Object Co-segmentation Via Discriminative Low Rank Matrix Recovery

Yong Li[†],Jing Liu[†],Zechao Li[†‡],Yang Liu[†],Hanqing Lu[†]
[†]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[‡]School of Computer Science, Nanjing University of Science and Technology
{yong.li,jliu,liuyang6,luhq}@nlpr.ia.ac.cn, zechao.li@gmail.com

ABSTRACT

The goal of this paper is to simultaneously segment the object regions appearing in a set of images of the same object class, known as object co-segmentation. Different from typical methods, simply assuming that the regions common among images are the object regions, we additionally consider the disturbance from consistent backgrounds, and indicate not only common regions but salient ones among images to be the object regions. To this end, we propose a Discriminative Low Rank matrix Recovery (DLRR) algorithm to divide the over-completely segmented regions (i.e., super-pixels) of a given image set into object and non-object ones. In DLRR, a low-rank matrix recovery term is adopted to detect salient regions in an image, while a discriminative learning term is used to distinguish the object regions from all the super-pixels. An additional regularized term is imported to jointly measure the disagreement between the predicted saliency and the objectiveness probability corresponding to each super-pixel of the image set. For the unified learning problem by connecting the above three terms, we design an efficient optimization procedure based on block-coordinate descent. Extensive experiments are conducted on two public datasets, i.e., MSRC and iCoseg, and the comparisons with some state-of-the-arts demonstrate the effectiveness of our work.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms, Experimentation, Theory

Keywords

Object Co-segmentation, Low Rank Matrix Recovery, Discriminative Learning

MM '13 , October 21 - 25 2013, Barcelona, Spain

Figure 1: Examples about the object region. The region inside red contour is the true object regions, while the one inside green contour is the non-object region although is salient.

1. INTRODUCTION

The object segmentation is a fundamental task in computer vision and multimedia, and it can benefit from many applications, e.g., object retrieval and image editing. Provided with some priors to indicate what or which the object regions are, some supervised or interactive object extraction approaches [9][6] can achieve good performance. However, they are hard to be extended to a large scale dataset due to expensive requirements on human interaction or manually labeled training data. Alexe et al. [1] proposed a general object detector, but it does not leverage the shared information among different images of the same or similar objects.

To alleviate such expensive requirements and leverage the shared information, recent researches focus on the task of object co-segmentation, which is to simultaneously segment the same or similar objects appearing in a set of images. A representative solution is to mine the similar object regions by a discriminative clustering framework [4], in which the clustering step is used to merge image pixels into two clusters while the discriminative learning step is to maximally distinguish the two clusters. Considering the diversity of background, Joulin et al. [5] proposed an improved multi-class co-segmentation method by combining spectral clustering and discriminative clustering. Besides, Mukherjee et al. [8] adopted a direct solution to make the possible object regions in images similar to each other. Most of previous approaches work on the assumption that the regions common among images are deemed as the object regions. However, the truth is not the case, in that the background regions may be consistent. For example, shown in Fig. 1, the 'grass' regions may be common within the 'baseball' images. How to

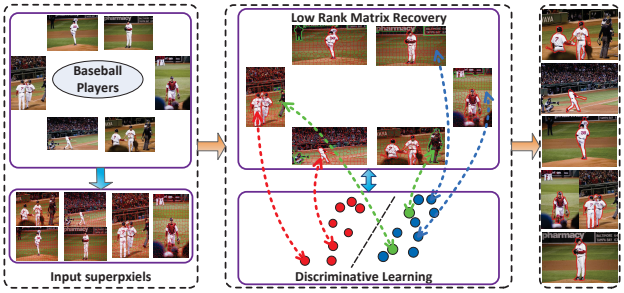

Figure 2: The flowchart of our method. In "discriminative learning" part, red points denote salient and common regions (i.e., object), green points denote salient but uncommon regions (i.e., non-object), and blue points denote non-object regions.

effectively resist the disturbance of such case, and further to precisely identify the true object regions become our focus in this paper.

To this end, we jointly exploit saliency detection and common region mining from a set of images to perform the task of object co-segmentation. That is, the object regions are assumed to be not only common among images but also salient in contrast with background regions. It can naturally eliminate the disturbance of those background regions consistent with each other. Just as shown in Fig. 1, we can easily catch the 'baseball player' regions as the true object regions because they are salient and simultaneously appear in these four images, while the common but non-salient regions (e.g., baseball field) and the salient but uncommon regions (e.g., baseball referee and billboard) are deemed as background.

