SALIENCY DETECTION USING TENSOR SPARSE RECONSTRUCTION RESIDUAL ANALYSIS

Yi Zeng, Yi Xu

Shanghai Key lab of Digital Media Processing and Transmission
Shanghai Jiao Tong University, Shanghai, China
zenghappy@126.com, xuyi@sjtu.edu.cn

ABSTRACT

In this paper, a visual saliency detection model based on tensor sparse reconstruction for images is proposed. This algorithm measures saliency value of image regions by the reconstruction residual and performs better on color images than current sparse models. Current sparse models treat a color image as multiple independent channel images and vectorise the image patches, ignoring interrelationship between color channels and spatial correlation between neighbouring pixels. In contrast, the proposed tensor sparse model treats a color image as a 3D array, retaining the spatial color structures entirely during the sparse coding. The proposed saliency detection method is tested on ASD dataset and OSIE dataset and compared with traditional sparse reconstruction based models. The experimental results show that our model achieves higher AUC scores than traditional sparse reconstruction based models.

Index Terms— tensor, tensor orthogonal matching pursuit (TOMP), saliency detection, sparse reconstruction, residual

1. INTRODUCTION

In dynamic visual scenes of complex environments, it is a very important mechanism for human beings to catch critical information effectively. In recent years, visual saliency has been studied by researchers in domains of psychology, neurophysiology and computer vision. Meanwhile, it becomes more significant to automatically extract the salient regions from images with an explosive growth of image information.

Some visual saliency detection models are proposed to be extensively used in object detection, target recognition and image comprehension. Most of these models take efforts to explain the cognitive process of humans [1], [2], [3]. Physiological experiments show that the neuron response is suppressive when the surrounding items are close to the center while the response is excitatory when they show a lot of difference from the center. Itti et al. [4] are motivated to define a visual attention model as center-surround contrast based on multi-scale image analysis, where a salient region pops up from a scene due to big difference from its neighbouring regions in the appearance of color, intensity and orientation.

Physiological data have suggested that primary visual cortex (area V1) uses a sparse code to efficiently represent natural scenes and the mechanisms in the area V1 contribute to the high saliency of pop-up objects. In recent years, the researchers are motivated to use sparse representation model for saliency computation, where the salient regions are extracted according to sparse reconstruction residuals since these regions cannot be well approximated using its neighbouing patches as dictionaries. Han et al. [5] proposed a weighted sparse coding residual model for bottom-up saliency detection, where the reconstruction residuals are weighted with the L_0 norm of sparse coefficients to produce the saliency map. In [6], the saliency value of each region is measured by the Incremental Coding Length (ICL), where the ICL is the description length of the sparse coding and increases when the center patch is more informative than its surrounding patches. All these methods used traditional sparse models to compute the reconstruction residuals. However, these traditional sparse models cannot provide a good approximation of the entire spatial color structures of the image.

In order to avoid color distortions during sparse representation, most recent works focus on establishing tensor-based sparse models to represent multi-channel images, e.g. RGB color images. As we know, tensor analysis is helpful in data structure preservation. Accordingly, tensor sparse models are expected to explore new solutions of classical problems of color image compression [7], 3D image reconstruction [12] and multispectral image denoising [8]. Tensor decomposition method had already been proposed by Tucker [10]. Based on Tucker decomposition technique, Caiafa et al. proposed TOMP [11] algorithm to compute sparse representations of a tensor. We use TOMP to calculate sparse representations of color images.

In this paper, we are motivated to propose a saliency detection model based on tensor sparse reconstruction method and center-surround mechanism of biological vision. To our knowledge, there are no prior works explicitly applying tensor sparse model in salient object detection. Meanwhile, we can expect that tensor sparse model will

provide a good solution of saliency detection problem due to well-preservation of local data structure.

