Co-segmentation of Image Pairs with Quadratic Global Constraint in MRFs

Yadong Mu and Bingfeng Zhou

Institute of Computer Science and Technology
Peking University, Beijing, 100871
{muyadong,zhoubingfeng}@icst.pku.edu.cn

Abstract. This paper provides a novel method for co-segmentation, namely simultaneously segmenting multiple images with same foreground and distinct backgrounds. Our contribution primarily lies in four-folds. First, image pairs are typically captured under different imaging conditions, which makes the color distribution of desired object shift greatly, hence it brings challenges to color-based co-segmentation. Here we propose a robust regression method to minimize color variances between corresponding image regions. Secondly, although having been intensively discussed, the exact meaning of the term "co-segmentation" is rather vague and importance of image background is previously neglected, this motivate us to provide a novel, clear and comprehensive definition for co-segmentation. Thirdly, it is an involved issue that specific regions tend to be categorized as foreground, so we introduce "risk term" to differentiate colors, which has not been discussed before in the literatures to our best knowledge. Lastly and most importantly, unlike conventional linear global terms in MRFs, we propose a sum-of-squared-difference (SSD) based global constraint and deduce its equivalent quadratic form which takes into account the pairwise relations in feature space. Reasonable assumptions are made and global optimal could be efficiently obtained via alternating Graph Cuts.

1 Introduction

Segmentation is a fundamental and challenging problem in computer vision. Automatic segmentation [1] is possible yet prone to error. After the well-known Graph Cuts algorithm is utilized in [2], there is a burst of interactive segmentation methods ([3], [4] and [5]). Also it is proven that fusing information from multiple modalities ([6], [7]) can improve segmentation quality. However, as argued in [8], segmentation from one single image is too difficult. Recently there is much research interest on multiple-image based approaches.

In this paper we focus on *co-segmentation*, namely simultaneously segmenting image pair containing identical objects and distinct backgrounds. The term "co-segmentation" is first introduced into the computer vision community by Carsten Rother [8] in 2006. Important areas where co-segmentation is potentially useful are broad: automatic image/video object extraction, image partial distance,

Y. Yagi et al. (Eds.): ACCV 2007, Part II, LNCS 4844, pp. 837–846, 2007.

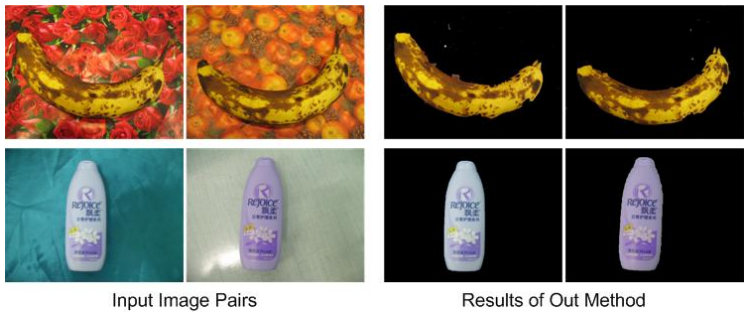

Fig. 1. Experimental results for our proposed co-segmentation approach

video summarization and tracking. Due to space consideration, we focus on the technique of co-segmentation itself, discussing little about its applications.

We try to solve several key issues in co-segmentation. Traditional global terms in MRF are typically linear function, and can be performed in polynomial time [9]. Unfortunately, such linear term is too limited. Highly non-linear, challenging global terms [8] are proposed for the goal of co-segmentation, whose optimization is NP-hard. Moreover, although having been intensively discussed, the exact meaning of the term "co-segmentation" is rather vague and importance of image background is previously neglected. In this paper, we present a more comprehensive definition and novel probabilistic model for co-segmentation, introduce a quadric global constraint which could be efficiently optimized and propose *Risk Term* which proves effective to boost segmentation quality.

2 Generative Model for Co-segmentation

2.1 Notations

The inputs for co-segmentation are image pairs, and it is usually required that each pair should contain image regions corresponding to identical objects or scenes. Let $K = \{1, 2\}$ and $I_c = \{1, \ldots, N\}$ are two index sets, ranging over images and pixels respectively. k and i are elements from them. Z_k and X_k are random vectors of image measurements and pixel categories (foreground/background in current task). z_{ki} or x_{ki} represents i-th element from k-th image. We assume images are generated according to some unknown distribution, and each pixel is sampled independently. Parameters for image generation model could be divided into two parts, related to foreground or background regions respectively. Let θ_{kf} and θ_{kb} denote object/background parameters for k-th image.

