Proceedings of the International Conference on Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 56

Structured Low-rank Coding for Top-down Saliency

Kun Mei, Weijia Zou, Lianyang Ma, Rui Zhang
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University, Shanghai, P.R. China
meikun1992@gmail.com; zouweijia@sjtu.edu.cn; malianyang2009@sjtu.edu.cn; zhang_rui@sjtu.edu.cn

Abstract – In this paper, we propose a structured low-rank coding method for top-down saliency detection. Both spatial consistency and structured information are considered in our proposed method. Spatial consistency encourages local image patches which are spatially close in an image to have similar representations. Structured information facilitates the patches from the target to have similar representations, while the patches from the background to have various representations. Furthermore, we perform a structured dictionary learning paradigm by integrating both cues. Robust structured and low-rank representations for image patches are obtained over the learned dictionary. The experiment results on two datasets demonstrate that our method can effectively exploit the spatial consistency and structured information and achieve the state-of-the-art performance on saliency detection.

Keywords: Low-rank, Saliency detection, Spatial consistency, Structured information, Dictionary learning.

1. Introduction

Recently development on the saliency detection reveals the significance for a lot of computer vision and multimedia applications, such as object recognition and detection, image quality assessment and video enhancement. Generally, the mechanism of the saliency detection can be divided into two categories: bottom-up and top-down. The bottom-up models are based on kinds of low-level visual information, which is an unsupervised process. While the top-down is a supervised process which utilise the prior knowledge from the training data.

Recently the top-down method has achieved a sustained development. Yang et al. (2012) proposed a top-down saliency algorithm to utilise local contextual information via jointly learning of conditional random field and a discriminative dictionary. Qiu et al. (2012) proposed a saliency detection model via contextual pooling which takes advantage of the spatial contexts in neighbourhood. Following the existing top-down methods, we regard the saliency detection as a binary classification model, in which a log-linear model is learned to separate the salient part from background regions. The main contribution in this paper is that we propose a structured low-rank coding method for top-down saliency detection by incorporating spatial consistency and structured information. Spatial consistency is taken into consideration by low-rank recovery. Structured information from training data is incorporated into a dictionary learning process by adding a suitable regularization term to the objective function. We can obtain structured and low-rank representations over a learned high-quality dictionary. The experimental results demonstrate that our method outperforms other state-of-the-art methods on two datasets.

The rest of the paper is organized as follows: Section 2 elaborates the structured low-rank coding method. Section 3 conducts experiments to evaluate our method and compares with other methods. Finally, the conclusion is presented in section 4.

2. The Structured Low-rank Coding and Representation

In this section, we present the proposed structured low-rank coding method for saliency detection in detail. As illustrated in Fig. 1, the general framework consists of four indispensable and successive steps: feature extraction, descriptors coding, contextual pooling and saliency prediction. For coding step, on the

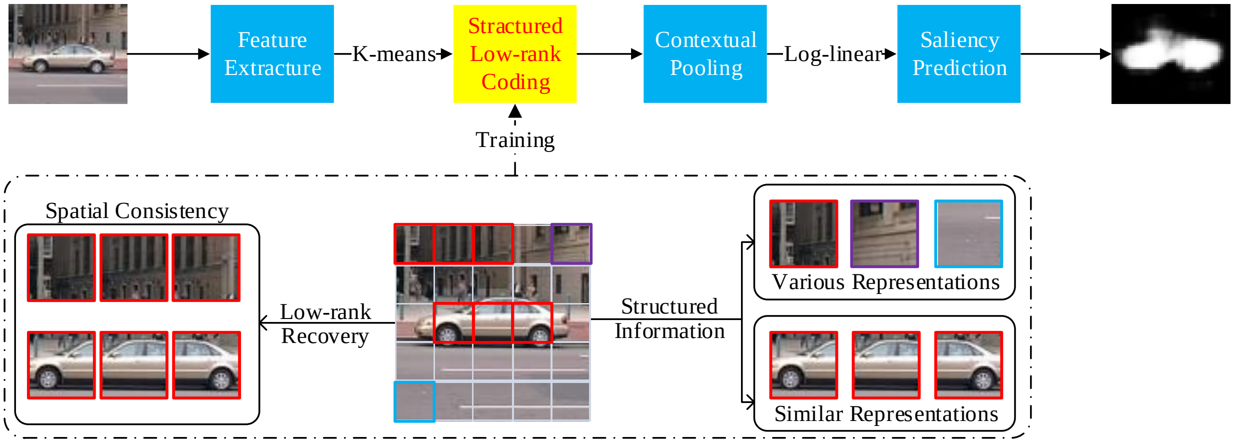

Fig. 1. Framework of proposed scheme (best viewed in colour).