The remainder of this paper is organized as follows. The theory of tensor sparse model including TOMP algorithm is presented in Section 2. A saliency detection scheme is designed to extract salient regions in Section 3. Experimental results and comparative analysis are shown in Section 4. Finally, we give some conclusion remarks in Section 5.

2. TENSOR SPARSE MODEL

In traditional sparse representation methods, a two dimensional (2D) or a three dimensional (3D) image patch is reformed to a long vector for sparse coding [5],[6],[13]. When the input is a color image patch, it would incur both color and spatial structure distortions since a 3D array is reduced to a 1D vector. As for a grey-scale image patch, it also loses spatial structure information since the 2D image matrix is reduced to a 1D vector. It is noted that structure distortions are introduced in abovementioned traditional sparse models during the reduced order approximation of the high-order array. To tackle this problem, we use tensor sparse reconstruction model to represent input images, which provides a useful analysis tool for high-order data without order reduction.

2.1. Basic concepts of tensor sparse representation

It is proved that tensor decomposition shows advantages of good preservation of local structures in handling color images, multispectral images, video sequences and other high-dimensional signals. Typical decomposition methods are PARAFAC [9] and TUCKER [10]. The PARAFAC decomposition model is to decompose the tensor as a sum of several rank-1 tensors with minimization of residuals, while the TUCKER decomposition decomposes a tensor into a set of 2D matrices and one small core tensor.

Throughout this paper, tensors are defined using bold handwritten letters, e.g. $\boldsymbol{\mathcal{A}} \in \mathrm{R}^{I_1 \times \ldots \times I_N}$, matrices using capital letters, e.g. $\mathrm{D} \in \mathrm{R}^{J \times I}$, and vectors using bold lowercase letters, e.g. **a**. The notations are listed as follows:

Table 1.
Notations

Notation	Explanation
$\|\|\boldsymbol{\mathcal{A}}\|\|$	$(\sum_{i_1,\ldots,i_N} a_{i_1,\ldots,i_N}^2)^{1/2}$
$\times_n$	mode-n product
$\otimes$	Kronecker product
$\circ$	outer product

Let's consider $\boldsymbol{\mathcal{A}} \in \mathrm{R}^{I_1 \times I_2 \ldots \times I_N}$ as an N-order coefficient tensor, then its mode-n unfolding matrix is a 2D matrix, i.e. $\boldsymbol{\mathcal{A}}_{(n)} \in \mathrm{R}^{I_n \times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N}$. Given a dictionary matrix $\mathrm{D}_n \in \mathrm{R}^{J_n \times I_n}$, then its mode-n product with $\boldsymbol{\mathcal{A}}$ is defined as follows,

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{A}} \times_n \mathrm{D}_n \qquad (1)$$
$$y_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_N} a_{i_1,\ldots,i_N} d_{n\, j_n i_n}$$

where $\boldsymbol{\mathcal{Y}} \in \mathrm{R}^{I_1 \times I_2 \ldots \times I_{n-1} \times J_n \times I_{n+1} \ldots \times I_N}$, $y_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N}$ is the element of $\boldsymbol{\mathcal{Y}}$, subscript represent its position in its dimension.

For a common case, we set N=3, i.e. $\boldsymbol{\mathcal{A}} \in \mathrm{R}^{I_1 \times I_2 \times I_3}$. According to the TUCKER decomposition model [14], tensor $\boldsymbol{\mathcal{Y}} \in \mathrm{R}^{J_1 \times J_2 \times J_3}$ is a multilinear transformation of a core tensor $\boldsymbol{\mathcal{A}}$ by the factor matrices $\mathrm{D}_i = \left[\mathbf{d}_{\mathbf{i}_1}, \mathbf{d}_{\mathbf{i}_2}, \ldots, \mathbf{d}_{\mathbf{i}_{I_i}}\right] \in \mathrm{R}^{J_i \times I_i}$, $i = 1,2,3$. $\mathbf{d}_{\mathbf{i_j}}$ is the column of $\mathrm{D_i}$ and $\mathbf{d}_{\mathbf{i_j}} \in \mathrm{R}^{\mathrm{J_i}}$, $\mathrm{j} = 1,2,\ldots,\mathrm{J_i}$. We can reformulate (1) as below:

