RELATIONAL ENTROPY-BASED SALIENCY DETECTION IN IMAGES AND VIDEOS

Kester Duncan, Sudeep Sarkar

Computer Science and Engineering Department, University of South Florida
Tampa FL, U.S.A.

ABSTRACT

Salient regions in an image facilitate the non-uniform allocation of computational resources to just the interesting parts of an image. In this paper, we present a saliency detection mechanism using relational distributions that capture geometric statistics based on distance and gradient direction relationships between pixels. The entropy of these normalized distributions is related to saliency. We employ an efficient technique for calculating the Rényi entropy of the probabilistic relational distributions using Parzen window weighted samples, thus eliminating the need for constructing intermediate histogram representations. We quantitatively demonstrate the biological plausibility of our method by showing how the saliency maps produced strongly correlate to human fixations in still images and to dominant objects in video. We find that our approach is better than six other saliency models.

Index Terms— Saliency detection, Parzen window density estimation, Rényi entropy, Motion saliency

1. INTRODUCTION

Saliency is the quality that enables image pixels to stand out relative to their neighbors, thereby permitting immediate concentration on candidate objects of interest in an image. It is believed that biological vision systems are robust to clutter due to inherent saliency mechanisms. Consequently, these vision systems seldom need to perform an extensive scan of a scene in order to detect an interesting object. In this paper, we focus on the automatic detection of salient regions in images and video. This has applications in object recognition [1], segmentation [2, 3], and background subtraction [4].

As a result, saliency has been well studied in the computer vision literature, but still remains an open problem. Most saliency detection approaches are either image-driven (bottom-up) [5, 6, 7, 8], task-dependent (top-down) [9], or a combination of both (integrated) [1]. Bottom-up models are independent of the knowledge of the content in an image whereas top-down approaches demand a more thorough understanding of the scene. For example, Itti and Koch presented a bottom-up saliency model in [5] that mimicked early vision in humans by normalizing and summing luminance, color, and orientation feature maps. Also, Hou and Zhang presented a saliency detection mechanism based solely on the log spectra representation of images [6]. Furthermore, Achanta et al. presented a saliency model that produced full-resolution saliency maps by exploiting the luminance and color properties of images [3]. Even though bottom up models are quite flexible and generic in nature, little work has been done whereby the geometry of image feature relationships is used as a saliency indicator. Inspired by this, we propose a novel bottom-up saliency model based on distance and gradient direction relationships between pixels.

Notably, most saliency approaches focus on single images but seldom on video. Motion cues in video can be strong saliency indicators. Taking this into consideration, we present a method that works for both image and video saliency detection without significant modification. In terms of algorithmic novelty, we present a pure sampling-based strategy for computing the entropy of a distribution that captures geometric statistics without the need for constructing histograms.

2. PROPOSED SALIENCY MEASURE

Our saliency measure is formulated on the entropy of geometric statistics between pixels. Geometrically organized regions in images, e.g., window blinds or symmetric arrangements of flower petals, will have geometric distributions possessing sharp peaks, i.e. low entropy, whereas the corresponding entropy for a disorganized image patch will be high. We exploit this phenomenon in our measure.

2.1. Relational Distributions

We adopt the notion that the structure perceived in an image is determined more by the relationships among image features rather than by the individual feature attributes. We capture these image structures by probability functions referred to as relational distributions. We construct these distributions using Parzen window density estimation. Relational distributions can be defined as follows [7]. Let 

\[ F = \{ f_1, ..., f_N \} \]

represent the set of N features in an image, 

\[ F_k \]

represent a random k-tuple of features, and let the relationship among these k-tuple features be denoted by 

\[ R_k. \]

Therefore, pairwise relationships between features are represented by 

\[ R_2, \]

which we consider in this paper. Low-order spatial dependencies are captured by small values of k whereas higher-order dependencies are captured by larger values of k.

Also, let the relationships 

\[ R_k \]

be characterized by a set of M attributes 

\[ A_k = \{ A_{k1}, ..., A_{kM} \}. \]

Hence, image structures can be represented by joint probability functions: 

\[ P(\mathbf{A}_k = \mathbf{a}_k), \]

also denoted by 

\[ P(a_{k1}, ..., a_{kM}) \]

or 

\[ P(\mathbf{a}_k), \]

where \( a_{ki} \) is the value taken by the relational attribute \( A_{ki}. \) These distributions can be interpreted as: given an image or video frame, if you pick k-tuples of features, what is the probability that it will exhibit the relational attributes \( a_{ki} \) or 

\[ P(\mathbf{A}_k = \mathbf{a}_k)? \]

2.1.1. Pixel-based features

Each pixel, \( f_i \), is associated with a gradient orientation, \( \theta_i. \) To capture some structure between two pixels, we use the difference between gradient orientations \( (\theta_i - \theta_j) \) and the Euclidean distance \( (d_i - d_j) \) between them as the attributes, represented as 

\[ \{ A_{21}, A_{22} \}. \]
of $R_2$. These attributes are ideal because they are invariant with respect to image plane rotation and translation. We also utilize the gradient magnitude differences between pixels as weights $w_i$ for the probability density estimation using Parzen windows. This effectively reduces the contribution of pixels in regions of constant intensity and increases the contribution of pixels in textured regions.

