IMAGE CO-SALIENCY DETECTION VIA LOCALLY ADAPTIVE SALIENCY MAP FUSION

Chung-Chi Tsai[1,2] *Xiaoning Qian*[1] *Yen-Yu Lin*[2]

[1] Texas A&M University [2]Academia Sinica

ABSTRACT

Co-saliency detection aims at discovering the common and salient objects in multiple images. It explores not only intra-image but extra inter-image visual cues, and hence compensates the shortages in single-image saliency detection. The performance of co-saliency detection substantially relies on the explored visual cues. However, the optimal cues typically vary from region to region. To address this issue, we develop an approach that detects co-salient objects by region-wise saliency map fusion. Specifically, our approach takes intra-image appearance, inter-image correspondence, and spatial consistence into account, and accomplishes saliency detection with locally adaptive saliency map fusion via solving an energy optimization problem over a graph. It is evaluated on a benchmark dataset and compared to the state-of-the-art methods. Promising results demonstrate its effectiveness and superiority.

Index Terms— Co-saliency detection, graph-based optimization, energy minimization, locally adaptive fusion

1. INTRODUCTION

Saliency detection attempts to unsupervisedly identify the salient pixels in an image. It is an active and fundamental topic in image processing, since it can help automate many applications such as image segmentation [1] and video compression [2]. Despite the significant progress, e.g. [3, 4, 5, 6, 7, 8, 9], the performance of single-image saliency detection is still restricted by its unsupervised nature, especially when with complex image content. Co-saliency detection, e.g. [10, 11, 12], is introduced to address the difficulties inherent in single-image saliency detection. It aims to locate the common salient objects. The information used in most approaches for co-saliency detection can be divided into two categories, i.e. *intra-image* and *inter-image* evidences. The former is extracted based on appearance contrast and spatial cues in a single image. The latter is obtained by detecting the correspondences between a group of images.

A single type of evidences in general is insufficient for handling complex co-saliency detection problems. Most modern approaches carry out co-saliency detection by fusing multiple saliency maps. For instance, the approaches in [10, 11] adopt *fixed-weight summation* for map fusion, while the one in [12] uses *fixed-weight multiplication*. Cao *et al.* [13] instead proposed a *self-adaptive* framework where the weights for map fusion are dynamically generated according to the input images.

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

Fig. 1: Image co-saliency detection. (a) & (b) An image pair and the ground truth. (c) $\sim$ (f) Saliency maps produced by using (c) the intra-image evidence [9], (d) the inter-image evidence [10], (e) the method in [13], and (f) our method.

The aforementioned approaches [10, 11, 12, 13] fuse saliency maps in a *map-wise* manner. Namely, a weight is given for the whole-image saliency map. These approaches neglect the phenomenon that the goodness of a saliency map is often *region-dependent*. As an illustration, Fig. 1 shows an image pair and the saliency maps generated by using the intra-image evidence [9], the inter-image evidence [10], the method of self-adaptive fusion [13], and our proposed method. It can be observed that using a single type of evidences doesn't suffice for this case. While using only the intra-image evidence [9] leads to the false alarm in the text part of the second image, using only the inter-image evidence [10] fails to detect a penguin in the first image. The method [13] combines both types of evidences. It gives better results, but it also inherits both the shortcomings of false alarms and misses.

To tackle these challenges of co-saliency detection, we propose an approach that can jointly consider both intra-image and inter-image evidences, and carry out region-wise saliency map fusion. As shown in Fig. 1f, our approach effectively alleviates the unfavorable effects of false alarms and misses, and results in the saliency maps of higher quality.

