Building Detection in Aerial Images Based on Watershed and Visual Attention Feature Descriptors

Ana-Maria Cretu
Department of Computer Science and Engineering
Université du Québec en Outaouais
Gatineau, Canada
ana-maria.cretu@uqo.ca

Pierre Payeur
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, Canada
ppayeur@eecs.uottawa.ca

Abstract—**This paper investigates a novel solution for the recognition of objects of interest in aerial images. The solution builds on a combination of algorithms inspired from the human visual system with classical and modern algorithms. The goal is to achieve intelligent and powerful approaches that allow for fast and automatic treatment of complex images. The methodology that is proposed innovatively combines a variation of the classical watershed segmentation algorithm with a series of feature descriptors derived from a computational model of visual attention. The feature descriptors are tuned with a machine learning approach for the task of detecting buildings in aerial images. The experimental evaluation that is conducted demonstrates that objects recognition with features derived from human visual attention performs better than when only traditional features, such as statistical texture descriptors and shape descriptors, are used. As well, the proposed solution obtains better classification rates than those reported on image processing-based recognition of buildings in the remote sensing literature.**

Keywords-aerial images; feature descriptors; visual attention; object recognition; segmentation; remote sensing.

I. Introduction

Computer vision algorithms are continuously being improved for visual categorization and recognition tasks. The research on this topic has advanced over the years, resulting in more powerful algorithms that achieve better performance and increased processing speed. On the other hand, humans still show a significantly superior performance in extracting and interpreting visual information to any state-of-the-art artificial vision model. The exploitation of biological and psychological knowledge derived from human visual mechanisms was shown to contribute to the improvement of computational vision systems [1]. Early vision-inspired algorithms for object recognition, in spite of their relative novelty, have already reached performance comparable to the best computer vision systems [2] for certain domains of application and biologically-inspired visual features have been successfully applied for various tasks in image processing [1-6]. Based on the recent achievements in these two different areas of research, it is realistic to anticipate that an efficient combination of algorithms inspired from human visual system with algorithms developed in the computer vision research community will lead to the implementation of more powerful solutions for the identification of objects of interest in complex images, such as aerial and satellite data.

This paper explores an innovative combination of features extracted from a visual attention model, the classical watershed segmentation algorithm, and a machine learning approach for the detection of buildings in aerial images. The latter inherently contain a complex array of features. Multiple objects of interest with different visual properties are generally situated against a cluttered background and affected by lighting conditions and strong shadowing effects. Moreover the objects of interest often share similar characteristics with other objects that are not of interest: e.g. trees or roads often share similar color with building roofs. 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 remote sensing applications nor fully automated. As a result, the false positive rate of decision is very high and several features of interest remain undetected. The low accuracy of such algorithms when running on large collections of aerial images leaves a substantial opportunity for new approaches to improve upon the current state-of-the-art techniques in terms of performance, as those aimed in this paper.

II. Literature Review

An element that has a significant impact on the detection and recognition of objects is the determination of an appropriate set of features for a certain object or region of interest to be non-ambiguously identified, in spite of changes in its posture, scale, illumination, and background that often occur in practical applications. Several feature extraction algorithms have been proposed in the literature for the purpose of object detection and recognition including: statistical texture descriptors, such as the mean of average intensity, the smoothness of intensity, the uniformity or the skewness of the histogram [7], local-binary patterns (LBP), shape descriptors based on moment invariants [8, 9], elliptical Fourier descriptors for shape boundary description [8, 10], and other general-purpose features such as SIFT key points, Harris corners, Gabor features, Difference-of-

978-0-7695-4983-5/13 $26.00 © 2013 IEEE
DOI 10.1109/CRV.2013.8

Gaussian features, just to mention a few. The related detectors are capable to various extents to deal with changes in the object shape, characteristics and environment.

In terms of algorithms inspired from the human visual system, computational models of visual attention have been shown to ameliorate the speed of scene understanding and object recognition [4, 5, 11] by attending only the regions of interest in an image. They were identified to be capable to detect more repeatable discriminative features than other feature detectors such as corners or SIFT key points [4]. Previous work of the authors [12] showed as well that features derived from a computational model of visual attention achieve better performance for pattern recognition than Harris corners, Gabor features, and Difference-of-Gaussian features.

In terms of approaches proposed in the remote sensing literature for building identification based on computer vision, Persson *et al.* [13] train an ensemble of self-organizing maps to recognize red, copper, light and dark building roofs. Their experimentation is limited, with results being reported on 17 buildings only. In [14], buildings are detected based on color invariants and shadow information, but the illumination angle is calculated based on the assumption that the roofs have a red-brown color, which is not always the case. The approach of Liu and Prinet [15] uses a set of features, such as shadow ratio, shape feature, distance to straight lines and entropy, and a probability function to identify buildings in high resolution satellite images. Their shadow model however is limited since it cannot be generalized from one image to another. In [16], vegetation, building and non-building objects are identified based on the assumption that buildings have convex rooftop sections. The solution is based on color segmentation and color invariants for the identification of shadow and vegetation areas.

