INCREMENTAL SPARSE SALIENCY DETECTION

Yin Li, Yue Zhou, Lei Xu, Xiaochao Yang, Jie Yang

Institute of Image Processing and Pattern Recognition,
Shanghai Jiaotong University, China

ABSTRACT

By the guidance of attention, human visual system is able to locate objects of interest in complex scene. We propose a new visual saliency detection model for both image and video. Inspired by biological vision, saliency is defined locally. Lossy compression is adopted, where the saliency of a location is measured by the Incremental Coding Length(ICL). The ICL is computed by presenting the center patch as the sparsest linear representation of its surroundings. The final saliency map is generated by accumulating the coding length. The model is tested on both images and videos. The results indicate a reliable and robust saliency of our method.

Index Terms— Saliency Detection, Incremental Coding length, Sparse Coding

1. INTRODUCTION

In the real world, the biology vision system has the ability to locate the object of interest in complex background. During this stage, attention provides a mechanism to rapidly identify subsets within the scene that contain important information. A computational approach to attention serves as a key to further object detection or scene analysis, and provides an insight into the human vision system

The human visual system follows a center-surround approach in early visual cortex [1]. When the surrounding items resemble closely the central one, the neuron response to the central location is suppressive modulated; but when they differ grossly, the response is excitatory modulated [2]. Thus, the regions in the visual field which most differ from their surroundings in certain low level features automatically pop up. The candidates popped up here are called proto-objects [3], which could be further grouped into coherent units to be perceived as real objects. How can a computer vision system benefits from this presentation and extracts the salient regions from an arbitrary scene?

Several computational approaches have been proposed to model visual attention [4, 5, 6, 7]. Itti and Koch proposed a saliency model based on "feature integration" that simulates the visual search process of human [4] and extended it to surprising events detection in video [6]. Bruce and Tsotsos proposed a model of overt attention with selection based on the self-information [7]. Hou and Zhang proposed a saliency detection method based on spectral residues [5].

Inspired by aforementioned research, we propose a novel sparse center-surround saliency model and apply it to images and videos. By the local definition of saliency [1], our model follows a simple principle: a center patch is considered salient if it is more informative than its spatial or spatio-temporal surrounding patches. The 'information' is measured by Incremental Coding Length (ICL) [8]. By encoding the center patch with its surrounding patches, the ICL in our approach is approximated by finding the sparsest linear representation via l_1 minimization. The final saliency map is calculated by the coding length.

A local saliency model based on Incremental Coding Length is presented in Section 2; In Section 3, a sparse coding scheme is designed specifically to compute the incremental coding length; Experimental evaluations and result analysis are performed in Section 4; Finally, Section 5 concludes the paper.

2. INCREMENTAL SALIENCY MODEL

One basic principle in visual system is to suppress the response to frequently occurring input, while at the same time keeps sensitive to novel input. Based on this fact, we try to find a measure to decide whether one area differs significantly from its surrounding areas. Incremental coding length (ICL) is proposed as a criterion for classification and clustering in [8, 9]. The principle of ICL is that similarity could be represented as ability to compress. We propose a modified version of ICL to make it more suitable for the sense of saliency.

Denote I as an image and $c \in I$ as a patch of image I. Let $S(c)$ be the spatial or spatio-temporal surrounding patches of patch c, overlapping is allowed here in order to capture the structure information of the image (Fig.1).

Assume an arbitrary lossy coding method with fixed distortion ε is adopted to encode the patches of image I, then the coding length of patches $S(c)$ and $S(c) \bigcup c$ are denoted as $L_\varepsilon(S(c))$ and $L_\varepsilon(S(c) \bigcup c)$, respectively. If $L_\varepsilon(S(c))$ ap-

Supported by the National Natural Science Foundation of China, Grant NO.60772097, NO.60675023; China 863 High-Tech Plan No2007AA01Z164.

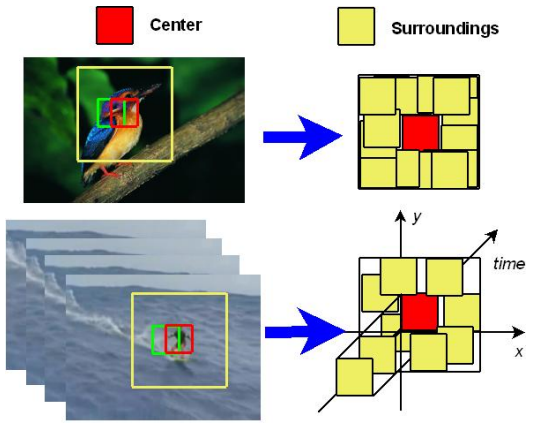

Fig. 1. Overlapping surrounding patches of the center patch

proaches to $L_\varepsilon(S(c)\bigcup c)$, we can infer that the center patch c is similar to its surrounding patches $S(c)$, vice versa. Thus we use their difference δL to represent the similarity (or novelty) between patch c and its surroundings $S(c)$, which forms the definition of Incremental Coding Length in [8, 9]:

$$\delta L_\varepsilon(c) = L_\varepsilon(S(c)\bigcup c) - L_\varepsilon(S(c)) \tag{1}$$

Large value of $\delta L_\varepsilon(c)$ indicates less similarity and more saliency between patch c and its surroundings $S(c)$.

