TELKOMNIKA, Vol. 11, No. 10, October 2013, pp. 6143 ~ 6150
ISSN: 2302-4046

An Image Sparse Representation for Saliency Detection

Jun Yang*[1], Tusheng Lin[1], Xiaoli Jin[1]
School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China
*Corresponding author, e-mail: yangjun9118@gmail.com

Abstract

This paper presents a novel method for detecting saliency in static images based on image sparse representation. For each color channel, first, the image is partitioned into non-overlapping patches and each patch is represented by the way of sparse coding from a learned dictionary of patches from natural scenes. Then, global saliency and local saliency are calculated and fused to attain saliency of each patch. Local saliency is shown by popping out a patch from its surrounding patches. Global saliency is indicated by the rarity of a patch in the overall patches of the image. The final saliency map is attained by normalizing and fusing local and global saliency maps of all color channels. Experimental results in the benchmark image dataset demonstrate that the proposed method achieves a superior performance compared with most of state-of-the-art methods. Furthermore, both robustness and the low computational complexities make the presented algorithm feasible for subsequent applications.

Keywords*: global saliency, local saliency, sparse coding, saliency detection*

1. Introduction

Computational visual saliency model for human visual attention fascinates a lot of researchers in recent years. The human visual system has a remarkable ability to quickly grasp salient regions in nature scenes without training. How to simulate human vision is a very important goal in computer vision community. In the field of computer vision, many models have been proposed to perform this task automatically over the past few decades [1]-[3]. Most of the models used in this study transform a given input image into a two-dimensional intensity distribution that represents the saliency distribution over the image support, according to Koch and Ullman's original saliency map concept [4]. The saliency map can be used for a wide range of applications such as object detection [5], object tracking [6], image retrieval [7], and so on.

The approaches for determining low-level saliency can be based on biological models, purely computational ones, or a combination of both [8]. Some approaches detect saliency over multiple spatial scales [9]-[11], while others operate on a single scale [12]. In general, all methods use some means of determining local contrast of image regions with their surroundings using one or more of the features of color, intensity, and orientation. Usually, separate feature maps are created for each of the features used and then combined to obtain the final saliency map. Itti et al. [13] have built a computational model of saliency-based spatial attention derived from a biologically plausible architecture. Guo and Zhang [14] calculate the saliency map of an image from the phase spectrum of its quaternion Fourier transform (PQFT). Recently, Kim et al. [15] present a method for detecting salient regions in both images and videos based on a discriminated center-surrounding hypothesis that the salient region stands out from its surroundings.

Although previous approaches for obtaining saliency maps are very diverse, most of them fail to minimize false positives which occur in the big salient object or in the highly textured background areas. For example, an face in the garden should be salient, but some saliency detection methods only take the margin of the face as salient region just as shown in Figure 1.

In this paper, to overcome the problems mentioned above, we propose a novel framework for detecting the botton-up saliency maps in the static images. The novel framework for detecting saliency combines the global fashion with the local fashion. In the global fashion, our algorithm provides a statistical method for global saliency based on sparse representation of image blocks. At the same time, in the local fashion, we calculate the dissimilarity between a

Received April 2, 2013; Revised July 14, 2013; Accepted July 24, 2013

center patch and its surrounding patches to attain the local saliency. Compared with other state-of-the-art algorithms, an remarkable advantage of our algorithm is its robustness and its low computational complexity that make it feasible for subsequent applications.

The remainder of this paper is organized as follows. Section 2 addresses the technical details about our method. In Section 3, we demonstrate the effectiveness of the proposed approach by providing some experimental results. Finally, we conclude the paper in Section 4 with some relevant discussion.

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

Figure 1. Comparison of saliency maps generated by using SUN [9], IS [11] and our approach. (a) original images; (b) eye movement fixation desity map [16]; (c) saliency maps obtained using SUN; (d) saliency maps generated using IS; (e)saliency maps generated using our approach.

2. Proposed Method

Our proposed framework is presented in Figure 2. First, an input image is transformed into RGB format. For each color channel, the image is partitioned into nonoverlapping patches and each patch is represented by the way of sparse coding from a learned dictionary of patches from natural scenes. And then a global saliency based on image sparse representation, and a local saliency map based on the dissimilarity between a patch and its surrounding window, are computed, normalized, and combined. Corresponding global saliency map and local saliency map are normalized and combined to form the final saliency map of the model. The whole process can be performed to generate the scale-invariant saliency map. The details of the model are as follows.