$$\begin{aligned} \boldsymbol{\mathcal{Y}} &= \boldsymbol{\mathcal{A}} \times_1 \mathrm{D}_1 \times_2 \mathrm{D}_2 \times_3 \mathrm{D}_3 \\ &= \sum_{i_1=1}^{L_1} \sum_{i_2=1}^{L_2} \sum_{i_3=1}^{L_3} a_{i_1 i_2 i_3} \mathbf{d}_{\mathbf{1}_{i_1}} \circ \mathbf{d}_{\mathbf{2}_{i_2}} \circ \mathbf{d}_{\mathbf{3}_{i_3}} \end{aligned} \qquad (2)$$

Then it can be deduced in a matrix and vector form as,

$$\begin{aligned} \boldsymbol{\mathcal{Y}}_{(n)} &= \boldsymbol{\mathcal{A}}_{(n)} \mathbf{D}_n (\mathbf{D}_N \otimes \ldots \mathbf{D}_{n+1} \otimes \mathbf{D}_{n-1} \ldots \otimes \mathbf{D}_1)^T \\ vec(\boldsymbol{\mathcal{Y}}) &= (\mathrm{D}_3 \otimes \mathrm{D}_2 \otimes \mathrm{D}_1) vec(\boldsymbol{\mathcal{A}}) \\ \mathbf{y} &= (\mathrm{D}_3 \otimes \mathrm{D}_2 \otimes \mathrm{D}_1)\mathbf{a} \end{aligned} \qquad (3)$$

where $vec(\boldsymbol{\mathcal{Y}}) = \mathbf{y}$. Operator $vec(\cdot)$ is used to stack all columns of 3D array $\boldsymbol{\mathcal{Y}} \in \mathrm{R}^{J_1 \times J_2 \times J_3}$ as vector $\mathbf{y} \in \mathrm{R}^{J_1 J_2 J_3}$, and stack all columns of 3D array $\boldsymbol{\mathcal{A}} \in \mathrm{R}^{I_1 \times I_2 \times I_3}$ as vector $\mathbf{a} \in \mathrm{R}^{I_1 I_2 I_3}$.

As for PARAFAC decomposition method, it decomposes tensors as a linear summation of outer products of rank-1 vectors :

$$\boldsymbol{\mathcal{Y}} = \sum_i \boldsymbol{\lambda}_i \mathbf{u}_\mathbf{i} \circ \mathbf{v}_\mathbf{i} \circ \mathbf{w}_\mathbf{i} \qquad (4)$$

where vectors $\mathbf{u}_\mathbf{i} \in \mathrm{R}^{J_1}, \mathbf{v}_\mathbf{i} \in \mathrm{R}^{J_2}, \mathbf{w}_\mathbf{i} \in \mathrm{R}^{J_3}$ compose of the eigen subspace of $\boldsymbol{\mathcal{Y}}$ and λ_i is a scalar factor.

The process for tensor sparse representation of input 3-order signal $\boldsymbol{\mathcal{Y}}$ can be formulated as:

$$\boldsymbol{\mathcal{A}} = \mathbf{arg}\,\boldsymbol{min}_{\boldsymbol{\mathcal{A}}} \|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{A}} \times_\mathbf{1} \mathbf{D_1} \times_\mathbf{2} \mathbf{D_2} \times_\mathbf{3} \mathbf{D_3}\|_F + \boldsymbol{\lambda}\|\boldsymbol{\mathcal{A}}\|_\mathbf{0} \qquad (5)$$

where λ is the regularization parameter to achieve trade-off between the two cost terms. When dictionaries $\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3$ are known, it is more convenient to establish tensor sparse representation using TUCKER decomposition mode than using PARAFAC decomposition method.

