Statistical Textural Distinctiveness for Salient Region Detection in Natural Images

Christian Scharfenberger, Alexander Wong, Khalil Fergani, John S. Zelek and David A. Clausi
University of Waterloo, Vision and Image Processing (VIP) Research Group
Waterloo, Ontario, Canada
{cscharfenberger, a28wong, kfergani, jzelek, dclausi} @uwaterloo.ca

Abstract

A novel statistical textural distinctiveness approach for robustly detecting salient regions in natural images is proposed. Rotational-invariant neighborhood-based textural representations are extracted and used to learn a set of representative texture atoms for defining a sparse texture model for the image. Based on the learnt sparse texture model, a weighted graphical model is constructed to characterize the statistical textural distinctiveness between all representative texture atom pairs. Finally, the saliency of each pixel in the image is computed based on the probability of occurrence of the representative texture atoms, their respective statistical textural distinctiveness based on the constructed graphical model, and general visual attentive constraints. Experimental results using a public natural image dataset and a variety of performance evaluation metrics show that the proposed approach provides interesting and promising results when compared to existing saliency detection methods.

1. Introduction

The underlying goal of saliency detection in natural images is to identify and localize objects of interest that attract the visual attention of a human observer compared to the rest of the scene. For example, Fig. 1 shows examples of natural images, where the garden gnome or the flower are visually unique and draw a viewer's attention from the surrounding environment. The research area of saliency detection from natural images has gained tremendous interest in the field of computer vision given its wide applicability for many computer vision tasks such as image segmentation [9], image retargeting [3], object detection [20], and object recognition [24].

To achieve saliency detection in an automatic manner, one must define what constitutes as a salient object based on some quantifiable visual attributes such as intensity, color, structure, texture, size, or shape that makes that object appear visually distinct and attractive to the observer's attention compared to the rest of the scene. A particularly interesting visual attribute that deserves deeper exploration for the purpose of automatic saliency detection in natural images is texture, which reveals significant information about not only local spatial-color relationships, but also the global compositional characteristics of an image given that natural images exhibit heterogeneous textural characteristics. In the context of saliency in natural images, one can then view salient objects of interest as objects that possess textural characteristics that are highly distinctive from a human observer perspective when compared with that of the rest of the scene. As such, we are interested in explicitly taking advantage of textural characteristics in a quantitative manner to detect saliency objects of interest within a scene. Two important challenging aspects associated with explicitly accounting for textural characteristics are:

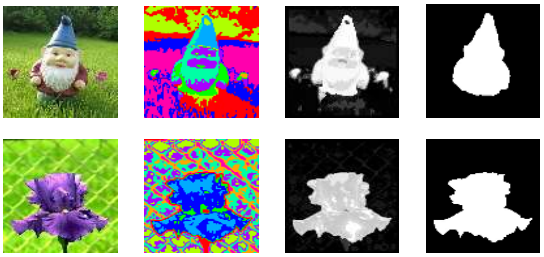


Figure 1: From left to right: Salient objects with texture patterns that are visually significantly different from those of the rest of the scene. Pixels associated with the corresponding atom of a learned texture model. Computed saliency map and ground truth mask.

1. the choice of appropriate texture representations for distinguishing between salient and non-salient regions,

2. the added computational complexity associated with textural characteristics compared to simpler visual attributes such as color and intensity, particularly if one were to analyze and compare all possible texture pattern pairings in the image in a direct fashion.

Prior work that incorporated textural characteristics [25] has attempted to address these two issues by making use of low-level filter-based texture features and relied on image segmentation to reduce computational complexity while enforcing feature coherence within local regions. However, the reliance on advanced pre-processing algorithms such as image segmentation means that the computational complexity and performance of the saliency detection method depends heavily on the properties of the segmentation method used, even if oversegmentation is performed. Therefore, an efficient method for performing saliency detection based explicitly on descriptive textural characteristics that does not rely on additional pre-processing would be much desired.

The main contribution of this paper is the introduction of a novel approach to saliency detection based on the concept of statistical textural distinctiveness. Rotational-invariant neighborhood-based texture representations are extracted and used to learn a set of representative texture atoms for defining a sparse texture model for the image. Based on the learnt sparse texture model, a statistical textural distinctiveness graphical model is constructed to characterize the distinctiveness between all texture atom pairs. Finally, the saliency of each pixel in the image is computed based on the probability of occurrence of the representative texture atoms within the image, their respective statistical textural distinctiveness based on the constructed graphical model, and general visual attentive constraints. By incorporating sparse texture modeling within a statistical textural distinctiveness framework, the proposed approach is designed to take explicit advantage of the textural characteristics in the image to detect salient regions in an efficient yet characteristic manner. To the best of the authors' knowledge, the use of sparse texture modeling within a statistical textural distinctiveness framework to characterize and compare textural characteristics within an image for the purpose of saliency detection has not been previously proposed or investigated.