2.2 Graphical Models for Co-segmentation

Choosing appropriate image generation models is the most crucial step in co-segmentation. However, such models are not obvious. As in the previous work in

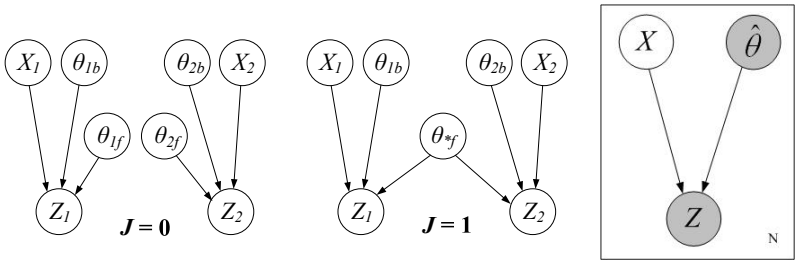

Fig. 2. Generative models for co-segmentation. **(a)** Rother's model (refer to [8] for details) based on hypothesis evaluation. $J = 1$ and $J = 0$ correspond to the hypothesis that image pairs are generated with/without common foreground model respectively. **(b)** Generative model proposed in this paper for co-segmenting.

[8], Rother etc. selected 1D histogram based image probabilistic models, whose graphical models are drawn in Figure 2(a). As can be seen, Rother's approach relies on hypothesis evaluation, namely choose the parameters maximizing the desired hypothesis that two images are generated in the manner of sharing non-trivial common parts. It could also be equivalently viewed as maximizing the joint probability of observed image pairs and hidden random vectors (specifically speaking, these are θ_{kf}, θ_{kb} and X_k in Figure 2(a), where k ranges over $\{1, 2\}$).

However, the above-mentioned generative models are not practical although flexible. The drawbacks lie in several aspects. Firstly, Rother's model makes too many assumptions for the purpose of feasibility, which complicates parameter estimation and makes this model sensitive about noises. Some model parameters even could not unbiasedly estimated due to lack of sufficient training samplings. For some parameters there is only one sampling could be found. An example for this point is that, image likelihoods under hypothesis $J = 0$ are always almost equal to 1, which is certainly not the true case.

Secondly, the final deduced global term in [8] is highly non-linear. In fact it could be regarded as the classical 1-norm if we treat each histogram as a single vector, which complicated optimization for optimal pixel labeling.

Lastly and most importantly, the authors did not seriously take into account the relation between background models in the image pair. Let h_{kf} and h_{kb} denote image measurement histograms (typically color or texture) of foreground/background for k-th image. The final energy function to be minimized in [8] only contains an image generation term proportional to $\sum_z |h_{1f}(z) - h_{2f}(z)|$, while background parameters disappear. This greedy strategy sometimes brings mistake. Here we argue that the effect of background could not be neglected. An example to illuminate our idea is given in Figure 3, where two segmenting results are shown for comparison. In case 1, the extracted foregrounds match each other perfectly if just comparing their color histogram. However, it seems the segmentation in case 2 is more preferable, although the purple regions in

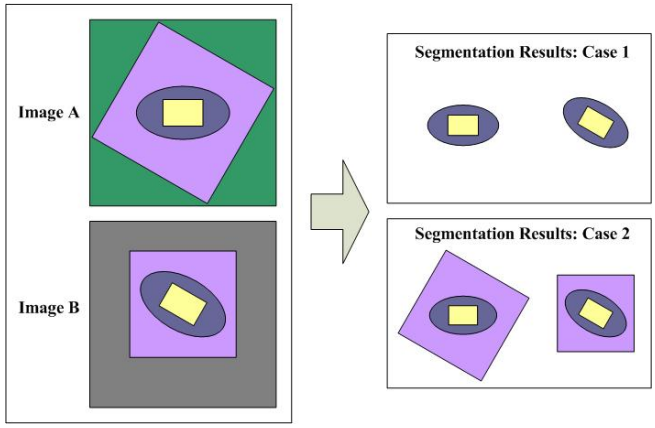

Fig. 3. An example to illustrate the relation between "optimality" and "maximality". The purple region in the bottom image is slightly larger than the top image's. If we only consider foreground models as in [8], case 1 is optimal. However, it is not maximal, since the purple regions are supposed to be labeled as foreground as in case 2.

the two images differ greatly in size. In other words, we should consider both "optimality" and "maximality". Case 1 is an extreme example, which is optimal according to the aforementioned criteria, yet not maximal. We argue that the task of co-segmentation could be regarded as finding the *maximal common parts* between two feature sets together with spatial consistency. Unlike [8], we obtain maximality by introducing large penalties if the backgrounds contain similar contents. A novel energy term about image backgrounds is proposed and detailed in Section 4.