III. Proposed Approach for Building Detection in Aerial Images

The proposed approach for building versus non-building categorization in aerial images, illustrated in Fig. 1, can be summarized as follows: a reduced dataset of aerial images is initially presented to the system for training. A watershed algorithm is used for the initial segmentation of each image. Areas corresponding to vegetation and shadow are then eliminated based on color invariants, using an approach similar to [14, 16]. This procedure is further described in section III. A. The remaining watershed segments and parts of watershed segments are then classified as buildings, streets or distractors, based on manually segmented masks for buildings and streets, as those illustrated in the second row of images of Fig. 1, which are made available to the training stage. However, aerial images of residential areas contain multiple distractors that are neither streets, nor buildings, nor shadows or vegetation that can be eliminated based on color invariants. Such distractors can include pools, vegetation of different color, driveways, etc. Such distractors are included in the training of the system to improve the identification of buildings. In order to identify distractors in images, after the removal of vegetation and shadow is performed in a given watershed segment, the latter is masked with both the building and the street masks used together, as shown in the first row of images in Fig. 1. All what remains after this operation is considered a distractor.

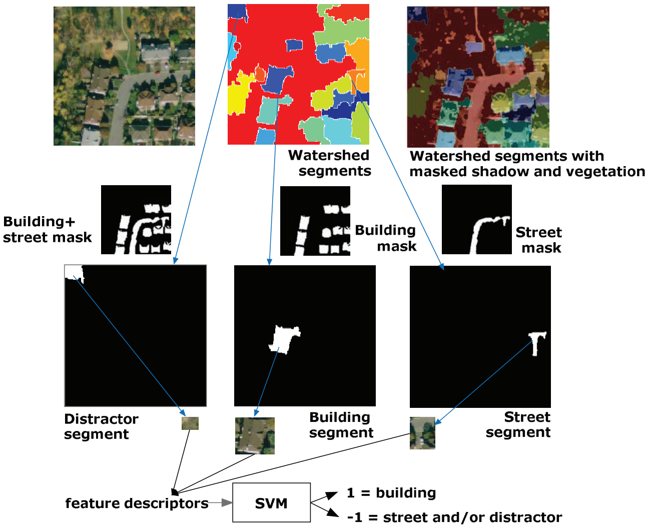

Figure 1. (a) Proposed approach using watershed segmentation and a binary SVM for building versus non-building (street and/or distractor) identification in aerial images.

A bounding box is built around each remaining watershed segment and/or fragment of a watershed segment (that can result after the removal of vegetation and shadow areas) from the building, street and distractor category respectively. The corresponding area (sub-image) of each of these bounding boxes is recuperated from the initial image. The sub-images are therefore rectangular areas of the initial color image defined by the coordinates of the bounding boxes surrounding the watershed segments and/or fragments. This allows for the computation of a feature descriptor that characterizes the content of each sub-image, which actually represents one of the different categories of objects of interest: building, street or distractor. Descriptors derived from visual attention, but also a series of statistical texture descriptors and shape descriptors are calculated on each sub-image as further detailed in section III.B, to enable a comparison with the proposed descriptors. Such a descriptor is assembled in a vector. A support vector machine, detailed in section III.C, is finally used to classify each feature descriptor vector as describing a building or a non-building.

A. *Segmentation Based on the Watershed Algorithm*

The marker-controlled watershed segmentation algorithm is adapted from [7]. Local regional maxima are used as internal markers, along with a series of morphological opening and closing operations to remove small disconnected areas. The use of the local intensity standard deviation of the image, a measure of average contrast, as an external marker allows a better definition of contours around the objects. This marker-controlled algorithm is used for the initial segmentation of the grayscale equivalent image of each aerial image.

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 aerial images, as proposed in [14, 16]. The color invariant, *v*, which defines vegetation areas, builds upon the green (*G*) and blue (*B*) channels of the color aerial image and is defined as:

$$v = \frac{4}{\pi} \cdot \arctan\left(\frac{G-B}{G+B}\right) \tag{1}$$

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

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

The resulting images of these invariants are then Otsu thresholded to decide what pixels belong to vegetation and shadow areas respectively. The resulting pixels representing vegetation and shadows are used to mask the watershed segments, as detailed below, for each of building, street and distractor category. Fig. 2b shows an example of watershed segments resulting after the application of the watershed segmentation, over the grayscale transformed initial color image. Fig. 2d shows the segments after the vegetation and shadow removal.