2. Related Work

Existing saliency are either biologically motivated, computational oriented, or perform local or global analysis of contrast using intensity only, and/or different colorspaces. Biologically inspired techniques [13, 10] for saliency detection are commonly based on the approach of Koch *et al.* [16] and rely on low-level features such as edges, orientation of edges, motion and color in natural images. Itti *et al.* [13] extended the approach of Koch *et al.* [16] by implementing a Difference of Gaussian (DoG) approach to better evaluate the features. All these approaches are designed to identify salient regions with high visual stimuli, but tend to blur saliency maps and to highlight local features such as small objects. They are useful for applications in robotics, but challenging for image-based segmentation or object detection.

To better preserve the structure of salient regions, Hou *et al.* [11] and Guo *et al.* [8] proposed to extract the residuals of input images in either the amplitude or phase spectrum of input images, and to use the residuals to construct saliency maps in the spatial domain. However, these methods highlight boundaries of salient regions rather than their entire region. However, extracting salient regions in images with textured background properly is challenging for these methods.

Saliency detection approaches considering colorspaces analyze the local or global contrast. Local saliency detection methods usually evaluate saliency of input image with respect to small neighborhoods. Examples include dissimilarity at pixel level [19], and the analysis of histograms [18] and Difference of Gaussians at multiple scales [12]. As shown in [1], these approaches do not consider global relationships between regions or pixels and emphasize edges or noise. In addition, they also tend to highlight cluttered and textured non-salient regions in images.

Global approaches consider contrast relationships over the entire image. Patch-based approaches determine salient regions by computing dissimilarities between images patches, e.g., [18, 7, 26]. These methods have high computational complexity and are applicable to images with low resolution only. In order to make patch-based dissimilarity approaches applicable on high resolution images, Duan *et al.* [6] suggested to reduce the dimensionality of the patches using PCA. However, down sampling or reducing the dimensionality of the patches may lead to a loss of small salient regions. Achanta *et al.* [1] overcome high computational complexity by computing color dissimilarities to the mean image color on a per-pixel basis. In [2], they extended their concept to take into account the spatial relationship inside the image. In addition, the approach of Cheng *et al.* [5] generates a color histogram of the entire image, and compute the saliency based on the dissimilarity between the histogram bins, and also use image segmentation for improving saliency estimation. To better handle images with cluttered or textured background, Perazzi *et al.* [21] abstract input images into homogeneous elements, and determine salient regions by applying two contrast measures based on the spatial distribution and uniqueness of elements. However, both image segmentation and abstraction remove textural information which might indicate salient regions in images.

Finally, the approach of Shen *et al.* [25] (LR) explicitly incorporated textural characteristics obtained from low-level filter-based texture features of segmented regions for saliency detection. The textural characteristic of a region is represented by a feature vector, which all together build a feature matrix. Matrix decomposition based on a previously trained background model is then performed to identify salient regions, which have to be refined using strong

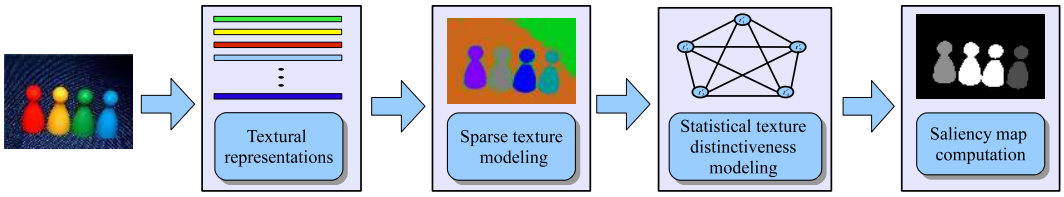


Figure 2: Architecture for salient region detection based on sparse texture modeling and statistical textural distinctiveness.

priors such as spatial, color and semantic priors to obtain good performance results.