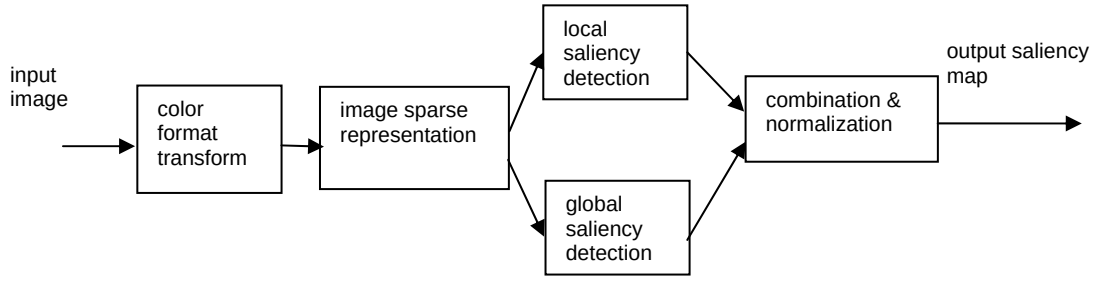

Figure 2. The framework of our proposed model

2.1. Image Color Format Transform

In the human brain, there exists a “color opponent-component system”. In the center of receptive fields, neurons which are excited by one color are inhibited by another color. Red/green, green/red, blue/yellow, and yellow/blue are color pairs which exsit in human visual cortex [17]. For example, color opponent cells and intensity cells in human cortex can detect color, intensity features at their receptive fields respectively. Therefore, the RGB color image can be transformed as follows:

$$\begin{aligned} I &= R + G + B \\ RG &= R - G \\ BY &= G - \frac{R+G}{2} - \frac{\min(R,G)}{2} \end{aligned} \tag{1}$$

where R,G and B denote the red, green, and blue components of an input image. *I* refers to intensity channel, *RG* and *BY* denote red/green, and blue/yellow channels respectively.

2.2. Image Sparse Representation

Sparse and redundant representation modeling of signals is a very effective way to describe the inner-structure of signal sources . Sparse coding along with dictionary learning has proven to be very successful in many image processing tasks such as face recognition [18], image denosing [19], and saliency detecion [9]. The underlying idea behind sparse coding is that a vision system should be adapted based on statistics of the visual enviornment where it is supposed to operate [20]. As a supporting evidence for this theroy, it has been shown that receptive fields(RF) of some neurons in V1 cortex resembel those RFs that are learned by sparse coding algorithm [21].

Using an $M \times c$ overcomplete dictionary matrix $\mathrm{D}=\left[\mathrm{d}_1,\cdots,\mathrm{d}_c\right]$ that contains c atoms, d_i, as its columns, it is assumed that an M-dimensional signal ***x*** can be represented as a sparse linear combination of these atoms. Usually, the representation of ***x*** is approximate, $\mathrm{X} \approx \mathrm{D}\alpha$, satisfying $\|X - D\alpha\| \le \varepsilon$ where $\|\bullet\|$ is the l^2 norm, ε is a prescribed tiny positive number, and the α denotes the representation coefficients of the signal X. If $M \square c$ and *D* is full-rank matrix, an infinite number of solutions are available for the representation problem, so constraints on the solution must be set. The solution with the fewest number of non-zero coefficients is certainly an appealing representation. Therefore, the sparsest representation is the solution of

$$\min \frac{1}{2}\|X - D\alpha\|_2^2 + \lambda_1 \|\alpha\|_1 \tag{2}$$

where $\|\alpha\|_1$ is the l^1 norm and λ_1 is a regularization parameter. To learn the dictionary *D*, considering a training set of n data samples $Y = [y_1, y_2, \cdots, y_n]$ in $\square^{m \times n}$ an empirical cost function $g_n(D) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, D)$ is minimized, where $l(y_i, D)$ is:

$$l(y_i, D) = \min_{A,\{\alpha_i\}_{i=1}^N} \sum_{i=1}^{N} \frac{1}{2}\|y_i - D\alpha_i\|_2^2 + \lambda_1 \|\alpha\|_1 \tag{3}$$

There are many optimization methods to resolve the question (3) such as the maximum likelihood, method of optimal directions, maximum a-posterior [22], and k-singular vector decomposition [23]. Further details can be referred to the papers mentioned above.

