# IMAGE COMPRESSION USING OBJECT-BASED REGIONS OF INTEREST

*Sunhyoung Han*  *Nuno Vasconcelos*[*]
Department of Electrical and Computer Engineering
University of California, San Diego

## ABSTRACT

A new architecture for region of interest (ROI) image coding is proposed. ROIs are defined as image regions containing objects of interest, and an efficient algorithm proposed for the detection of such regions. This algorithm is based on the principle of discriminant saliency, under which salient regions are the image regions of strongest response for a set of features that discriminate the object class of interest from all others. The resulting ROI masks are fully compatible with the JPEG2000 standard. Experimental results are presented for images of complex scenes, which contain both objects and background clutter, demonstrating significant gains for object-based ROI coding, in terms of both subjective image quality and SNR. The proposed ROI-based coder is also shown to be trainable with small, informally collected, image collections (e.g. by simple web search). This suggests the possibility of user-trained image coders.

***Index Terms***— ROI coding, Object detection, discriminant saliency

## 1. INTRODUCTION

The problem of using regions of interest (ROI) to increase the efficiency of image/video transmission over constrained-bandwidth channels has received considerable attention in the image processing literature. This is due to the fact that various applications can strongly benefit from the uneven distribution of resources (bits, error protection, resolution, etc) through different image regions. For example, images can be robustly packetized by assigning higher priority to packets that cover a region deemed to require higher resiliency to transmission errors due to a noisy network [1, 2]. Also, in very low bit rate coding (e.g. video conferencing over cell phones) it is important to encode some image areas (e.g. faces) with higher fidelity than others (e.g. tree leaves moving in the background). Object-based ROIs can also be useful in applications such as web browsing, or image retrieval.

While the detection of ROIs has been studied by various researchers, existing solutions have significant practical limitations. These include ROI definitions based on low-level image attributes (e.g. edges [3, 5] or spatial homogeneity [8]) of small semantic significance, or the requirement for manual specification of ROI shape (or other geometrical properties) by eventual users [4]. More recently there have been efforts to formalize ROIs as perceptually salient regions [9, 7] but these methods are still based on bottom-up definitions of saliency, that cannot account for high-level image interpretation.

In this work, we propose an alternative formulation of ROI detection, based on top-down, object-based, saliency. Under this formulation, saliency is associated with object recognition, and defined in a discriminant sense; the salient attributes of an object are those that most distinguish it from all other objects that may be of interest. We build on the results of [10], namely the fact that it is possible to design discriminant saliency detectors of great computational efficiency, and introduce a number of extensions that enable the design of image coders with support for object-based ROIs. The proposed ROI detection architecture places no limitations on the size, or other geometrical properties, of the detected ROIs, and is fully compatible with existing ROI-based compression standards, such as JPEG2000. Experimental results show that object-based ROI-compression can 1) lead to significant coding gains, under both objective and subjective evaluations of coding fidelity, and 2) be trained with small collections of example images, informally collected through web search.

## 2. DISCRIMINANT SALIENCY

The proposed architecture for object-based ROI detection is based on the principles of discriminant saliency detection [10].

### 2.1. Feature selection

Computationally, discriminant saliency is implemented by identifying image features which are discriminant for the *1-vs-all* classification problem that opposes an object class of interest to all remaining classes. Defining a binary random variable $Y$ such that $Y = 0$ for the *all* class and $Y = 1$ for class of interest and assuming that the feature vectors are drawn from a random process $\mathbf{X} = (X_1, \ldots, X_n)$, the saliency of each feature is measured by the mutual information between the feature and the class label

$$I(X_k; Y) = < KL[P_{X_k|Y}(x|i) || P_{X_k}(x)] >_Y \qquad (1)$$

[*]This research was supported by NSF Career award IIS-0448609, NSF award IIS-0534985, and a gift from Google Inc.

where $KL[p||q] = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the Kullback-Leibler divergence between the distributions $p(x)$ and $q(x)$ and $< f(i) >_Y = \sum_i P_Y(i) f(i)$. The salient features for the class of interest are those that maximize this mutual information. The procedure is repeated for all classes for which saliency detectors must be designed, by making each class the class of interest. The whole process can be performed quite efficiently, since the bulk of the computation can be re-used from one class to the next. See [10] for details.