However, none of these approaches explicitly consider rotational-invariant neighborhood-based texture representations (atoms) for salient region detection. In contrast to approaches that rely on image segmentation and image abstraction where each region can only characterize a small area in the image, each learnt sparse texture atom can represent large or disjoint regions without explicit spatial context. Although our method and the LR [25] have comparable performance based on experimental results (see Section 4), there are some very important differences between the two methods. The overall good performance of LR is the result of three strong priors applied to saliency computation. Our approach makes limited use of one prior (location) only and achieves comparable performance. In addition, the LR requires a previously trained background model for matrix decomposition whereas our approach does not rely on any previously trained data at all. This makes our method more robust and suitable for applications where a priori background information is not available.

3. Statistical Textural Distinctiveness Model

The underlying goal of the proposed statistical textural distinctiveness approach is to explicitly take advantage of inherent heterogenous textural characteristics in the image in an efficient manner for quantifying the saliency of regions within the image. The overall architecture of the proposed approach can be broken down into four main stages – as shown in Fig. 2: i) rotational-invariant neighborhood-based textural representation, ii) sparse texture modeling via representative texture atom learning, iii) statistical textural distinctiveness graphical model construction, and iv) saliency map computation based on occurrence probabilities of representative texture atoms, statistical textural distinctiveness, and general visual attentive constraints. A detailed description of each stage is provided in the following sections.

3.1. Rotational-invariant neighborhood-based textural representations

In order to learn a texture model for natural images with heterogeneous textural characteristics, we must first define a texture feature model to represent the underlying textural characteristics of the image in a local manner to account for this heterogeneity. In this work, a compact rotational-invariant neighborhood-based texture feature model is utilized in the form of a sparsified radially-sorted textural representation based upon the work by Li *et al.* [17]. This form of textural representation has been found to be beneficial in striking a balance between robustness to distortional variations and preservation of spatial-intensity context, making it well-suited for local textural representation in the proposed work (see Section 4, Fig. 5c and Fig. 6a for a comparison with unsorted textural representations).

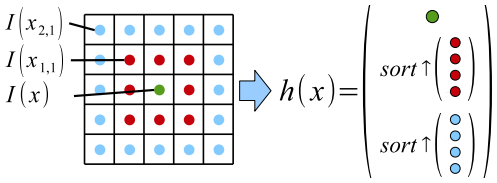


Figure 3: Sorted textural representation of a 5×5 pixel neighborhood centered at pixel x.

The sparsified radially-sorted textural representation can be described as follows. Let $I(x)$ be the $M \times N$ image that we wish to analyze. Given a neighborhood $\aleph$ centered at pixel location x in the image $I(x)$, the corresponding local textural representation $h_c(x)$ for each color channel c can be defined as:

$$h_c(x) = \langle I_c(x) \;\; \text{sort}_{\uparrow}\{I_c(x_{1,j})\} \;\; \text{sort}_{\uparrow}\{I_c(x_{2,j})\} \;\; \ldots \;\; \text{sort}_{\uparrow}\{I_c(x_{n,j})\}\rangle \quad (1)$$

where $x_{i,j}$ denote the pixel in the j^{th} position of the i^{th} radial layer about pixel location x, and $sort_{\uparrow}$ denotes sorting in ascending order. An illustration of this local textural representation for single channel images is shown in Fig. 3. Internal experiments with different square neighborhoods showed that a 5×5 square neighborhood is a good choice for sparse texture model learning. Given the local textural representation $h(x)$, e.g., $h(x) = \langle\, h_L(x)\; h_a(x)\; h_b(x) \,\rangle$ with 75 element for Lab images, we wish to produce a compact version of this local textural representation to increase the variance between the elements of the texture descriptor and to improve the efficiency of the subsequent sparse texture model and statistical textural distinctive model stages. In this work, a sparsified textural representation $t(x)$ is produced by taking the u principal components of the local textural representation $h(x)$ with the highest variance using PCA:

$$t(x) = \langle \Phi_i(h(x)) \mid 1 \leq i \leq u \,\rangle, \quad (2)$$

where Φ_i is the i^{th} principal component of $h(x)$. The choice of u is based on a selection criteria related to a variance

compaction. We selected the u principal components of $h(x)$ that represent 95% of the variance of all textural representations – as suggested for many machine learning approaches [4].

3.2. Sparse texture model via texture learning

Given the set of $M \times N$ local texture feature representations extracted from the image $f(x)$:

$$T = \{t_1, t_2, t_3, \ldots, t_{M \times N}\}, \tag{3}$$

let us now define a global texture model to represent the heterogeneous textural characteristics for the entire image $f(x)$. One simple strategy to construct such a global texture model is to simply utilize the entire set of extracted local textural representations. However, this strategy to global texture modeling is highly computational- and memory-intensive for the purpose of texture-based saliency detection given the use of pair-wise textural representation analysis to establish a quantitative relationship between the different texture patterns within the image.