2.2. Tensor Orthogonal Matching Pursuit （TOMP）

Saliency detection methods based on sparse representation show that the residual increases when a salient region is represented using surrounding neighbors as compared with a non-salient region. In this paper, we are motivated to use tensor sparse representation to represent color image patches. The dictionaries are selected as the 3×3 image patches adjacent to the center patch. Once the dictionary is given, the formula (5) can be solved by TOMP method [11]. TOMP is developed as an extension of OMP (Orthogonal Matching Pursuit) method to tensor space. The implementation details are given in Algorithm 1.

Algorithm 1: TOMP

Input: A color image patch $\mathcal{Y} \in \mathrm{R}^{J_1 \times J_2 \times J_3}$, 3-mode dictionaries $\{\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3\}$ with $\mathrm{D}_i \in \mathrm{R}^{J_i \times I_i}$, the predefined maximum number of non-zero coefficients k_{max} and reconstruction residual tolerance σ.

Output: sparse coefficients $\mathcal{A}, \mathcal{A} \in \mathrm{R}^{I_1 \times I_2 \times I_3}$

1. $\mathrm{M}_n = [\emptyset](n = 1,2,3)$, $\mathcal{R} = \mathcal{Y}$, $\mathcal{A}$=0, $\mathcal{X} \in \mathrm{R}^{I_1 \times I_2 \times I_3}$ $vec(\mathcal{X}) = \mathbf{x}$;
2. **While** $|\mathrm{M}_1||\mathrm{M}_2||\mathrm{M}_3| < \mathrm{k}_{max}$ && $\left|\left|\mathcal{R}\right|\right|_F > \sigma$
3. $[\mathrm{m}_1^k \mathrm{m}_2^k \mathrm{m}_3^k] = \arg\min_{[m_1 m_2 m_3]} |\mathcal{R} \times_1 \mathrm{D}_1{}^T(:, m_1) \times_2 \mathrm{D}_2{}^T(:, m_2) \times_3 \mathrm{D}_3{}^T(:, m_3)|$;
4. $\mathrm{M}_n = \mathrm{M}_n \cup [\mathrm{m}_n^k]$, $\mathrm{B}_n = \mathrm{D}_n(:, \mathrm{M}_n)$;
5. $\mathbf{x} = \arg\min_{\mathbf{x}} ||\mathbf{y} - (\mathrm{B}_3 \otimes \mathrm{B}_2 \otimes \mathrm{B}_1)\mathbf{x}||_2^2$;
6. $\mathcal{R} = \mathcal{Y} - \mathcal{X} \times_1 \mathrm{B}_1 \times_2 \mathrm{B}_2 \times_3 \mathrm{B}_3$;
7. **end while**
8. $\mathcal{A}(\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3) = \mathcal{X}$
9. return $\mathcal{A}$

3. SALIENCY DETECTION BASED ON RECONSTRUCTION RESIDUALS

As for those research works of saliency detection based on traditional sparse models, they treat RGB channels separately or stack RGB channel as a long vector. However, it is not consistent with the mechanism of human visual system, which processes the color channels parallelly.

In order to tackle this problem, we propose to use tensor sparse reconstruction residuals to measure the saliency of each image region. The framework is shown in Fig.1. We use sliding windows to get image patch. For a color image patch, we reconstructed it using its surrounding patches as the dictionary atoms and normalize the reconstruction residual to the value range of [0, 1].

