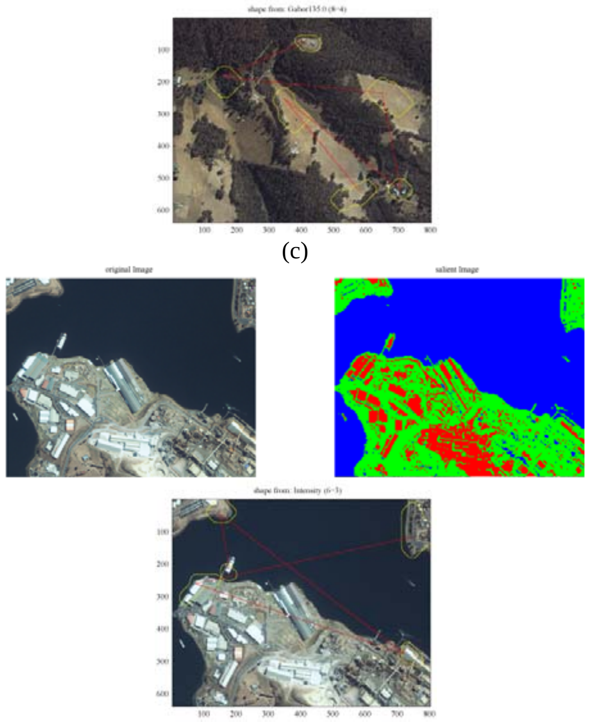
Our results suggest that our model outperform Itti and Koch's model in emulating interpretation. First, our results show a set of ground objects; in comparison, it is common that something trivial or less geographical meaning can be inferred from centres of their closed curves. Second, their results confine to the neurobiology and fail to figure out the more regions' information such as regions' category. Our model also inspires from neurobiology, but comprises more ground objects' detailed information by extracting textons. Third, Itti and Koch's results only give out the position and the possible regions instead of more detailed information such as shape and size. Therefore, sometimes their results contains a lot of redundant or heterogeneous ground objects. Our results contain a lot of important information beneficial to future interpretation such as size and shape. Finally, their model does not detect homogeneous objects in turn, and this seems not accord with our interpreting custom; our model treats homogeneous objects at the same hierarchy.

It should be noted that this study has focused on retrieving more geographic information instead of attention sequence. We get rarity hierarchy not saliency hierarchy. In this point, Itti and Koch's model is better than ours. However, the accuracy of their model's attention shift is still under suspicion as told before. The inner reason is the mechanism of visual attention is still uncertain, thus any pure simulation based on present knowledge is far from satisfaction. The possible outlet is to combine the latest discovery from relative discipline, such as neurobiology and psychology, with the specific problems.



(a)

(b)



(c)

(d)

Figure 5. Some results of our method. (a) and (b) are results of aerial photos; (c) and (d) are results of IKONOS data. In each sub-group, the second image is our result and the third on is Itti and Koch's. In our results, the most scarce regions, the red regions, are the most attractive parts; then the green regions; finally the blue ones.

5. CONCLUSION

We present a bottom-up model to emulate interpretation of remote sensing photos. We introduce textons to analyse images to apply neurobiological discovery. The results indicate our model is more likely to interpret remote sensing than Itti and Koch's model. This model is capable of detecting ground objects and providing rarity for computing saliency hierarchy. According to the experiments, we confirm Parkhurst and Niebur's viewpoint that texture attribute can contribute to the attention in a bottom-up fashion in neuroscience (Parkhurst 2004).

Because we implement a bottom-up model, the prior knowledge and context is unnecessary in this stage. We are now approaching to combine context with the bottom-up model to make sense the mutual relationship between the attentive ground objects in order to recognize them.

References

Chen, X., Huo, H., Fang, T., Li D., 2007. New approach to saliency based on intrinsic relationship among texture features. *SPIE--The International Society for Optical Engineering*, Wuhan, China.

Chun, M.M., Wolfe, J.M., 2001. *Visual attention*. In: Goldstein, E.B. (Ed.), Blackwell's Handbook of Perception, Chapter 9 Blackwell, Oxford, UK, pp. 272–310.

Davies, C.; Tompkinson, W.; Donnelly, N.; Gordon, L.; Cave, K., 2006. Visual saliency as an aid to updating digital maps. *Computers in Human Behavior*, 22, pp. 672–684.

Fabrikant, S.I. and Goldsberry, K., 2005. Thematic relevance and perceptual salience of dynamic geovisualization displays. *Proceedings of the 22nd ICA international cartographic*.

Hou X., Zhang L., 2007. Saliency Detection: A Spectral Residual Approach. *IEEE Conference on Computer Vision and Pattern Recognition 2007*.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20(11), pp.1254-1259.

Itti, L., Koch, C., 2001. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience,* 2(3), pp. 194-203.

Itti, L, P. B., 2005. A Principled Approach to Detecting Surprising Events in Video. *IEEE Conference on Computer Vision and Pattern Recognition*.

Itti, L., Koch, C., & Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 1254 – 1259.

Jäger M., O. H., 2005. Saliency and Salient Region Detection in SAR Polarimetry. *Proceedings of IGARSS 2005*.

Julesz, B., 1981. Textons, the elements of texture perception, and their interactions. *Nature*, 290, 91-97.

Leung, T. and Malik, J., 1999. Recognizing surface using threedimensional textons. *In Proc. of 7th ICCV, Corfu, Greece, 1999.*

Lloyd, R., Hodgson, M., and Stokes, M., 2002. Visual Categorization with Aerial Photographs. *Annals of the Association of American Geographers*, 92(2), pp. 241-266.

MANCAS, M., Gosselin B., MACQ B., 2007. A Three-Level Computational Attention Model. *Proceedings of ICVS Workshop on Computational Attention & Applications (WCAA-2007).*

Parkhurst, D., J., and Niebur, E., 2004. Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience,* 19(3), pp. 783-789.

Reznik, H. M. a. S., 2007. Building facade interpretation from uncalibrated wide-baseline image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing,* 61(6), pp. 371-380.

Scassellati, F. S. a. B., 2007. A Behavioral Analysis of Computational Models of Visual Attention. *International Journal of Computer Vision*, 73(2), pp. 159–177.

Sergio, E.; Petia, R.; Oriol, P., 2007. Complex Salient Regions for Computer Vision Problems. *IEEE Conference on Computer Vision and Pattern Recognition 2007*.

Siagian, C. I., L., 2007. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2), pp. 300 - 312

Trias-Sanz, R., 2006 . Texture Orientation and Period Estimator for Discriminating Between Forests, Orchards, Vineyards, and Tilled Fields. *IEEE Transactions on Geoscience and Remote Sensing*, 44(10), pp.2755 - 2760

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F., 1995. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78, 507 – 545.

Yang, J. W., R. S., 2007. Classified road detection from satellite images based on perceptual organization. *International Journal Of Remote Sensing*, 28(20), pp. 4653-4669.

Zhu SC, C. G., Y Wang, Xu Z., 2005. What are textons? *International Journal of Computer Vision*, 62(1/2), pp. 121-143.

Acknowledgements

This research was supported by National Key Basic Research and Development Program of China (Grant No. 2006CB701303). The authors are grateful to Jianjun Qin, Chunting Liu and Qinqin Shi for their zealous help, Songhao Zhu and Sheng Xu for precious suggestions. The authors would like to thank Frederick Shic for his ardent help on the evaluation of results. The authors would also like to thank Imagery courtesy of Space Imaging LLC for providing IKONOS data.