one hand, we utilise the low-rank recovery to exploit the spatial consistency of the local image patches. On the other hand, the structured information are exploited sufficiently by dictionary learning, which encourages the patches extracted from the target to have similar representations, while the representations of the patches extracted from the background are various. Saliency detection is carried out directly on these robust and discriminative representations.

2.1. Problem Statement and Discussion

After the feature extraction, we can get a data matrix $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N}]$ with $\mathbf{N}$ patches where $\mathbf{X_i}$ corresponds to patch *i*. We should reconstruct the $\mathbf{X}$ by the optimum representation in the coding step. While with the past coding methods which solve the coding problem for each feature independently, we have inconsistent codes even for the similar features. Inspired by the work of Liu et.al (2010), we adopt low-rank matrix recovery to decompose a corrupted matrix $\mathbf{X}$ into a low-rank representation $\mathbf{D} \times \mathbf{Z}$ and a sparse noise component $\mathbf{E}$. So given an initial dictionary $\mathbf{D}$, the objective function is formulated as:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_{2,1} \qquad s.t \;\; \mathbf{X} = \mathbf{D} \times \mathbf{Z} + \mathbf{E} \tag{1}$$

where $\mathbf{Z}$ is the representation matrix, the parameter λ controls the sparsities of the noise matrix $\mathbf{E}$. $\|.\|_*$ and $\|.\|_{2,1}$ denote the nuclear norm and the $l_{2,1}$-norm of a matrix respectively.

Based on above the formulation (1), the spatial consistency is taken into consideration. Besides, the quality of dictionary also influences the representations. Similar to (Liu et al., 2013), we expect the representation matrix $\mathbf{Z}$ to be block-diagonal with a semantic dictionary. Specifically we construct a matrix $\mathbf{W}$ in block-diagonal to redefine a dictionary leaning formulation, which will learn a structured dictionary to facilitate $\mathbf{Z}$ to be close to $\mathbf{W}$. The objective function is defined as:

$$\min_{\mathbf{Z},\mathbf{E},\mathbf{D}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_{2,1} + \beta \|\mathbf{Z} - \mathbf{W}\|_F^2 \qquad s.t \quad \mathbf{X} = \mathbf{D} \times \mathbf{Z} + \mathbf{E} \tag{2}$$

where parameter β controls the contribution of the regularization term. Based on the formulation (2), we integrate the structure information into the dictionary learning process with regularization term$\|\mathbf{Z} - \mathbf{W}\|_F^2$. Because some of the data matrix $\mathbf{X}$ are corrupted and the others are clean, the $l_{2,1}$-norm can be better for saliency detection which encourages the columns of $\mathbf{E}$ to be zero.

2.2 Optimization of the Objective Function

We use the inexact alternating direction method (**ADM**) to solve the above optimization problem. The augmented Lagrangian function $\mathbf{L}$ of (2) is defined as:

$$\mathbf{L}(\mathbf{Z},\mathbf{J},\mathbf{E},\mathbf{D},\mathbf{Y_1},\mathbf{Y_2},\theta) = \|\mathbf{J}\|_* + \lambda\|\mathbf{E}\|_{2,1} + \beta\|\mathbf{J}-\mathbf{W}\|_\mathrm{F}^2 + \langle\mathbf{Y_1},\mathbf{X}-\mathbf{D}\times\mathbf{Z}-\mathbf{E}\rangle + \langle\mathbf{Y_2},\mathbf{Z}-\mathbf{J}\rangle + \frac{\theta}{2}\left(\|\mathbf{X}-\mathbf{D}\times\mathbf{Z}-\mathbf{E}\|_\mathrm{F}^2 + \|\mathbf{Z}-\mathbf{J}\|_\mathrm{F}^2\right) \tag{3}$$

where $\langle\mathbf{A},\mathbf{B}\rangle = \mathrm{trace}(\mathbf{A}^{\boldsymbol{T}}\times\mathbf{B})$, $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are Lagrangian multipliers and θ is penalty factor. The process of optimization is presented in Algorithm 1.

