Decomposed Learning for Joint Object Segmentation and Categorization

Yi-Hsuan Tsai
ytsai2@ucmerced.edu
Jimei Yang
jyang44@ucmerced.edu
Ming-Hsuan Yang
mhyang@ucmerced.edu

Electrical Engineering and Computer Science
University of California
Merced, USA

Abstract

We present a learning algorithm for joint object segmentation and categorization that decomposes the original problem into two sub-tasks and admits their bidirectional interaction. In the first stage, in order to decompose output space, we train category-specific segmentation models to generate figure-ground hypotheses. In the second stage, by taking advantage of object figure-ground information, we train a multi-class segment-based categorization model to determine the object class. A re-ranking strategy is then applied to classified segments to obtain the final category-level segmentation results. Experiments on the Graz-02 and Caltech-101 datasets show that the proposed algorithm performs favorably against the state-of-the-art methods.

1 Introduction

The problems of image segmentation and categorization, although closely related, have been tackled as two independent tasks [1, 21]. Recent findings show that both segmentation and categorization significantly improve the performance of each other [22, 24, 31]. In this paper, we consider the interactions of these two tasks and propose an algorithm for joint object segmentation and categorization.

When image regions are labeled as figure and ground through segmentation, such labeling makes it feasible to incorporate contour and shape features in more effective representations as well as local contexts for object recognition. Carreira *et al*. [6] generate multiple hypothesized figure-ground segmentation results by using the constrained parametric min cut (CPMC) so that object recognition can be carried out by ranking the hypotheses. As this approach usually generates a large set of redundant segmentations, segment filtering is essential for efficient hypothesis verification in the recognition phase. On the other hand, object category information provides global constraints of visual elements on which the segmentation task operates (pixels or super pixels) such that ambiguities of constituent components can be minimized. Recently, Jain *et al*. [17] exploit a high-order Conditional Random Field (CRF) for joint segmentation and categorization, in which object categorization is modeled

as the global constraint on the bag-of-words (BoW) representation. However, as the number of object categories increases, the inference and learning processes on a high-order CRF become computationally expensive.

Considering the above challenges, we present an algorithm that decomposes the joint segmentation and categorization problem into two sub-tasks:

1. category-specific figure-ground segmentation, and

2. segment-based object categorization,

such that both learning and inference processes can be carried out efficiently and effectively.

Most of category-specific object segmentation algorithms [9, 20, 23] generate one segmentation based on maximum a posteriori (MAP) inference. Such approaches are likely to miss small objects or generate incomplete masks, although they may have high precision segmentation results on the pixel level. We take the classic hypothesize-and-test approach and generate multiple object segmentations with a high recall rate since it is rather difficult to infer the category information if numerous constituent segments are missing. Inspired by the CPMC method, we generate multiple plausible segmentation hypotheses to increase the chances of finding all the true segments of objects.

In the segmentation stage, we train category-specific classifiers for figure-ground segmentation based on the pylon model [23]. By introducing the parametric min cut [19] into the pylon model during the inference stage, we are able to generate multiple segmentation hypotheses. We aim to generate object segmentation hypotheses with a high recall rate on the positive images (with target objects), and meanwhile allow false positives on the negative images (without target objects). Therefore, we train a pylon model only with the positive images for each object category. When applied to negative images, the learned pylon model will identify image regions that look similar to the target object category, which is essentially the process of mining hard negative examples. Having generated object segmentations from both positive and negative images, we train support vector machine (SVM) classifiers on figure-ground representations for categorization. For each figure-ground hypotheses, we extract bag-of-words features on the foreground and the background regions separately, and stack them to represent this image. Thus, foreground features encode the structure of an object while background features serve as the corresponding context information. For each test image, we evaluate hypotheses in a class-wise manner. That is, we do not need to evaluate the hypotheses from class A by the SVM classifier for class B. This operation improves the efficiency of hypothesis verification for categorization and also allows parallelization. We select the highest classification score as the image label prediction and re-rank the hypotheses of the predicted class as the final segmentation for an input image.

Our algorithm enjoys bidirectional interactions between segmentation and categorization. In the segmentation phase, category information facilitates breaking down the multi-class segmentation problem into class-wise sub-problems such that high-quality figure-ground separation can be generated in a reduced labeling space. In the categorization phase, segmentation information helps identifying object locations, shapes as well as context, and hence objects can be precisely represented in the feature space and improve the categorization performance. For concreteness, we demonstrate the merits of the proposed algorithm on the Graz-02 and Caltech 101 data sets. Experimental results show the proposed algorithm performs favorably against the state-of-the-art methods in both segmentation and categorization tasks.