In this paper, to learn a dictionary for image representation, we extracted 800,000 image patches, 8×8 resolutions and each sub channel of I,RG, and BY, from 2000 randomly selected color images from various nature scenes. Each atom of the dictionary is a 64 dominations vector.

After dictionary learning, image sparse representation is as follows: Given an input image, it is first resized to $2^w \times 2^w$ pixels where patch size w is selected in a way that 2^w is divisible to w [20]. Let B={b1,b2,…,bn} denote the set of linearized image patches with nonoverlapping. And then, the sparse codes of each patch are computed with the learned dictionary above using OMP algorithm [24]. Finally, all of the sparse codes of each patch can make up the sparse representation of the image. For example, an image with 256×256 pixels can be divided into 1024 8×8 image patches with nonoverlapping, and a 200×1024 matrix can be attained for image sparse representation in one color channel.

2.3. Local Saliency Detection

Local visual saliency computes local contrast among pixels of the image. The underlying hypothesis is that fixation is attracted by high-contrast image details. Guided by the well-established computational architecture of Koch and Ullman et al. [4], we adopt the center-surrounding framework to compute the local saliency of each patch in the image. A problem with the computation of local contrast is that its value is highly scale-dependent. This yields the

unwanted side effect that textured regions may have high local contrast on a small scale, whereas they are not salient when observed at larger scale due to their highly predictable regular structure. To avoid this problem, mean local saliency is computed over a range of spatial scales. Local saliency in our model is the average dissimilarity between a center patch *i* and its M patches in a surrounding rectangular neighborhood. Considering various spatial scale, we calculate the local saliency of down-sampled images from the original image and then take the average as the final local saliency:

$$SL_i^c(x_i) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{\rho_{ij}^c}$$
$$SL^c(x_i) = \frac{1}{L}\sum_{l=1}^{L} SL_l^c(x_i) \quad (4)$$

where M is the number of surrounding patches of center patch *i*. L denotes the scale factor and L usually is set to 3. ρ_{ij} is the Euclidean distance between the center patch *i* and the surrounding patch *j.*

2.4. Global Saliency detection

It often happens that some uniformly textured salient objects would appear in the image. Although, most of traditional methods which use center-surrounding framework are able to detect the small salient object, but they are apt to fail in a homogeneous region so that they result in blank holes and only borders of salient objects. To solve this problem, we propose our global saliency detection method. There are two factors which are considered for evaluating the global saliency: the dissimilarities between image patches just like in local saliency detection, and their spatial distance of each patch from the center of the image because of the central bias as stated in [25]. With the increasing of the distance between a patch and the center, the saliency of the patch should be appropriately decreased. Global saliency is then defined as follows:

$$SG_i^c(x_i) = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{\rho_{ij}^c}\times\frac{1}{1+dist(x_i, x_{center})}\right)$$
$$SG^c(x_i) = \frac{1}{L}\sum_{l=1}^{L} SL_l^c(x_i) \quad (5)$$

where N is the number of all patches of the image. L denotes the scale factor and L usually is set to 3. ρ_{ij} is the Euclidean distance between the patch *i* and the another patch *j* in the image. $dist(x_i, x_{center})$ denotes the distance between patch *i* and the center patch of the image.

2.5. Fusion of Saliency maps

Gloabl saliency and local saliency are normalized samilar to [20] and combined as follows:

$$SL(x_i) = \sum_{c\in\{I,RG,BY\}} \Box(SL^c(x_i))$$
$$SG(x_i) = \sum_{c\in\{I,RG,BY\}} \Box(SG^c(x_i)) \quad (6)$$

where $\Box(\bullet)$ denotes normalized operation. $SL(x_i)$ and $SG(x_i)$ refer to local saliency and global saliency respectively. The final visual saliency can be defined as

$$Sal(x_i) = alpha\times SL(x_i) + (1-alpha)\times SG(x_i) \quad (7)$$

where alpha denotes a constant from 0 to 1.

3. Results and Analysis

The experiments are conducted on the ImgSal dataset of about 235 images provided by Lijian [16] and the AIM dataset provided by Bruce [26]. Images in the ImgSal database are divided into 6 categories which include 50 images with large salient regions, 80 images with intermediate salient regions, 60 images with small salient regions, 15 images with cluttered backgrounds, 15 images with repeating distractors, 15 images with both large and small salient regions, and their resolution is 480×640 pixels. The AIM dataset contains 120 color images and their resolution is 511×681 pixels. In our experiments, for computing visual saliency the size of image patches is commonly 8×8 pixels and alpha is 0.5 for the final saliency.