2. RELATED WORK

The literature of saliency detection is extensive. Most of them target at *human eye fixation prediction* [3, 4] or *salient object*

This work was partially supported by Award #1547557 from the National Science Foundation, Grants MOST 104-2628-E-001-001-MY2 and MOST 105-2221-E-001-030-MY2 from the Ministry of Science and Technology.

detection [5, 6, 7, 8, 9]. Approaches to eye fixation prediction are inspired by the primitive human visual system. For example, Itti *et al.* [3] computed center-surround differences across multi-scale image features for detecting saliency. Despite the novelty, this method poorly detects object borders. On the contrary, Hou *et al.* [4] defined the saliency through the residual on the log-frequency domain. Although their method is computationally efficient, it mostly discovers object boundaries rather than the whole salient regions. Both methods [3, 4] involve image resizing process, which probably causes the loss of frequency content.

In the category of salient object detection, Achanta *et al.* [5] devised a full resolution method by which more uniformly highlighted salient regions as well as more precise object boundaries can be obtained. However, their method neglected the spatial layout of objects in images, so it tends to predict background regions as salient. Perazzi *et al.* [7] improved Achanta *et al.*'s model by further considering the appearance contrast and the spatial distribution in saliency detection. In addition to the low-level features, Shen and Wu [6] further integrated higher level prior knowledge, such as the center or semantic prior, into detecting salient objects. Yang *et al.* [8] used the background priors inferred from object boundaries as well as the foreground proposals to rank the saliency degrees of superpixels. Following [8], Zhu *et al.* [9] proposed a more robust method for background prior generation. Their method coupled with other contrast cues achieves the state-of-the-art performance in the single-image saliency detection.

Stemming from the unsupervised nature, the performance of the aforementioned approaches to single-image saliency detection is still restricted. Co-saliency detection is introduced to further improve the performance. The shared visual cues obtained across images facilitate foreground location and background removal. For instance, Li and Ngan [10] utilized the *SimRank* algorithm on a co-multilayer superpixel tree, and detected the color and texture similarity between superpixels across images. Meng *et al.* [11] improved the SimRank matching method by further taking geometric constraints into account. Fu *et al.* [12] proposed a clustering based process to learn inter-image correspondence. To effectively integrate multiple cues, Cao *et al.* [13, 14] employed a low-rank constraint on the salient regions of multiple saliency map proposals, and adaptively determined the fusion weight of each map proposal. Inspired by the fact that the optimal saliency map proposal is often region-dependent, our approach adaptively seeks the weights for saliency fusion in a region-wise manner, thus leading to more favorable results.

3. THE PROPOSED APPROACH

Given a pair of images I_1 and I_2 for co-saliency detection, we apply M existing (co-)saliency detection algorithms, e.g. [3, 4, 5, 10], and get M saliency maps for each image. For locally adaptive saliency map fusion, images I_1 and I_2 are respectively decomposed into N_1 and N_2 *superpixels*, which serve as the domain of region-wise fusion. Our approach aims to seek a weight vector $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \dots\ y_{i,M}]^\top \in \mathbb{R}^M$ for each superpixel i, where $i \in \{1, 2, .., N_1 + N_2\}$. The co-saliency detection is accomplished by superpixel-wise fusing the M saliency maps. Our approach formulates this task of region-wise fusion as an energy minimization problem over a graph. In the following, the image pre-processing and the graph construction are introduced first. The proposed energy function and its optimization are then described.

3.1. Image Pre-processing

The *SLIC* algorithm [15] is used for deriving superpixels, because it effectively preserves inherent structures while abstracts unnecessary details. We set the numbers of superpixels to $N_1 = N_2 = 200$ in this work.

Two types of visual features, color and texture, are extracted for each superpixel. For color features, each pixel in the three color spaces, `RGB`, `L*a*b*`, and `YCbCr`, is represented by a 9-dimensional vector. Using the *bag-of-words* model, all pixels in the image pair are quantized into clusters by using the k-means algorithm. Each superpixel is then represented as a $k = 100$-dimensional histogram. For texture features, Gabor filter responses with eight orientations, three scales and two phase offsets are extracted for each pixel. The texture features of a superpixel are similarly encoded as a 100-dimensional histogram by using the bag-of-words model.