To address this issue, we first generalize a natural image as being composed of a set of areas where a particular texture pattern is repeated over each area, where the number of areas with unique texture patterns is much smaller than the total number of pixels within the image. Based on this generalization of a natural image, we can then establish a textural sparsity assumption for natural images, where the global textural characteristics of an image can be well-represented by a small set of distinctive local textural representations. This compact, sparse representation of the global, heterogeneous textural characteristics of an image motivates the use of a sparse texture model. In this work, the sparse texture model can be defined as a set of $m << M \times N$ representative texture atoms:

$$T^r = \{t_i^r | 1 \leq i \leq m\}, \tag{4}$$

where the L_p-norm between the first u principal components of each of the representative texture atom t_i^r and that of its corresponding set of local textural representations (denoted by S_i) is minimized:

$$T^r = arg\,min \sum_{i=1}^{m} \sum_{t_j^r \in S_i} \|t_j^r - t_i^r\|_p. \tag{5}$$

Given the aforementioned model, a simple and efficient strategy employed in the proposed method to learning the sparse texture model of an image is to assert a L_2-norm criteria and solve for T_r using the k-means algorithm [15]. By employing a sparse texture model for representing the heterogeneous textural characteristics of the entire image, the computational and memory requirements for representing and quantifying the relationships between each texture pair is significantly reduced since only the representative texture atoms need to be analyzed (e.g., $1/2 \cdot m(m-1)$ relationships as opposed to $1/2 \cdot M \cdot N(M \cdot N - 1)$ relationships). As later presented in Section 4, a set of $m = 20$ representative texture atoms is an appropriate choice to represent the global textural characteristics of natural images.

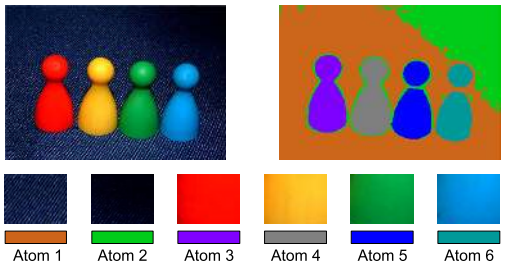


Figure 4: Example image containing salient objects, and the learned texture model with pixels associated with the corresponding atoms (six for illustrative purposes).

3.3. Statistical textural distinctiveness graphical model construction

In natural images, salient regions of interest can be characterized as regions that are visually distinct from the rest of the scene in terms of their visual attributes. In this work, we first consider a salient region of interest as regions that have highly unique and distinctive textural characteristics when compared to the rest of the scene (see Fig. 4). As such, we are motivated to introduce a metric for quantifying the uniqueness and distinctiveness of texture patterns within an image relative to each other. Here, we introduce the concept of statistical textural distinctiveness, where an area of interest is salient if it has low textural pattern commonality compared to the rest of the scene. As such, the concept of statistical textural distinctiveness takes explicit advantage of the statistical relationships between texture patterns within an image to discern underlying saliency.

Given the learnt sparse texture model, let us first define the statistical textural distinctiveness between two texture patterns. Let t_i^r and t_j^r denote a pair of representative texture atoms in the sparse texture model. Suppose that t_i^r can be seen as a realization of t_j^r in the presence of noise:

$$t_j^r = t_i^r + \eta_{i,j}, \tag{6}$$

where $\eta_{i,j}$ is a noise process between the representative texture atoms t_i^r and t_j^r following some distribution $P(\eta_{i,j})$. If the noise process $\eta_{i,j}$ is assumed to be independent and identically distributed, the probability of t_i^r being a realization of t_j^r can be written as:

$$P(t_i^r | t_j^r) = \prod_k P(t_{i,k}^r | t_{j,k}^r), \tag{7}$$

where $t_{i,k}^r$ is the k^{th} element in the texture atom t_i^r. As such, $P(t_i^r | t_j^r)$ can be viewed as the statistical commonality

between the two representative texture atoms. In this work, $P(\eta_{ij})$ is modeled as an independent and identically distributed zero-mean noise Gaussian process with a variance corresponding to the variance of the L_p-norm between the representative texture atoms, i.e., $\mathrm{var}(\|\mathrm{t_j^r} - \mathrm{t_i^r}\|_\mathrm{p})$, as it was found to provide strong saliency detection performance.

