SPARSE LIKELIHOOD SALIENCY DETECTION

Minh Chau Hoang and Deepu Rajan

School of Computer Engineering
Nanyang Technological University
Singapore

ABSTRACT

This paper addresses the problem of detection salient regions in images by exploiting the redundancy in image patches. We assume that redundant patches are more likely to be sparsely represented by other patches in the image while salient patches are not. Such sparse likelihood can be measured via L1-minimization by finding the sparse representation of an image patch based on a dictionary constructed using all other patches from the input image. We show that this approach leads to a robust saliency algorithm and the evaluation based on a database of 1000 images demonstrates that our algorithm achieves significant improvement over existing methods.

***Index Terms*—** sparse representation, saliency, L1-minimization

1. INTRODUCTION

Saliency detection involves finding regions in an image that capture the attention of the human visual system. It is useful in applications such as object recognition and image retargeting. Since it is well-known that V1 primary visual cortex can be efficiently represented by a sparse code based on an over-complete dictionary, sparse representation has been exploited for saliency detection. Such sparse representation is obtained from Independent Component Analysis (ICA) in [1] and [2]. Sun [1] proposed a saliency algorithm based on the difference-to-average approach using sparse representation learned by ICA for each image patch. Using a trained dictionary via ICA as the features and the sparse coding of an image patch as the response to such bases, Hou [2] measured saliency value based on the amount of entropy an image patch introduced to the system. The sparse representation found in these methods is often not very sparse and will not work when the dictionary size becomes large and contains more redundancy, in which case the problem of finding sparse representation becomes NP-hard. Luckily, recent advance in compressed sensing have provided robust tools such as [3] to tackle the problem. Applying these tools, commonly known as the L1-minimization approach has resulted in many successful works in a wide range of applications [4]. The most recent application of L1-minimization for saliency detection is the algorithm proposed by Li [5], in which saliency value can is measured by the length of the sparse coding of a center patch using the surrounding as a dictionary. However, because the coding length of a image patch may vary depending on the rank of the low-dimensional subspace it lies in regardless of whether it is salient or not, this measured saliency value may not be stable.

In this paper we develop a statistical approach to L1-minimization in the sparse coding framework for robust detection of salient regions. By assuming redundant image patches are more likely to have sparse representation based on a dictionary constructed by other patches in the image, we show how a L1-minimization-based framework can naturally lead to a robust algorithm which outperforms other existing methods.

2. SALIENCY MODEL

2.1. L1-minimization review

Given a signal $\mathbf{y} \in \mathbb{R}^n$ which can be sparsely represented by some over-complete dictionary $D \in \mathbb{R}^{n \times M}, M > n$ which contains M atoms $\{\mathbf{d}_i\}_{i=1}^{M}$. Suppose $\mathbf{y}$ can be expressed as a linear combination of dictionary atoms i.e. $\mathbf{y} = \sum_{i=1}^{M} \mathbf{d}_i \alpha_i$, $\alpha_i \in \mathbb{R}$. The coefficient $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_M]^T$ can be found by solving the system of equations $\mathbf{D}\alpha = \mathbf{y}$. Since $\mathbf{D}$ is over-complete, this system is under-determined hence there are infinite solutions. One might be interested in only the sparsest the solution which can be obtained by solving a L1-minimization problem given by:

$$(P1): \quad \hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \quad ||\mathbf{D}\alpha - \mathbf{y}||^2 + \lambda||\alpha||_1 \qquad (1)$$

where λ is a parameter that controls the trade-off between reconstruction error and sparsity. This problem has drawn much attention recently and can be solved efficiently by linear programming methods such as [3].As shown later, one may learn very interesting information about natural image patches using L1-minimization.

2.2. L1 approximation with non-negativity constraint for natural image patches

Let $\mathbf{S} = \{\mathbf{y}_i \in \mathbb{R}^n, i = 1 \dots N\}$ be a collection of natural image signals sampled from an input image simply by stacking pixels of image patches of size $\sqrt{n} \times \sqrt{n}$ in a lexicographic manner and suppose they are normalized to unit norm. We observe that a set of signals obtained from similar patches in this manner are highly correlated. For instance, in figure 1 similar blue and brown patches having varying brightness and pattern have an astonishingly high minimum dot product with their corresponding normalized mean - 0.9842 and 0.9945, respectively [1]. Suppose all signals collected from the input

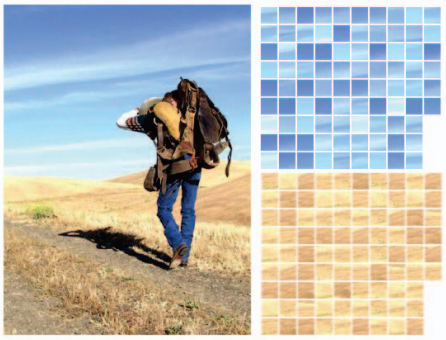

Fig. 1. Left: original image. Top right: patches sampled from the sky region with minimum dot product of 0.9842 to their normalized mean. Down right: patches sampled from the grass region with minimum dot product of 9.9945 to their normalized mean. The minimum dot product between all patches sampled from the image is 0.2354.