If the result of this masking operation results in bounding box areas larger than a certain threshold, it means that the watershed segment could contain objects of interest and is further processed. For example, for buildings and streets, for each resulting watershed segment and fragment, from the independent components that results after the vegetation and shadow removal, only the ones that have the area of the bounding box between 150 and 3000 pixels are retained. The two values correspond to the minimum and maximum areas of the bounding boxes of buildings, as computed from the manually segmented masks, and they would have to be adjusted for another dataset. This is nevertheless a very simple operation once the masks are available. In Fig. 2, all the independent components remaining after the removal of vegetation and shadow areas, marked by green bounding boxes in Fig. 2e have an area between 150 and 3000 pixels, and are therefore further processed. To identify building areas, these components are masked with the building mask, shown in Fig. 2f. If more than 35 pixels (experimentally determined value) remain in the newly masked image, it means that the component contains a building and its bounding box coordinates are retained. The building areas are marked by red rectangles in Fig. 2h. A similar approach is used for the streets as for the buildings, by using instead a manually segmented mask for streets. The street mask for the image in Fig. 2a is illustrated in Fig. 2g and the area corresponding to streets in Fig. 2 is marked by a blue rectangle in Fig. 2h. The other larger part of the street is not marked in the image because its area is larger than the 3000 pixels threshold.

In order to identify the distractors in images, after the removal of vegetation and shadow in a given watershed segment, the segment is masked with both the building and the street masks used together. All the remaining components are considered distractors. These are marked by yellow bounding boxes in Fig. 2h and the coordinates of their bounding boxes are retained as well.

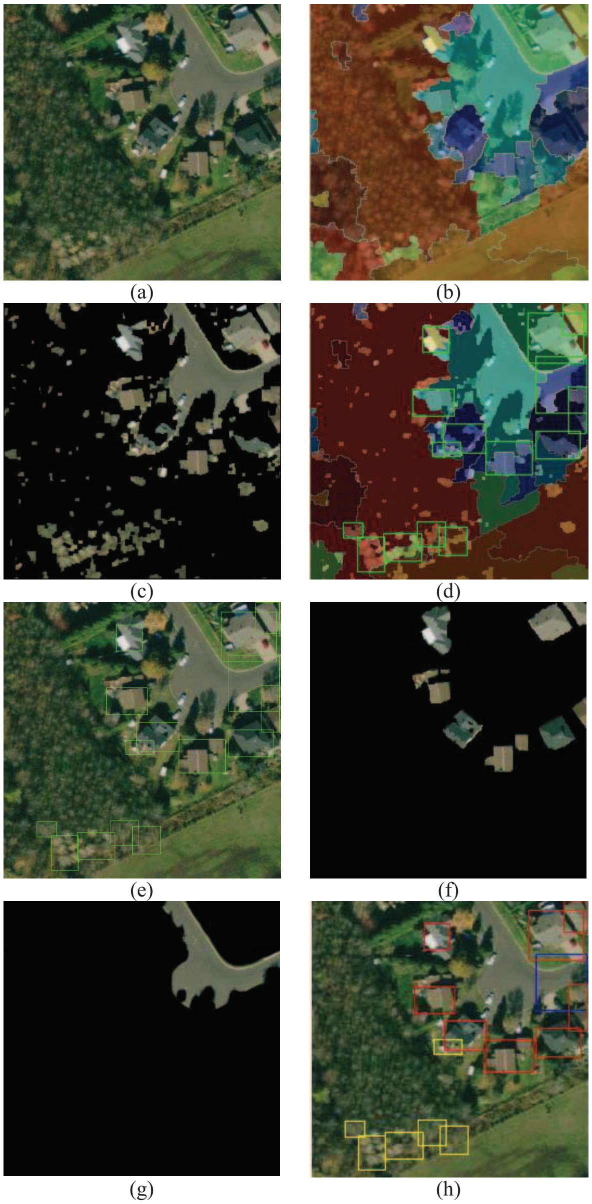

Figure 2. (a) Initial image (b) watershed segments, (c) image after the removal of vegetation and shadow, (d) watershed segments with vegetation and shadows masked, (e) resulting areas to be used in training, (f) building mask, (g) street mask and (h) resulting areas to be used in training and their corresponding class (red=building, blue=streets, yellow=distractors) based on masking with the building and street masks.

B. Descriptors for Watershed Segments

Visual attention-inspired descriptors and also descriptors based on texture and shape information are used to encode the properties of the extracted rectangular regions in section III.A that represent different objects of interest in aerial images.