Given that we are interested in the distinctiveness of a texture atom relative to the other texture atoms in the sparse texture model for the purpose of saliency detection, a more meaningful metric for quantifying textural distinctiveness is the probability of t_i^r not being a realization of t_j^r:

$$\begin{aligned} \beta_{i,j} &= 1 - P(t_i^r|t_j^r) \\ &= 1 - \textstyle\prod_k P(t_{i,k}^r|t_{j,k}^r) \end{aligned} \tag{8}$$

Based on this definition, $\beta_{i,j}$ increases as the two texture atoms becomes more distinct from each other.

Given the aforementioned definition of statistical textural distinctiveness, one can then construct a weighted graphical model to characterize all pair-wise statistical textural distinctiveness within the sparse texture model of the image, which can be described as follows. Let G be a weighted complete graph defined by $G = \{V, E\}$, where V is the set of m vertices representing the representative texture atoms and E is the set of $\frac{m(m-1)}{2}$ edges representing every pair of representative texture atoms in the sparse texture model. Each edge $e_{i,j}$ is associated with a weight equal to the statistical textural distinctiveness ($\beta_{i,j}$) between a pair of representative texture atoms t_i^r and t_j^r.

3.4. Saliency map computation

Using the aforementioned statistical textural distinctiveness graphical model, and complimented by general visual attentive constraints, the saliency map for an image $I(x)$ can now be computed based on the following extended assumptions:

1. Salient objects are associated with texture patterns that are highly distinct from that of the rest of the scene **(statistical textural distinctiveness)**.

2. Salient objects are associated with texture patterns that are in closer spatial proximity to the center of the scene **(visual attentive constraints)**.

Given these two key assumptions, the saliency of a representative texture atom t_i^r (which we will denote as α_i) can be computed as the product of:

1. the expected statistical textural distinctiveness of t_i^r given the image $I(x)$, and

2. the weighted spatial proximity of pixels whose texture patterns represented in the sparse texture model by t_i^r (i.e., S_i) to the center of the image (denoted by x_c) as suggested by [14],

As such, the saliency α_i can be defined in the context of the proposed work:

$$\alpha_i = \left(\sum_{j=1}^{m} \beta_{i,j} P\left(t_i^r|I(x)\right)\right)\left(\exp\left(-\frac{1}{n_{t_i^r}}\sum_{x\in S_i}\frac{(x-x_c)^2}{\sigma^2}\right)\right) \tag{9}$$

where $P\left(t_i^r|I(x)\right)$ is the occurrence probability of t_i^r in the image $I(x)$, and $n_{t_i^r}$ is the number of pixels associated with t_i^r in the image $I(x)$. Given the saliency α computed for each of the m representative texture atoms in the sparse texture model, one can easily compute the saliency for each pixel x in the image $I(x)$ (denoted here by $\Psi(x)$) based on the representative texture atom in the sparse texture model that the pixel maps to:

$$\Psi(x) = \alpha_i, \text{ if } \mathrm{x} \in \mathrm{S_i} \tag{10}$$

There are two key benefits to this approach to computing the saliency map based on the constructed statistical textural distinctiveness model:

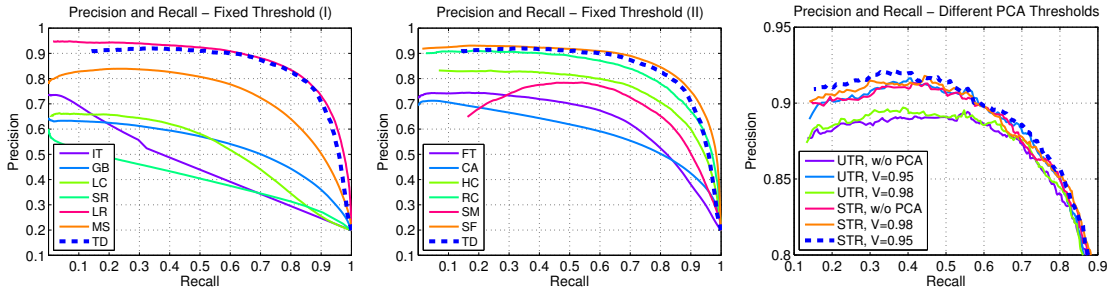
1. Only m saliency computations are needed, one for each representative texture atom in the sparse texture model. As such, the computational complexity of the saliency computations is independent of the size of the image and thus scales linearly (i.e., $O(m)$) as the number of texture atom in the sparse texture model increases, not as the image size increases.

2. The occurrence probability of texture atoms $P\left(t_i^r|I(x)\right)$ used to compute the saliency of each representative texture atom only needs to be computed once per image.

