www.s2is.org
International Journal on Smart Sensing and Intelligent Systems
ISSN 1178-5608

VISUAL ATTENTION MODEL WITH ADAPTIVE WEIGHTING OF CONSPICUITY MAPS FOR BUILDING DETECTION IN SATELLITE IMAGES

A.-M. Cretu[1] and P. Payeur[2]
[1]Department of Computer Science and Engineering, Université du Québec en Outaouais, Canada
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Canada
Emails: ana-maria.cretu@uqo.ca, ppayeur@site.uottawa.ca

Submitted: Aug. 24, 2012 Accepted: Nov. 5, 2012 Published: Dec. 1, 2012

Abstract - The lack of automation and the limited performance of current image processing techniques pose critical challenges to the efficient and timely use of the large amount of data made available by aerial and space based assets. The imitation of fast adaptation and inference capability of human visual system appears to be a promising research direction for the development of computational algorithms able to deal with large variations in image content, characteristics and scale as those encountered in satellite imaging. The paper explores the potential use of an improved computational model of visual attention for the complex task of building identification in satellite images. It contributes to extend the envelope of application areas of such models and also to expand their current use from single object to multiple object detection. A set of original weighting schemes based on the contribution of different features to the identification of building and non-building areas is first proposed and evaluated against existing solutions in the literature. A novel adaptive algorithm then chooses the best weighting scheme

based on a similarity error to ensure the best performance of the attention model in a given context. Finally, a neural network is trained to predict the set of weights provided by the best weighting scheme for the context of the image in which buildings are to be detected. The solution provides encouraging results on a set of 50 satellite images.

Index terms*:* Visual attention, weighting schemes, neural networks, building detection, satellite images.

I. INTRODUCTION

The mainstream adoption of the geospatially-enabled web and of other spatial information production tools for digital maps and geographical information systems is attested by the increasing availability of high resolution digital satellite and aerial images. 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 spatial images in a timely manner. As the current generation of algorithms for feature extraction and recognition seems to have somewhat reached its limits, the recent years marked an increasing interest in algorithms inspired from biological human vision. While the research in this direction evolved significantly over time, most of the proposed biological visual-inspired systems have been tested for a limited number of images in simplistic scenarios. Up to now, none of the computational attention models was demonstrated to work under such a wide range of image contents, characteristics and scales as those encountered in satellite imaging.

This paper tackles the issue of building recognition in satellite images. It explores the potential use of an improved computational model of visual attention for this task. Visual attention models highlight areas of high interest based on a set of features such as color, intensity, and orientation. The current model also accounts for texture as an additional feature, as texture is believed to play a critical role for object detection in human vision, particularly in cluttered scenes. A set of original weighting schemes is proposed to enhance the detection of desired objects in an image, based on the contribution of different features to the desired (e.g. buildings) versus non-desired (e.g. non-building) areas. Since it is known that the human visual system is effective due to its capability to adaptively determine which features to use in a given context, the paper also proposes an original adaptive algorithm to choose the best weighting scheme (and concomitantly the best set of weights) to highlight desired areas and a mechanism to predict the best scheme in a given context.

II. RELATED WORK

Recent years witnessed an increased interest in computational models that mimic the human sensing perception [1], the human visual system, and particularly the human visual attention. The main idea of such systems inspired from visual attention is to compute several features derived from a color image and fuse their saliencies into a representation called saliency map. The strategy to combine these feature maps into the saliency map is still a challenge. Minimalist approaches select one feature map to represent saliency. Walther *et al.* [2] select the feature map that contributes the most to the saliency map, while Gopalakrishnan *et al.* [3] choose the saliency map as either the color or the orientation map. Kim and Kim [4] use a variance measure to choose the most discriminative feature among a set of linear and non-linear color combinations. In [5] only spatial color contrast is used as a feature. Other approaches study appropriate weighting schemes to enhance the saliency map. The model of Itti and Koch [6] integrates color, orientation, and intensity features using non-iterative or iterative normalization, or using feature weights learned from a training set. Zhao and Liu [7] propose a sparse embedding feature combination strategy for the color, intensity and orientation feature maps. In [8], spatial compactness and saliency density are used to weight dynamically the color, intensity and orientation feature maps. Hu et *al.* [9] propose a local context suppression strategy to adaptively combine intensity, color, and texture attention cues. In [10], a complex unsupervised filter method is proposed for the identification of relevant color and texture features. Vu and Chandler's algorithm [11] adaptively selects low-level features such as contrast, sharpness, and edge strength for main subject detection in images based on statistics. Frintrop [12] uses a learning mechanism for storing target-relevant features and excites or inhibits features to obtain a target-dependent saliency map. The approach of Goferman et *al.* [13] aims at detecting the image regions that are representative for a scene, instead of a single subject. Murray *et al.* [14] expand a low-level visual model that predicts color appearance for saliency estimation. A different stream of saliency detectors does not take inspiration from visual attention but rather capitalizes on image frequency [15-18], on spatiotemporal cues [19], on regional contrast [20] or on natural image statistics [21]. Most of these solutions are tested in simplistic scenarios, on images containing usually one object. In our previous work, a model inspired from visual attention was used with success for a real-world task, in the context of multi-view vehicle category identification [22]. This paper explores the use of visual attention models in the more complex context of satellite imaging, characterized by a wide range of image content variations and by multiple small objects against a cluttered background.

III. PROPOSED APPROACH

The computation of feature and conspicuity maps follows largely the bottom-up model of Itti [6]. Initially, one or several image pyramids are created from the input image to enable the computation at different scales. Several features are then computed in parallel and feature-dependent saliencies are computed for each channel. Aside from the intensity, color and orientation feature maps of Itti, the proposed model also incorporates an additional color feature, the DKL color code, and a contrast map. Center-surround operations, modeled as a difference between fine and coarse scales, are applied on all features. Each set of features is stored in feature dependent saliency maps, called conspicuity maps, in form of grayscale images where the intensity of each pixel is proportional to its saliency. At this level, two other conspicuity maps are introduced in the model, both being measures of texture. The first one is the local standard deviation and the second one is the local entropy, a statistical measure of randomness that characterizes the texture of an image. After normalization, all these maps are combined in the final saliency map.

a. Computation of Conspicuity Maps

The RG-BY color opponency feature maps and the corresponding conspicuity map, *CC,* are computed as in [6], for the center-scales, $c=\{2,3,4\}$ and the surround scales, $s=c+\delta$, $\delta=\{3,4\}$. In the biologically-inspired DKL color code, colors are represented using a luminance axis and two opponent chromatic axes: one axis along which chromaticity varies without changing the excitation of blue-sensitive cones, and one axis along which chromaticity varies without change in the excitation of red-sensitive or green-sensitive cones. Details of the RGB to DKL conversion are available in [23]. Typical center-surround operations are applied on the 3 color components prior to the computation of the DKL color code conspicuity map, *DC*. The intensity is computed as $I = (R+G+B)/3$ where R, G and B are the red, green and blue color channels respectively and the orientation information is obtained based on Gabor pyramids for the orientations $\{0^o, 45^o, 90^o, 135^o\}$. The intensity and orientation conspicuity maps are denoted *IC*, and *OC*, respectively. The contrast map, *RC*, is computed based on the global standard deviation of pixel intensities. Two more conspicuity maps are introduced to bring additional texture information that is believed to play a critical role to capture visual attention in images that contain many small objects present in a cluttered background [8], as is the case in satellite imaging applications. This extra texture information is not computed on a pyramid of images, but directly on the input image. The first texture conspicuity map, *TC*, is based on the local standard deviation of an image in a 3-by-3 neighborhood, while the second map, *EC*, is composed of the local entropy value, *e,* of the 9-by-9 neighborhood around the corresponding pixel in the input image as described in [24]. The final saliency map, *S*, is computed as a weighted sum of the conspicuity maps, such as:

$$S = \frac{w_c \cdot CC + w_d \cdot DC + w_i \cdot IC + w_o \cdot OC + w_r \cdot RC + w_t \cdot TC + w_e \cdot EC}{7} \tag{1}$$

For the standard bottom-up saliency map [6], the weighting coefficients, w_α, $\alpha = \{c, d, i, o, r, t, e\}$ are all set to 1. The seven conspicuity maps are therefore contributing equally to the saliency map.