In our saliency detection framework, an input color image patch $\mathcal{Y}$ can be represented using the tensor sparse representation as:

$$\mathcal{Y} = \mathcal{A} \times_1 \mathrm{D}_1 \times_2 \mathrm{D}_2 \times_3 \mathrm{D}_3 + \varepsilon \quad (6)$$

where $\mathcal{A}$ is the sparse coefficients while ε is the reconstruction residual. $\mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3$ are the dictionaries. If we use $\mathcal{D} \in \mathrm{R}^{\mathrm{M} \times \mathrm{N} \times 3 \times \mathrm{K}}$(K is the number of surrounding patches) to represent surrounding patches, then $\mathrm{D}_i = \mathcal{D}_{(i)}$ .During saliency detection, the goal of sparse coding is to find a sparse coefficient with the least reconstruction residual, as given in (5).

As we know, the term of ε in (6) indicates the prediction uncertainty of $\mathcal{Y}$ when surrounding image patches and sparse coefficient $\mathcal{A}$ can be obtained. The unpredictability of $\mathcal{Y}$ will increase with the higher value of ε. Accordingly, we define the saliency value Sc of image patch $\mathcal{Y}$ as the L2 norm of the reconstruction residual:

$$Sc(\mathcal{Y}) = ||\mathcal{Y} - \mathcal{A} \times_1 \mathrm{D}_1 \times_2 \mathrm{D}_2 \times_3 \mathrm{D}_3||_2^2 \quad (7)$$

The saliency computation algorithm is given in Algorithm 2.

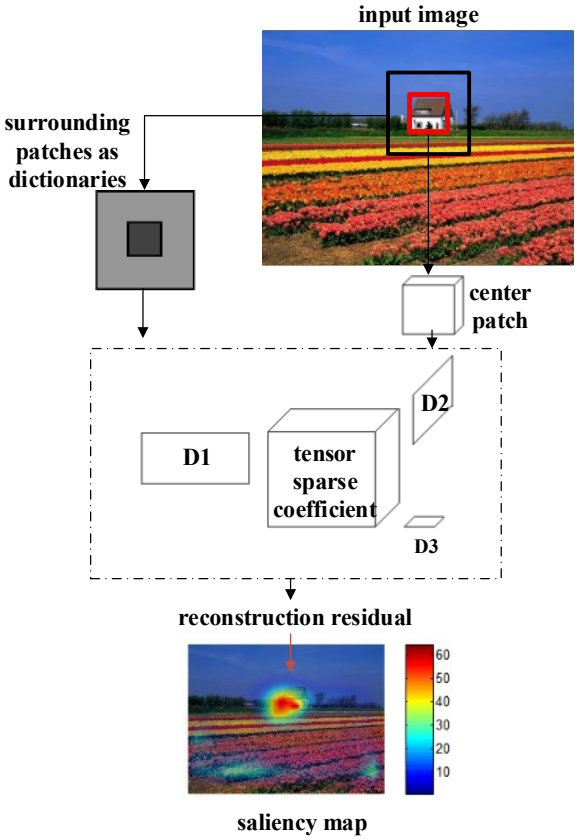


Fig. 1 Framework of our method

Algorithm 2: Saliency computation based on tensor sparse reconstruction residual

Require: Given color image I

output: The saliency map Sc

1. for each image patch $\mathcal{Y}$ of the image I, set dictionary $\mathrm{D}_i(\mathrm{i} = 1,2,3)$ from its surrounding patches
2. Use TOMP algorithm to obtain the sparse representation of image patch $\mathcal{Y}$
3. The saliency value of image patch $\mathcal{Y}$ is calculated by: $Sc(\mathcal{Y}) = ||\mathcal{Y} - \mathcal{A} \times_1 \mathrm{D}_1 \times_2 \mathrm{D}_2 \times_3 \mathrm{D}_3||_2^2$
4. Compute the saliency value of all the image patches, return Sc

4. EXPERIMENTAL RESULTS

In this section, we randomly selected some of the images over ASD dataset [14] and OSIE (Object and semantic images and eye-tracking) dataset [15] to evaluate the performance of our saliency detection method. ASD dataset is widely used as a benchmark in salient object detection with the ground truth of accurate object segmentation. OSIE dataset provides object and semantic saliency, including 700 images and 5551 objects with contour outlined and semantic attribute annotated.