We evaluate the performance of saliency detection algorithms both qualitatively and quantitatively by comparison to human observers. For the former, we essentially compare the saliency map to the original image with eye movement tracking density map. For the latter, we have used freely available human fixation data as ground truth to quantitatively evaluate the algorithms. ROC curve and ROC score (area under the ROC curve, AUC) are adopted to measure their performance.

In order to validate the superiority of our proposed method, we compare our method with the state-of-the-art methods, which are hypercomplex Fourier transform (HFT) based method [16], saliency using natural statistics (SUN) base model [9], image signature (IS) based method [11], and the method exploiting image patch rarities (IPR) for saliency detection [20].

For qualitative assessment, we show saliency maps of methods mentioned above and our algorithm in Figure 3.

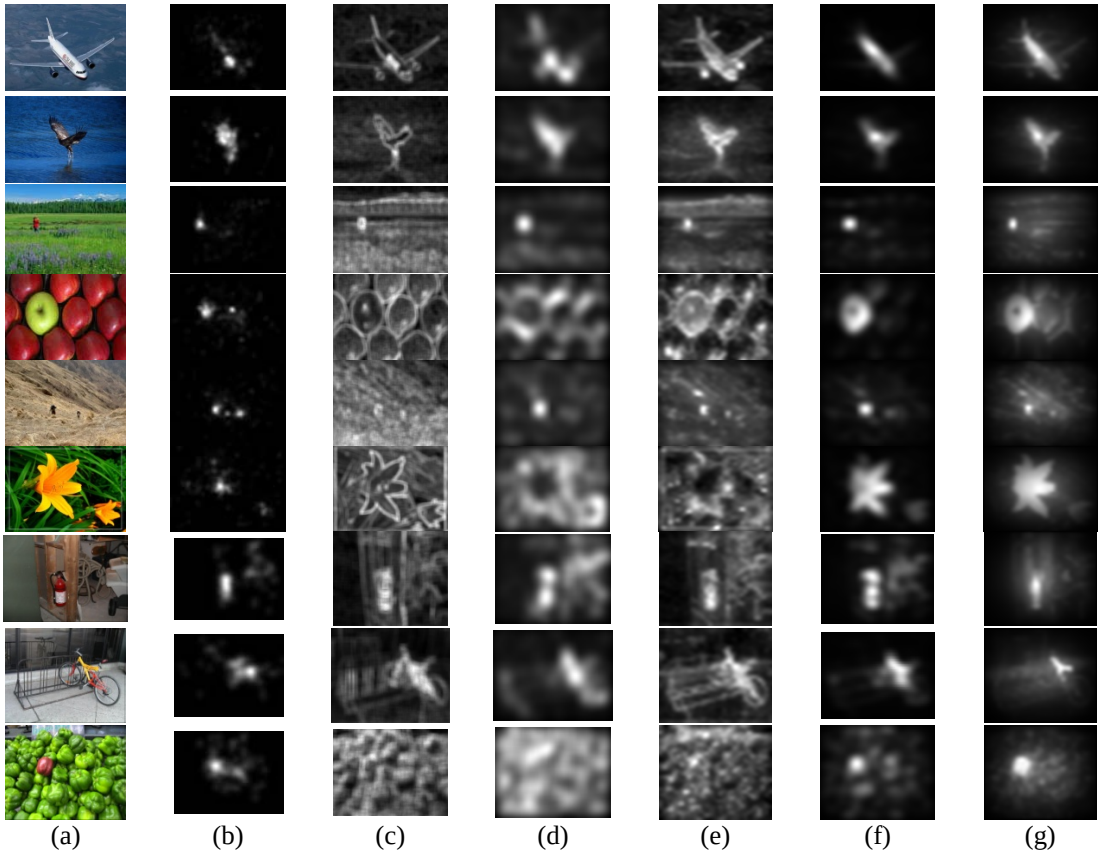

Figure 3. Comparison of saliency maps generated by using SUN, IS, IPR, HFT and our approach. (a) original images; (b) eye movement fixation desity map; (c) saliency maps obtained using SUN; (d) saliency maps generated using IS; (e) saliency maps obtained using IPR; (f) saliency maps obtained using HFT; (g) saliency maps generated using our approach.