1) Descriptors Derived from Visual Attention: The descriptors derived from visual attention are built based on the local saliency map obtained by the classical computational attention model of Itti *et al.* [17]. Such a saliency map is computed on each of the defined sub-images of objects extracted from an aerial image as explained in section III.A. The main idea behind all the bottom-up computational systems proposed in the literature in general, and for Itti *et al.*'s system in particular, is to compute several features derived from a color image provided as input and fuse their saliencies into a representation called saliency map [17]. 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. Itti's computational attention model considers as features the intensity I = (R+G+B)/3 where R, G and B are the red, green and blue color channels respectively; color (color maps are represented by the RG and BY color opponency); and orientation (local orientation information is obtained from the intensity image I using oriented Gabor pyramids of different scales and different preferred orientations, e.g. 0^o, 45^o, 90^o and 135^o in the current work). 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. After normalization, these maps are summed up linearly in the final saliency map. The full implementation details are available in [17].

Due to the fact that buildings are usually much smaller than the size of an entire aerial image, the sub-images are forced to higher resolution (e.g. 128×128) prior to the application of computational visual attention model to ensure that their characteristics are properly encoded. The resulting saliency map is then Otsu thresholded, and downsampled to a map of size 16×16 to ensure better classification rates. The size of 16×16 is chosen based on trial-and-error. The effect of alternative downsampling sizes is reported along with the results of the classification in section IV. Finally, the downsampled map is transformed into a feature vector that contains, for each image, 16×16 = 256 binary values.

2) Statistical Texture Descriptors: A set of 6 descriptors of texture based on the intensity histogram of local regions [7] are examined, namely the mean of average intensity, the standard deviation, the relative smoothness of the intensity in a region, the third moment, which a measure of the skewness of the histogram, the uniformity, and the entropy, a measure of randomness. The mean of average intensity is computed as:

$$m = \sum_{i=0}^{L-1} z_i p(z_i) \tag{3}$$

where z_i is a random variable indicating intensity, $p(z)$ is the histogram of the intensity levels in a region, L is the number of possible intensity levels. The standard deviation is used as a measure of average contrast:

$$\sigma = \sqrt{\mu_2(z)} = \sum_{i=0}^{L-1} (z_i - m)^2 p(z_i) \tag{4}$$

The relative smoothness is 0 for constant intensity and approaches 1 for a region with large variations in the values of its intensity level, and is computed as:

$$R = 1 - 1/(1+\sigma^2) \tag{5}$$

The third moment measures the skewness of a histogram:

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i) \tag{6}$$

The measure is 0 for symmetric histograms, positive for histograms skewed to the right (about the mean) and negative for histograms skewed to the left. The previous two measures in eq. (5) and (6) are both divided by $(L-1)^2$ in order to be brought into a range of values comparable to the other five measures.

The uniformity measure, computed as:

$$U = \sum_{i=0}^{L-1} p^2(z_i) \tag{7}$$

is maximum when all gray levels are equal.

The final measure of randomness is the local entropy value, e, computed in a 9-by-9 neighborhood around the corresponding pixel in the input image as in [4].

$$e = -\sum_{i=0}^{L-1} p(I_i) \log_2 p(I_i) \tag{8}$$

where $p(I_i)$ is the histogram of the intensity levels in the region and L the number of possible intensity levels (e.g. L=256 for experimentation).

The six measures are concatenated in a texture descriptor vector.

3) Shape Descriptors: The shape can be encoded in the arrangement of its pixels and captured by moment invariants or in its boundary description by means of elliptical Fourier descriptors. Both types of descriptors are considered.

a) Region descriptors based on moment invariants: Moments describe the shape's layout or the arrangement of its pixels and are global descriptors of a shape. Legendre moments [9] and the Hu moment invariants [10] are tested as pattern descriptors for each sub-image. The Legendre moment invariants are uniform contrast invariant. They are not rotation invariant, but they can be made affine invariant. A value of 45 Legendre moments was identified during experimentation to provide the best results in term of rates of precision, recall, specificity and accuracy. Hu invariants are invariant to rotation, translation and uniform scaling. A number of 7 Hu moments are used in the experimentation. Each sub-image is therefore encoded in a 45 value vector by Legendre moments and in a 7 value vector by the Hu moments.

b) Elliptical Fourier descriptors for boundary description: Fourier descriptors allow for the characterization of the contour of a shape by a set of numbers that represent the frequency content of the whole shape [8]. Due to their ability to map arbitrary shaped contours, elliptical Fourier descriptors are chosen in the context of this application to represent the shape of objects in aerial images. For the computation of elliptical Fourier shape descriptors, a more elaborate procedure is followed, because they require the contour of the corresponding object in each sub-image extracted from the initial image. After the substraction of the vegetation and shadow areas, the sub-image is binarized and from the components, if more than one exists (e.g. when a building is close to other buildings whose parts are included in the same bounding box), the one with the largest area is retained only. The boundary of this largest area is then used as an input contour for the elliptical Fourier descriptors. A number of 7 harmonics are used to describe this contour and the results are concatenated in a vector of 28 (7×4) values.