Inspired by the center-surround modulation in biology vision [2], we propose a simple scheme to compute $\delta L_\varepsilon(c))$ directly, which is given by.

$$\delta L_\varepsilon(c) = L_\varepsilon(c|S(c)) \tag{2}$$

where $c|S(c)$ represents a coding scheme for the center patch using its surrounding patches, which will be further explained in section 3.

Assume the distortion is tolerable and the Incremental coding length $\delta L_\varepsilon(c)$ is computed for patch c, we can measure the saliency based on $\delta L_\varepsilon(c)$. For simplicity, we define the saliency $Sa(c)$ of patch c as

$$Sa(c) \;=\; \delta L_\varepsilon(c) \tag{3}$$

Furthermore, patch level saliency is accumulated to form an image saliency map. If the size of the center and its surroundings is given, the fixed distortion ε becomes the only user defined parameter of our model.

This incremental saliency model is simple in concept, intuitive, and highly flexible. Actually, any lossy coding scheme can be used to measure the saliency of a patch c. However, effective saliency measurement demands that the chosen coding scheme be approximately optimal for the given patches.

3. SPARSE CODING SCHEME

The incremental coding length requires the coding scheme be approximately optimal. In [9], a theoretic coding length bound is proposed based on the assumption of gaussian distribution. Due to the complex structure of the image or video and the specific situation in our problem, we design a sparse coding scheme for the problem in this section.

3.1. Sparse Representation

The simplest and most common coding scheme is the linear coding scheme. Assume each center patch c is represented by a column vector $\boldsymbol{x} = Fc \in R^m$, and the surrounding patches $S(c)$ is represented by a set of vectors $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2 ... \boldsymbol{s}_N] = FS(c)$ where $s_i \in R^m$. F is some kind of (linear) feature extraction transform. We represent the center $\boldsymbol{x}$ as the linear combination of its surroundings $\boldsymbol{S}$

$$\boldsymbol{x} \approx \sum_{i=1}^{N} w_i \boldsymbol{s}_i = \boldsymbol{S}\boldsymbol{w} \tag{4}$$

where $\boldsymbol{S} \in R^{m\times N}$ and $\boldsymbol{w}$ is a vector of weights.

In most cases, the overlapping surrounding patches result in $m \ll N$. That is why we could always expect a sparse representation of $\boldsymbol{x}$. To be more precise, $\boldsymbol{S} \in R^{m\times N}$ is an natural overcomplete dictionary of N patch atoms since $m \ll N$, and the the center patch $\boldsymbol{x}$ can be represented as a sparse linear combination of these atoms. That is, the center $\boldsymbol{x}$ can be written $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{w}$ where $\boldsymbol{w} \in R^N$ is a vector with very few ($\ll N$) nonzero weights.

3.2. Solution via l_1 Minimization

Given $\boldsymbol{x}$ and $\boldsymbol{S}$, we need to solve the vector of weights $\boldsymbol{w}$. However, for $m < N$ the system of equations $\boldsymbol{x} = \boldsymbol{S}\boldsymbol{w}$ is typically underdetermined and its solution is not unique. Traditional l_2 norm minimization produces a dense solution. To effectively measure the saliency, we need to find the optimal coding length which is closely related to the sparsity.

Suppose the vector of weights $\boldsymbol{w}$ is achieved, every nonzero entry of $\boldsymbol{w}$ should be coded in the same length. Thus, the coding length $\delta L_\varepsilon(c)$ for patches c is proportional to the number of nonzero entries of $\boldsymbol{w}$. The more sparse $\boldsymbol{w}$ is, the shorter the coding length will be. Thus, we address the optimal coding scheme as to finding the sparsest representation of $\boldsymbol{x}$, which can be formulated as the l_0 norm optimization:

$$min \; \|\boldsymbol{w}\|_0 \quad s.t. \quad \|\boldsymbol{x} - \boldsymbol{S}\boldsymbol{w}\|_2^2 \le \varepsilon \tag{5}$$