### 2.2. Features

In our experience, the precise choice of feature dictionary does not have a major impact on saliency judgments. We have tested various frequency decompositions including Gabor and Haar wavelets, and the discrete cosine transform (DCT), with similar results[1]. More important is to collect features at various image scales, since this enables the automatic determination of both the *location* and *scale* of salient image points. This is implemented by preliminary decomposition of the image into a Gaussian pyramid, and application of the feature transformation to each of the resulting pyramid layers. Color information is captured by measuring the hue, which is relatively unaffected by shadows and image variability due to illumination changes [6], at each image location. We currently do not use spatially-supported color features, but intend to investigate their use in the future. The feature vector $\mathbf{x}(l)$ at image location $l$ includes the hue and the multiresolution coefficients measured at $l$ for all pyramid scales.

### 2.3. Salient point detection

The detection of image points that are salient for a given class is implemented with the minimum probability of error rule for the *1-vs-all* problem associated with that class. This consists of a likelihood ratio test based on the class distribution and that of the *all* class, where the different features are weighted according to their saliency, and produces a measure of the discriminant saliency of each image location

$$S(l) = \sum_k I(X_k; Y) \log \frac{P_{X_k|Y}(x_k(l)|1)}{P_{X_k|Y}(x_k(l)|0)}. \qquad (2)$$

We refer to $S(l)$ as the *saliency map* with respect to the class of interest.

Salient points are local maxima of the saliency map. They are identified by feeding the latter to a peak detection module, based on non-maximum suppression implemented with a winner-take-all (WTA) network [11]: the location of largest saliency is identified, the saliency map suppressed (set to zero) in a neighborhood of diameter equal to the saliency scale at that location, the next most salient location is found, its neighborhood suppressed, and so on. The saliency scale of a given location is the scale (radius of the region of support) of the most salient luminance feature at that location. The process is iterated until the entire saliency map is suppressed. The result is a table of salient points characterized by location $l_k$, scale $s_k$, and amplitude $S(l_k)$. Note that salient locations are ordered by decreasing saliency.

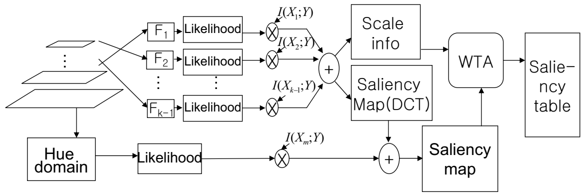

**Fig. 1**. Salient point detection.

According to the likelihood ratio test based on (2), only those $l_k$ for which $S(l_k)$ is larger than a threshold $T$ (which depends on the prior probabilities $P_Y(i)$) are truly salient. Instead of specifying this threshold directly, we rely on the conservative assumption that there cannot be more salient points than those required to cover the entire image. This gives an upper bound on the number of salient points $N$ of the form

$$\sum_{i=1}^{N} \pi s_i^2 \leq \mathcal{A},$$

where $\mathcal{A}$ is the image area. Finally, the amplitude of the salient points is normalized

$$\pi_k = \frac{S(l_k)}{\sum_{j=1}^{N} S(l_j)}$$

and the detector outputs the list of points $\mathbf{z}_k = (\pi_k, l_k, s_k)^T, k = 1, \ldots, N$. The various modules are summarized in Figure 1.

## 3. GENERATION OF ROI MASKS

Due to the conservative nature of our salient point selection, it is likely that some of the salient points detected may not belong to the object class of interest. Further elimination should not consider points in isolation, since they tend to cluster in object regions. Typically, the region of support of the object of interest is covered by a cluster of points that do not necessarily have the largest saliency amplitudes among the entire collection. The saliency of the whole cluster tends, nevertheless, to be highly correlated with the saliency of the region.