C. Training of a SVM for Building versus Non-building Recognition

The set of input vectors is assembled by concatenating all the building, street and distractor vectors for each type of descriptor separately and for all the images in the training dataset. The inputs are then classified using a least-squares support vector machine (LSSVM) [18]. A LSSVM classifier with a Gaussian RBF kernel, the regularization parameter γ=10 and the squared bandwidth σ_2=0.4 is used. One binary SVM is trained for each type of descriptor: visual attention based, texture descriptors, Legendre moments, Hu moments, and Elliptical Fourier, to classify between buildings on one side (the corresponding output is 1) and streets and distractors on the other side (the corresponding output is -1). The input and output datasets are shuffled in random order to improve the training. The training for buildings, running on Matlab, takes about 0.09s per image.

A five fold cross-validation procedure is used train and to compute the performance measures. 80% of the dataset in each fold is used for training and 20% for testing. The performance is reported in terms of precision (completeness, *TP/(TP+FN)*), recall (correctness, *TP/(TP+FP)*), specificity *TN/(TP+TN)* and accuracy *(TP+TN)/(TP+TN+FP+FN)*, where *TP* = true positives are buildings identified as buildings, *FP* = false positives are distractors or streets identified as buildings, *TN* = true negatives are distractor and street identified as such and *FN* = false negatives are buildings that are erroneously identified as street or distractor.

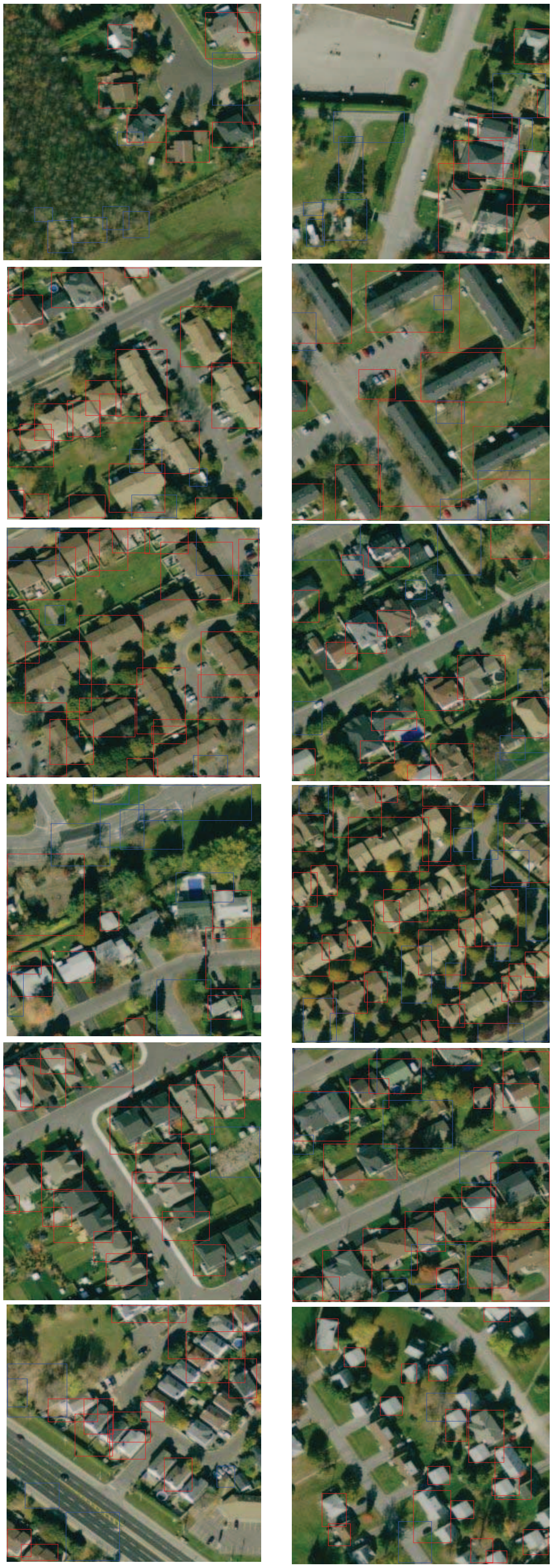

Figure 3. Detected buildings (in red) for images in the test set.