b. Alternative Weighting Schemes

The saliency map computed with all unitary weights using eq. (1) with $w_\alpha=1$, $\forall \alpha \in \{c, d, i, o, r, t, e\}$ typically highlights all the areas of high interest. However, if one wants to identify only one particular type of object in an image, for example buildings, the areas corresponding to buildings in the saliency map should be the most prominent. In the human visual system this is achieved by top-down visual attention, which is a goal-oriented attention. To mimic such goal-oriented behavior, the conspicuity maps that contribute most to make object areas prominent should be further excited (i.e. weighted stronger), while those that contribute more to non-desired areas in an image should be inhibited (or eliminated altogether). A novel approach is proposed in this paper to achieve this sort of behavior. Desired object areas are first manually extracted in a sample set of images. The segmented image is binarized to create a mask, *MS*, in which the areas of interest (e.g. buildings) are white (1). The purpose of this mask is twofold: it helps gauge the contribution of conspicuity maps and it serves as a reference for the evaluation of the quality of different saliency maps resulting from different weighting schemes introduced in eq. (1). The contribution, $b_{d(FM)}$, of each conspicuity map, *FM*, to the saliency map is computed first for the desired areas as the mean value of each conspicuity map within the desired areas:

$$b_{d(FM)} = mean(FM_d(i,j)),\ FM_d = \{FM(i,j) \mid MS(i,j) = 1\} \tag{2}$$

Similarly, for the non-desired areas, the contribution is:

$$b_{nd(FM)} = mean(FM_{nd}(i,j)), FM_{nd} = \{FM(i,j) \mid MS(i,j) = 0\} \tag{3}$$

with *FM* being each of the conspicuity feature maps computed in section III.a, that is *CC, DC, IC, OC, RC, TC,* and *EC* respectively. Three novel weighting schemes are proposed below which, based on the contributions of the respective conspicuity maps computed in eq. (2) and (3), enhance those areas representing objects of interest, here buildings.

Differential Weighting *(E-I D-ND or E D-ND)***:** In this approach, the weights associated with each conspicuity map are set based on the difference between the contribution of desired areas and the contribution of non-desired areas. Therefore the scheme excites ("E") those conspicuity maps for which the contribution to the desired areas is larger than to non-desired areas, that is $b_{d(FM)} > b_{nd(FM)}$. Reversely, the feature maps for which the contribution to the non-desired areas is larger, that is $b_{d(FM)} < b_{nd(FM)}$, are inhibited ("I"). The weight for each conspicuity map is defined as:

$$w_{FM(E-I\ D-ND)} = b_{d(FM)} - b_{nd(FM)},\ FM \in \{CC, DC, IC, OC, RC, TC, EC\} \tag{4}$$

with the possibility that some of the weights $w_{FM(E-I\ D-ND)}$ are negative.

Alternatively, the conspicuity maps for which the contribution to non-desired areas is higher can be ignored by setting the corresponding weights to 0:

$$w_{FM(E\ D-ND)} = \begin{cases} b_{d(FM)} - b_{nd(FM)}, \textit{if } b_{d(FM)} > b_{nd(FM)} \\ 0, \quad \textit{otherwise} \end{cases} \tag{5}$$

This latter scheme excites only, as the inhibitory weights are set to 0. The final saliency map, *S*, is computed using eq. (1), but with the seven weights respectively computed in eq. (4), or eq. (5).

Ratio Weighting *(E-I D/ND)***:** This scheme is an exciting and inhibiting scheme in which the weights of the feature maps are scaled according to the ratio between the contribution of desired and non-desired areas:

$$w_{FM(E-I\ D/ND)} = b_{d(FM)} / b_{nd(FM)} \tag{6}$$

The conspicuity maps that contribute more to non-desired areas are inhibited in a similar fashion to what is proposed by Frintrop [12], that is with an amount equal to the inverse of the corresponding weight computed using eq. (6) if the respective weight is smaller than 1. In this case, instead of using eq. (1), the final saliency map, *S*, is computed as:

$$S = \sum_{FM} w_{FM(E-I\ D/ND)} \cdot FM|_{w_{FM(E-I\ D/ND)}>1} - \sum_{FM} \frac{1}{w_{FM(E-I\ D/ND)}} \cdot FM|_{w_{FM(E-I\ D/ND)}<1} \tag{7}$$

where $FM \in \{CC, DC, IC, OC, RC, TC, EC\}$. The pixels of *S* with negative values are set to 0.

Differential Ratio Weighting *(E (D-ND)*D/ND* or *E-I (D-ND)*D/ND)***:** This scheme is a combination between the two previous schemes. The weights of the feature maps that contribute more to desired areas are made stronger due to the multiplication operation between the differential and ratio components and those that contribute less are either ignored (as in eq. (5)):

$$w_{FM(E(D-ND)*D/ND)} = w_{FM(E\ D-ND)} \cdot w_{FM(E-I\ D/ND)} \tag{8}$$

or inhibited (as per eq. (4)):

$$w_{FM(E-I(D-ND)*D/ND)} = w_{FM(E-I(D-ND))} \cdot w_{FM(E-I\ D/ND)} \tag{9}$$

For the ratio component, $w_{FM(E-I(D/ND))}$, the weights are computed as in eq. (6). Finally, eq. (1) is used to obtain the final saliency map, *S*, but with the seven weights respectively computed in eq. (8) or eq. (9).

The following two schemes are based on solutions already proposed in the literature and are included here for enabling a formal comparison with the novel weighting schemes:

Ratio Weighting Based on Most Salient Region (*MSR*): This scheme is similar to the one proposed in [12], with the exception that the uniqueness function is ignored in order to allow for the detection of multiple objects in an image. In this case only the top 5% of the most salient points, *MSR*, are used in the computation of the weights as opposed to eq. (6) where the mean of the entire saliency map is used:

$$w_{FM(MSR)} = MSR_{d(FM)} / MSR_{nd(FM)} \tag{10}$$

$$S = \sum_{FM} w_{FM(MSR)} \cdot FM|_{w_{FM(MSR)}>1} - \sum_{FM} \frac{1}{w_{FM(MSR)}} \cdot FM|_{w_{FM(MSR)}<1} \tag{11}$$

where $FM \in \{CC, DC, IC, OC, RC, TC, EC\}$. As in eq. (7), the pixels of *S* with negative value are all set to 0.

Statistics-Based Weighting *(Statistics)***:** This scheme is similar to the one proposed in [11]. It does not take into account the contribution of each conspicuity map to the final saliency map, but rather computes a statistic measure of its utility, α_i, as a sum of its variance and kurtosis. These measures are sorted in descending order and a weight of 1 is assigned to the conspicuity map with greatest statistic, 2/3 to the second greatest statistics, 1/3 to third greatest statistics and 0 to all the other ones.

$$w_{FM(Statistics)} = \begin{cases} 1, & if\ \alpha_i = \overline{\alpha_1} \\ 2/3, & if\ \alpha_i = \overline{\alpha_2} \\ 1/3, & if\ \alpha_i = \overline{\alpha_3} \\ 0, & otherwise \end{cases}, \quad \overline{\alpha} = sort(\alpha_i) \tag{12}$$

As in [11], the saliency map is computed as defined in eq. (1) by replacing the values of weights with those computed in eq. (12). The saliency map is then normalized and rescaled to [0, 1].

c. Vegetation and Shadow Removal

To further improve the detection of building areas, domain-specific knowledge is introduced by means of two color invariants for the detection of vegetation areas and of shadowed areas in satellite images, as proposed in [25]. The color invariant, *v*, which defines vegetation areas, builds upon the green (*G*) and blue (*B*) channels of the color satellite image and is defined as:

$$v = \frac{4}{\pi} \cdot \arctan\left(\frac{G - B}{G + B}\right) \tag{13}$$

The invariant, *s*, which defines shadow areas, also includes the red (*R*) color channel and the global intensity (*I*) value. It is formulated as:

$$s = \frac{4}{\pi} \cdot \arctan\left(\frac{I - \sqrt{R^2 + G^2 + B^2}}{I + \sqrt{R^2 + G^2 + B^2}}\right) \tag{14}$$

The resulting images of these invariants are then Otsu thresholded to decide what pixels belong to vegetation and shadow areas. The resulting pixels representing vegetation and shadows are used to mask the corresponding areas in the saliency map to further improve building detection.

d. Adaptive Selection of Best Weighting Scheme

In order to evaluate the most appropriate weighting scheme for an image, a measure of similarity is computed between the manually segmented mask image, *MS*, (considered as a reference) and any of the weighted saliency maps, *S*, with or without the vegetation and shadow removal. The similarity is computed as a normalized cross-correlation (*NCC*):

$$NCC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(MS(i,j) - \overline{MS})(S(i,j) - \overline{S})}{\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(MS(i,j) - \overline{MS})^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(S(i,j) - \overline{S})^2\right)}} \tag{15}$$

where $\overline{MS}$ and $\overline{S}$ represent the average values of pixels over the entire *MS* and *S* maps respectively. The best weighting scheme is selected as the one that obtains the largest *NCC* value, as a larger value denotes a higher similarity.

e. Learning of Winning Weights

Up to now, the proposed solution determines the best weighting scheme based on a set of manually segmented images. In order to ensure the generalization capability and remove the constraint imposed by the need of a prior manual segmentation, a neural network architecture is used to learn the association between the best weights as determined in section III.d with the corresponding image content. To quantify the content of an image, a measure of the energy of the conspicuity map is used, given that the energy inherently depends on the content. At the presentation of a similar context, that is a previously unseen satellite image containing buildings, the network is able to provide an estimate for the set of optimal weights to ensure the detection of buildings or other objects that the system would be trained to detect. A feedforward neural architecture with 17 hidden neurons (H_1, …, H_{17}) and sigmoid activation function, as illustrated in Fig. 1, is trained with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton variant of backpropagation [26] for 1500 epochs, on 75% of the images in the dataset. The other 25% of images are used for testing.