Coincidentally, theoretical studies suggest that primary visual cortex (area V1) uses a sparse code to efficiently represent natural scenes [2]

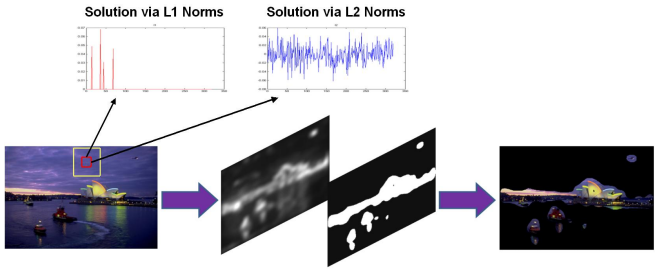

Fig. 2. The procedure of our method: l_1 norms v.s. l_2 norms

The optimization (5) is an NP-hard problem, but recent studies [10, 11] indicate that as long as the vector of weights $\boldsymbol{w}$ are sufficiently sparse(always satisfied when $m \ll N$), the problem can be efficiently solved by instead minimizing the l_1-norm:

$$min \ \|\boldsymbol{w}\|_1 \quad s.t. \quad \|\boldsymbol{x} - \boldsymbol{S}\boldsymbol{w}\|_2^2 \leq \varepsilon \tag{6}$$

Due to the Convex Analysis, an equivalent formulation could be achieved as

$$min \ \lambda\|\boldsymbol{w}\|_1 \ + \ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{S}\boldsymbol{w}\|_2^2 \quad (\lambda > 0) \tag{7}$$

Since the fixed distortion assumption could not be satisfied here, λ becomes the only user defined parameter and balances between the sparsity and distortion. The optimization problem (7) is essentially a linear regression known in statistical literature as the Lasso [12].

Solving (7) for each patch provides a sparse coding scheme for our incremental saliency model. The parallel nature of the coding scheme is similar to the early human visual system.

3.3. Saliency Map

Suppose a sparse solution $\boldsymbol{w}$ of the (7) is achieved, the coding length $\delta L_\varepsilon(c)$ for patches c is proportional to $\|\boldsymbol{w}\|_0$. Therefore, the saliency of patch c could be calculated as

$$Sa(c) \ = \ \delta L_\varepsilon(c) = \ \|w\|_0 \tag{8}$$

Since overlapping is allowed, the final saliency map is generated by accumulating the saliency per pixel. The proto-object could be further separated from the saliency map by threshold. And the entire model is summarized as Algorithm 1.

4. EXPERIMENT AND ANALYSIS

To evaluate the performance of our model, experiments on both images and videos are conducted. For simplicity, we set the saliency map at the resolution of 320×240 in all experiments. And we choose the center patch to be a block of 8×8 pixels. For better visual effect, a 2D gaussian filter with $\sigma = 4$ is performed on all the results.

Algorithm1 (Incremental Sparse Saliency)

1.***Input*** **:** given image I
2.***for*** each patch c of the image I, calculate $\boldsymbol{x} = Fc$ and take patches from its surroundings to form $\boldsymbol{S}$
- solve the optimization problem $min \ \lambda\|\boldsymbol{w}\|_1 \ + \ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{S}\boldsymbol{w}\|_2^2$
- given the sparse solution $\boldsymbol{w}$, calculate the patch saliency $Sa(c)$ by $Sa(c) = \|w\|_0$, and accumulate the saliency by pixels

3.***end***
4.***Output*** **:** the saliency map of I

4.1. Natural Images

To fairly test the method, 80 hand labeled natural images with resolution around 800×600 are used as a test set, which are also used in [5, 4]. The surrounding region of a center patch is set 9 times the size of the center block. Thus, the only parameter for our method is the λ, Fig.3 shows its effect on the results.

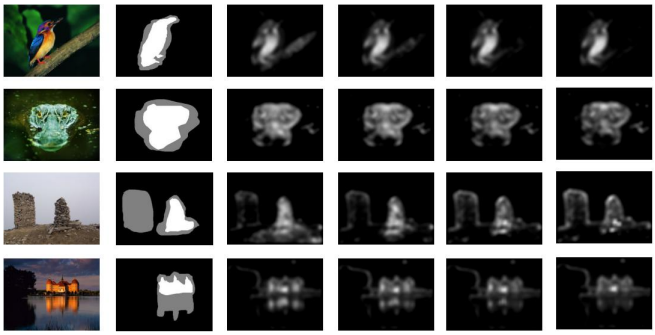

Fig. 3. Saliency map produce by hand label, and our method with different $\lambda = 0.1, 0.2, 0.3, 0.4$(from left to right) respectively. Larger λ shows more discriminant saliency.