Let $\mathbf{p}_i$ and $\mathbf{q}_i$ denote the color and texture representations of superpixel i respectively. The similarity between superpixel i and superpixel j is defined as

$$A(i,j) = \exp(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\sigma_c} - \gamma\frac{d(\mathbf{q}_i, \mathbf{q}_j)}{\sigma_g}), \qquad (1)$$

where $d(\cdot)$ is the χ^2 distance. We set $\gamma = 1.5$ to put more emphasis on Gabor features. The value of constant σ_c is set to the average pair-wise distance between all superpixels under their color features. The value of σ_g is similarly set.

3.2. Graph Construction

We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2)$. In $\mathcal{G}$, each vertex $v_i \in \mathcal{V}$ corresponds to superpixel i, thus $|\mathcal{V}| = N_1 + N_2$. The edge $e_{ij} \in \mathcal{E}_1$ is added to link v_i and v_j if superpixels i and j are spatially connected in an image. The edge $e_{ij} \in \mathcal{E}_2$ is included to connect v_i and v_j if superpixel j is one of the ℓ nearest neighbors of superpixel i in the opposite image according to the similarity in Eq. (1). We set $\ell = 1$ to simulate the one-to-one superpixel matching scenario. Edge weights for both types of edges are assigned by (1) to get the affinity matrix A for $\mathcal{G}$. We also construct the corresponding Laplacian matrix $L \in \mathbb{R}^{N\times N}$, where $N = N_1 + N_2$.

3.3. Energy Function

We seek the optimal weights $Y = [\mathbf{y}_1\ \mathbf{y}_2\ \dots\ \mathbf{y}_N] \in \mathbb{R}^{M\times N}$, where M is the number of saliency maps, and N is the total number of superpixels of I_1 and I_2, for superpixel-wise map fusion by minimizing the proposed energy function

$$\begin{aligned}\min_{Y}\quad & \lambda_1 \sum_{v_i\in\mathcal{V}} U(\mathbf{y}_i) + \lambda_2 \sum_{v_i\in\mathcal{V}} V(\mathbf{y}_i) \\ & + \lambda_3 \sum_{e_{ij}\in\mathcal{E}} B(\mathbf{y}_i,\mathbf{y}_j) + ||Y||_2^2 \qquad (2)\\ \text{s.t.}\quad & \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \mathbf{0},\ \text{for } 1\leq i\leq N,\end{aligned}$$

where $\mathbf{0}$ is a vector whose elements are zero, and λ_1, λ_2 and λ_3 are three positive constants. There are four terms introduced in Eq. (2). The first two unary terms, $U(\mathbf{y}_i)$ and $V(\mathbf{y}_i)$, respectively leverage intra-image and inter-image evidences to estimate the power of each saliency map on superpixel i. The pairwise term $B(\mathbf{y}_i,\mathbf{y}_j)$ encourages the smoothness of the derived weights on superpixel pairs connected in the graph $\mathcal{G}$. The last term $||Y||_2^2$ is included for regularization.

3.3.1. On Designing Unary Term $U(\mathbf{y}_i)$

We intend to assign a higher weight to a saliency map that is consistent with other saliency maps on superpixel i. It helps exclude distinct biases in individual maps. Inspired by [16], we employ a low-rank constraint for this task, but we further generalize the method in [16] to *locally* estimate the goodness of each saliency map. For superpixel i, we find its n spatially nearest superpixels. Let $\mathbf{x}_{i,m} \in \mathbb{R}^{256}$ be a 256-dimensional histogram representing the intensity distribution of saliency values of saliency map m on these n superpixels. By stacking the M different vectors for all saliency maps, $X_i = [\mathbf{x}_{i,1}\ \mathbf{x}_{i,2}\ \dots\ \mathbf{x}_{i,M}] \in \mathbb{R}^{256\times M}$, we infer the consistent part by seeking a low-rank surrogate of X_i. Specifically, *robust PCA* [17] is adopted to decompose X_i into a low-rank approximation L_i plus a residual matrix E_i by solving