Algorithm 1 Structured Low-rank Representation via Inexact **ADM** and Dictionary Learning

```
Input: Data X, Initial Dictionary D, parameter $\lambda, \beta$
Output: Z, E, D
Initialize: $\mathbf{Z_0} = \mathbf{J_0} = \mathbf{E_0} = \mathbf{Y_1^0} = \mathbf{Y_2^0} = 0, \varepsilon = 10^{-8}, \delta = 10^{-7}, \theta = 10^5, \rho = 1.1, \theta_{\max} = 10^{10}, \eta = 0.1, \mathrm{maxIter} = 10^6$
while  not converged, $\mathrm{k} \le \mathrm{maxIter}$  do
    1. fix Z, E and update J by  $\mathbf{J} = \mathit{arg}\min\|\mathbf{J}\|_* + \beta\|\mathbf{J}-\mathbf{W}\|_\mathrm{F}^2 + \frac{\theta}{2}\left\|\mathbf{J}-(\mathbf{Z}+\frac{1}{\theta}\mathbf{Y_1})\right\|_\mathrm{F}^2$
    2. fix J, E and update Z by  $\mathbf{Z} = \left(\mathbf{I}+\mathbf{D}^{\boldsymbol{T}}\times\mathbf{D}\right)^{-1}(\mathbf{D}^{\boldsymbol{T}}\times\mathbf{D}-\mathbf{D}^{\boldsymbol{T}}\times\mathbf{E}+\mathbf{J}+\frac{1}{\theta}(\mathbf{X}^{\boldsymbol{T}}\times\mathbf{Y_1}-\mathbf{Y_2}))$
    3. fix J, Z and update E by  $\mathbf{E} = \mathit{arg}\min\lambda\|\mathbf{E}\|_{2,1} + \frac{\theta}{2}\left\|\mathbf{E}-(\mathbf{X}-\mathbf{D}\times\mathbf{Z}+\frac{1}{\theta}\mathbf{Y_1}\right\|_\mathrm{F}^2$
    4. fix J, Z, E and update D by  $\mathbf{D}^{k+1} = \eta\mathbf{D}^k + (1-\eta) * \frac{1}{\theta}\left(\mathbf{Y_1}+\theta(\mathbf{X}-\mathbf{E})\right)\times\mathbf{Z}^{\boldsymbol{T}}\times(\mathbf{Z}\times\mathbf{Z}^{\boldsymbol{T}})^{-1}$
    5. update the Lagrangian multipliers:
       $\mathbf{Y_1} = \mathbf{Y_1} + \theta(\mathbf{X}-\mathbf{D}\times\mathbf{Z}-\mathbf{E})$   $\mathbf{Y_2} = \mathbf{Y_2} + \theta(\mathbf{Z}-\mathbf{J})$
    6. update the penalty factor $\theta$ by $\theta = \min(\rho\theta, \theta_{\max})$
    7. check the convergence conditions:
       $\|\mathbf{X}-\mathbf{D}\times\mathbf{Z}-\mathbf{E}\|_\infty < \varepsilon, \|\mathbf{Z}-\mathbf{J}\|_\infty < \varepsilon\ \mathit{and}\ \left\|\mathbf{D}^{k+1}-\mathbf{D}^k\right\|_\infty < \delta$
end  while
```

3. Experiments

We evaluate the proposed top-down saliency method based on structured low-rank coding on two datasets: INRIA-Car and Graz-02.