Our proposed graphical model could be found in Figure 2(b). At each phase, we optimize over X on one image by assuming parameters of the other image are known (Note $\hat{\theta}$ in Figure 2(b) is colored in gray since its value is known.). We solve this optimization using alternating Graph Cuts, which is illustated in Figure 4. The joint probability to be maximized could be written as:

$$X^* = \arg\max_X P(X)P(Z|X,\hat{\theta}) \tag{1}$$

To solve this optimization problem is equivalent to find the minima of its negative logarithm. Denote $E_1 = -\log P(X)$ and $E_2 = -\log P(Z|X,\hat{\theta})$. For convenience we use the latter log form.

3 Preprocessing by Color Rectification

It is well known that RGB color space is not uniform, and each of the three channel is not independent. It is previously argued in [10] that proper color coordinate transformations are able to partition RGB-space differently. Similar

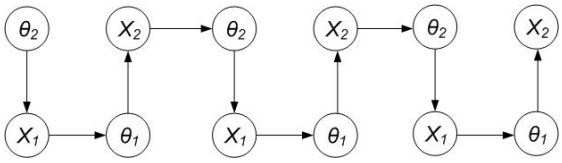

Fig. 4. Illustration for alternating Graph Cuts. The optimization is performed in an alternative style. The vertical arrows denote optimization with graph cuts, while the horizontal arrows indicate building color histograms from segmentation X and pixel measurements Z.

to the ideas used in intrinsic images [11], we abandon the intensity channel, keeping solely color information. In practice, we first transform images from RGB-space to CIE-LAB space, where the L-channel reprensents lightness of the color and the other two channels are about color. After that, we perform color rectification in two steps:

- Step One: Extract local feature points from each image, and find their correspondences in the other image if existing.
- Step Two: Sample colors from a small neighborhood of matching points, use linear regression to minimize color variance.

In step one, we adopts SIFT [12] to detect feature points. SIFT points are invariant to rotation, translation, scaling and partly robust to affine distortion. Also it shows high repeatability and distinctiveness in various applications and works well for our task. Typically we can extract hundreds of SIFT points from each image, while the number of matching point pairs varies according to current inputs. An example for SIFT matching procedure could be found in Figure 8. In the middle column of Figure 8, matching pixels are connected with red lines. These matching points are further used to perform linear regression [13] within each color channel. Colors are scaled and translated to match theirs correspondences, so that color variances between image pair are minimized in a sense of least squared error (LSE). An example can be found in Figure 5. Robust methods such as RANSAC could be exploited to remove outliers.

4 Incorporating Global Constraint into MRFs

4.1 Notations

In this section we provide definitions for E_1 and E_2, which are the negative log of image prior and likelihood respectively. Since we focus on only one image each time, we will drop the k subscript and use it to index histogram bins. We adopted the following notations for convenience:

- $x_i \in \{1, -1\}$, where $x_i = 1$ implies "object", otherwise background.
- $I_h = \{1, \ldots, M\}$ and $I_c = \{1, \ldots, N\}$ are index sets for histogram bins and image pixels.