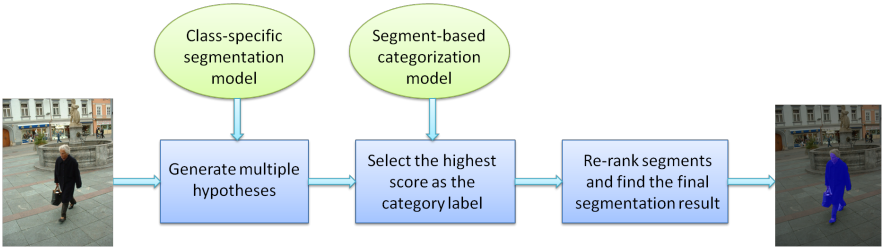

Figure 1: Overview of the algorithm.

2 Related Work and Problem Context

Image segmentation for classification. Numerous algorithms [6, 7, 8, 14, 26, 30, 33] have been proposed to exploit image segmentation for classification. Existing methods mainly use bottom-up approaches to generate redundant image segmentations by using multiple algorithms [26], hierarchical grouping [14], different location seeds [11] and parameters [6]. While these methods only investigate the influence of segmentation on categorization, the proposed algorithm allows bidirectional interactions between segmentation and categorization. In the segmentation stage, we use the categorization information to generate category-specific object hypotheses. Compared to category-independent methods, our method generates a smaller set of high plausible segmentations for each class, which facilitates the categorization process. Recently, Chai *et al*. [7, 8] use category-level information for segmentation in a weakly supervised manner. Image co-segmentation is carried out from image level to dataset level to determine the best figure-ground mask for each image. Their co-segmentation results are then used for fine-grained object recognition.

Joint object detection and classification. Similar to joint segmentation and categorization, numerous methods have been proposed to exploit the relationships between object detection and classification [16, 27, 29, 32]. The key idea is that object location information helps the categorization task and vice versa. In [27, 29], a joint learning algorithm is formulated in a weakly supervised fashion without bounding box annotations. Therefore, the object locations are treated as latent variables and learned jointly with categorization in latent SVM models. In [16, 32], the output of object detection and categorization are used as their mutual context information. Song *et al*. [32] propose an iterative approach to boost the performance of detection and categorization while Harzallah *et al*. [16] present a cascade approach by accommodating detection and categorization at different stages. In our work, we also take the location of objects into account by supervised segmentation and perform classifier training for categorization based on the previously generated segmentation hypotheses. In this sense, our method bears some similarity to the cascade model [16].

3 Joint Segmentation and Categorization

3.1 Overview of the algorithm

The first step of our algorithm is to train a figure-ground segmentation model for each class. We train the pylon model [23] based on the segmentation tree generated by the global probability of boundary (gPb) method [1]. In the test stage, we solve multiple parametric graph-cut [18] to generate multiple figure-ground hypotheses. Therefore, for each image we can

Figure 2: Examples of hypotheses generated from different segmentation models. The **left** column is generated by the car model, referring to the positive bag. The **middle** and the **right** column are from the bike and person models as negative bags. Usually the hypotheses from negative bags include noises and incomplete object masks.

obtain a hypothesis set from each class-wise segmentation model. We treat the hypothesis set constructed by the positive class model as the positive bag which contains the best segmentation result, and others are negative bags with all negative samples. All of these bags will be used for training in the object categorization phase. Note that our framework is similar to Multiple Instance Learning (MIL) [9], but there is no uncertainty in the positive bag since we know which segmentation is the best hypothesis (see Section 3.3).

The second step is to learn a multi-class categorization model based on the positive and negative figure-ground samples obtained from the first stage. In this work, we train a SVM classifier in a one-vs-all manner. For each class, the ground truth segmentation and the best segmentation by our algorithm are used as positive samples while all the samples in the negative bags are used as negative samples. To generate the final segmentation and categorization results, we first determine the image category label by selecting the highest classification score. Note that the segmentation hypothesis with the highest classification score does not necessarily correspond to the best segmentation result, as the classification model may select the hypothesis with the most salient parts instead of the entire foreground region. Therefore, for each class we also train a Support Vector Regressor (SVR) using all the positive samples and their segmentation scores based on their overlap with the ground truth. From the predicted class label, we re-rank all the hypotheses by the SVR of that class and choose the top one as the segmentation result. Figure 1 shows the main steps of our algorithm and we present the details of each step in the following sections.