### 3.1. Saliency density

To address this problem, we interpret the saliency map as a *probability density of salient points*, i.e. a probability density from which salient points are drawn. This density is compactly characterized as a mixture of Gaussians with parameters determined by the amplitude, location, and scale of the

[1] All results reported in this work were obtained with the DCT

salient points $\mathbf{z}_k$, namely

$$p(l) = \sum_{k=1}^{N} \pi_k g(l, l_k, s_k^2 \mathbf{I}) \qquad (3)$$

where $g(x, \mu, \mathbf{\Sigma})$ is the density at point $x$ of a two-dimensional Gaussian of mean $\mu$ and covariance $\mathbf{\Sigma}$. This probability density can also be expressed as a sum of two components

$$p(l) = p_s(l) + p_o(l)$$

where $p_s$ is the distribution of truly salient points and $p_o$ that of outliers. The goal is to identify which components of (3) are associated with $p_s$ and which correspond to $p_o$.

### 3.2. ROI mask for single object

For this, we consider the spatial adjacency of the salient points. We assume that large connected components are associated with truly salient regions, while isolated saliency points are likely to be outliers, independently of their amplitude. To detect connected regions we use the following simple procedure

1. threshold $p(l)$, i.e. let

$$P(l) = \begin{cases} p(l) & p(l) \geq \pi_N \\ 0 & p(l) < \pi_N \end{cases}$$

2. successively identify $\mathcal{C}_k$, the largest connected component of $P(l)$, and remove it from $P(l)$, until $area(\mathcal{C}_k) \leq \alpha \ area(\mathcal{C}_{k-1})$.

3. set the ROI mask to $\mathcal{R} = 1_{\cup_k \mathcal{C}_k}$, where $1_\mathcal{C}$ is the indicator function of $\mathcal{C}$.

The first step eliminates probability tails due to the infinite support of the Gaussian, and the last sets the ROI to the union of the connected components. $\alpha$ is a user-defined parameter that enables control over the relative sizes of the connected components in the ROI. In all experiments we used $\alpha = 0.35$.

### 3.3. ROI mask for multiple objects

For scenes with multiple objects, the procedure is repeated for each of the object classes, typically producing some overlap between individual ROIs. This is addressed as is usual for the implementation of multiclass problems as a sequence of *1-vs-all* rules: conflicting decisions are disambiguated by the strength of the likelihood ratio of (2). In particular, if $p_c(l)$ is the saliency density of the $c^{th}$ class and $\mathcal{R}_c$ the associated ROI, location $l$ is classified as salient for class $c^*(l)$ such that

$$c^*(l) = \arg \max_{c \,|\, l \in \mathcal{R}_c} p_c(l).$$

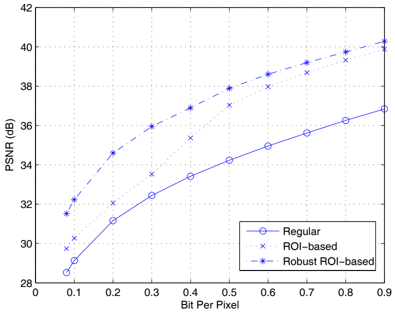

**Fig. 2**. Coding PSNR (dB) for normal vs. ROI-based JPEG compression.

## 4. ROBUSTNESS

Since image understanding is always error prone, it is important to include in the coder mechanisms that can recover from saliency detection mistakes. Our experience with the saliency detectors described above is that there is a significant difference between the appearance of the saliency map of images where saliency is correctly detected and those where this is not the case. While correct detection tends to originate a small number of clusters of strong saliency, the detection failures exhibit a much more uniform distribution of salient locations, with smaller clusters of lower amplitude.