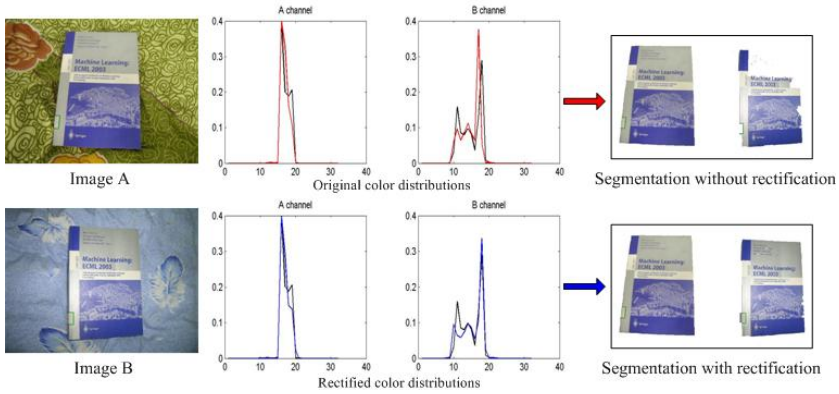

Fig. 5. Illustration for color rectification. Variances of foreground colors affect final segmentation results notably (see the top images in the third column, compare it with the bottom segmentations). We operate in CIE-LAB color space. After color rectification, 1-norm of distribution difference in A-channel is reduced to 0.2245, compared with original 0.2552. And the results in B-channel are more promising, from 0.6531 to 0.3115. We plot color distribution curves in the middle column. Color curve for image A remains unchanged as groundtruth and plotted in black, while color curves for image B before/after rectification are plotted in red and blue respectively. Note that the peaks in B-channel approach groundtruth perfectly after transformation. The two experiments in rightmost column share same parameters.

- $S(k)$ is the set of pixels that lies in histogram bin k.
- $F(k)$ and $B(k)$ denote the number of pixels belonging to foreground/background in bin k. Specifically speaking, $F(k) = \frac{1}{2}(|S(k)| + \sum_{i \in S(k)} x_i)$, $B(k) = \frac{1}{2}(|S(k)| - \sum_{i \in S(k)} x_i)$, where $|\cdot|$ means the cardinality of a set.
- N_f and N_b denotes pixel counts labeled as foreground/background across the whole image. $N_f = \frac{1}{2}(N + \sum_{i \in I_c} x_i), N_b = \frac{1}{2}(N - \sum_{i \in I_c} x_i)$.
- $DIST(h_1, h_2)$ is a metric defined on histograms. We adopt a sum-of-squared-difference (SSD) form, namely $DIST(h_1, h_2) = \sum_k (h_1(k) - h_2(k))^2$.

4.2 Ising Prior for $P(X)$

We adopt the well-known Ising prior for $P(X)$. Similar to [8], a *preference term* is added to encourage larger foreground regions, whose strength is controlled by a positive constant α. A second term is over neighboring pixels. This energy term could be summarized as follows:

$$E_1 = -\alpha \sum_i x_i + \lambda \sum_{i,j} c_{ij} x_i x_j \tag{2}$$

where $c_{ij} = \exp(-||z_i - z_j||^2/\sigma^2)$ are coefficients accounting for similarity between pixel pairs.

4.3 Global Term for $P(Z|X,\hat{\theta})$

As argued before, the global constraint should take into account both the effects of foreground/background. We adopt a simple linear combination of the two, that is:

$$E_2 = w_f DIST(\hat{h}_f, h_f) - w_b DIST(\hat{h}_b, h_b) \tag{3}$$

where $\hat{h}$ denotes known histograms of the reference image, while h represents histograms to be estimated. In practice we build 2D histogram from the two color channels in LAB-space. It is obvious that this global term favors *maximal common parts*: similar foregrounds, and backgrounds that are different from each other as much as possible. For the purpose of tractability we assume $w_f = \gamma_1 N_f^2$ and $w_b = \gamma_2 N_b^2$, then E_2 could be written as:

$$\begin{aligned}
E_2 &= w_f DIST(h_f, \hat{h}_f) - w_b DIST(h_b, \hat{h}_b) \\
&= \gamma_1 N_f^2 \sum_k (\frac{F(k)}{N_f} - \hat{h}_f(k))^2 - \gamma_2 N_b^2 \sum_k (\frac{B(k)}{N_b} - \hat{h}_b(k))^2 \\
&= \gamma_1 \sum_k F^2(k) - \gamma_2 \sum_k B^2(k) + \gamma_1 \sum_k N_f^2 \hat{h}_f^2(k) - \gamma_2 \sum_k N_b^2 \hat{h}_b^2(k) \\
&\quad -2\gamma_1 \sum_k N_f F(k) \hat{h}_f(k) + 2\gamma_2 \sum_k N_b B(k) \hat{h}_b(k)
\end{aligned} \tag{4}$$

Now we will prove Equation 4 is actually quadric function about X. Denote the first two terms in Equation 4 as T_1, middle two as T_2, the last two as T_3, thus $E_2 = T_1 + T_2 + T_3$. Recall that in Equation 2, parameter α indicates user's preference for the ratio $\frac{\text{Foreground size}}{\text{Image size}}$ (typically set to 0.3 in our experiments), thus we could deduce that $\sum_{i \in I_c} x_i = (2\alpha - 1)N$. Basing on this observation, it is easy to prove that:

- $T_1 = \frac{1}{2}(\gamma_1 - \gamma_2) \sum_{\exists k, i, j \in S(k)} x_i x_j + \sum_{i \in I_c} p_i x_i + const$, where p_i is coefficient unrelated to X.
- T_2 is unrelated to X.
- $T_3 = \sum_{i \in I_c} q_i x_i + const$, where q_i is coefficient concerning i-th pixel.