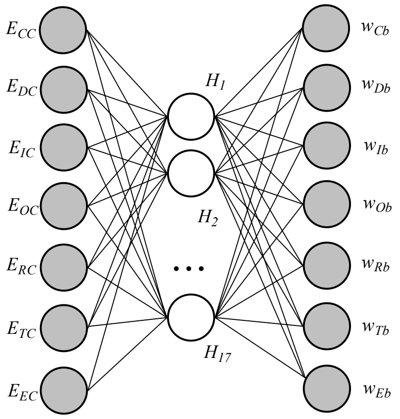


Figure 1. Neural network architecture.

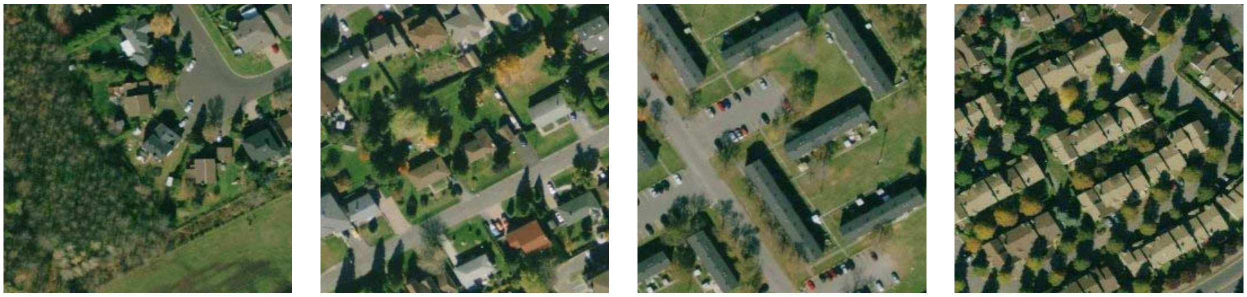
The network receives at the input a vector [E_{CC}, E_{DC}, E_{IC}, E_{OC}, E_{RC}, E_{TC}, E_{EC}] composed of the energy computed on each conspicuity map, *CC, DC, IC, OC, RC, TC,* and *EC* respectively, which is computed as:

$$E_{FM} = \sum_{i} \sum_{j} FM(i, j), \quad FM \in \{CC, DC, IC, OC, RC, TC, EC\} \tag{16}$$

The network outputs the set of weights [$w_{Cb}, w_{Db}, w_{Ib}, w_{Ob}, w_{Rb}, w_{Tb}, w_{Eb}$] that correspond to the best weighting scheme to apply with the corresponding context of the images on which the conspicuity maps were computed. These weights correspond to the winning scheme derived in section III.d and are used for training. The training error achieved is of order 0.0006, while the testing error is of order 0.015. The use of such a neural architecture provides a straightforward and efficient way to assign the optimal weighting scheme from a rapid examination of the energy contained in each of the conspicuity maps of a given image, energy that inherently depends on the image content and indirectly captures its context. This extra stage generalizes the use of the proposed visual attention model with a learning technique for application to a broad set of images that are not used for initial tuning or training.

IV. EVALUATION ON SATELLITE IMAGES

The proposed solution is tested on a dataset of 50 satellite images [27].



(a) (b) (c) (d)

Figure 2. Sample of satellite images from the dataset.

Each has a resolution of 256×256 pixels and contains residential areas with different topologies and complexities. Fig. 2 shows samples from the dataset, while Fig. 3 illustrates an example of an image and the corresponding seven conspicuity maps computed in section III.a. It can be seen that different maps highlight different characteristics of the initial image. While one can notice that there is a significant difference between the RG-BY color conspicuity map (Fig. 3b) and the DKL color conspicuity map (Fig. 3c), it is difficult to judge which one is better unless the purpose of the application is known.

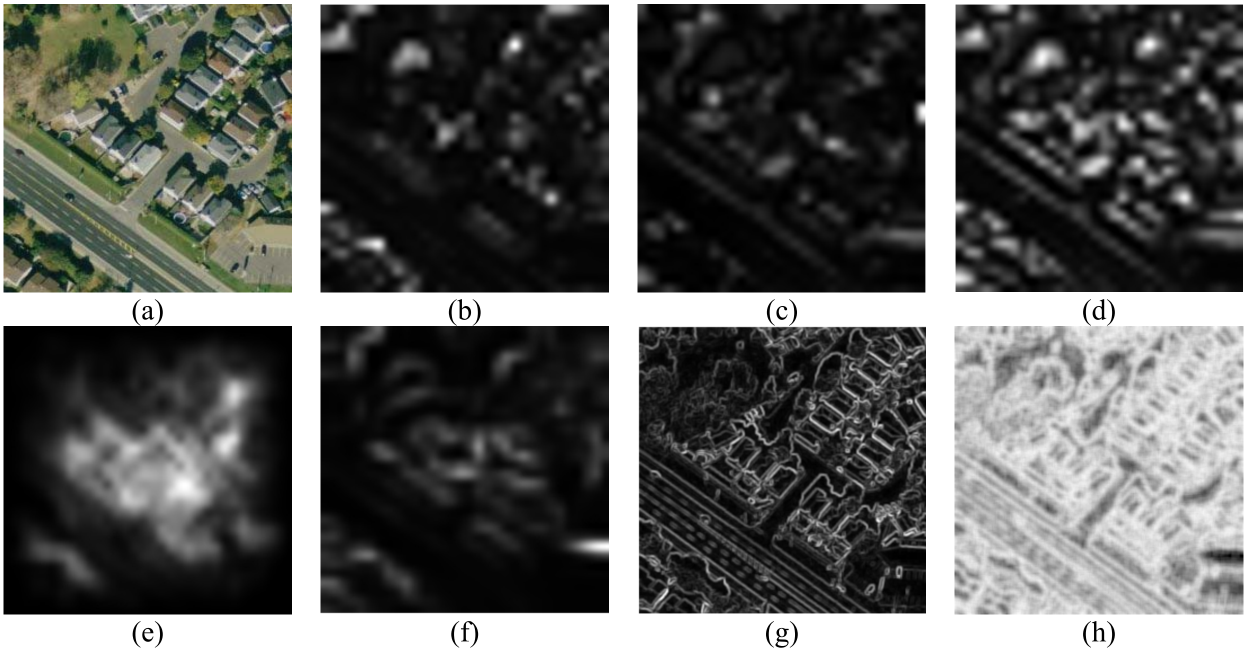


Figure 3. (a) Image and corresponding feature conspicuity maps: (b) RG-BY color, (c) DKL color, (d) intensity, (e) orientation, (f) contrast, (g) texture, and (h) entropy.

If one wants, for example, to identify buildings in an image, the areas corresponding to buildings in the saliency map should be the most prominent, as explained in section III.b. To achieve this, desired object areas are manually extracted in a sample set of images. An example is shown in Fig. 4.

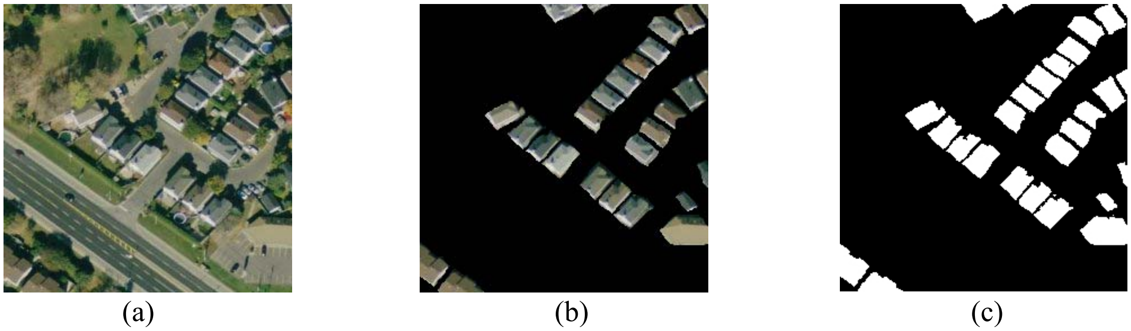


Figure 4. (a) Initial image, (b) manually segmented buildings and (c) corresponding mask image.