IV. Experimental Results And Evaluation

The proposed approach is tested on a dataset of 50 aerial images [19]. Each has a resolution of 256×256 pixels and contains residential areas with different topologies and complexities. Overall, the dataset contains 845 buildings. Fig. 3 shows samples from the test set with the recognized buildings shown in red and the streets and distractors in blue, when visual-attention descriptors of 256 values (resulting from a saliency map of 16×16) are used to encode the object information. One can notice that the approach provides good results in most of the cases. A limited number of buildings are not separated by the watershed segmentation algorithm and therefore fail to be identified, as it can be observed in the last row of Fig. 3.

Table I reports the average performance over the five folds for different sizes of the saliency map.

TABLE I. Performance For Different Saliency Map Sizes

Descriptors	Recall	Precision	Specificity	Accuracy
Visual Attention 4×4	86.5%	83.45%	89.43%	88.32%
Visual Attention 8×8	97.18%	91.37%	94.33%	95.41%
Visual Attention 16×16	97.37%	93.90%	95.80%	95.99%
Visual Attention 32×32	97.20%	93.62%	95.84%	96.35%
Visual Attention 64×64	97.02%	94.03%	96.17%	96.5%

It can be observed that overall the performance does not significantly increase with larger saliency maps after the size of 16×16 for the downsampled saliency map. Until that value there is an increase of more than 2% in at least one measure of performance (e.g. the precision). For larger sizes than 16×16 a very slight decrease in the recall can be noticed, due to larger sizes of input vectors while no significant improvement occurs in the other measures. Table II and Fig. 4 illustrate the comparative performance when different descriptors are used.

TABLE II. Performance of Different descriptors

Descriptors	Recall	Precision	Specificity	Accuracy
Statistical texture	71.51%	85.24%	91.48%	83.3%
45 Legendre Moments	73.06%	91.02%	95.16%	86.07%
Hu Moments	60.99%	78.53%	88.61%	77.4%
Elliptical Fourier	99.36%	90.85%	90.85%	94%
Visual Attention 16×16	97.37%	93.9%	95.8%	95.99%

It can be noticed that the elliptical Fourier descriptors and the visual attention descriptors provide the best results, reflected by the high percentages for recall, precision, specificity and accuracy.

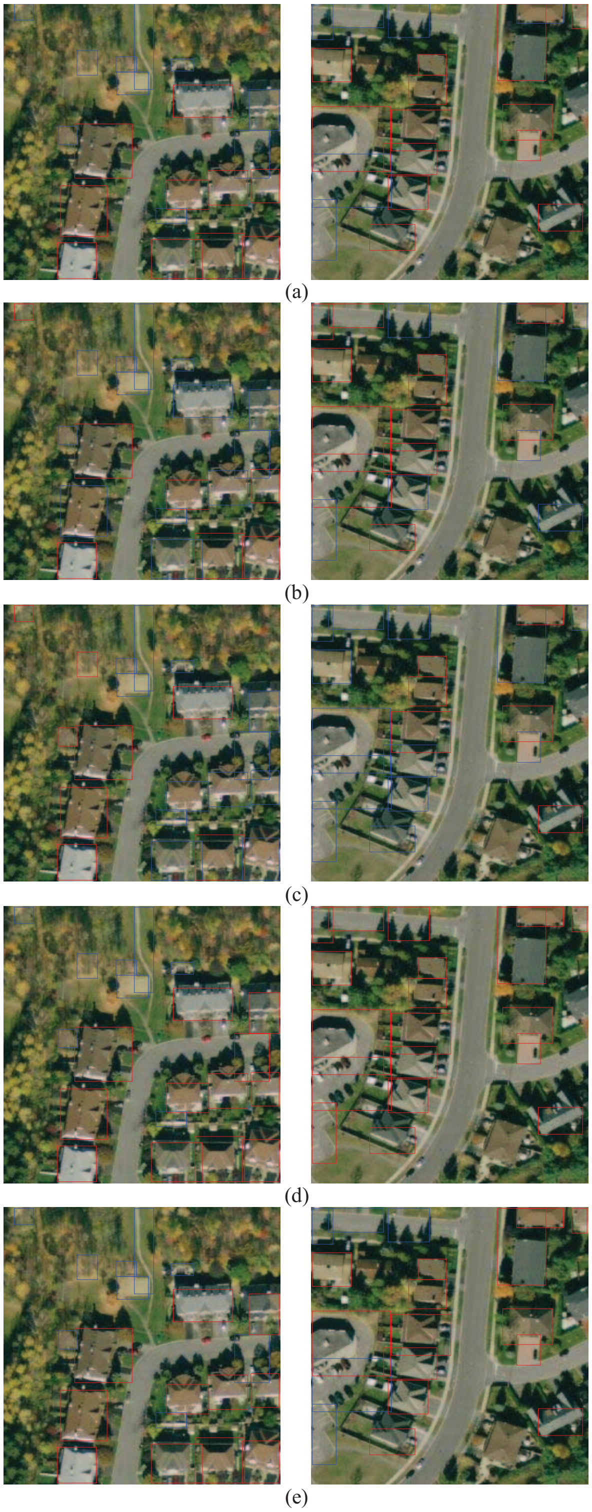

Figure 4. Detected buildings, shown in red, and detected streets and distractors, shown in blue, using different descriptors: (a) statistical texture descriptors, (b) Legendre moments, (c) Hu moments, (d) elliptical Fourier descriptors and (e) visual attention-based descriptors.