We build on this observation to design a classifier of saliency detection success. This classifier takes as input the saliency map, and computes several features (such as $\pi_1$, $\pi_1 - \pi_2$, the average saliency difference between the ROI and remaining image area, etc.) which are then fed to a support vector machine trained from a collection of saliency detection successes and failures. This classifier is applied after ROI detection, to determine whether ROI coding should be used. If the image is classified as a saliency detection failure, the encoder reverts to non-ROI based compression, minimizing the likelihood of severe degradation of the ROIs due to saliency misjudgments.

## 5. EXPERIMENTAL RESULTS

We conducted a number of experiments to evaluate the performance of image compression using object-based ROIs, on two different image collections. The first was designed to provide an objective evaluation of the gains of ROI-based coding. The second was designed to test the viability of this type of coding under informal training conditions.

### 5.1. Objective evaluation

The first set of experiments was performed on the Caltech database of object classes [12], a database commonly used for the evaluation of object recognition algorithms. We only considered the "faces" class, which contains images of human faces presented against cluttered backgrounds[2]. The main reasons for considering this class were 1) its obvious interest

[2] Images have 896 × 592 pixels and faces about one quarter of this size.

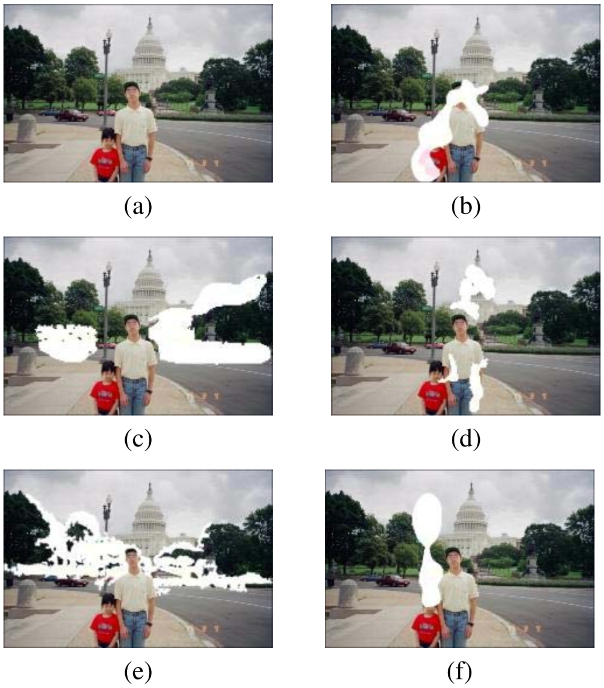
(a) (b) (c) (d) (e) (f)

**Fig. 3**. Original image (a) and ROI mask for class “faces” (b), “cars” (c), “Capitol” (d), “trees” (e), and “lamp” (f).

for very video-conference, and 2) the availability of ground-truth segmentation (a bounding box for each face). We note that segmentation information was not used to facilitate training of the saliency detector, which was learned from the cluttered images, but we relied on the segmentation ground truth to evaluate coding performance (by restricting the comparison to the face region). The 435 face images were divided into a training and test set, each containing half of the examples. The training set was used to design the saliency detector, and coding fidelity (PSNR after JPEG2000 encoding and decoding) was measured on test images.

The PSNRs measured on the face region, as a function to the average bit-rate spent to transmit the whole image, are presented in Figure 2. Results are presented for three JPEG coders: regular, ROI-based, and robust ROI-based. Regular indicates coding without ROIs, ROI-based refers to a coder that always employs ROI-based coding, and robust to a coder that relies on the classifier of Section 4 to detect saliency failures (in which case images are encoded in the normal mode). Note the consistent gain of close to 4dB of the robust coder over regular JPEG. Perceptually the gains are also significant, as illustrated in the following section.