The segmented image is binarized (a threshold of 0.18 was identified by trial and error) to create a mask, *MS*, in which the building areas are white (1), as illustrated in Fig. 4c. The purpose of this mask is twofold: it helps gauge the contribution of conspicuity maps and also serves as a reference for the evaluation of the quality of different saliency maps resulting from different weighting schemes applied in eq. (1) or eq. (7). Using as reference the mask, *MS*, shown in Fig. 4c, the average contribution of each conspicuity map to building areas is computed using eq. (2) and to non-building areas using eq. (3). The contributions to the desired areas, $b_{d(FM)}$, and to non-desired areas, $b_{nd(FM)}$, for the image in Fig. 4a are shown in Fig. 5.

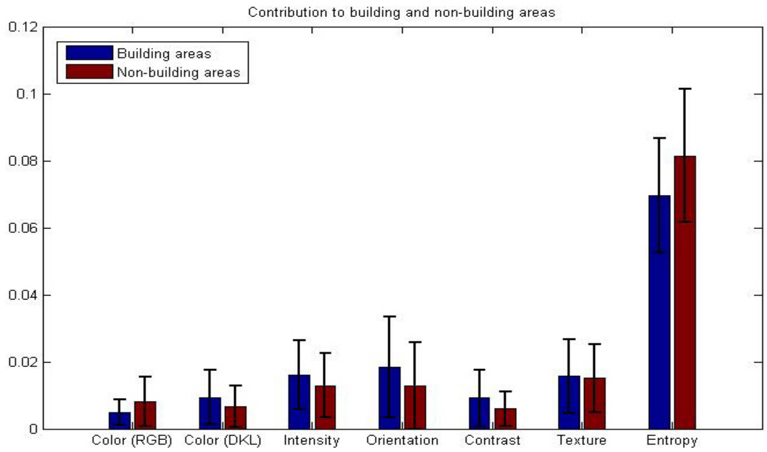


Figure 5. Contribution and standard deviation of contribution of conspicuity maps to building (desired) and non-building (non-desired) areas for the image in Fig. 4a.

As the contribution of different conspicuity maps to building and non-building areas varies with the content of the image, the adjustment of weights is performed separately on each image from the entire dataset before the trends are captured by the neural architecture introduced in section III.e. An analysis of the standard deviation of the contribution graphs, similar to Fig. 5, for every image in the dataset leads to the conclusion that no conspicuity maps should be eliminated from the model due to a high deviation (marked by segments in Fig. 5) and a low contribution.

Fig. 6b shows a saliency map obtained using all equal unitary weights in eq. (1) as discussed in section III.a. Fig. 6c to Fig. 6i present the saliency maps computed over the image of Fig. 6a for all the proposed and comparative weighting schemes defined in section III.b. For each image in the dataset, the best weighting scheme is identified based on the normalized cross-correlation, defined in section III.d. For example, in this case, the E-I D/ND weighting scheme illustrated in Fig. 6e obtains the highest normalized cross-correlation value and is selected as the winner.

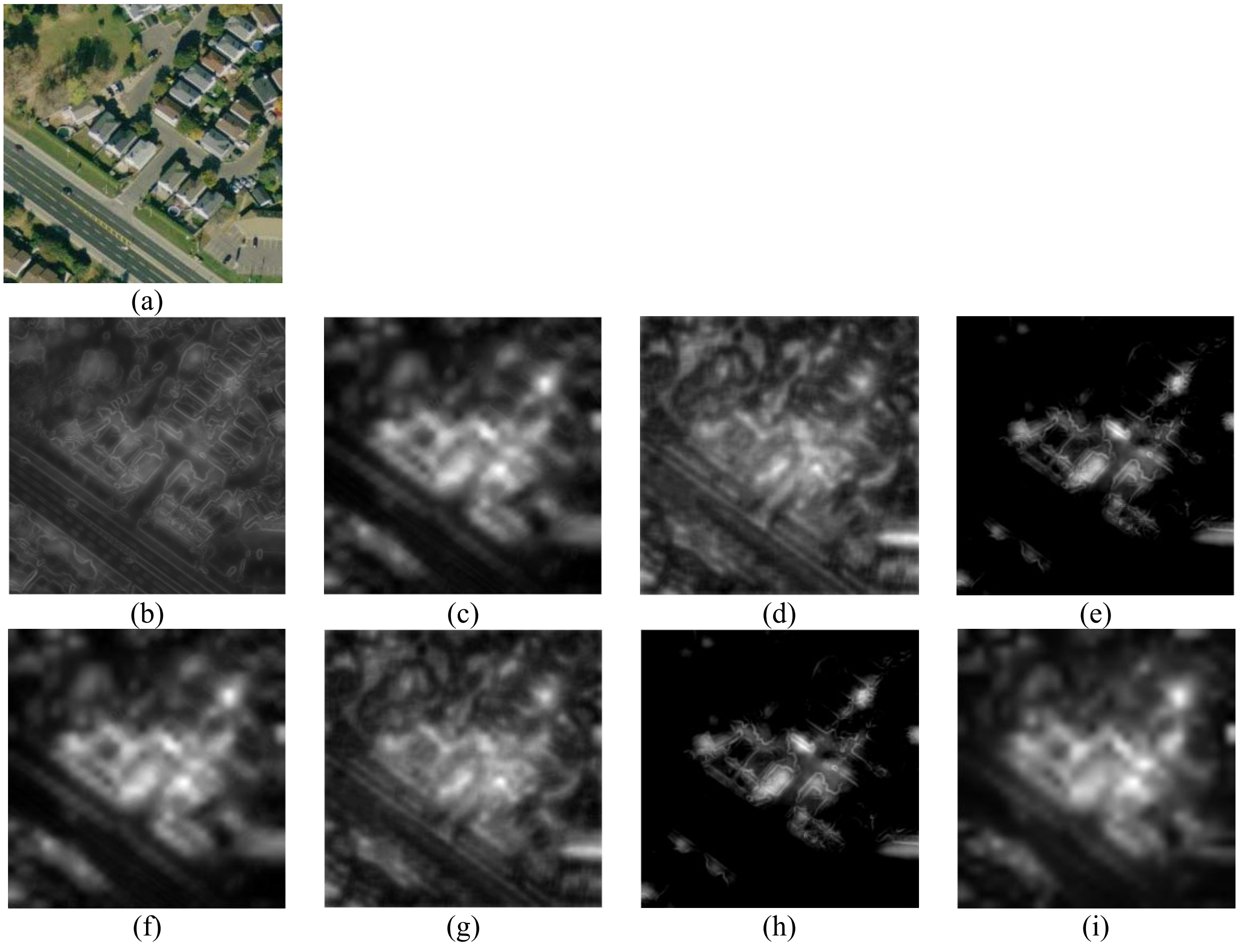


Figure 6. (a) Original image and saliency maps for (b) unitary weights, (c) E-I D-ND, (d) E D-ND, (e) E-I D/ND, (f) E (D-ND)*(D/ND), (g) E-I (D-ND)*(D/ND), (h) MSR and (i) Statistics.

Furthermore, using the color invariants defined in section III.c and Otsu thresholding, the vegetation and shadow areas can be easily identified, as illustrated in Fig. 7.