At the same time, the elliptical Fourier descriptors tend to provide more false positives, a fact demonstrated both by the lower percentage in the specificity column of Table II and by comparing visually the results illustrated in Fig. 4d and Fig. 4e. In Fig. 4d, many bounding boxes belonging to streets are classified as buildings.

Overall, for all sample images considered in this work, the best performance is achieved by the visual attention-based descriptors. It is nevertheless worth mentioning that the values shown in Table I and Table II report on the performance on the training and testing dataset derived from the watershed segments, and are therefore biased by the effectiveness of the segmentation stage. Additional testing is performed to further study the average performance of building detection with visual attention descriptors when the number of detected buildings is not computed solely from the segments of the watershed algorithm. Instead, the number of buildings is first computed for each image in the test set, based on the corresponding and manually extracted building mask. The number of streets and distractors is computed as well, based on the street, and the building and street masks, respectively. These numbers that take into account finer definitions of the buildings are compared with the results in terms of buildings versus streets and distractors obtained by the LSSVM. The performance is estimated and reported in Table III as an average over the five folds.

Table III shows that the average performance is only slightly altered by the watershed segmentation procedure, when it is considered for providing the reference number of buildings detected. This is reflected by the slightly lower percentages in Table III as compared to the last row of Table II. To further evaluate the proposed approach, Table IV compares the performance for building detection of the proposed approach with other results reported in the remote sensing literature for building detection based on computer vision solutions. While it is quite difficult to perform a comparison as no standard dataset for remote sensing data exists and the aerial images on which these solutions are tested are different, one can notice that the results obtained with the proposed visual-attention descriptors are very promising.

TABLE III. PERFORMANCE FOR BUILDING DETECTION WITH VISUAL ATTENTION DESCRIPTORS

	Recall	Precision	Specificity	Accuracy
Proposed approach with visual attention descriptors 16×16	92.62%	96.43%	92.26%	92.09%

Additional testing is performed to check if a recursively applied watershed segmentation could further improve the performance of the proposed system in an attempt to address the issue where some buildings are not separated in individual entities, as illustrated in the last row of Fig. 3. In this case, all resulting watershed segments are considered, even those larger than 3000 pixels that are not treated in section III.C.

TABLE IV. PERFORMANCE COMPARISON FOR BUILDING DETECTION FROM OTHER APPROACHES IN THE LITERATURE

Building detection average performance reported	Recall	Precision	Number of buildings
Persson et al. [13]	82.0%	N/A	17
Sirmacek and Unsalan [14]	86.6%	N/A	177
Liu and Prinet [15]	94.5%	83.4%	277
Shorter and Kasparis [16]: all buildings	55.4%	48.2%	2643
Shorter and Kasparis [16]: buildings of area 50sq.m or more	77.3%	64.4%	1414
Shorter and Kasparis [16] : buildings of area 210sq.m or more	91.8%	44.5%	306
Proposed approach with visual attention descriptors	92.6%	96.4%	845

