Scene Classification with a Sparse Set of Salient Regions

Ali Borji Laurent Itti

Abstract— **This work proposes an approach for scene classification by extracting and matching visual features only at the focuses of visual attention instead of the entire scene. Analysis over a database of natural scenes demonstrates that regions proposed by the saliency-based model of visual attention are robust to image transformations. Using a nearest neighbor classifier and a distance measure defined over the salient regions, we obtained 97.35% and 78.28% classification rates with SIFT and C2 features from the HMAX model at 5 salient regions covering at most 31% of the image. Classification with features extracted from the entire image results in 99.3% and 82.32% using SIFT and C2 features, respectively. Comparing attentional and adhoc approaches shows that classification rate of the first approach is 0.95 of the second. Overall, our results prove that efficient scene classification, in terms of reducing the complexity of feature extraction is possible without a significant drop in performance.**

I. Introduction

Scene classification is a fundamental problem in image understanding. Automatic techniques for associating scenes with semantic labels have a high potential for improving the performance of other computer vision applications such as browsing, retrieval and object recognition. To do robust scene classification, two steps are necessary. The first one concerns the scene representation, that is, how to efficiently extract effective representations from visual input whereas second step focuses on algorithms and classifiers to process these representations. Regarding the first step, it is much desired to develop highly robust features with least computational complexity.

Many approaches for scene classification have been developed which can be classified into the following three categories. 1) Low-level feature based schemes: which represent scenes by global visual information [7], including color, texture, and shape have been successfully utilized in indoor/outdoor, city/landscape and forest/mountain applications. 2) Local feature based schemes: represent scene images with detected interest points (or regions) based on some descriptors [33] [11] [5]. Local-global features [6] based schemes utilize both the global spatial information and the local descriptors of interest points (or regions) to represent scene images to achieve robust classification. 3) Biologically inspired feature based schemes: classify scenes by mimicking the process of visual cortex in recognition tasks. Recent reports from both neuroscience and computer vision have demonstrated that biologically plausible features [20] are attractive in visual recognition. In [21] [13], authors have proposed scene classification methods based on the idea of "gist" which is exquisite ability of humans at instantly recognizing a scene; for example, following a presentation of a photograph for just a fraction of a second, an observer correctly report that it is an indoor kitchen scene with numerous colorful objects on the countertop.

Authors are with Department of Computer Science, University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, California, 90089-2520, USA. {borji,itti}@usc.edu

Fig. 1. An example image with its saliency map overlaid (bottom) (derived from basic saliency model [18]).

Our work in this paper falls in the third category and aims to apply the inherent capability of human brain known as attention to outdoor scene classification. Visual attention is the capability of biological vision systems by which humans and animals select most salient regions of a scene to later concentrate higher vision tasks on those areas. This explains to some extent high efficiency of humans in every day detection and recognition tasks. Saliency means that a subset of image regions conveys more information for a biological creature, to optimize its behavior (see Fig. I). Saliency can be determined by bottom-up, image-based characteristics which are mainly derived by the early visual areas like V1 and V2 and also top-down task-driven cues from higher areas like LIP, V4 and PFC.

From a theoretical perspective, attention is a solution for the problem of information overloading and computational complexity. This problem has attracted many concerns in

computer vision and cognitive robotics where large amount of data has to be processed efficiently to guarantee a limited response time. It is also of interest to produce human-like behavior in human-computer interaction applications [2].

While a large body of research has been focused on understanding biological underpinnings of visual attention and modeling it, only recently applications have started to emerge in computer vision and robotics. Attention has been incorporated to solve complex visual tasks like 3D reconstruction [32], robot navigation and localization [22] [23], scene [13], object and face recognition [12] [11], image rendering [25], image thumbnailing [27][28], video shot detection [29] and attention-based object detection where saliency has been extensively used to improve efficiency [35] [36] [37].

Several computational implementations of models of visual saliency have been published in last decade [23]. Example are AIM (Attention based on Information Maximization) [15], Incremental Coding Length (ICL) model [16], Surprise model [30], Saliency Using Natural statistics (SUN) model [31], Extended saliency (E-Saliency) model [26]. However, little research has been reported in investigating benefits of saliency concept for scene and object classification in natural environments [24]. In this paper, applicability of saliency-based model of visual attention is examined for scene classification. We first investigate to what extent images contain similar visual content at salient regions. We furthermore aim to demonstrate how this repeatability can be utilized for outdoor scene classification.