To verify the benefits of tensor sparse model in saliency detection, we compare the proposed saliency detection framework with three saliency detection methods based on traditional sparse Incremental Coding Length (ICL) [6], traditional sparse representation residual model (TSRR) [16] and bottom-up saliency based on weighted sparse coding residual (WSCR) [5].

We used the area under the ROC curve (AUC), precision (Pre), recall (Rec) and F-measure (F_m) [14] values to quantitatively evaluate the performance of these four saliency detection methods. The AUC, Pre, Rec and F_m values are widely-used metrics for performance evaluation of saliency detection. $F_m = \frac{(1+\lambda)\times Pre\times \mathrm{Rec}}{\lambda\times \mathrm{Pre}+\mathrm{Rec}}$, where we set $\lambda = 0.3$ to emphasize precision. We listed mean AUC scores in Table 2 for a statistical analysis from two datasets and mean precision, recall and F measure value in Table 3 from ASD dataset.

The parameters of our method are set as below: patch size as $6\times 6\times 3$, k_{max}=3, $\sigma = 10^{-6}$. For surrounding patches, we used an 18×18 pixel window (consisting of 3×3 patches, 6×6 pixels per patch), so that we got 8 surrounding patches as dictionaries.

Table 2.
COMPARISON OF THE MEAN AUC SCORES

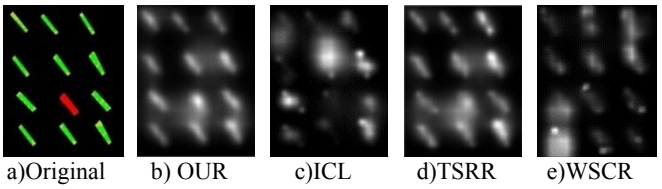
Method	Dataset ASD	Dataset OSIE
ICL[6]	0.8254	0.7314
TSRR[16]	0.8259	0.7198
WSCR[5]	0.8316	0.7384
OUR	0.8407	0.7417

Table 3.
COMPARISON OF PRE, REC, F-MEASURE ON DATASET ASD

Method	Pre	Rec	*F*-measure
ICL[6]	0.4679	0.5066	0.4763
TSRR[16]	0.5344	0.4269	0.5050
WSCR[5]	0.5163	0.7059	0.5504
OUR	0.5827	0.8050	0.6223

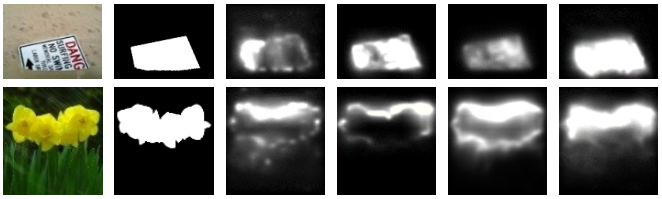
The AUC reflects the prediction accuracy of the saliency map for the fixation point of human eyes. If we get higher mean AUC score, then the algorithm can achieve more accurate prediction. From Table 2, we observe that our method achieves the highest mean AUC scores over two datasets. We also get the both high precision and high recall in Table 3, means that our algorithm can promotes the salient region and restrains unsalient region well. So, our proposed method presents the most promising performance of saliency detection in color images.

Specifically, we compared the four saliency detection methods on the images, which contain the pop-up objects with same shape but in different color. As shown in Figure 2, we can observe that our saliency detection method can figure out the pop-up object but the other three methods lost their detection efficiency. It is because that tensor sparse reconstruction is more accurate than traditional models.



a)Original b) OUR c)ICL d)TSRR e)WSCR

Fig. 2: Visual comparison of four saliency detection models



(a) Origin (b) Ground Truth (c) ICL (d) TSRR (e) WSCR (f) OUR

Fig. 3: Visual comparisons of saliency maps.