As a result, we could represent global term E_2 in the following form:

$$E_2 = \frac{1}{2}(\gamma_1 - \gamma_2) \sum_{\exists k, i, j \in S(k)} x_i x_j + \sum_{i \in I_c} (p_i + q_i) x_i + const \tag{5}$$

This novel quadratic energy term consists of both unary and binary constraints, thus fundamentally different from conventional ones used in [2], [3] and [4], where only linear constraints are utilized. Moreover, it also differs from the pairwise Ising term defined in Equation 2, since the latter performs on neighborhood system in spatial domain while the pairwise term in Equation 5 works in feature space. From a graph point of view, each adjacent pixel pair in feature space (that is, they fall into the same histogram bin) is connected by an edge, even if they are far away from each other in the spatial domain.

4.4 Computation

To optimize above-defined energy function is challenging due to the existence of quadric global constraint. Although optimization methods like graph cuts [14] or normalized cuts [1] could found its optimal, required memory space is too huge for current computer hardware. For image pairs with typical size of 800*600, the global term usually gives rise to more than 1G extra edges, which is intolerant. General inference algorithm like MCMC [15], hierarchical methods or iterative procedures [8] are more favorable for such optimizing task.

However, the common drawback for these methods lies in that they are too time-consuming, thus not suitable for real-time applications. To make a balance between efficiency and accuracy, we let γ_1 be equal to γ_2 in Equation 5, reducing the global term into a classical linear form. Experiments prove effectiveness of this approximation.

4.5 Risk Term

Another important issue is seldom considered in previous work. For an input image pair, small regions with unique color usually tend to be categorized as "foreground" (see Figure 6 for an concrete example). This is mainly because they affect E_2 much slighter than the preference term in E_1. To mitigate this problem, we propose a novel constraint named *Risk Term*, which reflects the risk to assign a pixel as foreground according to its color. h_1, h_2 denote 2D histograms for image pair. For histogram bin k, its risk value is defined as follows:

$$R(k) = \frac{|h_1(k) - h_2(k)|}{|h_1(k) + h_2(k)|} \tag{6}$$

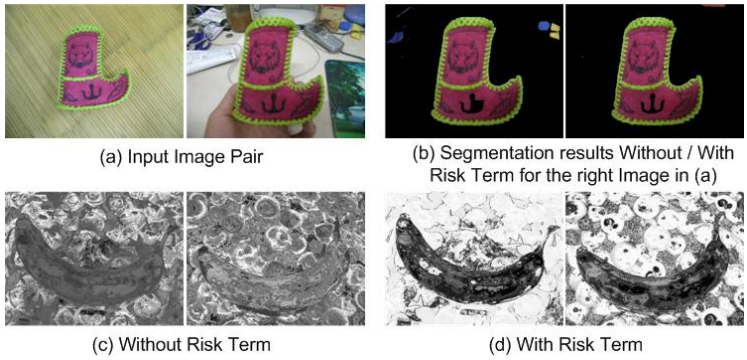

Fig. 6. Illustration for *Risk Term*. For the right image in (a), several small regions are labeled as foreground objects (see the left image in (b)), after introducing risk term they are removed. Also we draw the coefficients $p_i + q_i$ in Equation 5 (normalized to $[0, 255]$) in (c) and (d). Lower brightness implies more tendency to be foreground. The benefit of risk term is obvious.