As shown in Figure 3(a), we choose six representative images which are from six categories images in the ImgSal dataset (the first six rows) and three representative images from AIM dataset (the last three rows) respectively. Figure 3(b) shows the eye movement tracking desity maps of the nine images. For SUN based saliency maps as shown in Figure 3(c), some saliency objects can not be detected and some big salient object can not be detected completely, such as in the sixth row of Figure 3(c). For IS saliency maps as shown in the fourth row of Figure 3(d), some parts of the background usually show low contrast with the highlighted salient object region. As shown in Figure 3(e), saliency maps of IPR highlight the edges of salient objects. The same cases appear in Signature model as Figure 3(f). As compared to these results, it is easy to see that our method (Figure 3(g)) provides visually acceptable saliency, which is consistent with human visual attention. Compared with the other five approaches, our approaches show more robust performance on all of six images and more consistence with human fixation.

For quantitative assessment, the fixation ROC curves and ROC score for SUN, IS, IPR, HFT and our method are shown in Figure 4 and Table 1. In Figure 4, a comparison with ROC curves is illustrated. It is clear that the proposed method results in better performance than others. This is due to the fact that the local saliency and global saliency complement each other quite well. In Table 1, the ROC scores (area under the ROC curve, AUC) are shown, we can see that our proposed method attains the highest average AUC value in all the five methods. It is shown that our method is more consistent with human visual attention than others. However, for images with repeating distractors (C4 classification in Table 1), the AUC value of our proposed method is less than HFT mehod. The main reason is likely due to the poor sparse representation of repeating distractors in our proposed method.

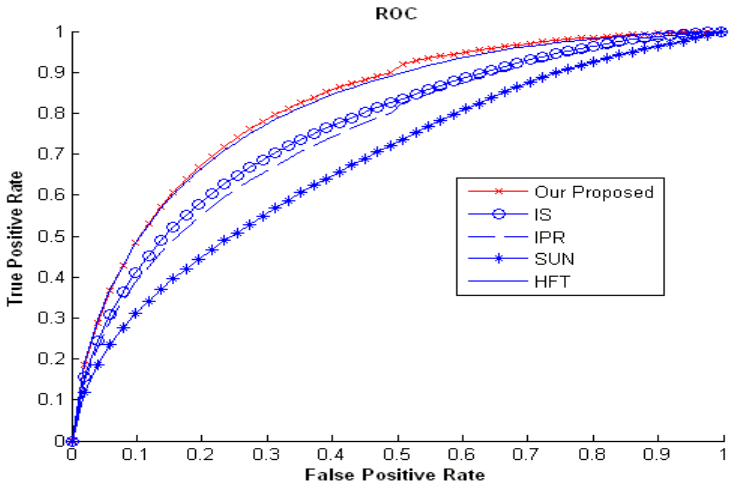

Figure 4. ROC curves of SUN, IS, IPR, HFT and our proposed method

Table 1. Comparison of the AUC using SUN, IS, IPR, HFT and our Approach

AUC \ Methods	SUN	IS	IPR	HFT	Our model
Overall ROC score	0.68026	0.74853	0.73426	0.80529	0.81003
C1 ROC score	0.67808	0.74301	0.74426	0.81388	0.82565
C2 ROC score	0.68148	0.77037	0.75142	0.81026	0.82084
C3 ROC score	0.70839	0.77148	0.74819	0.80561	0.81605
C4 ROC score	0.60584	0.70308	0.63733	0.79505	0.75537
C6 ROC score	0.63503	0.78289	0.78706	0.84449	0.85656
C5 ROC score	0.71279	0.80063	0.77491	0.84519	0.83916
AIM ROC score	0.67484	0.70927	0.71532	0.77678	0.79556

All experiments are performed on a personal computer with 2.66 GHz duo core CPU and 2GB RAM using MATLAB 2008a implementation of the five approaches. Table 2 shows the average processing time of single image on the two database using SUN, IS, IPR, HFT and our approach, respectively. We can see from Table 2 that our approach achieves the higher computational efficiency. Although our method is not the fastest, compared with Signature method and HFT method, our proposed algorithm is better in performance than others.