### 5.2. Informal training

One potential problem for object-based ROI compression is the requirement for a training set of images from each class of interest. To investigate the feasibility of user-trained coders, we designed an experiment where the detectors were trained in quite informal conditions: training was based on a small training set (40 examples per class) of images collected from the web. The idea was to simulate the scenario where a user wants to encode the image of Figure 3 a), but has not detectors for the desired class (possible classes for this image include “faces”, “cars”, “trees”, “Capitol”, or “Lamp post”). The user simply goes on the web and collects a small set of examples from that class.

Figures 3 b) to f) present the ROI masks for the different object classes in the scene. Figure 4 shows a comparison of the images compressed with and without ROIs for the “faces” and “Capitol” classes. Note the significant improvement in image fidelity, that makes fine details significantly more legible.

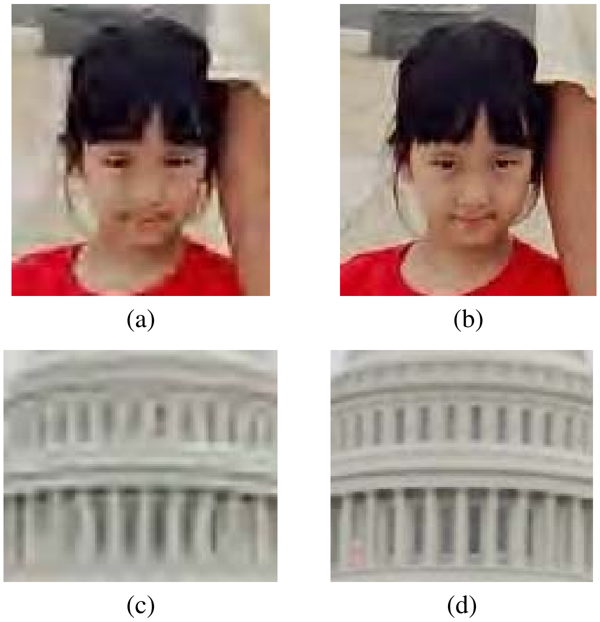
(a) (b) (c) (d)

**Fig. 4**. Regions compressed without (a,c), and with (b,d) ROI-based coding (in both cases, the average bit rate of the whole image is the same). (b) is compressed with ROI for the “face” class and (d) is compressed with ROI for the “Capitol” class

### 6. REFERENCES

[1] Sanchez, V.; Basu, A.; Mandal, M.K., “Prioritized region of interest coding in JPEG2000”, *IEEE Trans. on CSVT*, 14(9) 2004

[2] Sanchez, V.; Mandal, M.; Basu, A. “Robust, Wireless transmission of regions of interest in JPEG2000”, *ICIP* 2004

[3] K. An, M Lee, J. Shin, “Saliency map model based on the edge images of natural scenes”, *IJCNN* 2002

[4] Nguyen, G.P.; Worring, M. “An user based framework for salient detail extraction”, *ICME* 2004

[5] B Ko, S Kwak; H Byun, “SVM-based salient region(s) extraction method for image retrieval”, *ICPR* 2004

[6] Carron, T.; Lambert, P. “Color edge detector using jointly hue, saturation and intensity”, *ICIP* 1994

[7] C.M. Privitera, L.W. Stark, “Algorithm for defining visual Regions-of-Interest: Comparison with Eye Fixations”, *IEEE PAMI* 22(9), 2000

[8] Huibao Lin; Si, J.; Abousleman, G.P. “Knowledge-based hierarchical Region-of-Interest detection”, *ICASSP* 2002

[9] Itti, L. “Automatic foveation for video compresstion using a neurobiological model of visual attention”. *IEEE Trans. Im. Proc.*, 13(10), 2004

[10] D. Gao, N. Vasconcelos. “Discriminant Saliency for Visual Recognition from Cluttered Scenes”, *Neural Information Processing Systems*, 2004

[11] C. Koch, S. Ullman, “Shift in selective visual attention: towards the underlying neural circuitry,” *Neurobiol.*, vol. 4, pp. 219-227, 1985.

[12] R. Fergus, P. Perona, A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” *IEEE Conf. CVPR*, 2003.