are stacked together to form a dictionary $\mathbf{D} \in \mathbb{R}^{n \times N}$, which is partitioned into M sub-matrices $\mathbf{D} = \{\mathbf{C}_i\}_{i=1}^{M}$, $\mathbf{C}_i \in \mathbb{R}^{n \times N_i}$, each containing N_i similar signals which exhibits such degenerate structure. If $\mathbf{C}_i$ is redundant enough, a new signal $\mathbf{y}$ drawn from $\mathbf{C}_i$ can be linearly represented by $\mathbf{y} = \mathbf{C}_i \alpha_i$ where $\alpha_i \in \mathbb{R}^{N_i}$. Hence, $\mathbf{D}, \mathbf{y}$ can be linearly expressed by $\mathbf{D}$ i.e. $\mathbf{y} = \mathbf{D}\alpha$ such that $\alpha = [0, \dots, \alpha_i^T, \dots, 0]^T$, that is all indices of α are zero except those associated with $\mathbf{C}_i$. If $\{\mathbf{C}_i\}_{i=1}^{M}$ are separated enough, the sparse solution α can be found via L1-minimization approach. Based on the nature of the natural image patches, we further require that α is non-negative. The problem of interest hence turns to be

$$(P2): \quad \hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \quad ||\mathbf{D}\alpha - \mathbf{y}||^2 + \lambda||\alpha||_1, \quad \alpha \succeq 0. \tag{2}$$

This is reasonable since $\mathbf{D} \succeq 0$ and $\mathbf{y} \succeq 0$ and a contribution of *negative patch* to the target patch is hard to interpret [2]. Such sparse representation is sparser and more informative in comparison to representation learned by conventional methods like ICA, where the coefficient is often widespread across all bases. For instance, one may expect that with a high probability a sparse solution α can be found exactly, i.e. non-zero indices in α should only correspond to patches which are most similar to $\mathbf{y}$ (figure 2).

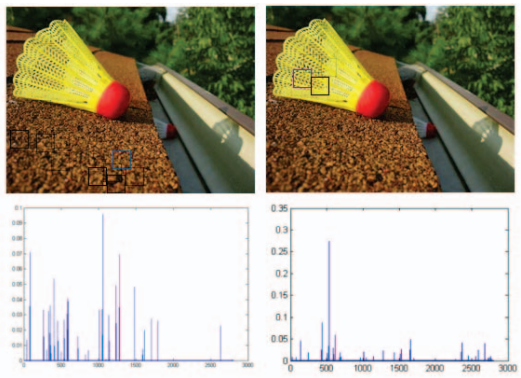

Fig. 2. Result of L1-minimization indicates similarity between patches. Top row: The target patch (blue rectangle) is approximated via L1-minimization using a dictionary $\mathbf{D}$ formed by sampling other patches from the image in overlapping manner. Black rectangles with varying transparency indicate how much weight is given to a patch in order to approximate the target patch. Bottom row: sparse coefficient learned by non-negative L1-minimization using algorithm from [7].

2.3. Saliency measurement via statistical perspective

With the constraint of non-negative coefficient, natural image signals are modeled in a way that signals from the same 'class' span a tight and highly concentrated convex cone. We have seen that such structure can be exploited by a L1-minimization approach using a dictionary $\mathbf{D}$ formed by all signals sampled from the input image. Since $\mathbf{D}$ contains all information about the input image, given a new input signal $\mathbf{y}$, one may be interested to learn some statistics information of $\mathbf{y}$ given $\mathbf{D}$. For instance, if signals which are similar to $\mathbf{y}$ appear to be redundant in $\mathbf{D}$, it is likely that a sparse approximation of $\mathbf{y}$ can be found. On the other hand, if $\mathbf{y}$ is rare and does not belong to any cone a sparse representation is very hard to achieve (figure 3). Hence we propose to use a statistical approach to measure how likely an input $\mathbf{y}$ can be sparsely represented by $\mathbf{D}$. It is known that minimizing the equation given in problem (P1) corresponds to a MAP inference in a probabilistic model with a Laplacian prior [8]. Let α have a Laplacian distribution, i.e $p(\alpha) = \frac{1}{2}e^{-|\alpha|_1}$. The

[1] This is to compare with observation done by [6] where highly correlated face images of the same person have a minimum dot product of 0.723 with their normalized mean.

[2] This assumption is also aligned with observation from Wright [6] that even without explicit constraints, the coefficient tends to be non-negative.