$$\min_{L_i,E_i} (||L_i||_* + \lambda||E_i||_1), \quad \text{s.t. } X_i = L_i + E_i, \qquad (3)$$

where $||L_i||_*$ is the nuclear norm of L_i, and λ is a constant. After solving Eq. (3), higher weights are assigned to saliency maps with lower residual errors $E_i = [\mathbf{e}_{i,1}\ \dots\ \mathbf{e}_{i,M}]$, i.e.,

$$w_{i,m} = \frac{\exp(-||\mathbf{e}_{i,m}||_2^2)}{\sum_{j=1}^{M}\exp(-||\mathbf{e}_{i,j}||_2^2)},\ \text{for } 1\leq m\leq M. \qquad (4)$$

The above procedure is repeated for each superpixel i. A penalty variable $z_{i,m} = \exp(1-w_{i,m})/\sum_{j=1}^{M}\exp(1-w_{i,j})$ is introduced to construct the first term in Eq. (2) by letting

$$\sum_{v_i\in\mathcal{V}} U(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{z}_i^\top \mathbf{y}_i = tr(Z^\top Y), \qquad (5)$$

where $\mathbf{z}_i = [z_{i,1}\ \dots\ z_{i,M}]^\top$ and $Z = [\mathbf{z}_1\ \dots\ \mathbf{z}_N]$.

3.3.2. On Designing Unary Term $V(\mathbf{y}_i)$

This term is designed to reduce the false saliency detection by exploring inter-image correspondences. Let e_i represent the similarity between superpixel i and its most similar superpixel in the other image. Let $s_{i,m}$ denote the mean saliency value of saliency map m on superpixel i. The larger the value of e_i is, the more likely superpixel i has a correspondence in the other image. Thus, we prefer saliency map m if the value of $s_{i,m}$ is proportional to that of e_i.

This unary term penalizes the case where only one of e_i and $s_{i,m}$ has large values, encouraging salient regions with matched regions in the other image. Penalizing variable $r_{i,m}$ is defined as

$$r_{i,m} = \frac{\exp((1-e_i)s_{i,m} + e_i(1-s_{i,m}))}{\sum_{j=1}^{M}\exp[(1-e_i)s_{i,j} + e_i(1-s_{i,j})]}. \qquad (6)$$

The denominator in Eq. (6) is for normalization. By considering all superpixels, the second term in Eq. (2) becomes

$$\sum_{v_i\in\mathcal{V}} V(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{r}_i^\top \mathbf{y}_i = tr(R^\top Y), \qquad (7)$$

where $\mathbf{r}_i = [r_{i,1}\ \dots\ r_{i,M}]^\top$ and $R = [\mathbf{r}_1\ \dots\ \mathbf{r}_N]$.

3.3.3. On Designing Pairwise Term $B(\mathbf{y}_i,\mathbf{y}_j)$

We impose this pairwise term to encourage the smoothness of the weight distribution Y between connected superpixels in the graph $\mathcal{G}$. The formulation of this term is defined as

$$\sum_{e_{ij}\in\mathcal{E}} B(\mathbf{y}_i,\mathbf{y}_j) = \sum_{e_{ij}\in\mathcal{E}} A(i,j)\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = tr(YLY^\top), \qquad (8)$$