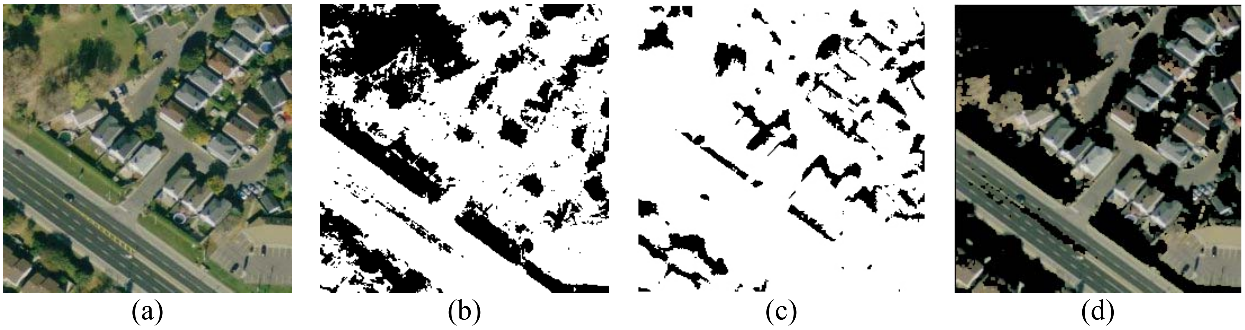


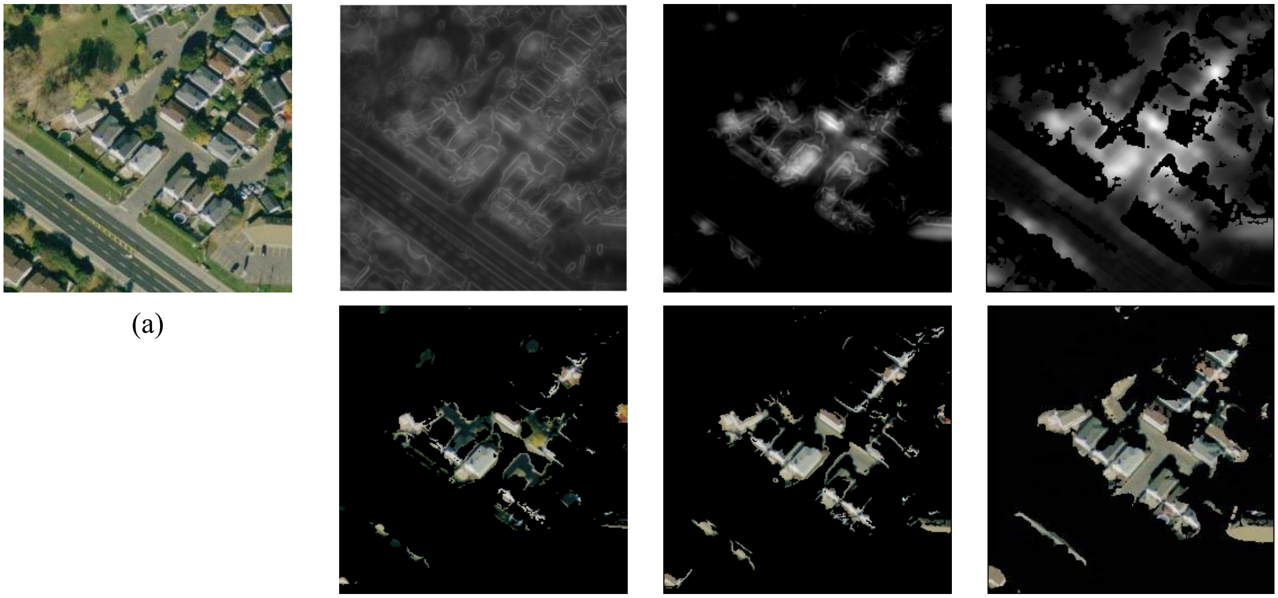
Figure 7. Identified (b) vegetation and (c) shadow areas (shown in black) and (d) image after removal of vegetation and shadow areas from the original image (a).

The figure shows the areas of vegetation and shadow in black, in Fig. 7b and 7c respectively, for the image in Fig. 7a. Fig. 7d shows the result of masking the initial image to remove the detected vegetation and shadow areas, demonstrating that color invariants are reliable to correctly identify these areas. Similarly, the saliency maps obtained using the various weighting schemes can be masked to improve the success rate of building detection by eliminating distractors in the vegetation and shadow areas. Table I presents for both the case with vegetation and shadow removal (VSR) and the case without vegetation and shadow removal (NVSR), the determination of the winning weighting schemes as a percentage of the number of selections over the entire image dataset in the first two colums, and the associated average NCC value for each of the winning schemes in the last two columns.

Table I. Frequency of use and average NCC value for the various weighting schemes.

Weighting Scheme	***Frequency of use***		***Average NCC***	
	VSR	***NVSR***	***VSR***	***NVSR***
Equal weights	0%	0%	N/A	N/A
E-I(D-ND)	4%	48%	**0.4058**	**0.2630**
E(D-ND)	28%	4%	**0.4364**	**0.2625**
E-I (D/ND)	6%	14%	**0.4536**	**0.3483**
E(D-ND)D/ND	52%	6%	**0.4908**	**0.3101**
E-I(D-ND)D/ND	0%	16%	N/A	**0.2776**
MSR	8%	12%	0.3664	0.2339
Statistics	2%	0%	0.1005	N/A

A N/A value in the average NCC columns indicates that the corresponding weighting scheme is never selected as a winning scheme. It can be first observed that the selection of the winning scheme varies significantly with and without the use of vegetation and shadow removal and that the removal of vegetation and shadow areas improves the overall matching between the detected building areas and the manually segmented mask images used as reference, as reflected by the higher average NCC values with VSR. The proposed weighting schemes, with highlighted NCC values in the table, are more frequently selected than the ones already proposed in the literature, denoted as “Equal weights”,“MSR” and “Statistics”, and also lead overall to better performance, demonstrated by the higher average NCC rates, with and without vegetation removal. Fig. 8 presents the saliency map for the same image, shown in Fig. 8a, obtained with equal weights in column (b), with the winning scheme when the shadow and vegetation are not removed in column (c), and with the winning scheme with shadow and vegetation removed in column (d).



(a)

(b) (c) (d)

Figure 8. (a) Original image and saliency maps (upper row) and top 5% salient points (bottom row) for (b) equally weighted map, and (c) winning weighting scheme with NVSR and (d) winning scheme with VSR.

The upper row shows the corresponding saliency map, while the second row shows the 5% most salient points in each of the saliency maps extracted from the initial image based on the corresponding location in the saliency map. It can be observed that the performance for building detection is improved when weighting schemes are used rather than all unitary weights, and also with the removal of shadows and vegetation. Additional results for other images extracted from the dataset, highlighting the top 25% salient points obtained using the proposed winning weighting scheme with VSR are shown in Fig. 9. The value of 25% is selected to permit the detection of multiple buildings. These cases demonstrate the capability of the proposed method to extract areas of potential interest for building detection.

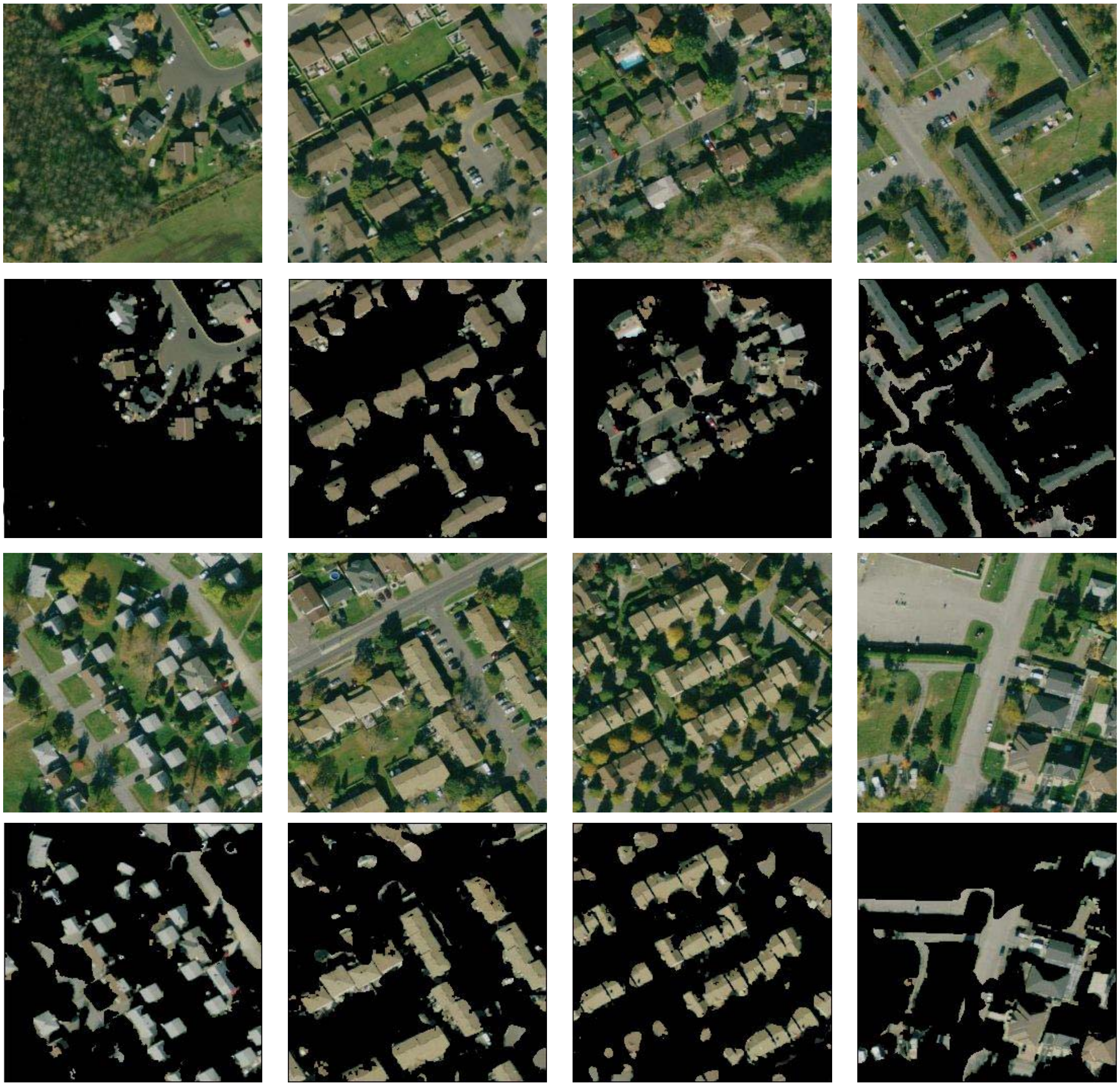


Figure 9. Results obtained using the winning schemes with VSR (even rows) over other image samples (odd rows) from the dataset for top 25% salient points.