Repeatability of a different class of saliency, discriminative saliency, defined by Gao *et al.* was studied in [4]. They speculate that saliency serves to maximize recognition. Using SVM classification, they have shown that discriminate saliency detector (DSD) leads to higher recognition rate than Scale Saliency Detector (SSD) [1] and Harris Saliency Detector (HSD) over face, motorbike and airplane datasets. They have also compared the repeatability of DSD with Hessian-Laplace (HesLap) [9] and Maximally Stable Extermal (MSER) [10] detectors.

Local image descriptors have become a very powerful representation of images in categorization and recognition tasks. Much of their success is due to their distinctiveness and to the fact, that this type of image representation is robust to occlusions and affine transformations. While in some applications like 3D reconstruction a large number of local features is needed, in some others only a few suffices. In many real-world applications where computational resources are limited, a small number of descriptors are preferred because it greatly reduces the complexity of matching algorithms. Walther et al. [24] have proposed a method for multiple object recognition in cluttered scenes using SIFT features. First a salient region is detected using the basic saliency model and then a patch is grown to fit the object extent. SIFT feature extraction is done at this patch for matching the object to the learned objects. Similar to this work, we also extract SIFT features at some salient spatial region but for the purpose of scene recognition.

From another perspective the proposed algorithm reduces the computational complexity of SIFT feature extraction algorithm. The SIFT algorithm generates a large number of features, e.g., about 2000 for a 500×500 image, which is very time consuming. There have been attempts to speed up this process by modifying the way features are calculated [3]. We claim that it is possible to extract SIFT features at a subset of spatial locations without a major drop in performance. Instead of feature extraction over the whole image, features are extracted over some moderately stable salient regions and are then matched.

Our scene classification algorithm consists of three integrated modules 1) a saliency detector for selecting a subset of scene regions 2) a matching algorithm which matches image content at the salient regions and returns a distance value between two images and 3) a classifier for class label prediction. All these stages to some extent agree with biological findings.

A. Saliency-based Model of Visual Attention

Several studies in computer vision have previously investigated the paradigm of visual attention and now there is an accepted model known as the basic saliency-based model [18] which copes with known biological concepts to some extent. This model generates a 2D topographical map that encodes stimulus conspicuity or saliency at every scene location. Saliency map is constructed as follows. The input image I is sub-sampled into a Gaussian pyramid [14], and each pyramid level is decomposed into channels for red (R), green (G), blue (B), yellow (Y), intensity (I), and local orientation (O_θ), then $R = r - (g + b)/2$, $G = g-(r+b)/2, B = b-(r+g)/2$, and $Y = r+g-2(|r-g|+b)$ (negative values are set to zero). Local orientations (O_θ) are obtained by applying steerable filters to the images in the intensity pyramid I. From these channels, center surround "feature maps" are constructed and normalized:

$$\begin{aligned} F_{I,c,s} &= N(|M_I(c) \ominus M_I(s)|), I = \{I\} \\ F_{C,c,s} &= N(|M_C(c) \ominus M_C(s)|), C = \{RG, BY\} \\ F_{O,c,s} &= N(|M_O(c) \ominus M_O(c)|), \\ O &= \{0^0, 45^0, 90^0, 135^0\} \end{aligned} \quad (1)$$

Where $\ominus$ denotes the across-scale difference between two maps at the center (c) and the surround (s) levels of the respective feature pyramids. $N(.)$ is an iterative normalization operator (refer to [19] for details). The feature maps are summed over the center-surround combinations using across-scale addition (Σ) and the sums are normalized again:

$$F_l = N(\sum_{c=2}^{4} \sum_{s=c+2}^{c+4} F_{l,c,s}), l \in \{I, C, O\} \quad (2)$$

For the general features color and orientation, the contributions of the feature dimensions are linearly summed and normalized once more to yield "conspicuity" maps:

$$C_I = F_I, C_C = N(\sum_{l \in C} F_l), C_O = N(\sum_{l \in O} F_l) \quad (3)$$

All conspicuity maps are combined into a final saliency map: $S = \frac{1}{3} \sum_{k \in \{I,C,O\}} C_k$.

Maxima of the final saliency map are the focuses of attention. Despite the main model where maxima are sequentially detected to generate a sequential attention behavior, we use a simple max operation to speed up the process. For a 640×480 image, it takes about 80ms to extract 5 salient points on average. Salient regions around local salient points are used for feature extraction and scene classification.

II. Repeatability of Salient Regions

In this section, we explore to what degree salient points repeat in different instances of the same scene class. To this purpose, we calculate the similarity of scene regions at the focuses of attention. This is indirectly in accordance with the repeatability criteria used to evaluate the quality of local feature detectors.