2.2. Saliency Measure

For video sequences, the saliency measure is based on relational distributions constructed from pixel pairs sampled from the $x$, $y$, and $t$ directions, whereas for single images sampling is done only in the $x$ and $y$ directions. We use the R´enyi entropy to measure the amount of uncertainty or disorder of the relational distribution $P(d, \theta)$ and this is represented as:

$$H_\alpha(P) = \frac{1}{1 - \alpha} \log_2 \left( \sum_{i=1}^{n} p(x_i)^\alpha \right).$$ (1)

This is known as R´enyi’s Entropy of order $\alpha$ where $\alpha \geq 0$. We utilized an $\alpha$ value of 2 (Collision entropy). We believe that finding the entropy of the relational distribution $P(d, \theta)$ is a good indicator of the salient image structures. Thus, the saliency function $\Phi(.)$ is defined as $\Phi = 1 - H_\alpha[P(d, \theta)]$, where $\Phi$ is a measure of pixel saliency considering the pairwise comparisons of pixels. Higher values of $\Phi$ indicate greater saliency.

3. INCREMENTAL ENTROPY ESTIMATION

Geometric relational distributions of local pixel neighborhoods must be constructed for every pixel in an image. Rather than adopting a brute-force method that uses all of the pixels in a local neighborhood, we employ a sampling technique that uses a subset of pixels to construct the distribution. For our experiments, sampling at 25% was more than sufficient for producing well-defined saliency maps. Samples are taken from $M \times M \times M$ volumes and $M \times M$ square neighborhoods (where $M$ is the window dimension), for videos and individual images respectively. We use Gaussian pyramids to allow detection at multiple scales.

The R´enyi Entropy is given by $H_\alpha(p) = -\log_2 \int p^\alpha(x) dx$, where $p(x)$ is the probability density function. We can concentrate on computing the quantity $V(p) = \int p^2(x) dx$ and compute the logarithm at the end. It is interesting to note that this quantity may be expressed as $V(p) = E(p(x))$, the expectation of the density, $p(x)$, with respect to itself [10]. The density function, $p(x)$ can be estimated using the Parzen window density as $\hat{p}(x) = \frac{1}{N} \sum_{x_i \in D} K_\alpha(x, x_i)$, where $D$ is the sample set. $K_\alpha(x, \cdot)$ is the kernel or Parzen window, centered at $x_i$, with a width controlled by the parameter $\alpha$. $K_\alpha(x, \cdot)$ is a density function itself, which in turn makes $\hat{p}(x)$ a proper density function. We use the Gaussian distribution as the kernel function. By employing the sample mean approximation of the expectation operator, we have

$$\hat{V}(p) = \hat{E}(p(x)) \approx \frac{1}{N} \sum_{x_i \in D} \hat{p}(x_i)$$
$$= \frac{1}{N} \sum_{x_i \in D} \frac{1}{N} \sum_{x_j \in D} K_\alpha(x_i, x_j)$$
$$= \frac{1}{N^2} \sum_{x_i \in D} \sum_{x_j \in D} K_\alpha(x_i, x_j),$$ (2)

which is the estimate of the R´enyi entropy in kernel form used in [10]. We consider a variation of this form where we weight each sample. This weight is based on the gradient magnitudes. The weighted Parzen window density estimator is given by,

$$\hat{p}_w(x) = \frac{1}{\sum_{x_i \in D} w_i} \sum_{x_i \in D} w_i K_\alpha(x, x_i),$$ (3)

where $0 \leq w_i \leq 1$. The weighted version of Equation 2 is then given by

$$\hat{V}_w(p) = \frac{1}{w_k} \sum_{x_i \in D} \sum_{x_j \in D} w_i w_j K_\alpha(x_i, x_j),$$ (4)

Equation 4 reveals a neat formulation for computational exploitation. It consists of summations which can be iteratively computed. In light of this, we have

$$\hat{V}^k_w(p) = \frac{\hat{V}^{k-1}_w(p) + w_i w_j K_\alpha(x_i, x_j)}{w_k},$$ (5)

With the iterative formula given in Equation 5, we are able to effectively reduce redundant calculations inherent with the brute-force method of [7]. The Parzen window summations of pixels falling within the local neighborhood bounding boxes defined by each of the four randomly selected pixels (for pairwise comparisons) are incrementally updated. This reduces redundant calculations while bringing the R´enyi entropy estimates of the pixels closer to their real value. Moreover, for video sequences the saliency of a pixel is determined by its spatial as well as its temporal gradients. We analyze the changes that occur over $M$ frames and produce a saliency map highlighting those regions of the video sequence that stand-out in relation to their surroundings. Thus, a saliency map is produced for every $M$ frames as shown in Figure 1.