4. Experimental Results

To investigate the potential of our proposed statistical texture distinctiveness approach (TD) for robustly detecting salient regions, we evaluated our method based on the public EPFL database [1]. It contains 1000 natural images with accurate human-marked labels as ground truth, and is widely used as a benchmark for comparing saliency approaches, e.g., in [1, 2, 5, 25, 21]. In this paper, we compared our approach with 12 state-of-the-art saliency detection methods. These methods have been selected based on the following criteria [1, 5]: number of citations (spectral-residual (SR [11]), visual attention (IT [13])), recency (luminance-contrast (LC [27]), frequency-tuned (FT [1]), saliency-measure (SM [22]), metric-surround (MS [2]), context-aware (CA [7], histogram and region contrast (HC, RC [5])), and being related to our approach (graph-based (GB [10]), low-rank (LR [25]), and saliency-filters (SF [21])).

For a fair comparison with other approaches, we used the two different objective comparison measures as suggested by Achanta *et al.* [1] for performance evaluation.



(a) (b) (c)

Figure 5: a) and b) Precision and recall rates for all approaches based on the public EPFL database [1]. Our textural distinctiveness approach (TD, dashed line) achieves the state-of-the-art precision and recall rates. c) Precision and recall rates (zoomed) without selecting u principal component (w/o PCA) and for two PCA coefficients, representing 95% ($V = 0.95$) or 98% ($V = 0.98$) of the variance of all textural representations, for sorted (SRT) and unsorted (UTR) textural representations.

These measures are based on two experiments, the *fixed* and *adaptive threshold experiment*. Following this scheme, we performed binary segmentation of saliency maps using each possible fixed threshold Δ to compute precision-recall curves in a first experiment. We also used precision-recall curves to determine an optimal parameter configuration for our approach. In the second experiment, we compare the performance of our approach with other approaches, with segmenting salient objects using both adaptive thresholds and GrabCut [23].

Fixed threshold experiment. In a first experiment, we segmented the saliency maps using a fixed threshold $\Delta_{fix} \in [0, 255]$ to obtain binary images, with highlighting regions with saliency values larger than Δ_{fix} as foreground. We compare the resulting images to ground truth mask to determine recall and precision. By varying Δ from 0 to 255, we get precision and recall pairs used for both drawing precision-recall curves and evaluation. Fig. 5 shows the resulting precision-recall curves which were averaged over 1000 images from the EPFL database [1], and Fig. 8 the visual comparison of our and other approaches with ground-truth data.

As shown in Fig. 5, our approach outperforms HC and RC, and has similar performance as the LR and SF approaches. It also has to be noticed that the low-rank saliency approach (LR) requires a variety of additional constraints (such as color, spatial and semantic priors) to achieve excellent precision-recall curves, whereas our method relies on the saliency values obtained from the sparse texture model, weighted by the spatial proximity of pixels. Our approach also benefits from the rotational-invariant sorted textural representations (STR) which help to better reduce the influence of cluttered or textured background on saliency computation, as compared to an implementation with unsorted textural representations (UTR) as shown in Fig. 5c and Fig. 6a. Fig. 5c and Fig. 6a also illustrate the precision and recall curves obtained for several parameter configurations. It can also be seen that selecting u principal components using PCA further improves precision and recall, and that TD consistently performs over a wide range of parameter settings. However, the best configuration can be found for the use of: 1) radially-sorted textural representations over 5×5 square neighborhoods, 2) the use of PCA coefficients representing 95% of the variance, and 3) 20 representative texture atoms in the sparse texture model. Increasing the number of atoms to 50 does not improve recall and precision, whereas 5 atoms might be to few for representing the texture characteristic of natural images.