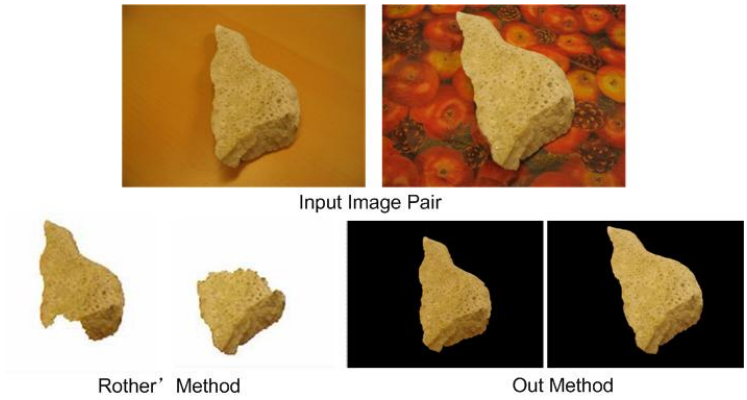

Fig. 7. Comparison with Rother's method. Parameters are identical in both experiments: $\alpha = 0.3$, $\lambda = 50$. Note that α corresponds to user's prior knowledge about the percentage of foreground in the whole image. It is shown that the way to choose α in our method is more consistent with user's intuition.

5 Experiments and Comparison

We apply the proposed method in a variety of image pairs from public image sets or captured by ourselves. Experiments show our method is superior to previous ones in aspects including accuracy, computing time and ease of use. Lacking color rectification makes previous methods such as in [8] couldn't handle input images captured under very different illuminating conditions or cluttered backgrounds (Figure 1, 5 and 6). Also, experiments shows the way to choose parameter in our method is more consistent with user's intuition (Figure 7). For typical 640*480

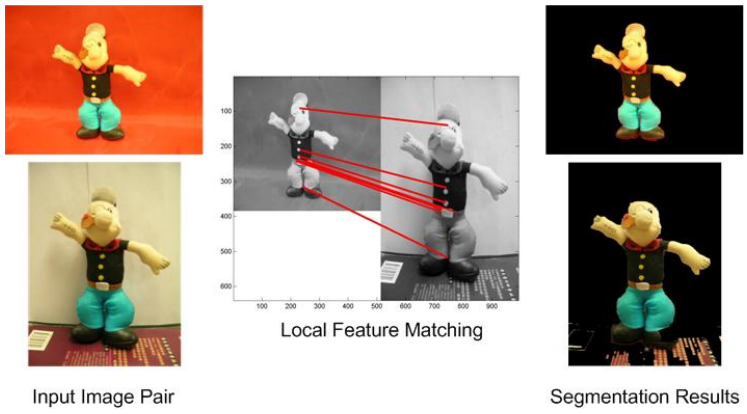

Fig. 8. A failure example due to confusion of foreground/background colors

image pairs, the algorithm usually converges in fewer than 4 cycles, and each iteration takes about 0.94 seconds on a Pentium-4 2.8G/512M RAM computer.

6 Conclusions and Future Work

We have presented a novel co-segmentation method. Various experiments demonstrated its superiority over the state-of-the-art work. Our result (Figure 8) also showed certain limitation of the algorithm due to only utilizing color information; and our future work will focus on how to effectively utilize more types of information such as shapes, textures and high-level semantics.

References

1. Yu, S.X., Shi, J.: Multiclass spectral clustering. In: ICCV, pp. 313–319 (2003)
2. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: ICCV, pp. 105–112 (2001)
3. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23(3), 309–314 (2004)
4. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy snapping. ACM Trans. Graph. 23(3), 303–308 (2004)
5. Wang, J., Cohen, M.F.: An iterative optimization approach for unified image segmentation and matting. In: ICCV, pp. 936–943 (2005)
6. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Bi-layer segmentation of binocular stereo video. CVPR (2), 407–414 (2005)
7. Sun, J., Kang, S.-B., Xu, Z., Tang, X., Shum, H.Y.: Flash cut: Foreground extraction with flash/no-falsh image pairs. In: CVPR (2007)
8. Rother, C., Minka, T.P., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. CVPR (1), 993–1000 (2006)
9. Narasimhan, M., Bilmes, J.: A submodular-supermodular procedure with applications to discriminative structure learning. In: UAI, pp. 404–441. AUAI Press (2005)
10. van de Weijer, J., Gevers, T.: Boosting saliency in color image features. CVPR (1), 365–372 (2005)
11. Weiss, Y.: Deriving intrinsic images from image sequences. In: ICCV, pp. 68–75 (2001)
12. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157 (1999)
13. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, Heidelberg (2001)
14. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? In: ECCV (3), pp. 65–81 (2002)
15. Barbu, A., Zhu, S.C.: Generalizing swendsen-wang to sampling arbitrary posterior probabilities. IEEE Trans. Pattern Anal. Mach. Intell. 27(8), 1239–1253 (2005)