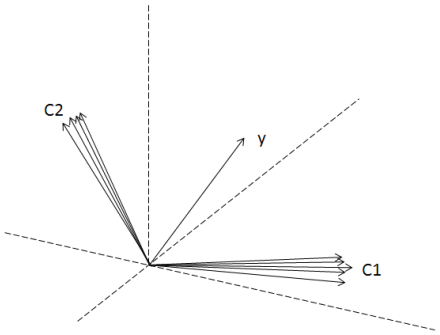

Fig. 3. Signals belong to the set spanned by 'bouquet' $\mathbf{C1}$ or $\mathbf{C2}$ are easy to be approximated with a sparse representation. It is hard to get a sparse representation for signal like $\mathbf{y}$ which does not belong to neither $\mathbf{C1}$ or $\mathbf{C2}$.

MAP estimate of α is:

$$\hat{\alpha} = \underset{\alpha}{argmin}\{-\log p(\alpha|\mathbf{y},\mathbf{D})\} \quad (3)$$

$$= \underset{\alpha}{argmin}\{-\log p(\mathbf{y}|\alpha,\mathbf{D}) - \log p(\alpha|\mathbf{D})\} \quad (4)$$

Assuming α is independent of $\mathbf{D}$, $-\log p(\alpha|\mathbf{D}) \sim |\alpha|_1 + c$ where c is some constant. With an appropriate Gaussian distribution model on $p(\mathbf{y}|\alpha,\mathbf{D})$ solving equation (4) is equivalent to the L1-minimization in the form of equation (1). The error of approximation $||\mathbf{D}\alpha - \mathbf{y}||^2$ is then a good indicator of how likely $\mathbf{y}$ can be sparsely represented by $\mathbf{D}$.

Let $p(i) = \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}||\mathbf{D}_i\alpha_i - \mathbf{y}_i||^2}$ be the likelihood measurement $p(\mathbf{y}_i|\alpha_i,\mathbf{D}_i)$ of the event patch i belonging to the sparse model with dictionary $\mathbf{D}_i$, then the rarity/saliency of patch i can be measured by:

$$s(i) = 1 - \bar{p}(i) \quad (5)$$

where $\bar{p}(i)$ is the normalized probability $p(i)$ to the range 0-1. $\mathbf{D}_i$ is a dictionary formed by all signals in $\mathbf{S}$ except for $\mathbf{y}_i$. To satisfy the non-negativity constraint, α_i is the coefficient learned by solving problem (P2).

2.4. Intensity integration

Using a L1-minimization approach will require that all input signals have to be normalize to L2-norm unit length to avoid the scaling problem where vectors with high magnitude tends to be penalized [9]. By doing so we lose the brightness information. Although some tolerance to the intensity is good, a very dark and very bright patches should not be treated similarly. Any modification made should maintain the convex cone structure and should not affect the input signals in term of magnitude. One solution is to map the intensity values to a set of polar vectors with radius 1 and angle varies from θ_{min} to θ_{max}. Let $\mathbf{y}_i$ be the original signal, the new vector $\mathbf{y} = [\mathbf{y}_i^T \mathbf{i}_i^T]^T$ can be formed by concatenating the original vector $\mathbf{y}_i$ with the intensity vector $\mathbf{i}_i$. The L2 norm of each vector $\mathbf{y}_i$ is then $||\mathbf{y}_i||^2 + ||\mathbf{i}_i||^2$ i.e. no discrimination is made in term of magnitude since for $\forall i$, $||\mathbf{i}_i||^2$ = 1. Furthermore, because the inner product between 2 vectors is then $\mathbf{y}_i^T\mathbf{y}_j + \mathbf{i}_i^T\mathbf{i}_j$, by varying the range $[\theta_{min},\theta_{max}]$ one can control the discrimination power of the intensity to the original input signals. The important factor here is how large the range $[\theta_{min},\theta_{max}]$ is, not θ_{min} and θ_{max} individually.

2.5. Adaptive dictionary

One common problem with saliency algorithms is that it is often hard to identify large size objects. Many algorithms based on surrounding contrast often highlight strong edges and miss the interior of the object. The convex cone model we propose can handle this situation very easily. In case of large size object, a salient patch may have its surrounding similar to itself, but yet in terms of a global context this patch is still very distinctive. The presence of similar patches in the dictionary results in a good approximation and hence low saliency value. Excluding the surroundings in the dictionary is not a good remedy since a patch which is different from the surrounding is definitely salient. Therefore, one may want to remove only similar patches which lie in the surrounding area of the target patch. Based on the proposed model, one simple solution is to eliminated any surrounding patch with an inner product with the center patch higher than a value β from the dictionary.