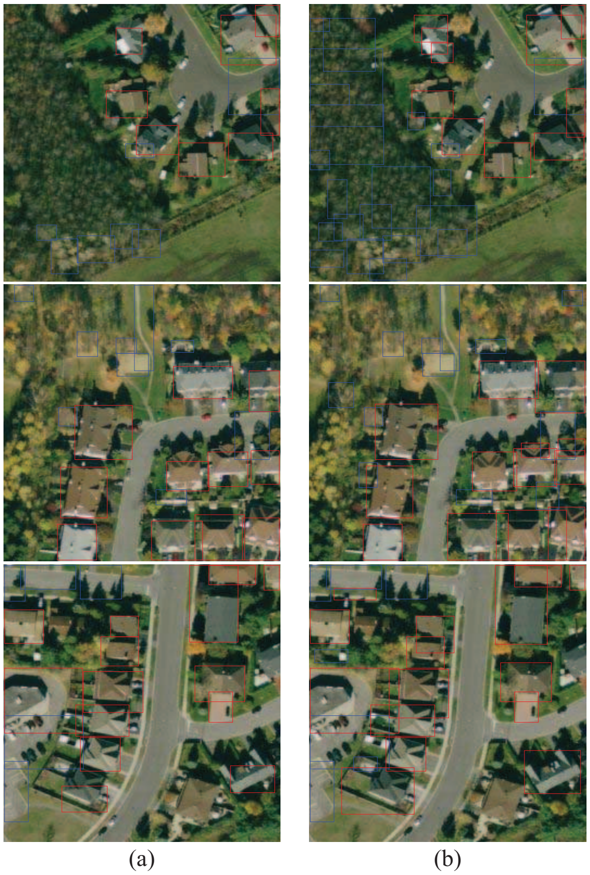

Figure 5. Detected buildings, shown in red, and detected streets and distractors, shown in blue for (a) simple watershed, (b) recursive watershed.

These larger segments are further examined by a recursive watershed algorithm, applied within the limited bounding box of the segment. While more computationally expensive, due to the testing of a larger number of boxes, the recursive application of the algorithm does not necessarily bring an improved performance for building detection. No previously missed buildings are identified as shown in Fig. 5 that compares on different images the results obtained using simple watershed (Fig. 5a) and recursive watershed (Fig. 5b). Moreover some buildings are identified in multiple boxes and some false positives appear as well, as illustrated in Fig. 5b. Therefore the recursive watershed is not a viable alternative for performance improvement.

Conclusion

This paper proposed an original combination of features extracted from a visual attention model with the classical watershed segmentation algorithm, and a support vector machine for the detection of buildings in aerial images. The experimental results showed that descriptors based on visual attention encode an appropriate set of features for the objects of interest to be non-ambiguously identified, in spite of changes in posture, scale, illumination, and background that occur in the dataset. These descriptors were also shown to lead to higher classification rates than other descriptors such as statistical texture descriptors, moment invariants and elliptical Fourier descriptors. Furthermore, a better performance was achieved when compared to reported solutions based on computer vision for remote sensing applications.

References

[1] T. C. Kietzmann, S. Lange, and M. Riedmiller, “Computational Object Recognition: A Biologically Motivated Approach”, *Biological Cybernetics*, vol. 100, pp. 59-79, 2009.

[2] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, “Robust Object Recognition with Cortex-Like Mechanisms”, *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 29, no. 3, pp. 411-426, 2007.

[3] E. Meyers and L. Wolf, “Using Biologically Inspired Features for Face Processing”, *Int. Journal Comput. Vis.*, vol. 76, pp. 93–104, 2008.

[4] S. Frintrop and P. Jensfelt, “Attentional Landmarks and Active Gaze Control for Visual SLAM”, *IEEE Trans. Robotics*, vol. 24, no. 5, pp. 1054-1065, 2008.

[5] C. Siagian and L. Itti, “Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300- 312, 2007.

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

[7] R.C. Gonzalez, R.E. Woods and S.L. Eddins, *Digital Image Processing Using Matlab*, Prentice Hall, 2004.

[8] M. Nixon and A. Aguado, *Feature Extraction & Image Processing*, Elsevier, 2009.

[9] J. Flusser, T. Suk, and B. Zitova, *Moments and Moment Invariants in Pattern Recognition*, John Wiley & Sons, 2009.

[10] F.P. Kuhl and C.R Giardina, “Elliptic Fourier Features of a Closed Contour”, *Computer Graphics and Image Processing*, vol. 18, pp. 236-258, 1982.

[11] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, “Attentional Selection for Object Recognition”, *Proc. Biologically Motivated Computer Vision, LNCS*2525, pp. 472-479, 2002.

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

[13] M. Persson, M. Sandvall, and T. Duckett, “Automatic Building Detection from Aerial Images for Mobile Robot Mapping”, *Proc. IEEE Int. Symposium Computational Intelligence in Robotics and Automation,* pp. 273-278, 2005.

[14] B. Sirmacek and C. Unsalan,“Building Detection from Aerial Images using Invariant Color Features and Shadow Information”, *Proc. Intl. Symposium Computer and Information Sciences*, Istanbul, Turkey, pp. 105-110, 2008.

[15] W. Liu and V. Prinet, “Building Detection from High-Resolution Satellite Images using Probability Model”, *Proc. IEEE Int. Geoscience and Remote Sensing Symposium*, Seoul, pp. 3888-3891, 2005.

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

[17] L. Itti, C. Koch, and E. Niebur, “A Model of Saliency-based Visual Attention for Rapid Scene Analysis”, *IEEE Trans. PAMI,* vol. 20, no. 11, pp. 1254-1259, 1998.

[18] Least-Squares Support Vector Machines (LSSVM) Matlab Toolbox, available online, http://www.esat.kuleuven.be/sista/lssvmlab/.

[19] *, [Online], Available: www.mapquest.com