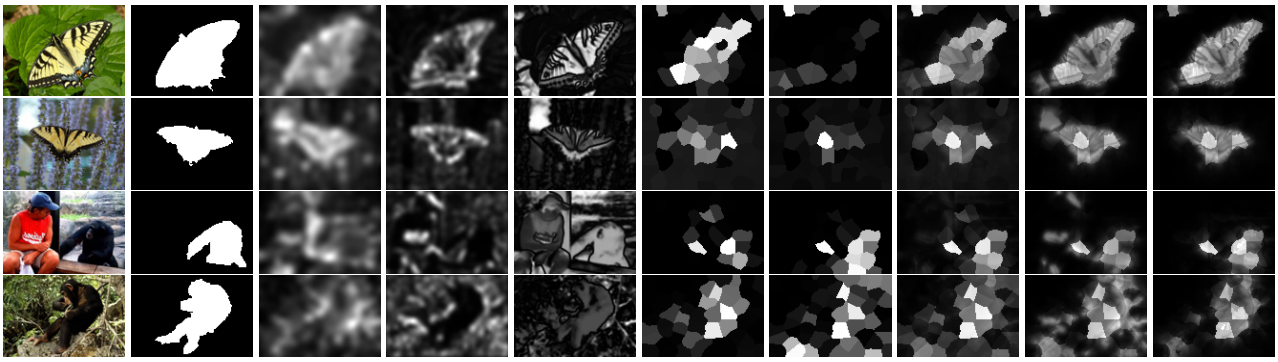
where L is the Laplacian matrix of $\mathcal{G}$.

3.4. Optimization Process and Spatial Refinement

With the definitions of the unary and pairwise terms in Eqs. (5), (7), and (8), the constrained optimization problem in Eq. (2) is a *quadratic programming* (QP) problem, and has a globally optimal solution. The asymptotic worst-case time complexity using the interior-point method for the convex QP is $\mathcal{O}((NM)^3)$ [18]. We adopt the CVX solver [19] on MATLAB to solve it, and the average running time for each image pair is around 13 seconds on a PC with an Intel i7 2.5GHz CPU and 16G RAM. After optimization, the saliency detection results can be compiled by superpixel-wise fusing the saliency maps with the solution Y. To further improve the performance, the spatial refinement process [13, 14] is applied to the yielded saliency map. It re-scales the saliency values by a combination of thresholding and normalization.

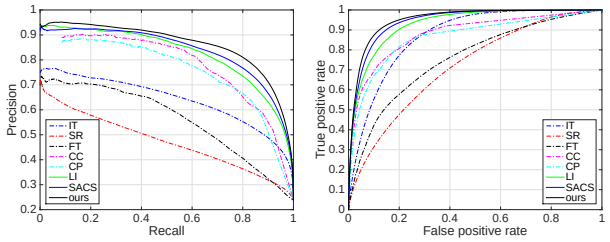
4. EXPERIMENTAL RESULTS

In this section, our approach is evaluated on the *Image Pair* dataset [10], which consists of 105 image pairs with manually labeled ground truth.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 2: (a) & (b) Two exemplar image pairs and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) IT [3], (d) SR [4], (e) FT [5], (f) CC [10], (g) CP [10], (h) LI [10], (i) SACS [13], and (j) Ours.



(a) (b)

Fig. 3: The performance of various approaches in (a) PR curves and (b) ROC curves.

Experimental Setup: Following [10], we compute five saliency map proposals by three saliency detection algorithms, IT [3], SR [4], and FT [5], and one co-saliency detection algorithm [10] with two features, color CC and texture CP. Except the five proposals, our approach is compared with two fusion-based approaches to co-saliency detection, including LI [10] and SACS [13]. Note that the approaches, LI, SACS, and ours, work by fusing the same five map proposals.

The performance of each evaluated approach is measured by the *precision-recall* (PR) curve, which is obtained by varying the saliency threshold. PR curves tend to favor methods that successfully detect the salient regions over methods that precisely locate the non-salient regions. Thus, *receiver operating characteristics* (ROC) curves and *mean absolute error* (MAE) based on the given ground truth are also included for performance evaluation. In our experiments, we have set $\lambda_1 = 8$, $\lambda_2 = 4$, $\lambda_3 = 1$ in (2) and $\lambda = 0.05$ in (3).

Result Analysis: The PR curves and the ROC curves from our approach and seven competing approaches are shown in Fig. 3a and Fig. 3b, respectively. We also report the *area under the curve* (AUC) of PR curves, the AUC of ROC curves, and MAE of these approaches in Table 1.