A. Features

In our approach, a within fixation processing extracts features at focuses of attention for scene representation. While any kind of features could be used, we employ two types of features, SIFT [11] and C2 features from the HMAX model [20], to compare Geometry-based and Appearance-based features for attentional classification.

SIFT features are distinctive features useful for reliable matching between different views of a scene or an object. They are invariant to image scaling and rotation, and are partially invariant to changes in illumination and 3D camera viewpoint. SIFT algorithm uses a Gaussian pyramid built from the original image to extract local features (keypoints) at the extreme points of difference between pyramid levels. Then, a descriptor is derived from the surrounding region of a keypoint using the histogram of edge responses.

C2 features are introduced by Serre *et al.* in [20]. They have developed a computational model based on the hierarchical organization of the visual ventral stream where shape, object, face and scene recognition is performed. Simply in these models, low-level features with high position and scale specificity are combined in a hierarchy to produce complex invariant features. This model starts with an image layer of gray-level pixels and builds simple (S) and complex (C) layers alternatively. Neurons in S layers convolve the image with a set of local filters to extract features and S units pool their inputs from previous layer to increase invariancy. (e.g. max operation). In this paper, we use the modified HMAX model [34] which learns a dictionary of prototype patches in the S2 layer from a set of sample images. Number of learned patches determines the dimensionality of the C2 feature vector. C2 features provide rich structural information useful for recognition and matching and incorporating them with visual attention may give insights of interaction of visual attention and object recognition in human visual system [24].

B. Distance Measures

Regarding the second step in scene classification, distance or a similarity measure between two images is necessary to construct a classifier. We define distance measures for each type of the above-mentioned features. Given a set of *n* images $I = \{I_1, I_2, \ldots, I_n\}$, a bag of features representation for the ith image with q SIFT features is:

$$F_i = \{f_{i1}, f_{i2}, \ldots, f_{iq}\}, f_{ij} = (x_{ij}, y_{ij}) \quad (4)$$

where x_{ij} is the value of the jth SIFT feature (a 128D vector) and y_{ij} is the position of this feature in image coordinate frame. The idea behind the state-of-the-art algorithms for matching and recognition, is that they measure the similarities between all local features within the compared images. Consider the following distance measure which is the average of the best matches for all features of ith and jth images:

$$D(F_i, F_j) = \frac{1}{2}(d(F_i, F_j) + d(F_j, F_i))$$
$$d(F_u, F_v) = \frac{1}{n_u} \sum_{l=1}^{n_u} \min_{t=1 \cdots n_v} norm(f_{ul} - f_{vt}) \quad (5)$$

where *norm* is the Euclidean distance between two SIFT features and n_u and n_v are the number of SIFT features of images u and v, respectively. Above distance measure does not consider the fact that features remain at spatial vicinity of each other most of the time when transformed. Adhering this constraint distance measure in eq. (6) enhances feature matching and hopefully recognition.

$$d(F_u, F_v) = \frac{1}{n_u} \sum_{l=1}^{n_u} \min_{t=1 \cdots n_v} norm(f_{ul} - f_{vt}).(H - e^G)$$
$$G = -\frac{1}{2\sigma^2}(norm(x_{uk} - x_{vp})^2) \quad (6)$$

In eq. (6), H is a large number, here 1000. For images represented by C2 feature vectors, the distance measure is simply the Euclidean norm of their difference as $D(C2_i, C2_j) = norm(C2_i - C2_j)$

The above definitions are for feature extraction over entire image. We now define distance of two images based on their features at salient regions. Assume $S_i = (s_{i1}, s_{i2}, \cdots, s_{ib})$, $saliency_{s_{ij}} > saliency_{s_{ij+1}}$ be the vector of salient points generated by a saliency detector ordered by their saliency value. Let

$$J = \begin{bmatrix} D(S_{i1}, S_{j1}) & \cdots & D(S_{i1}, S_{jb}) \\ \vdots & \ddots & \vdots \\ D(S_{ib}, S_{j1}) & \cdots & D(S_{ib}, S_{jb}) \end{bmatrix} \quad (7)$$

be the matrix of pairs of salient regions with $D(s_{i1}, s_{j1})$ as the distance between the first salient regions of the ith and jth images. In order to reduce the above distance matrix into a scalar, we need to match the salient regions. One way is to choose the minimum distance or the best match