3.1. The Datasets and Experimental Setup

The INRIA-Car contained 474 images. And we also use a subset of INRIA-Car named INRIA-Car-B containing 172 images, in which the ground-true maps are labelled accurately. The Graz-02 datasets consist of three categories: person, bicycle and car. Each contains 300 images. In the step of feature extraction, the SIFT features are extracted from densely sampled 16×16 pixel patches on a grid with the space of 8 pixels and an initial dictionary is gotten by K-means algorithm. Following the common settings in literature, we choose 40% images as the training data randomly in INRIA-Car dataset. And for Graz-02 datasets, we use 150 odd-numbered images as training data, while the rest as test data. We set $\lambda = 16, \beta = 15$ in our coding algorithm. For comparing to the method of Qiu et.al (2012), we also choose the optimal scale 3 in the contextual pooling step. All the same experiments are repeated five times.

3.2. Results and Evaluation Criterion

In order to evaluate the performance of our method, we adopt the average precision (**AP**), which is equal to the area under **P-R** curve, to assess the quality of saliency detection intuitively. The Fig. 2 show the quite clear saliency maps generated by our method on Graz-02 datasets. Table. 1 and Table. 2 show the **AP** (%) on Graz-02 datasets and INRIA-Car dataset respectively. We can clearly see that our method outperforms other state-of-the-art methods.

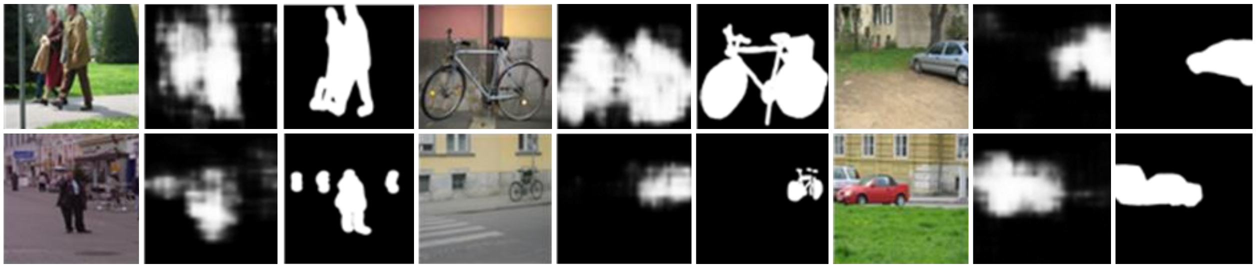

Fig. 2. The saliency detection results on Graz-02 datasets by our method.

Table 1. The **AP** (%) on Graz-02 datasets

Dataset	Kanan et.al (2009)	Yang et.al (2012)	Qiu et.al (2012)	our method
Graz-person	52.2	56.8	59.7	**62.5**
Graz-car	45.7	57.5	59.3	**64.9**
Graz-bike	61.9	71.9	71.6	**73.2**

Table 2. The **AP** (%) on INRIA-Car dataset

Dataset	Cheng et.al(2011)	Qiu et.al (2012)	our method
INRIA-Car	52.4	73.8	**80.5**
INRIA-Car-B	52.5	80.1	**84.0**

4. Conclusion

In this paper, we have proposed a structured low-rank coding method for top-down saliency detection by exploiting the spatial consistency and structured information. Particularly, we implement a dictionary learning paradigm to integrate both. Structured and low-rank representations for image patches are obtained over the learned dictionary. The experimental results demonstrate that our method achieve the state-of-the-art performance on saliency detection.

Acknowledgements

This work is supported by NSFC (61071155) and 973 Program (2010CB731401, 2010CB731406).

References

Cheng, M. M., Zhang, G. X., Mitra, N. J., Huang, X., & Hu, S. M. (2011, June). Global contrast based salient region detection. CVPR, 2011 IEEE Conference on (pp. 409-416). IEEE.

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In Proceedings of the 27th International Conference on Machine Learning, pp. 663-670.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., & Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. PAMI, IEEE Transactions on, 35(1), 171-184

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. Visual Cognition, 17(6-7), 979-1003.

Qiu, Y., Zhu, J., Zhang, R., & Huang, J. (2012). Top-Down saliency by multi-scale contextual pooling. In Advances in Multimedia Information Processing–PCM 2012 (pp. 294-305).

Yang, J., & Yang, M. H. (2012). Top-down visual saliency via joint CRF and dictionary learning. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2296-2303).