For this purpose, we propose a unified learning framework, named as discriminative low rank matrix recovery (DLRR) for object co-segmentation, as shown in Fig. 2. Given a set of images of an object class, we first over-completely segment each image into super-pixels, and then employ the proposed DLRR to identify the salient and common regions in the image set. Inspired by the work in [7], we adopt the basic idea of low-rank matrix recovery to detect the salient regions of an image, i.e., decomposing the super-pixel-wise representation of each image into a low-rank matrix and a sparse matrix, and using the l_1-norm of each column in the sparse matrix to measure the saliency of the corresponding superpixel. Besides, discriminative learning is incorporated on the image set to model the salient and common super-pixels. To promote the both terms each other, we import a regularized penalty to measure the disagreement between the predicted saliency and the discriminative output. By jointly considering the above three terms, a unified optimization problem is obtained, and a block-coordinate descent optimization algorithm is presented to solve the proposed problem. Extensive experiments on two publicly available benchmarks, i.e., MSRC and iCoseg, show the satisfied performance of our proposed method. Our main contributions are summarized as follows.

- To the best of our knowledge, we are first to consider the disturbance of consistent background regions for object co-segmentation.
- To overcome the disturbance, we propose to jointly perform saliency detection and common region mining among images to precisely identify the object regions.
- We propose a discriminative low-rank matrix recovery algorithm to solve the object co-segmentation problem, and an efficient optimizing process is also designed.

The rest of the paper is organized as follows. In Section 2, we elaborate our proposed model for object co-segmentation, and its optimization algorithm is presented in Section 3. The experimental evaluation is given in Section 4 followed with the conclusion in Section 5.

2. PROPOSED MODEL

Given a set of images τ with the same class label, image over-segmentation is performed by mean-shift clustering [3] to each image i based on extracted features including color, gabor feature and steerable pyramid feature, and N_i superpixels are obtained. For the j-th superpixel in the i-th image, we use the mean of the features in this superpixel as its feature representation $f_{ij} \in R^D$, then we get the feature representation of the i-th image $\mathbf{F}_i = [f_{i1}, f_{i2}, \cdots, f_{iN_i}]$. Let $y_i \in [0,1]^{1\times N_i}$ denote the probability vector of superpixels to be foreground in the i-th image. The larger y_{ij} is, the more likely for the j-th superpixel in the i-th image to be target object.

2.1 Low Rank Matrix Recovery

For an image, the background usually lies in a low dimensional space, while the salient regions are usually unique and are quite different from the rest [10]. Therefore, low-rank matrix recovery method is adopted to detect image saliency. An image is represented as a low-rank matrix plus sparse noises in the feature space, where the low-rank matrix explains the non-salient regions (or background), and the sparse noises indicate the salient regions. That is, $\mathbf{F}_i = \mathbf{L}_i + \mathbf{S}_i$, where $\mathbf{L}_i$ is the low rank matrix corresponding to the background and $\mathbf{S}_i$ is the sparse noise matrix corresponding to the salient regions. Since the rank norm and l_0 norm lead to an NP-hard problem and it has been shown that the nuclear norm and the l_1 norm is the tight convex approximation for the rank and the l_0 norm [11]. Thus, we obtain the following convex surrogate:

$$\begin{aligned}(\mathbf{L}_i^*, \mathbf{S}_i^*) &= \arg\min(\|\mathbf{L}_i\|_* + \lambda\|\mathbf{S}_i\|_1) \\ s.t. &\mathbf{F}_i = \mathbf{L}_i + \mathbf{S}_i\end{aligned} \tag{1}$$

where $\|\cdot\|_*$ is the nuclear norm. The l_1-norm of each column S_{ij} in $\mathbf{S}_i$ can be used to measure the saliency of the corresponding superpixel [10]. The larger $\|S_{ij}\|_1$ is, the more likely for the j-th superpixel in the i-th image to be salient region.

However, not all the salient regions are meaningful, and some small regions with high-contrast and uniqueness may be considered as meaningless noise by human. Hence we consider some priors to handle the meaningless noise, which is integrated to the low rank recovery framework as follows.

$$\begin{aligned}(\mathbf{L}_i^*, \mathbf{S}_i^*) &= \arg\min(\|\mathbf{L}_i\|_* + \lambda\|\mathbf{S}_i\|_1) \\ s.t. &\mathbf{F}_i\mathbf{P}_i = \mathbf{L}_i + \mathbf{S}_i\end{aligned} \tag{2}$$

where $\mathbf{P}_i$ is a diagonal matrix corresponding to the high level prior such as color or location prior. Since objects near the image center are more attractive to people, a Gaussian distribution based on the distance to the image center is chosen as a high level prior to reduce small salient regions near the image edge in our experiments.