TABLE I

ATTENTIVE SCENE CLASSIFICATION WITH KNN

1. Find salient regions for all images in the database
2. Extract features for all images at salient regions
3. Randomly split images in each scene class into P for train and remaining Q for test
4. For each test image
   a. For each train image
      i. Calculate the similarity matrix of salient regions of the test image to salient regions of this train image using the similarity measures in (6) or (7)
      ii. Calculate the overall similarity between two images by matching the salient regions using Hungarian algorithm
   b. Find the k nearest neighbors to this test image
   c. Assign the class label of this image as the label of the most frequent class
5. Calculate the correct classification rate

between salient regions, as was done for matching SIFT features in eq. (5) or eq. (6) by considering their pixel distances in the images. Here we use the optimal Hungarian method [38], which takes as input a distance matrix and returns a distance measure as well as the best match. This is a combinatorial optimization algorithm with the benefit of solving the assignment problem in polynomial time. A better match between two images means lower distance and hence higher similarity.

C. Repeatability Measure

Repeatability indicates how often features repeat in the transformed versions of an image. Inverse of above distances could be used for calculating similarity or repeatability of two images. Repeatability of a set of images (I) is the mean of repeatability of subsequent images (frames) and is defined as (also called tracking repeatability):

$$R_T(I) = \frac{n-1}{\sum_{i=1}^{n-1} D(I_i, I_{i+1})} \tag{8}$$

where D is one of distance measures defined in the previous section. Variable n is the number of images in a scene class. We compare the repeatability of saliency detector patches for increasing number of salient regions.

III. ATTENTIVE SCENE CLASSIFICATION

Having features and distance measures defined we are now ready to design a classifier. The k-nearest neighbor (KNN) algorithm is used in two cases when features are extracted 1) over the entire image or 2) only at salient regions. Since kNN is simple for the first case, we show how it could be used for classifying scenes in the second case. Attentive scene classification is shown in the pseudo code of Table I.

IV. EXPERIMENTAL RESULTS

Results of repeatability and attentive scene classification are shown in this section over a database of natural scenes. This database contains 44 classes each containing 24 scenes of size 640×480 with large viewpoint changes from locations in a university campus [39]. We restrict number of salient points to 5 since the concept of visual saliency is restricted to few features per image by definition.

A. Repeatability Results

Eight first instances of the seven first classes of the database in [39] with their salient regions marked with rectangles are shown in Fig. 2. As this figure shows, in many cases salient regions are nearly in the same regions in subsequent images but their orders differ. This shows the significance of region matching in attentional scene classification. Features are derived in a window of size W around each salient point. Tracking repeatability for salient regions for these classes using both features are shown in Fig. 3 (top row). Repeatability increases in each class as the number of salient points increases, for both SIFT and C2 features.

B. Classification Results

In the experiments in this section, we investigate how attention affects learning and recognition of cluttered scenes. Eighteen (P=18) images from each scene class were used for training and the remaining ones (Q=6) for testing. Results of attentive scene classification using C2 and SIFT features are shown in Fig. 3 (bottom row) using the distance measure in eq. (5) for different numbers of salient regions and different attention window sizes.

As shown in Fig. 3, recognition rate increases in all cases with the number of salient points. This is in accordance with the increase in repeatability with increasing the number of salient points. Larger window sizes result in higher recognition rates. To control whether high recognition rate is due to repeatability of salient regions or overlap in images, we also performed recognition using random and fixed image regions. Random patches were selected in random for every frame of a class. Fixed image patches had the same positions in image coordinates for all images of all classes. Fig. 4, left panel compares the classification results using different detectors with window size of 60 and different number of salient regions. As it shows saliency works better than random and fixed patches in all conditions over all 44 classes. It also shows that fixed patches work better than random patches in all cases. Table II, shows the classification results using KNN classifier with distance measure in eq. (6) and SIFT features for different values of σ. Results shows that σ equal to 3 leads to higher recognition in all cases but nothing can be said in general on setting this parameter.

We also performed classification when feature extraction was done over the entire scene. Classification results using the distance measure in (6) leads to 99.3 ± 0.3 %, 82.32 ± 2.1 % and 77.9 ± 1.2 % recognition rates using SIFT, C2($|C2| = 4096$) and C2($|C2| = 256$) features, respectively.