In order to show that our method has more accurate results in general saliency detection task in real scenes, we listed a set of subjective visual evaluation results in Figure 3. It is noted that our saliency maps are more consistent with the ground truth, providing contour outline of saliency regions more accurately. It should be noted that all of these four methods apply a final Gaussian blur filter on the constructed saliency maps to preserve piece-wise saliency smoothness.

5. CONCLUSIONS

In this paper, we proposed a method for saliency detection based on tensor sparse reconstruction residual and center-surround contrast model. Experimental results demonstrated that the proposed saliency detection framework can provide more consistent results with HVS than those methods based on traditional sparse models. The main reason is that the current sparse models lost spatial color structure information during the reduced order approximation of the high-order (order>2) signals. In contrast, we use tensor sparse model to represent high order signal without any information lost during sparse coding. Besides color images, it is significant that the proposed framework can be applied to the general saliency detection task in multidimensional signals.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant (61521062, 61201384 and 61521062) and the 111 Project B07022.

6. REFERENCES

[1] V. Navalpakkam and L. Itti. “An integrated model of top-down and bottom-up attention for optimizing detection speed”. In CVPR, pages 2049–2056, 2006.

[2] U. Rutishauser, D.Walther, C. Koch, and P. Perona. “Is bottomup attention useful for object recognition?” In CVPR, pages 37–44, 2004.

[3] L. Itti and C. Koch. “Computational modeling of visual attention.” Nature Reviews Neuroscience, 2(3):194–201, 2001.

[4] L. Itti, C. Koch, and E. Niebur. “A model of saliency-based visual attention for rapid scene analysis.” In PAMI, 20:1254–1259, 1998.

[5] Han, B., Zhu, H., Ding, Y.: “Bottom-up saliency based on weighted sparse coding residual.” In: Proceedings of the 19th ACM International Conference on Multimedia (MM), pp. 1117–1120 (2011)

[6] Li Y, Zhou Y, Xu L, et al. “INCREMENTAL SPARSE SALIENCY DETECTION”. IEEE International Conference on Image Processing, 2009:3093 - 3096.

[7] Ruiters R, Klein R. “BTF Compression via Sparse Tensor Decomposition”. Computer Graphics Forum, 2009, volume 28(4):1181-1188(8).

[8] Peng Y, Meng D, Xu Z, et al. “Decomposable Nonlocal Tensor Dictionary Learning for Multispectral Image Denoising”. In CVPR, 2014, pages 2949-2956.

[9] Bro R. PARAFAC: “tutorial and applications”. Chemometrics & Intelligent Laboratory Systems, 1997, 38(2):149–171.

[10] Tucker L R. “Some mathematical notes on three-mode factor analysis”. Psychometrika, 1966, 31(3):279-311.

[11] Caiafa C F, Plata C L, Cichocki A. “Block sparse representations of tensors using Kronecker bases”. In ICASSP, 2012, pages2709 - 2712.

[12] Zubair S, Wang W. “Tensor dictionary learning with sparse TUCKER decomposition”. International Conference on Digital Signal Processing, 2013:1 - 6.

[13] Rigas I, Economou G, Fotopoulos S. “Low-Level Visual Saliency With Application on Aerial Imagery”. Geoscience & Remote Sensing Letters IEEE, 2013, 10(6):1389 - 1393.

[14] R. Achanta, S. Hemami, F. Estrada and S. Süsstrunk,“Frequency-tuned Salient Region Detection”, in CVPR, pp. 1597 - 1604, 2009.

[15] J X, M J, S W, et al. “Predicting human gaze beyond pixels.” Journal of Vision, 2014, 14(1):97-97.

[16] Wright J, Ma Y, Mairal J, et al. “Sparse Representation for Computer Vision and Pattern Recognition”. Proceedings of the IEEE, 2010, 98(6):1031 - 1044.