Table 2. Comparison of Average Processing Time Using All Methods (Seconds)

Methods	SUN	IS	IPR	HFT	Our model
Average processing time	4.364	0.538	2.223	0.866	1.154

4. Conclusion

In this paper, we have presented a framework for saliency detection in static images. In our formulation, bottom up saliency models as the combination local saliency with global saliency. The proposed algorithm is simple and computationally efficienct, and proved to perform better compared with most of the state-of-the-art methods. We do not deny that for images with repeating distractors, the proposed method is not better performance than the saliency model HFT. This paper consideres only the static images without considering the saliency maps of video sequences. Future work will take moving information into this framework.

References

[1] Itti L, Koch C. A comparison of feature combination strategies for saliency-based visual attention systems. *Human Vision and Electronic Imaging Iv*. 1999; 3644: 473-482.
[2] Liu C, Yuen PC, Qiu GP. Object motion detection using information theoretic spatio-temporal saliency. *Pattern Recognition*. 2009; 42(11): 2897-2906.
[3] Rosin PL. A Simple Method for Detecting Salient Regions. *Pattern Recognition*. 2009; 42 (11): 2363-2371.
[4] Koch C, Ullman S. Shifts in Selective Visual-Attention-Towards the Underlying Neural Circuitry. *Human Neurobiology*. 1985; 4(4): 219-227.
[5] Gopalakrishnan V, Hu YQ, Rajan D. Salient Region Detection by Modeling Distributions of Color and Orientation. *IEEE Transactions on Multimedia*. 2009; 11(5): 892-905.
[6] Zhang G, Yuan ZJ, Zheng NN, et al. *Visual Saliency Based Object Tracking*. Computer Vision - Accv 2009, Pt Ii. 2010; 5995: 193-203.
[7] Yang LJ, Geng B, Cai Y, et al. Object Retrieval Using Visual Query Context. *IEEE Transactions on Multimedia*. 2011; 13(6): 1295-1307.
[8] Toet A. Computational versus Psychophysical Bottom-Up Image Saliency: A Comparative Evaluation Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2011; 33(11): 2131-2146.
[9] Zhang LY, Tong MH, Marks TK, et al. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*. 2008; 8(7): 1-15.
[10] Walther D, Koch C. Modeling Attention to Salient Proto-Objects. *Neural Networks*. 2006; 19(9): 1395-1470.
[11] Hou XD, Harel J, Koch C. Image Signature: Highlighting Sparse Salient Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012; 34(1): 194-201.
[12] Seo HJ, Milanfar P. Nonparametric Bottom-Up Saliency Detection by Self-Resemblance. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2009)*. 2009; 1 and 2: 950-957.
[13] Itti L, Koch C. Computational modelling of visual attention. *Nature Reviews Neuroscience*. 2001; 2(3): 194-203.
[14] Guo CL, Zhang LM. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Transactions on Image Processing*. 2010; 19(1): 185-198.
[15] Kim W, Jung C, Kim C. Spatiotemporal Saliency Detection and Its Applications in Static and Dynamic Scenes. *IEEE Transactions on Circuits and Systems for Video Technology*. 2011; 21(4): 446-456.
[16] Li J, Levine M, An X, et al. *Saliency Detection Based on Frequency and Spatial Domain Analyses*. Proceedings of the British Machine Vision Conference. 2011; 86.1-86.11.
[17] Engel S, Zhang X, Wandell B. Colour tuning in human vision cortex measured with functional magnetic resonance imaging. *Nature*. 1997; 388(6,637): 68-71.

[18] Wright J, Yang AY, Ganesh A, et al. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2009; 31(2): 210-227.
[19] Aharon M, Elad M, Bruckstein A. Image denosing via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Signal Processing*. 2006; 54(11): 4311-4322.
[20] Borji A, Itti L. Exploiting local and global patch rarities for saliency detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012; 478-485.
[21] Olshausen BA. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*. 1996; 381(6583): 607-609.
[22] Delgado KK, Murray JF, Rao BD, et al. Dictionary learning algorithms for sparse representation. *Neural computation*. 2003; 15(2): 349-396.
[23] Aharon M, Elad M, Bruckstein A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation*. IEEE Transactions on Signal Processing*. 2006; 54(11): 4311-4322.
[24] Tropp JA. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*. 2004; 50(10): 129-159.
[25] Tatler BW. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*. 2007; 7(14): 1-17.
[26] Bruce NDB, Tsotsos JK. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*. 9(3):5, 1–24.