Fig. 10 compares the results obtained with other salient region detectors from the literature and on a different image. It demonstrates how most methods do not deal well with multiple entities of objects of interest as those appearing in satellite imaging. This fact is, in general, related to the lack of context knowledge which is efficiently captured with the proposed approach. Here, the saliency is weighted adaptively in accordance with the context, defined by regions that show objects of interest during the tuning of the weights.

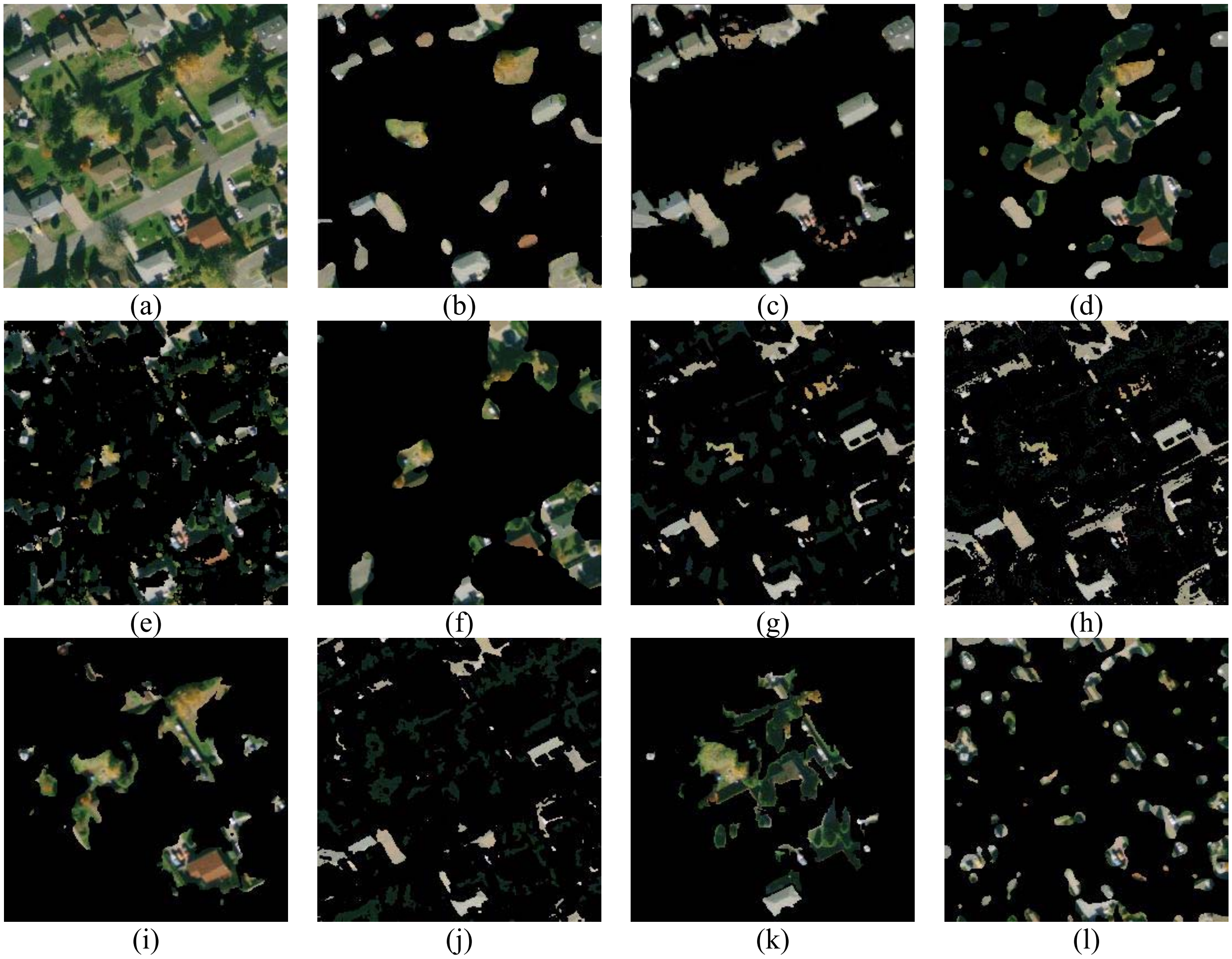


Figure 10. Comparison with other saliency detectors: (a) initial image and results for top 15% salient points using (b) proposed method with NVSR and (c) with VSR, (d) Itti's bottom-up model, (e) Frintrop's bottom-up model, (f) (Murray *et al.* 2011), (g) (Achanta and Susstrunk 2010), (h) (Zhai and Shah 2006), (i) (Goferman *et al*. 2010), (j)-(k) HC, RC (Cheng *et al.* 2011) and (l) (Hou and Zhang 2007).

To further quantify the results, the potential for building detection is evaluated for each of the methods. In order to count the number of buildings that can be detected, the top 25% saliency points are selected from the saliency map and the result is binarized (for a threshold = 0.18, that is the same value used for the identification of buildings in the reference mask image). An AND operation is performed between the reference mask image and the binarized top 25% saliency map that reveals that the buildings are successfully detected (and therefore have the potential to be recognized) in an image. All the resulting regions with an area smaller than a threshold (e.g.

150 pixels) are removed to eliminate the outliers. Some examples obtained using this procedure are shown in Fig. 11 for the same image as the one illustrated in Fig. 2d.