Algorithm 1: Object Co-segmentation by DLRR

Input: Feature Matrix $\mathbf{F}$, high level Prior $\mathbf{P}$ and the required parameters

Initialize: solve Equ.(2) by the method in [11]

1: **while** not converged do
2: $\alpha_i = \frac{1}{\max \|S_{ij}\|_1}$
3: $y_{ij} = \alpha_i \|S_{ij}\|_1$
4: **while** not converaged
5: $\theta^{t+1} = \theta^t - \alpha_{step1}\{[h(f_{ij}) - y_{ij}]f_{ij} + 2r\theta^t\}$
6: **end while**
7: **while** not converaged
8: $y^{t+1} = y^t - \alpha_{step2}[-\mu_1\theta^T\mathbf{F} + 2\mu_2(y^t - S)]$
9: **end while**
10: solve problem (6) in algorithm 2
11: **end while**

Output: foreground probability y

2.2 Discriminative Learning

The saliency detection is mainly evaluated from the view of a single image. However, it cannot exactly catch the object regions common among images. To this end, a logistic regression based discriminative learning is exploited to predict the probability of each superpixel to be the target object.The objective function is to minimize the following regularized function.

$$E_D = -\sum_{i=1}^{\tau}\sum_{j=1}^{N_i}[y_{ij}\log(h(f_{ij})) + (1-y_{ij})\log(1-h(f_{ij}))] + r\|\theta\|^2 \quad (3)$$

where θ is the model parameter to be learned. $h(f_{ij}) = \frac{1}{1+\exp(-\theta^T f_{ij}+b)}$ is the predictive result.

2.3 Proposed Formulation

As discussed in section 1, the ideal object regions as a result of co-segmentation are required to be both salient and common among a given set of images. Thus, we expect, the aforementioned two parts should be learned simultaneously and promote each other. We import a regularization penalty to measure the disagreement between their predicted results. Specifically, the l_1 norm of each column S_{ij} in $\mathbf{S}_i$ stands for the salient score of the j-th superpixel and y_{ij} in discriminative learning is the probability of the j-th superpixel to be target object. Consequently, the disagreement is measured by the following equation.

$$E_R = \sum_{i=1}^{\tau}\sum_{j=1}^{N_i}(y_{ij} - \alpha_i\|S_{ij}\|_1)^2 \quad (4)$$

where α_i is the normalized weight for the superpixel saliency in the i-th image. By jointly exploiting the above aspects, the proposed model is formulated as follows.

$$\min \sum_{i=1}^{\tau}(\|\mathbf{L}_i\|_* + \lambda\|\mathbf{S}_i\|_1) + \mu_1 E_D + \mu_2 E_R \quad s.t.\ \mathbf{F}_i\mathbf{P}_i = \mathbf{L}_i + \mathbf{S}_i, i \in \tau \quad (5)$$

where μ_1 and μ_2 are two non-negative trade-off parameters.

3. MODEL OPTIMIZATION

Considering the objective function is a difference of convex functions, a optimization procedure based on block-coordinate descent is designed and we summarize it in Algorithm 1, where $\mathbf{F} = [\mathbf{F}_1, \cdots, \mathbf{F}_\tau]$, $y = [y_1, \cdots, y_\tau]$, and $S = [\alpha_1\|S_{11}\|_1, \cdots, \alpha_\tau\|S_{\tau N_\tau}\|_1]$. Lines 4-6 is to learn model parameter θ in the discriminative algorithm. Lines 7-9 is to predict the probability to be target object based on saliency detection result and discriminative learning together. The low rank matrix recovery problem given the guidance information y can be solved by solving the following Augmented Lagrange Multiplier(ALM) problem.