4. RESULTS

This section demonstrates the results of our approach. Video sequence results are presented first in Section 4.1 and the results for single images are presented in Section 4.2.

![Fig. 1.](image-url) For video sequences, each saliency map corresponds to a $M \times M \times M$ volume. In this case, $M = 5$. From the changes that occur over $M$ frames and produce a saliency map highlighting those regions of the video sequence that stand-out in relation to their surroundings.
4.1. Video sequence results

We evaluate the effectiveness of our saliency map results with respect to the background subtraction task, which is often treated as the complement of saliency detection. It is important to note that image points considered salient via a bottom-up approach may not always coincide with what is subjectively considered foreground. We used sequences from the UCSD Background Subtraction dataset [4] for which there was appropriate ground-truth. We compare our results with those from the algorithms of Hou and Zhang [6], and Itti and Koch [5]. We show some of our results in Figures 2 and 3. In Figure 2, the sequence begins with one person skiing from the left of the captured scene to the right under heavy snowfall. Another skier enters the scene later on in the sequence. We do not expect any of the saliency methods to completely isolate the skiers since they are not the only salient candidates in the sequence. The falling snow contrasting with the mountain backdrop also ‘pops-out’ from this scene. The method of [5] highlights frame points in the vicinity of the skiers, but also highlights many points from the relatively uninteresting mountain backdrop. The method of [6] correctly highlights the snow, but misses substantial points corresponding to the prominent skier for the second input frame shown. Our approach highlights both the falling snow and the skiers. Interestingly though, it also highlights points corresponding to the motion path of the prominent skier over those $M$ respective frames. Figure 3 shows the results from two input frames of a sequence with birds walking along the seashore. Our method correctly highlights all the pixels corresponding to the birds for both frames as well as some pixels corresponding to their motion paths. The method of [5] does not highlight points corresponding to the tail of the bird moving towards the right of the first frame shown. Also, [6] does not fully express how conspicuous the birds are in these frames. For a more quantitative comparison, Table 1 shows the true positive and false positive rates for these sequences, demonstrating how our algorithm outperforms both [6] and [5].

4.2. Individual Image results

Another goal of this work was to analyze the biological plausibility of our algorithm. To ascertain this, we evaluated the performance of our measure on single images in relation to that obtained by the human visual attention mechanism. We compared our saliency maps with empirical human fixation maps [11] constructed by recording human eye fixations over an image which was displayed to test subjects for a limited amount of time. We demonstrate some of our results in Figure 4 along with results obtained from work done in [5, 6, 12, 2, 13] and [3]. To compare our results objectively, we utilized the 120 images from the dataset of [11]. The ground-truth images, in the form of human fixation maps, were binarized using a fixed threshold. The average true positive and false positive rates were then computed for each algorithm in an effort to create the ROC curve shown in Figure 5. Based on the curves in Figure 5, we were able to determine that our algorithm is comparable to the state of the art, handsomely outperforming most of the methods tested. Its performance is quite close to that of [13]. It is important to note that our algorithm does not take advantage of color information, even though

---

Table 1. Average hit and false alarm rates for the sequences of Figures 2 and 3

<table>
<thead>
<tr>
<th>Average HR</th>
<th>Our approach</th>
<th>Hou et al.</th>
<th>Itti et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.903</td>
<td>0.717</td>
<td>0.780</td>
</tr>
<tr>
<td>0.2</td>
<td>0.974</td>
<td>0.866</td>
<td>0.856</td>
</tr>
<tr>
<td>0.3</td>
<td>0.990</td>
<td>0.956</td>
<td>0.899</td>
</tr>
<tr>
<td>0.4</td>
<td>0.996</td>
<td>0.987</td>
<td>0.936</td>
</tr>
<tr>
<td>0.5</td>
<td>0.998</td>
<td>0.993</td>
<td>0.958</td>
</tr>
</tbody>
</table>

---

For some of the sequences of the dataset of [4], ground truth masks were not provided for all the frames in the sequence.

---

$^1$University of California San Diego

$^2$For some of the sequences of the dataset of [4], ground truth masks were not provided for all the frames in the sequence.
color is a useful cue for saliency detection. Our performance here demonstrates the biological plausibility of our algorithm.

5. CONCLUSION

We presented a bottom-up saliency detection mechanism based on geometric relationships exhibited between pixels using the Rényi entropy of probability density estimates captured using Parzen window weighted samples. We highlighted those frame and image points which stood out relative to their local pixel neighborhoods. We demonstrated how our results coincide with what is perceived as the “interesting” image regions by applying our method to the background subtraction task in video. We also presented still image results that were shown to be closer to human fixations than six other state of the art saliency algorithms.

6. REFERENCES