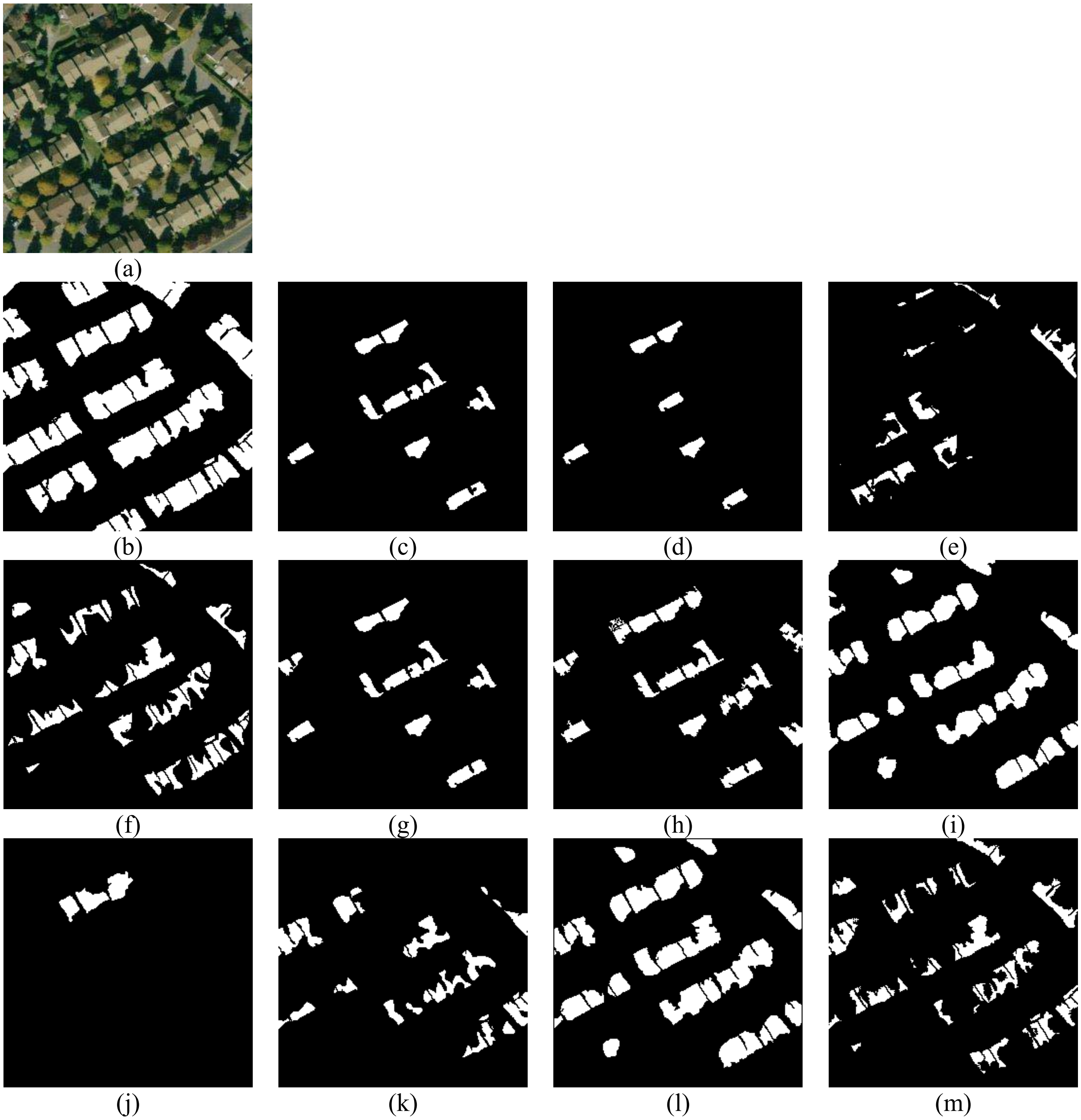


Figure 11. (a) Initial image, (b) mask of buildings in reference image and results obtained using the top 25% salient regions detected by: (c) (Achanta and Susstrunk 2010), (d) (Achanta 2009), (e) (Goferman *et al.* 2010), (f) (Hou 2008), (g) HC (Cheng *et al.* 2011), (h) (Zhai and Shah 2006), (i) proposed approach with NVSR, (j) RC (Cheng *et al.* 2011), (k) (Hou and Zhang 2007), (l) proposed approach with VSR, (m) (Zhang *et al.* 2008).

The buildings that are detected, and therefore have the potential to be further recognized within the top 25% salient points, are marked in yellow in Fig. 12 for another image extracted from the dataset. It is worth mentioning that the figure does not illustrate the results of detection, but only those parts of the maps that contain buildings, all other parts being removed due to the AND operation performed with the mask.

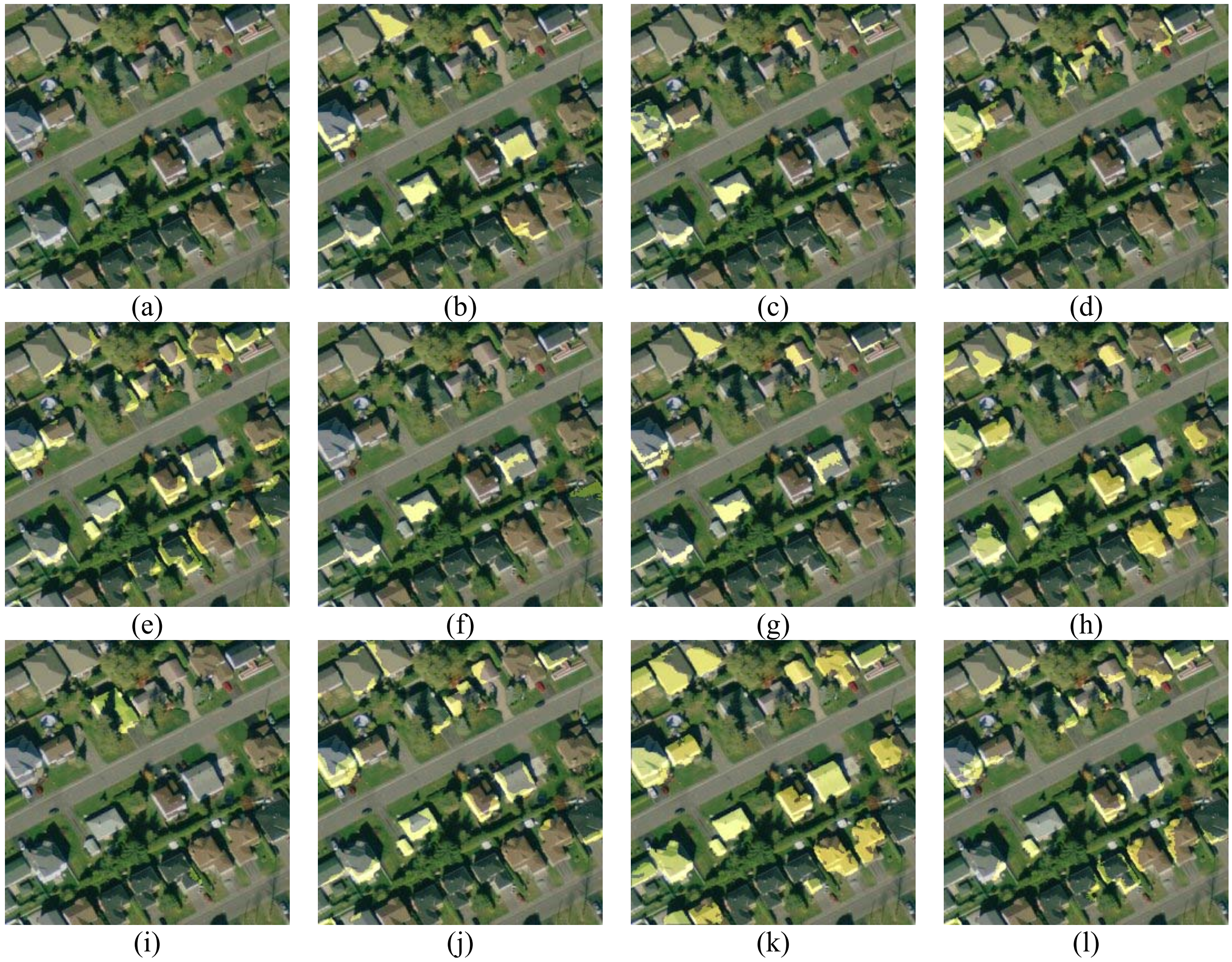


Figure 12. Buildings that can be identified in the image (a) using the top 25% salient regions detected by: (b) (Achanta and Susstrunk 2010), (c) (Achanta 2009), (d) (Goferman *et al.* 2010), (e) (Hou 2008), (f) HC (Cheng *et al.* 2011), (g) (Zhai and Shah 2006), (h) proposed approach with NVSR, (i) RC (Cheng *et al.* 2011), (j) (Hou 2007), (k) proposed approach with VSR, and (l) (Hou and Zhang 2007)

Fig. 11 and Fig. 12 show that proposed methods, particularly the ones using vegetation and shadow removal (VSR) illustrated in Fig. 11l and Fig. 12k, perform the best. To further demonstrate this, the percentage of buildings that can be successfully detected by each of the methods, based on a count of buildings in the reference mask image and in the experimental saliency map, is reported in Fig. 13.

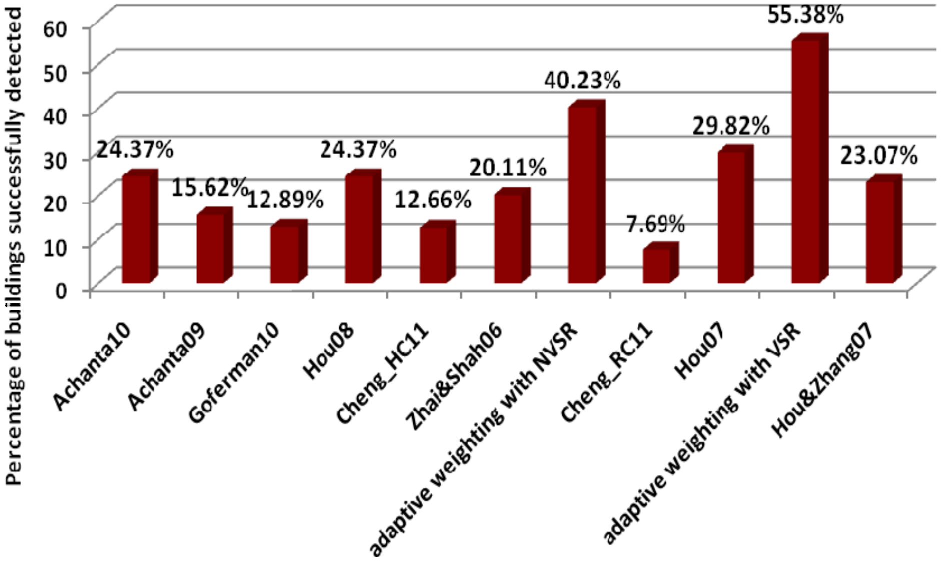


Figure 13. Percentage of buildings that are successfully detected over all images in the dataset using different saliency detectors.

In the previous results, the identification process considers that the mask of buildings is available. To monitor the applicability of the proposed solution to other images for which the related mask is unknown, the neural network strategy described in section III.e is applied to map the context of an image with an appropriate set of weights that ensures the detection of buildings. Fig. 14 considers images (shown in column 1) from the testing set. The top 25% saliency points are extracted from the initial images using the proposed adaptive selection of the best weighting scheme (with VSR) detailed in section III.d. The second column shows the results obtained when the identification of the best weighting scheme relies on the available mask of buildings. These results are compared with the regions extracted when using the neural network described in section III.e to predict the best weights for the given context, shown in the third column.