Algorithm 2: Solving problem (6) by inexact ALM

Input: matrix $\mathbf{F}_i$, $\mathbf{P}_i$, y_i; parameters λ, β, μ_2, α_i, D

Initialize: $\mathbf{Y}^0 = \mathbf{F}_i\mathbf{P}_i / J(\mathbf{F}_i\mathbf{P}_i)$; $\mathbf{S}^0 = \mathbf{0}$; $\beta^0 = 10^{-6}$; $\beta_{max} = 10^6$;$\rho = 1.5$; $k = 0$

1: **while** not converged do
2: $(\mathbf{U}, \Sigma, \mathbf{V}) = \text{svd}(\mathbf{F}_i\mathbf{P}_i - \mathbf{S}_i^k + (\beta^k)^{-1}\mathbf{Y}^k)$
3: $\mathbf{L}_i^{k+1} = \mathbf{U}T_{(\beta^k)^{-1}}[\Sigma]\mathbf{V}^T$
4: $\varepsilon = \frac{\lambda - 2\alpha_i\mu_2 y_n + 2\alpha_i^2\mu_2\sum_{t=1,\neq m}^{D}|S_i(t,n)|}{\beta + 2\alpha_i^2\mu_2}$
5: $\mathbf{x} = \frac{\beta}{\beta+2\alpha_i^2\mu_2}(\mathbf{F}_i\mathbf{P}_i - \mathbf{L}_i^{k+1} + (\beta^k)^{-1}\mathbf{Y}^k)$
6: $S_i^{k+1}(m,n) = T_\varepsilon[x(m,n)]$
7: $\mathbf{Y}^{k+1} = \mathbf{Y}^k + \beta^k(\mathbf{F}_i\mathbf{P}_i - \mathbf{L}_i^{k+1} - \mathbf{S}_i^{k+1})$
8: $\beta^{k+1} = \min(\rho\beta^k, \beta_{max})$
9: $k \leftarrow k+1$
10: **end while**

Output: $(\mathbf{L}_i^k, \mathbf{S}_i^k)$

$$(\mathbf{L}_i^*, \mathbf{S}_i^*) = \arg\min(\|\mathbf{L}_i\|_* + \lambda\|\mathbf{S}_i\|_1) + tr(\mathbf{Y}^T(\mathbf{F}_i\mathbf{P}_i - \mathbf{L}_i - \mathbf{S}_i)) + \frac{\beta}{2}\|\mathbf{F}_i\mathbf{P}_i - \mathbf{L}_i - \mathbf{S}_i\|_F^2 + \mu_2\sum_{j=1}^{N_i}(y_{ij} - \alpha_i\|S_{ij}\|_1)^2 \quad (6)$$

where $tr(\cdot)$ is the trace of a matrix, $\mathbf{Y}$ is the Lagrange multiplier and $\beta > 0$ is a penalty parameter. The inexact ALM method in [7] is used for efficiency and outlined in Algorithm 2, and $T_\varepsilon[\cdot]$ is the soft-thresholding (shrinkage) operator.

4. EXPERIMENTS AND RESULTS

To validate the effectiveness of the proposed method, we conduct experiments on two publicly available benchmarks, i.e., MSRC-v2[1] and iCoseg [2]. And we compare the proposed DLRR with three state-of-the-art approaches [4][5][8], and two special cases of DLRR. The first special case is denoted as LRR, which directly uses the the saliency detection result based on low rank matrix recovery as the object probability. That is, μ_1 and μ_2 are both set to 0 in the proposed model (5). The second one is denoted as DIS, which is a two-step method. The low rank matrix recovery is first performed to get the image saliency $\|S_{ij}\|_1$, and then we learn the discriminative model with the fixed S_{ij}. The visual features used for over-segmentation include color features, gabor features, and steerable pyramid features, which are same to the work in [10]. The segmentation performance is measured by the *intersection-over-union* score and defined by $\frac{1}{|\tau|}\sum_{i\in\tau}\frac{R_i \cap GT_i}{R_i \cup GT_i}$, where R_i is the segmentation result of image i and GT_i is the ground truth. This evaluation metric is standard in PASCAL challenges.

MSRC-v2 consists of 14 classes of images, and each class contains 30 images, except that the 'cat' class contains 24 images. The comparison of different methods on the dataset is listed in Table 1, where we directly cite the results of [4] [5]