C. Computational Complexity

In this section, we analyze to what degree attention reduces computational complexity of classification. Assume an image of size m×n then using b SIFT windows of size W, the ratio of feature extraction over salient regions to

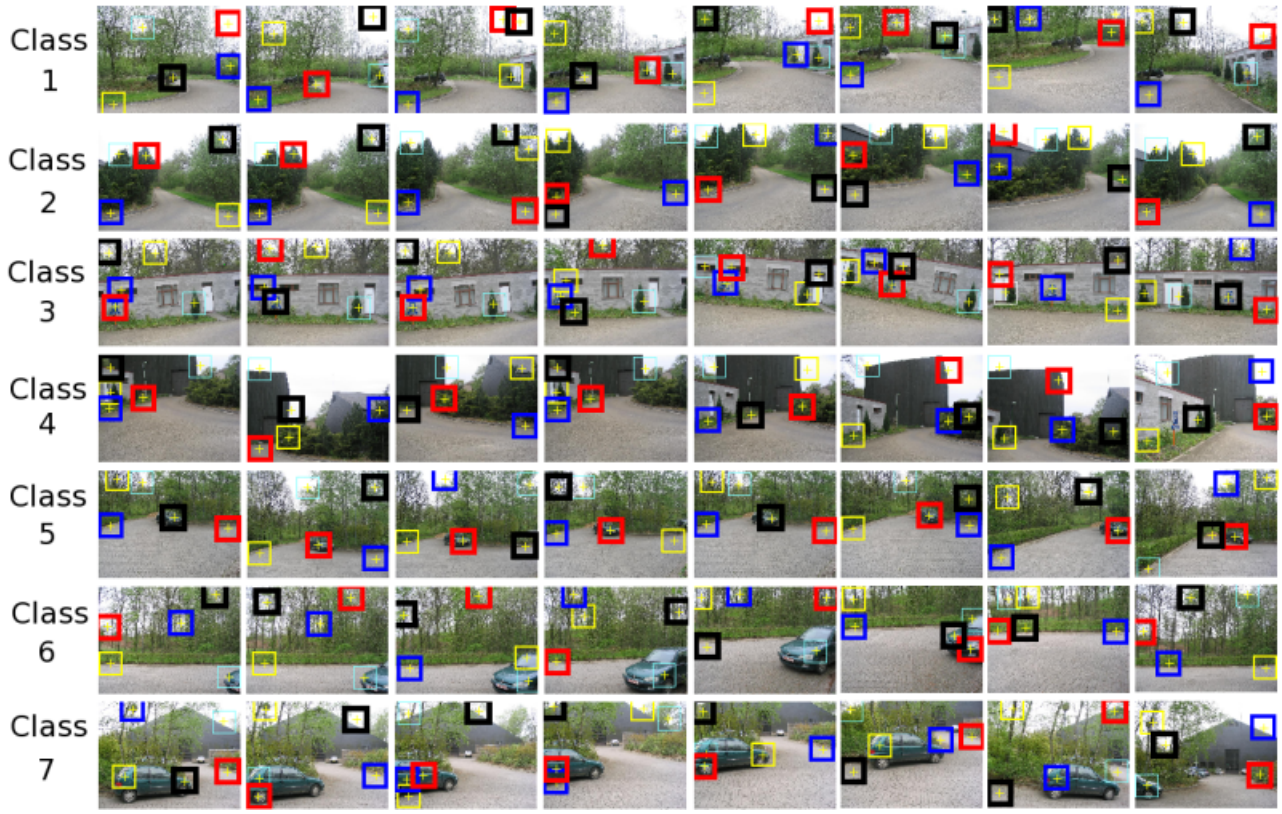

Fig. 2. Salient regions of 7 natural scenes. Each row is for a different class. Order of salient points is coded by colors: pink, yellow, blue, red and black. Narrowest rectangle is the first and so on. Window size (W) is equal to 60 pixels.

TABLE II
ATTENTIVE SCENE CLASSIFICATION USING SIFT AND THE SECOND DISTANCE MEASURE

σ	0	1	2	3	10
Rec(%)	95.59	97.1	96.77	98.1	93.8
Std(%)	(2.3)	(1.1)	(1.4)	(0.37)	(3.6)

entire image is $C = \frac{b \times W^2}{m \times n}$. Since W is usually small compared to m and n this leads to significant computational saving. Fig. 4, right panel shows computational saving versus recognition rates. The vertical axis shows the ratios of recognition and computational complexity in attentional and without-attention cases. This diagram is plotted for SIFT and C2 features for only first salient region as well as all of them for different window sizes. For example it shows that with a window size of 140 and 5 salient regions 95% of the recognition rate obtained in the without-attention case (processing the entire scene) can be achieved while only extracting features over 31% of the image.