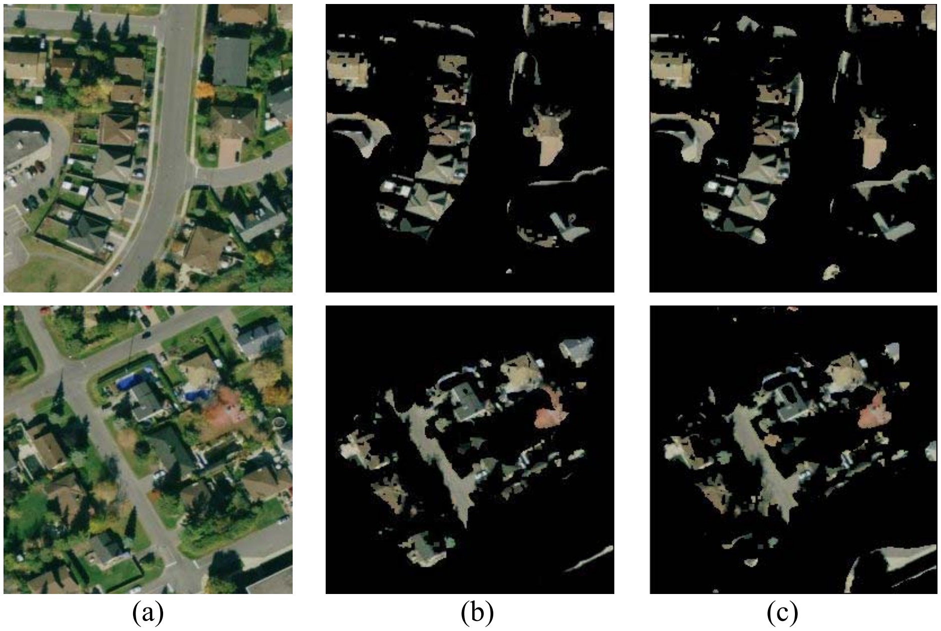


Figure 14. Examples of results obtained for top 25% salient points with best scheme selection (column 2) and when using the weights provided by the neural network (column 3).

Similar results were obtained for all tested images in the dataset. Two representative examples are provided here for compactness. One can notice that, while slightly noisier, the results provided by the neural network architecture are similar to the ones obtained using the adaptive identification of weights. The use of the neural network inference approach to dynamically adjust the respective weights applied on each feature conspicuity map allows for a reliable detection of buildings in previously unseen satellite images, therefore eliminating the need for a known mask of building locations, after a representative set of such images and masks has been provided for initial training, which can be performed offline.

V. CONCLUSIONS

The paper proposes an improved computational model of visual attention for the complex task of building detection in satellite images. A variety of original weighting schemes that capitalize on the contribution of different conspicuity maps to the desired (e.g. buildings) versus non-desired areas is initially proposed to compute an adaptively weighted saliency map which ensures that areas containing buildings or other regions of interest are highlighted. A novel adaptive algorithm is then implemented to choose the best weighting scheme, and concomitantly the best set of weights, based on a similarity error. The proposed weighting schemes are more frequently identified to be the best performing schemes when compared to schemes already proposed in the literature and also lead overall to better performance demonstrated by higher similarity with manually segmented areas of interest for a given context. Additionally, a neural network is defined and trained to predict the set of weights provided by the best weighting scheme based on the context of the image which is captured via the intrinsic energy contained in its conspicuity maps. The proposed solution demonstrates superior experimental performance to several approaches from the visual attention literature in identifying a limited set of areas of interest for building detection. Building upon the proposed strategy, and similar to the human visual system, an additional module can be introduced to further refine the recognition of buildings by eliminating remaining distractors such as trees or roads that often share similar color with the roofs. While the latter remains beyond the scope of this paper, the proposed attention model demonstrated its ability to restrict the recognition effort to only a limited number of areas of interest, which greatly speeds up the analysis of large satellite images.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the collaboration of PCI Geomatics and the financial support for the work presented in this paper from FedDev Ontario under the SME4SME program.

REFERENCES

[1] P. Wide, “Human-Based Sensing – Sensor Systems to Complement Human Perception”, *Int. Journal Smart Sensing and Intelligent Systems*, vol. 1, no.1, pp. 57 – 69, 2008.

[2] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, “Attentional Selection for Object Recognition – A Gentle Way”, *Int. Workshop Biologically-Motivated Computer Vision*, LNCS 2525, pp. 472 – 479, Springer, 2002.

[3] V. Gopalakrishnan*,* Y. Hu, and D. Rajan*,* “Salient Region Detection by Modeling Distributions of Color and Orientation”, *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892 – 905, 2009.

[4] H. Kim and W. Kim, “Salient Region Detection Using Discriminative Feature Selection”, *Advanced Concepts for Intelligent Vision Systems*, J. Blanc-Talon *et al.* (Eds.): LNCS 6915, pp. 305 – 315, Springer, 2011.

[5] Y.F. Ma and H.J. Zhang, “Contrast-Based Image Attention Analysis by Using Fuzzy Growing”, *Int. Conf. Multimedia*, vol. 1, pp. 374 – 381, 2003.

[6] L. Itti and C. Koch, “Feature Combination Strategies for Saliency-Based Visual Attention Systems”, *Electronic Imaging*, vol. 10, no. 1, pp. 161 – 169, 2001.

[7] C. Zhao and C. Liu, “Sparse Embedding Feature Combination Strategy for Saliency-Based Visual Attention System”, *Journal of Comp. Inf. Syst.*, vol. 6, no. 9, pp. 2831 – 2838, 2010.

[8] Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, and D. Rajan, “Salient Region Detection Using Weighted Feature Maps Based on the Visual Attention Model”, *Advances in Multimedia Information Processing*, LNCS 3332, pp. 993 – 1000, 2004.

[9] Y. Hu, D. Rajan, and L.-T. Chia, “Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection”, *Int. Conf. Multimedia and Expo*, pp. 1 – 4, 2005.

[10] V. Gopalakrishnan, Y. Hu, and D. Rajan, “Unsupervised Feature Selection for Salient Object Detection”, *Asian Conference on Computer Vision*, R. Kimmel, and A. Sugimoto (Eds.): LNCS 6493, pp. 15 – 26, 2011.

[11] C.T. Vu and D.M. Chandler, “Main Subject Detection Via Adaptive Feature Selection”, *Int. Conf. Image Processing*, pp. 3101 – 3104, Cairo, 2009.

[12] S. Frintrop, “VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search,” Ph.D. Thesis, Germany, 2006.

[13] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-Aware Saliency Detection“, *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 2376 – 2383, 2010.

[14] N. Murray, M. Vanrell, X. Otazu, and A. Parraga, “Saliency Estimation Using a Non-Parametric Low-Level Vision Model”, *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 433 – 440, 2011.

[15] R. Achanta and S. Susstrunk, “Saliency Detection Using Maximum Symmetric Surround”, *Int. Conf. on Image Processing*, pp. 2653 – 2656, Hong Kong, 2010.

[16] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, “ Frequency-Tuned Salient Region Detection”, *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1597 – 1604, 2009.

[17] X. Hou and L. Zhang, “Saliency Detection: A Spectral Residual Approach”, *IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 17 – 22, USA, 2007.

[18] X. Hou and L. Zhang, “Dynamic Visual Attention: Searching for Coding Length Increments”, *Conf. Neural Information Processing Systems*, pp. 681 – 688, 2008.

[19] Y. Zhai and M. Shah, “Visual Attention Detection in Video Sequences Using Spatiotemporal Cues”, *ACM Multimedia*, pp. 815 – 824, 2006.

[20] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “Global Contrast Based Salient Region Detection”, *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 409 – 416, 2011.

[21] L. Zhang, M. H. Tong, T.K. Marks, H. Shan, and G. W. Cotrell, “SUN: A Bayesian Framework for Saliency Using Natural Statistics”, *Journal of Vision,* vol. 8, no. 7, pp. 1 – 20, 2008.

[22] A.-M. Cretu and P. Payeur, “Biologically-Inspired Visual Attention Features for a Vehicle Classification Task”, *Int. Journal Smart Sensing and Intelligent Systems*, vol. 4, no. 3, pp. 402 – 423, 2011.

[23] P.K. Kaiser and R.M. Boynton, *Human Color Vision*, Washington DC, Optical Society of America, 1996.

[24] R.C. Gonzalez, R.E. Woods, and S.L. Eddins, *Digital Image Processing Using Matlab*, Upper Saddle River, NJ, Prentice Hall, 2004.

[25] N. Shorter and T. Kasparis, “Automatic Vegetation Identification and Building Detection from a Single Nadir Aerial Image”, *Remote Sensing Journal*, vol. 1, pp. 731 – 757, 2009.

[26] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design,* PWS Publishing Co., 1996.

[27] www.mapquest.com.