From Fig.3, we see that larger λ generally produces more discriminant results. Ideally, if $\lambda = 0$ the optimization would produce a dense solution; if $\lambda \to \infty$ the optimization only have the zero solution. In real implementation, we first normalize the pixel value of the image and choose $\lambda = 0.25$ according to our experience. Moreover, the 5% boundary of the saliency map is neglected due to the boundary effect of image. We compare our results with previous methods in the field [4, 5]. For our method, only the raw data is used($F = I$) and all color channels are combined as a single vector. Fig. 4 shows some results of our methods. Though slower than the spectral residues, our method provides a more robust solution.

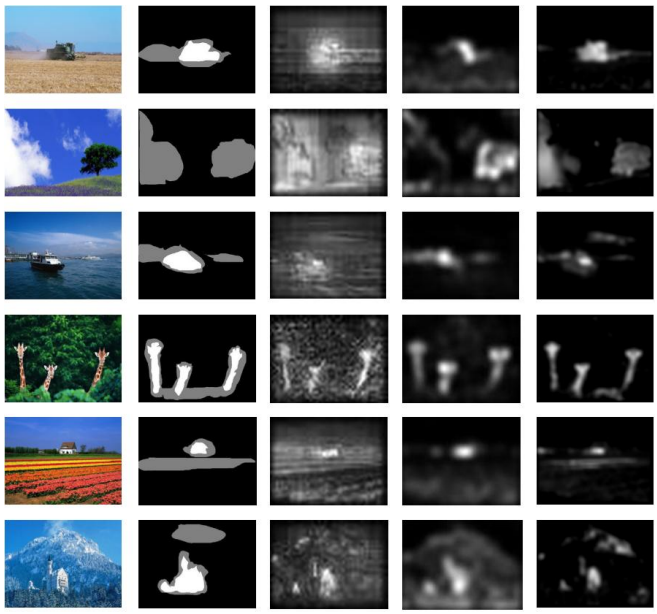

Fig. 4. Salient map produced by hand label, Itti's method [4], spectral residues [5] and our method.

4.2. Video Sequences

We use two videos collected from the internet with the resolution 320×240 to test the performance of our method. One has has a a stationary background with moving pedestrians and vehicle; the other one has a highly dynamic background with surfers riding a wave. The center-surround architecture provides our method insensitivity to ego-motion. The parameter is similar to the natural image, except that 5 frames are used as the spatio-temporal surroundings of the center patch. Fig.5 and Fig.6 present the results of our methods.

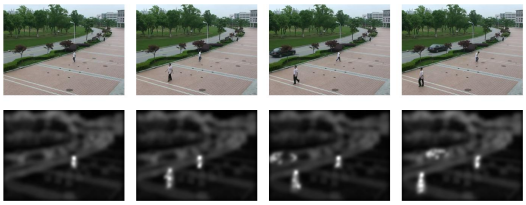

Fig. 5. Salient map in video with static background.

5. CONCLUSION AND FUTURE WORK

We proposed a method called Incremental Sparse Saliency to model both the spatial and spatio-temporal saliency. Further work may include the investigation into the role of feature extraction and the application of our method in computer vision.

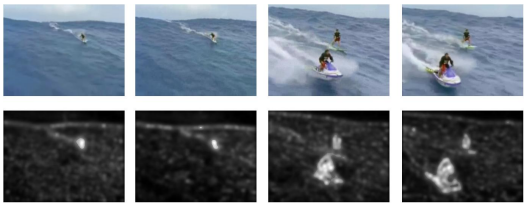

Fig. 6. Salient map in video with dynamic background.

6. REFERENCES

[1] T.N. Wiesel D.H. Hubel, "Receptive fields and functional architecture in two nonstriate visual areas," *Journal of Nerophysiol*, vol. 28, pp. 229–289, 1965.

[2] William E. Vinje and Jack L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[3] Ronald A Rensink, "Seeing, sensing, and scrutinizing," 2000.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[5] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, June 2007.

[6] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 631–637 vol. 1, June 2005.

[7] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18*, pp. 155–162. MIT Press, Cambridge, MA, 2006.

[8] John Wright, Yangyu Tao, Zhouchen Lin, Yi Ma, and Heung-Yeung Shum, "Classification via minimum incremental coding length (micl)," in *Advances in Neural Information Processing Systems 20*, pp. 1633–1640. MIT Press, Cambridge, MA, 2008.

[9] Yi Ma, H. Derksen, Wei Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1546–1562, Sept. 2007.

[10] David L. Donoho., "For most large underdetermined systems of equations, the minimal l^1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[11] D.L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[12] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.