Increasing attention window size results in higher recognition rate and higher computational complexity. An optimal value for window size depends on application and on available computational resources and minimum recognition rate needed. This analysis can also help determine minimum computation necessary to achieve a recognition rate.

V. DISCUSSION AND CONCLUSIONS

This paper has studied the problem of defining and estimating descriptive and compact visual models of scene classes for efficient scene recognition. In accordance with human recognition behavior first a fast, parallel and pre-attentive mechanism processes the entire image and selects some salient regions and then a complex and slow mechanism is restricted on these areas to extract more details and match them with a database of learned representatives. In the proposed approach, to overcome the information overloading bottleneck, feature extraction and matching is performed at few spatial locations proposed by visual attention which acts as a front-end to accelerate speed and reduce complexity.

Results show that salient regions have high repeatability over image transformations. Higher classification results of salient regions compared with random and fixed salient regions mean that it is not the overlap among regions which is responsible for discrimination, but it is because salient regions are at nearly the same locations. However, since rank of saliency points differs it is necessary to match them. As window size increases, similarity and hence recognition rate over both feature types increase. SIFT features are more successful in average for scene classification from our results. Compared with the situation when features are extracted over the entire image, we were able to achieve nearly the same recognition rates with much fewer computations.

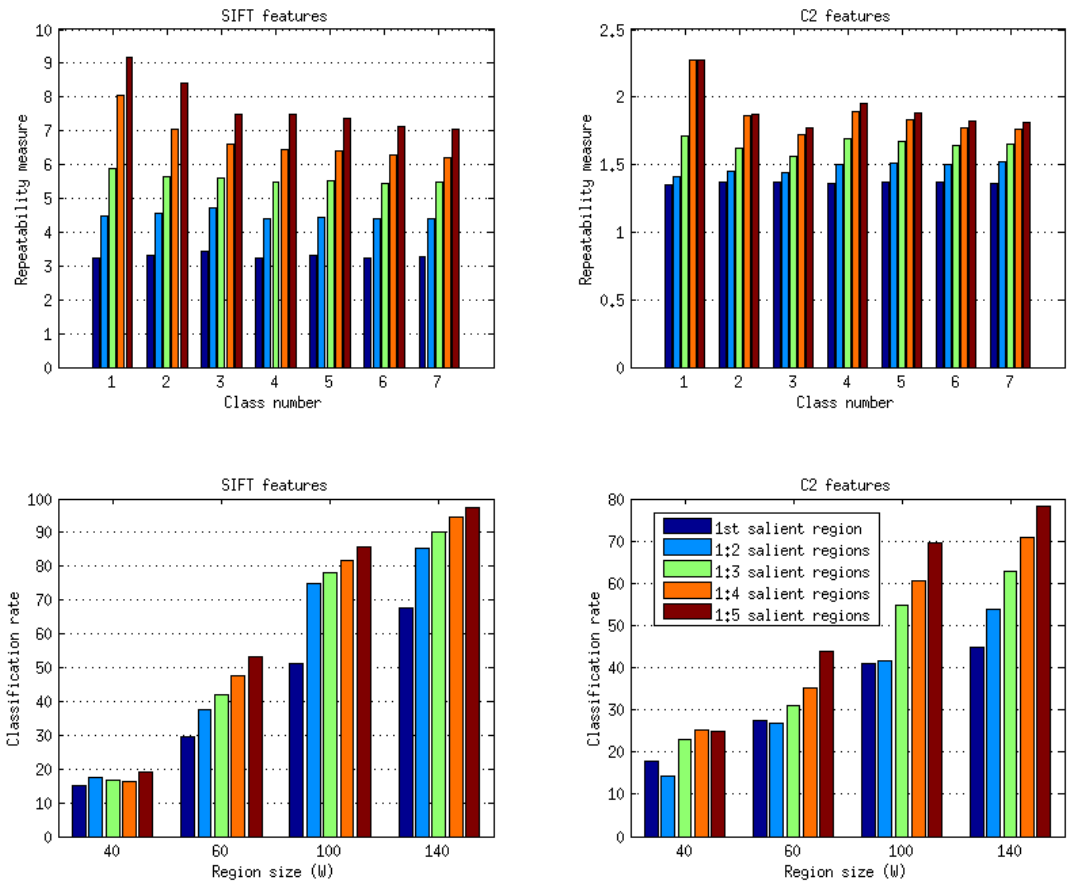

Fig. 3. Top: Tracking repeatability of 7 scene classes shown in Fig. 2 for different numbers of salient regions . First panels in each column are repeatabilities for SIFT and HMAX C2 features, respectively. W is 60 and dimensionality of C2 feature vector ($|C2|$) is equal to 4096. Bottom: Attentive scene classification using SIFT and C2($|C2| = 4096$) features with distance measure in eq. (5) and for different sizes of salient regions. 1:n means salient regions 1 to n.