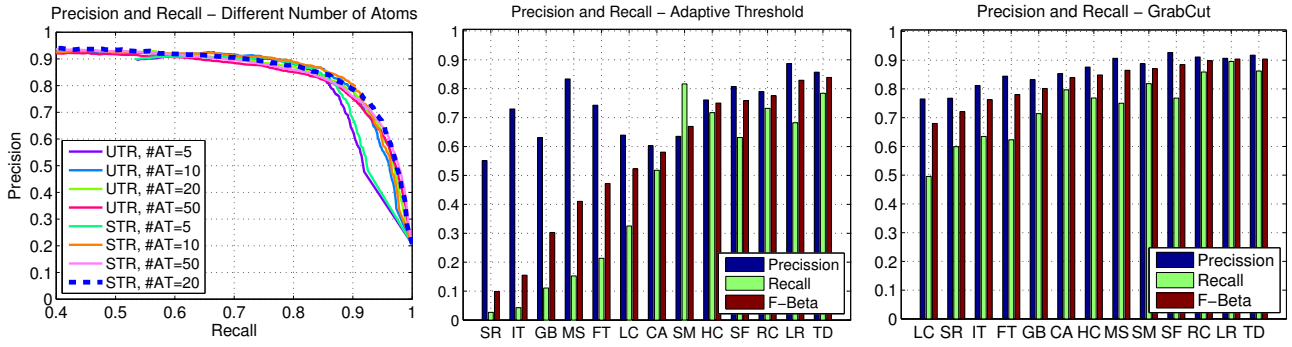
Adaptive threshold and GrabCut experiment. In the second experiment, we applied an image dependent threshold on the saliency maps to segment salient regions. Achanta *et al.* [1] defined this threshold as twice the mean of saliency maps $S(x)$, i.e., $\Delta_{ada} = 2 \cdot E(S(x))$. However, a closer analysis of the saliency maps obtained showed that the distribution of saliency values follows a Gaussian mixture model, with non-salient values having larger probabilities than salient values. To better appreciate this model, we define the adaptive threshold Δ_{ada} as follows:

$$\Delta_{ada} = E(S(x)) + STD(S(x)) \,, \tag{11}$$

taking into account the mean $E(S(x))$ and standard deviation $STD(S(x))$. After generating object images using Δ_{ada}, we computed precision (P), recall (R), and their harmonic mean measure F_β-measure for evaluation as follows:

$$F_\beta = \frac{(\beta^2 + 1)P \cdot R}{\beta^2 \cdot P + R} \tag{12}$$

Similar to [1, 5], we use $\beta^2 = 0.3$ to weight precision more than recall. The resulting curves show that our method (TD) achieves the best F-measure and recall (see Fig. 6b). In comparison to other approaches, the textural distinctiveness scheme can detect more salient regions with high precision. Except for the SF approach [21], TD achieves the similar



(a) (b) (c)

Figure 6: a) Precision and recall (zoomed) for different numbers of texture atoms (#AT), using a 5×5 square neighborhood and sorted (STR) and unsorted textural representation (UTR). b) Precision, recall and F-measure for adaptive thresholding. c) Precision, recall and F-measure for cut-based (GrabCut [23]) segmentation of salient objects, initialized with saliency maps from all tested saliency approaches.

Figure 7: GrabCut segmentation [23] based on statistical textural distinctiveness. From left to right: Input image, saliency map computed with our approach, segmented image after adaptive thresholding, and GrabCut segmentation.

precision such as the region contrast (RC) saliency approach [5] and low-rank (LR) saliency approach [25].

However, performing simple adaptive thresholding procedures on images containing differently colored and textured objects is challenging and might result in noisy segmentations (see Fig. 7, 3rd column). To increase the robustness of segmenting salient regions, Cheng *et al.* [5] suggested to perform GrabCut [23] as a post processing step on thresholded saliency maps. They use empirically chosen thresholds that give 95% recall rate. However, this depends on the chosen saliency appraoch, and requires prior knowlegde which is difficult to extract from unknown images. Hence, we used the adaptive threshold Δ_{ada} (see Eq. 11) to produce binary images, and refined the results obtained using GrabCut. Fig. 7, 4th column illustrates that the saliency-guided GrabCut approach produces good masks even for challenging images, and significantly improves precision, f-measure and recall of all approaches (see Fig. 6c). Fig. 6c also shows that our method (TD) achieves the best precision and F-measure due to the consideration of texture and the sparse texture model for saliency computation, which can help to reduce the influence of cluttered background on saliency computation.

5. Conclusions

In this paper, a novel saliency detection approach for natural images based on the concept of statistical texture distinctiveness was presented. Experimental results using a public natural image dataset demonstrated strong potential for identifying salient regions in images in an efficient manner, thus illustrating the usefulness of explicitly incorporating textural characteristics. Future work involves investigating alternative sparse textural representation and textural models to evaluate whether improvements in saliency detection can be achieved. This also involves investigating schemes for automatically determining the optimal number of textural representations, i.e., number of atoms, which explicitly take into account textural relationships between individual representations for better sparse texture model learning. Furthermore, it would also be of great interest in exploring the extension of the proposed statistical textural distinctiveness approach to higher-dimensional data such as volumetric data as well as video data.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 1597 –1604, 2009.