It can be observed in Fig. 3 and Table 1 that the methods LI [10] and SACS [13] can effectively leverage the mutual signal strengths among the five saliency proposals, IT, SR, FT, CC, and CP, and remarkably outperform all the five proposals. Our approach takes region-wise fusion into account, and can make the most of the five *locally complementary* saliency maps. As shown in Fig. 3, our approach consistently achieves better performance than all the competing approaches. In Table 1, the performance gain over SACS, the best competing approach, is significant, including 2.5% in AUC of the PR curve, 0.9% in the MAE and 0.8% in the AUC of the ROC curve.

Method	IT [3]	SR [4]	FT [5]	CC [10]	CP [10]	LI [10]	SACS [13]	Ours
PR AUC	0.640	0.471	0.559	0.702	0.681	0.824	0.836	**0.861**
ROC AUC	0.872	0.718	0.756	0.881	0.865	0.930	0.944	**0.952**
MAE	0.259	0.269	0.253	0.163	0.173	0.173	0.172	**0.163**

Table 1: The performance of various approaches in 1) AUC of PR, 2) AUC of ROC, and 3) MAE. The higher the better in the first two measures. The lower the better in MAE.

To gain insight into the quantitative results, Fig. 2 shows the detected saliency maps on two image pairs by using the seven competing approaches and ours. The saliency proposals that use intra-image evidences, including IT, SR and FT, produce many severe false salient regions. Meanwhile, the saliency proposals that use inter-image evidences, such as CC and CP, detect salient regions with lower confidence. Methods LI and SACS indeed give better results by fusion. Our approach with the aid of region-wise fusion complies the saliency maps that are *perceptually* the closest to the ground truth. Furthermore, the saliency maps by our approach are sharper, namely detection with higher confidence.

5. CONCLUSIONS

We have presented a saliency detection approach that carries out locally adaptive saliency map fusion. It is formulated as a quadratic programming problem and can be efficiently optimized by off-the-shelf solvers. It makes the most of multiple locally complementary saliency proposals and generates both quantitatively and perceptually high-quality saliency maps. In future, we plan to evaluate our approach with more benchmark datasets and generalize it to jointly work with related tasks, such as co-segmentation, sparse image matching, and dense image alignment.

6. REFERENCES

[1] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011.

[2] Zhicheng Li, Shiyin Qin, and Laurent Itti, "Visual attention guided bit allocation in video compression," *J. Image and Vision Computing*, 2011.

[3] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1998.

[4] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.

[5] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009.

[6] Xiaohui Shen and Ying Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[7] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[8] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[9] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, "Saliency optimization from robust background detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

[10] Hongliang Li and King Ngi Ngan, "A co-saliency model of image pairs," *IEEE Trans. on Image Processing*, 2011.

[11] Fanman Meng, Hongliang Li, and Guanghui Liu, "A new co-saliency model via pairwise constraint graph matching," in *IEEE Int'l Symposium, Intelligent Signal Processing and Communications Systems*, 2012.

[12] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu, "Cluster-based co-saliency detection," *IEEE Trans. on Image Processing*, 2013.

[13] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. on Image Processing*, 2014.

[14] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Xuewei Li, "Saliency map fusion based on rank-one constraint," in *Proc. Int'l Conf. Multimedia and Expo*, 2013.

[15] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "SLIC superpixels," Tech. Rep., 2010.

[16] Junxia Li, Jundi Ding, and Jian Yang, "Visual salience learning via low rank matrix recovery," in *Proc. Asian Conf. on Computer Vision*, 2014.

[17] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma, "Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Neural Information Processing Systems*, 2009.

[18] Florian A Potra and Stephen J Wright, "Interior-point methods," *J. Computational and Applied Mathematics*, 2000.

[19] Michael Grant and Stephen Boyd, "cvx users guide for cvx version 1.22," 2012.