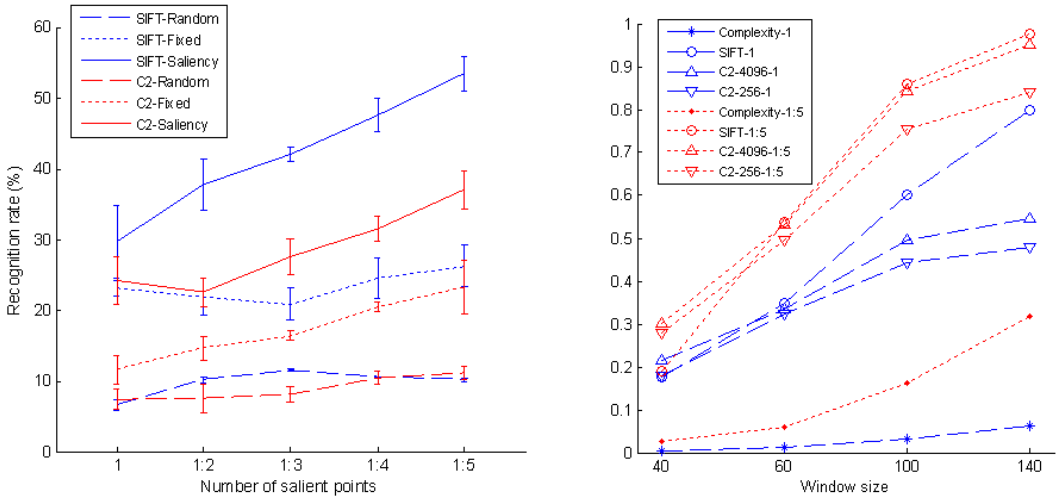

Fig. 4. Left: Comparison of recognition rate of salient regions as well as other detectors. W = 60 and $|C2| = 256$. Right: Recognition rate vs. computational complexity

We showed that high scene recognition rates are still possible without processing the entire images. The saliency model is purely data-driven and simply selects some spatial regions without using any feedback mechanism or top-down gains.

Modifying attentional systems to generate robust salient points when viewpoints change is a future research area. For example adding top-down extensions to the saliency model restricts it to select the same spatial areas over transformed images and therefore leads to higher feature stability in images. It is worth checking that other saliency measures as the first step of our approach in this paper. For example, while [4] has compared its saliency measure against Harris, MSER and SSD and has showed that it is more stable, it may results to better recognition results than basic saliency model. Applying this approach to robot control is also another interesting area since in robotic applications viewpoints change gradually and this may lead to higher repeatability of salient regions. Repeatability of salient points under different image transformations and other databases should also be considered in future works in this direction.

Acknowledgements: Supported by DARPA. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

REFERENCES

[1] T. Kadir and M. Brady, Unsupervlssed non-parametric region segmentation using level sets, ICCV, 2003.

[2] M. Staudte and M.W. Crocker, Visual attention in spoken human-robot interaction, HRI, 2009.

[3] M. Grabner and H. Grabner and H. Bischof, Fast approximated SIFT, ACCV, 2006.

[4] D. Gao and N. Vasconcelos, Discriminant Interest Points are Stable, CVPR, 2007.

[5] C. Harris and M. Stephens, A Combined Corner and Edge Detector, Alvey Vision Conference, 1988.

[6] S. Lazebnik and C. Schmid, and J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, IEEE CVPR, 2006.

[7] M. Boutell and C. Brown, and J. Luo, Review of the State of the Art in Semantic Scene Classification, Technical Report, Univ. Rochester, 2002.

[8] A. Vedaldi, An open implementation of the SIFT detector and descriptor, UCLA CSD Tech. Report 070012, 2006.

[9] K. Mikolajczyk and C. Schmid, Scale & Affine Invariant Interest Point Detectors, International Journal of Computer Vision, v 60, n 1, p:63-86,2004.

[10] J. Matas, O. Chum, M. Urban and T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image and Vision Computing, v 22, n 10, p 761-767, 2004.

[11] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, v 60, n 2, p 91-110, 2004.

[12] C. Schmid and R. Mohr, Local Grayvalue Invariants for Image Retrieval, IEEE T-PAMI, v 19, n 5, p 530-535, 1997.