[2] R. Achanta and S. Suesstrunk. Saliency detection using maximum symmetric surround. In *Int. Conf. on Image Processing*, pages 2653 –2656, 2010.

[3] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. on Graphics*, 26(3):10, 2007.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] M. M. Cheng, G. X. Zhang, N. Mitra, X. Huang, and S. H. Hu. Global contrast based salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 409 – 416, 2011.

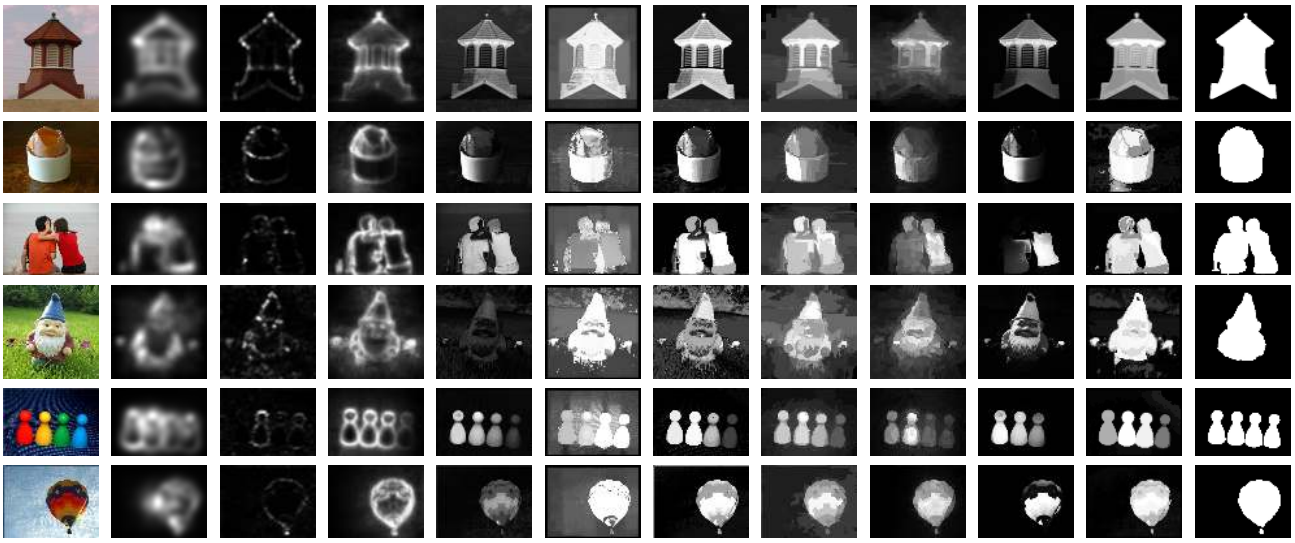


Figure 8: Visual comparison of our approach (TD) with other saliency approaches and ground truth (GT).

[6] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. In *Conf. on Computer Vision and Pattern Recognition*, pages 473 –480, 2011.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 2376 –2383, 2010.

[8] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Conf. on Computer Vision and Pattern Recognition*, pages 1 –8, 2008.

[9] J. Han, K. Ngan, M. Li, and H. Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE Trans. Circuits Syst. Video Technol.*, 16(1):141 –145, 2006.

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545 –552, 2007.

[11] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Conf. on Computer Vision and Pattern Recognition*, pages 1 –8, 2007.

[12] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *NIPS*, 2005.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11):1254 –1259, 1998.

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Int. Conf. on Computer Vision*, 2009.

[15] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):881 –892, Jul 2002.

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

[17] L. Li, P. Fieguth, G. Kuang, and H. Zha. Sorted random projections for robust texture classification. In *Int. Conf. on Computer Vision*, pages 391 –398, 2011.

[18] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. In *Conf. on Computer Vision and Pattern Recognition*, pages 1 –8, 2007.

[19] Y. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Int. Conf. on Multimedia*, ACM MULTIMEDIA, pages 374 –381, 2003.

[20] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. In *Int. Conf. on Image Processing*, volume 1, pages I–253, 2003.

[21] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 733 –740, 2012.

[22] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *Europe. Conf. on Computer Vision*, pages 366 –379, 2010.

[23] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graphics*, 23(3):309 –314, 2004.

[24] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 2 –37, 2004.

[25] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *Conf. on Computer Vision and Pattern Recognition*, 2012.

[26] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley. Image saliency: From intrinsic to extrinsic context. In *Conf. on Computer Vision and Pattern Recognition*, pages 417 –424, 2011.

[27] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. *ACM Multimedia*, page 815 824, 2006.