[1] http://research.microsoft.com/en-us/projects/ObjectClassRecognition/

Table 1: Results on MSRC Dataset

class	DLRR	LRR	DIS	[5]	[8]	[4]
Bike	48.8	45.3	**54.9**	43.3	42.8	42.3
Bird	44.4	45.0	32.3	**47.7**	-	33.2
Car	53.3	53.0	48.6	**59.7**	52.5	59.0
Cat	58.6	**59.4**	47.8	31.9	5.6	30.1
Chair	50.5	**52.1**	41.3	39.6	39.4	37.6
Cow	**63.8**	63.0	51.0	52.7	26.1	45.0
Dog	50.1	**51.3**	36.9	41.8	-	41.3
Face	52.1	52.7	44.2	**70.0**	40.8	66.2
Flower	**56.3**	54.2	53.3	51.9	-	50.9
House	51.1	49.3	45.4	51.0	**66.4**	50.5
Plane	**46.1**	43.8	44.2	21.6	33.4	21.7
Sheep	64.9	64.5	56.7	**66.3**	45.7	60.4
Sign	**62.4**	61.1	55.2	58.9	-	55.2
Tree	57.1	52.1	64.5	**67.0**	55.9	60.0
Avg.	**54.2**	53.3	48.3	50.2	40.9	46.7
Std	6.5	6.6	8.5	13.9	16.9	13.1

Table 2: Results on ICoseg Dataset

class	No.	DLRR	LRR	DIS	[5]	[4]
Baseball	25	**62.8**	55.2	55.8	13.6	31.4
Football	33	**39.3**	36.9	35.1	38.7	14.9
Monk	17	40.4	36.4	28.2	**73.8**	68.4
BrownBear	5	39.5	40.6	33.7	**57.5**	49.4
Ferrari	11	**54.5**	48.5	49.8	38.7	26.4
Skating	11	54.2	45.3	67.0	**72.7**	38.1
AlasBear	19	41.6	40.2	38.8	41.6	**46.1**
TajlMahal	5	**46.6**	45.4	41.1	37.1	38.4
Helicopter	12	62.4	**64.7**	50.8	33.3	61.0
Kite	18	45.8	43.9	37.3	22.1	**57.8**
Avg	16	**48.7**	45.7	43.8	42.9	43.2
Std	-	9.2	8.7	11.8	19.7	16.6

[8] reported in [5]. From the results in Table 1, DLRR outperforms all the compared methods in terms of the average performance, and specially achieves the best performance for 8 of 14 classes. In addition, our method is more robust than the others because of the minimum standard deviation. Compared with LRR, DLRR improves obviously for the classes with high coherence(e.g., sign, flower and tree). For DIS, using predicted saliency as the prior for discriminative learning in a progressive manner, it decreases the performance by a large margin compared with DLRR.

Table 2 gives a quantitative comparison with [4] [5] on the iCoseg dataset. Since we mainly focus on the segmentation of the common object regions from the background, then the class number in [5] is set to 2. We use the publicly versions of [5, 4] to get the co-segmentation results and prefer the regions with larger overlap with ground truth as the target object regions. From results in Table 2, we can find that DLRR achieves the best performance for 5 of 10 image classes, and the average segmentation score increases 12.7% relatively to [4]. Conclusions conducted above are further verified. DLRR outperforms the state-of-the-art methods for most classes and are more robust to class changes, and DLRR shares the advantages of LRR and DIS. Compared with LRR, DLRR improves obviously for the classes with high class coherence (e.g., baseball, skating and Tai Mahal).

Figure 3 presents some object co-segmentation results of the proposed method on the both datasets, in which the successful and unsuccessful examples are both shown.

Figure 3: Some object co-segmentation results.

The failed results may be due to the diversity of the object appearance.

5. CONCLUSIONS

In this paper, a novel discriminative low rank matrix recovery algorithm is proposed to perform object co-segmentation. Our method works on the assumption that object region should be not only common among images but salient one in an image. This is the first to be used in the task of object segmentation. We import the low rank matrix recovery term to measure the saliency of super-pixels so as to eliminate the disturbance from those consistent backgrounds, while a discriminative learning term is used to model the true object region simultaneously. Besides, a regularized penalty is employed to promote the both terms each other. A joint optimization algorithm is designed to solve the proposed formulation. Extensive experiments have shown the outperforming performance compared with some state-of-the-arts.

6. ACKNOWLEDGMENTS

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61272329, 61070104, and 60905008).

7. REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object ? In *CVPR*, 2010.

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. Interactively co-segmentating topically related images with intelligent scribble guidance. *IJCV*, 93(3):273–292, 2011.

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *TPAMI*, 24(5):603–619, 2002.

[4] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[5] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012.

[6] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012.

[7] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2214*, October 2010.

[8] L. Mukherjee, V. Singh, and J. Peng. Scale invariant cosegmentation for image groups. In *CVPR*, 2011.

[9] C. Rother, V. Kolmogorov, and A. Blake. "grabcut" - interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, Aug 2004.

[10] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012.

[11] J. Wright, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.