3. EXPERIMENTS

We conduct the experiments with our proposed algorithm using the database of 1000 color images with ground-truth masked by human, provided by [10]. Patches of size 8x8x3 are sampled from the input image with overlapping of 4 pixels to form a vector of size 196. This vector is concatenated with a 2D intensity vector carrying the average brightness of the patch to form our input signals set. For a limited effect of varying intensity, we choose the parameter $[\theta_{min}\theta_{max}]$ mentioned in section 2.4 to be $[0, \frac{\pi}{4}]$. For each signal, a dictionary is constructed by discarding the target signal and similar signals in a surrounding area of 5 times the patch size, in which the parameter β is set to 0.7. For each pair of signal and dictionary, the problem in equation (2) is solved using algorithm provided by [7] with parameter λ is set to 0.05. To evaluate the performance, we use the Receiver Operating Characteristic (ROC) method and compare our algorithms with ISS [5], short-term ICA [1], ICL [2] and Itti's [11]. Figure 4 shows that our algorithm outperforms all state-of-the-art algorithms, showing better consistency with the ground-truth. In terms of average area under the curve, our algorithms also yields the best result (table 1). Due to limited space we only show some samples of our saliency in comparison with saliency map of ISS (the next best algorithm in ROC evaluation). Unlike ISS, our saliency map is not attracted to strong edges. Due to the

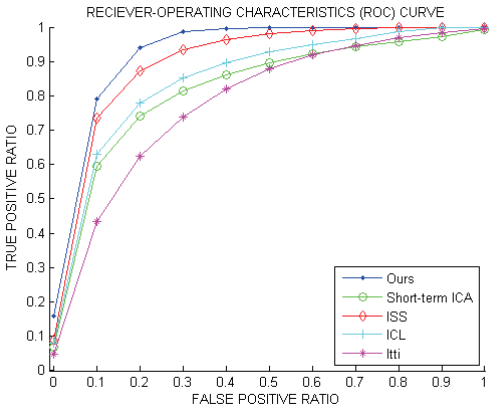

Fig. 4. Average ROC curves of all methods on 1000 images with human-masked ground-truth.

Ours	short-term ICA	Itti	ISS	ICL
0.9293	0.82375	0.78377	0.90167	0.8527

Table 1. Average area under the ROC curve of various methods

global excluding surrounding approach, our algorithm works best with salient object with relatively large size (figure 5).

4. CONCLUSION

In this paper we propose an algorithm which leverage the power of L1-minimization approach in saliency detection. Although the framework is relatively simple and saliency calculation is straight-forward, it is very easy to extend and integrate new information to improve the result.

5. REFERENCES

[1] X. Sun, H. Yao, R. Ji, P. Xu, X. Liu, and S. Liu, "Saliency detection based on short-term sparse representation," in *Internaltion Conference on Image Processing*. 2010, pp. 1101–1104, IEEE.

[2] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Advances in neural information processing systems*, vol. 21, no. 800, pp. 681–688, 2008.

[3] H Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, pp. 801, 2007.

[4] J. Wright, J. Mairal, G. Sapiro, and T.S Huang, "Sparse Representation for Computer Vision and Pattern Recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.

[5] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, "Incremental Sparse Saliency Detection," *Pattern Recognition*, 2009.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–27, Feb. 2009.

[7] S. Kim, "An Interior-Point Method for Large-Scale Logistic Regression," *Journal of Machine Learning Research*, vol. 8, pp. 1519–1555, 2007.

[8] P.J. Garrigues and B.A. Olshausen, "Group sparse coding with a laplacian scale mixture prior," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1–9, 2010.

[9] A.M. Bruckstein, D.L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[10] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Conference on Computer Vision and Pattern Recognition*. June 2009, number Ic, pp. 1597–1604, IEEE.

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

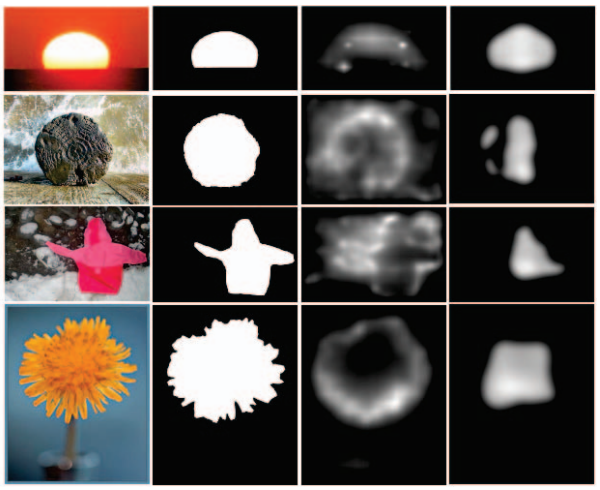

Fig. 5. Some examples of saliency maps generated. From left to right: input image, ground-truth saliency map, ISS method [5] and our method.