[13] C. Siagian and L. Itti, Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention, IEEE T-PAMI, v 29, n 2, p 300-312, 2007.

[14] P. J. Burt and E. H. Adelson, The Laplacian pyramid as a compact image code, IEEE Transactions on Communications, 31, n 4, p 532-540, 1983.

[15] Neil D. B. Bruce, John K. Tsotsos: Saliency Based on Information Maximization, NIPS, 2005.

[16] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments. NIPS, 681-688, 2008.

[17] N. Ouerhani and R. V. Wartburg, H. Hügli and R.M. Müri, Empirical validation of Saliency-based model of visual attention, Electronic Letters on Computer Vision and Image Analysis, v 3, n 1, p 13–24, 2003.

[18] L. Itti, C. Koch and E. Niebur, A Model of Saliency-based Visual Attention for Rapid Scene Analysis, IEEE T-PAMI, v 20, n 11, p 1254–1259 , 1998.

[19] L. Itti and C. Koch, Feature combination strategies for saliencybased visual attention systems, J. Electron. Imaging, v 10, n 1, p 161-169, 2001.

[20] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio , Object Recognition with Cortex-like mechanisms, IEEE T-PAMI, v 29, n 3, p 411–426, 2007.

[21] A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, International Journal of Computer Vision, v 42, n 3, p 145–175, 2001.

[22] S. Se. D. G. Lowe and J. J. Little, Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks, International Journal of Robotic Research, v 21, n 8, p 735-760, 2002.

[23] S. Frintrop and P. Jensfelt, Attentional Landmarks and Active Gaze Control for Visual SLAM, IEEE Transactions on Robotics, special Issue on Visual SLAM, v 24, n 5, p 735-760, 2008.

[24] D. Walther, U. Rutishauser, C. Koch and P. Perona, Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, Computer Vision and Image Understanding, v 100, n 1-2, p 41–63, 2005.

[25] A. Santella, D. DeCarlo, Abstracted painterly renderings using eye-tracking data. NPAR, 75-82, 2002.

[26] T. Avraham, M. Lindenbaum, Esaliency (Extended Saliency), meaningful attention using stochastic image modeling. IEEE Trans. Pattern Anal. Mach. Intell. 32(4), 693-708, 2010.

[27] L. Marchesotti, C. Cifarelli, G. Csurka, A framework for visual saliency detection with applications to image thumbnailing, ICCV 2009 (12th IEEE International Conference on Computer Vision) , 2009.

[28] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, A Coherent computational approach to model bottom-Up visual attention. IEEE Trans. Pattern Anal. Mach. Intell. 28(5), 802-817, 2006.

[29] G. Boccignone, A. Chianese, V. Moscato, A. Picariello, Foveated shot detection for video segmentation. IEEE Trans. Circuits Syst. Video Techn. 15(3): 365-377, 2005.

[30] L. Itti, P. Baldi: Bayesian surprise attracts human attention, NIPS, 2005.

[31] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, G. W. Cottrell, SUN: A bayesian framework for saliency using natural statistics, Journal of Vision 8(7):32, 1-20, 2008.

[32] M. Pollefeys, L.J.V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, Jan Tops and Reinhard Koch, Visual Modeling with a Hand-Held Camera, International Journal of Computer Vision, v 59, n 3, p 207-232, 2004.

[33] J. G. Daugman, Two-Dimensional Spectral Analysis of Cortical Receptive Field Profile, Vision Research, v 20, n 1, 847-856, 1980.

[34] J. Mutch and D.G. Lowe, Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields, International Journal of Computer Vision, v 80, n 1, p 45-57, 2008.

[35] G. Fritz, C. Seifert, L. Paletta and H. Bichof. Attentive Object Detection Using an Information Theoretic Saliency Measure, In Attention and Performance in Computational Vision, LNCS, vol. 3368, SPringer, 2005

[36] P. E. Forssen, D. Meger, K. Lai, S. Helmer, J. J. Little and D.G. Lowe. Informed Visual Search:Combining Attention and Object Recognition, ICRA, 2008

[37] T. Xu, T. Zhang, K. Kuhnlenz, and M. Buss , Attentional Object Detection with An Active Multi-Focal Vision System, Int J of Humanoid Robotics 7(2), 2010.

[38] R. Ahuja, T. Magnanti and J. Orlin, Network Flows, Prentice Hall, 1993.

[39] http://www.montefiore.ulg.ac.be/ jodogne/phd-database.