3.2 Category-specific segmentation

Pylon model. We solve the category-specific figure-ground segmentation problem by using the two-class pylon model [23] based on the segmentation tree. Suppose that the image I can be partitioned into hierarchical regions $\mathbf{S} = \{S_1, S_2, ..., S_{2N-1}\}$, the region from 1 to N are leaf regions and other regions including the root node (the whole image) are from $N+1$ to $2N-1$. We assign a figure-ground label, $f_i = 1$ or 2 for each region, respectively. Therefore, we formulate the conventional CRF energy function:

$$E(\mathbf{f}) = \sum_{i=1}^{2N-1} U(f_i) + \sum_{(i,j)\in\mathcal{N}} V(f_i, f_j). \quad (1)$$

The unary energy $U(f_i)$ indicates the cost of assigning f_i to the segment S_i and $V(f_i, f_j)$ is the smoothness term of the boundary cost for two neighboring segments S_i and S_j, and $\mathcal{N}$ is the set of adjacent segments. Furthermore, we define the unary energy:

$$U(f_i) = \begin{cases} |S_i| \cdot <\mathbf{w}_1, \mathbf{h}(\mathrm{S_i})>, & \text{for } f_i = 1, \\ |S_i| \cdot <\mathbf{w}_2, \mathbf{h}(\mathrm{S_i})>, & \text{for } f_i = 2, \end{cases} \tag{2}$$

where $\mathbf{w}_1, \mathbf{w}_2$ are the unary parameters and $\mathbf{h}(S_i)$ denotes the feature vectors extracted from each segment S_i. Note that the weighting factor $|S_i|$ is the size of the segment, which encourages the model to prefer larger regions. For the smoothness term, we define the energy by a weighted Potts model:

$$V(f_i, f_j) = <\mathbf{w}_3, \mathbf{b}(\mathrm{S_i}, \mathrm{S_j})> \cdot \delta[\mathrm{f_i} \neq \mathrm{f_j}], \tag{3}$$

where $\mathbf{w}_3$ is the smoothness parameter and $\mathbf{b}(S_i, S_j)$ denotes the boundary strength.

Inference and Learning. As pixels are included in multiple nodes of the segmentation tree, we introduce additional constraints between any pair of child and parent nodes in the tree to enforce that every pixel is only assigned to a single label. The constrained energy function can still be solved by a graph cut with some manipulations. The details of constructing a sub-modular energy function can be found in [23]. In this work, we develop a stochastic gradient descent algorithm to learn the pylon parameters in a max-margin fashion.

Inference for multiple hypotheses. With the learned energy function for $E(\mathbf{f})$, instead of finding only the MAP solution, we introduce a parameter λ into our unary function in Equation 1:

$$U(f_i, \lambda) = \begin{cases} U(f_i) + |S_i| \cdot \lambda, & \text{for } f_i = 1, \\ U(f_i) - |S_i| \cdot \lambda, & \text{for } f_i = 2. \end{cases} \tag{4}$$

To keep the consistency of the weighting factor in the original unary energy, we also multiply λ by $|S_i|$. Different values of λ provide our model a bias to generate parametrized results (similar to parametric min cut [19]) between the MAP solution and the ground truth such that the model capacity problem is alleviated. Therefore, we can adjust the hyperplane $\mathbf{w}$ and generate multiple segmentation hypotheses by solving a series of graph cuts with different λ values. Figure 2 presents some examples of segmentation hypotheses.

Note that the generated hypotheses may be redundant. In order to alleviate the computational load for the training process for object categorization, we filter out the duplicated hypotheses by checking if the overlapping ratio of two hypotheses is larger than a threshold.

Feature representation. We use four different types of features for each region S_i in the segmentation step. We extract a BoW SIFT histogram and a color histogram to represent the region appearance. We also extract a location histogram and a contour descriptor to capture the object shape information. The contour shape is computed by a spatial pyramid of oriented *gPb* edge responses [14]. After concatenating all these four features into a vector, we map this vector to a high-dimensional space with the explicit χ^2 kernel [34]. Detailed parameter settings can be found in Section 4.

3.3 Segment-based categorization

Learning. In the previous stage, we train category-specific classifiers for segmentation. Given an image, we apply each segmentation model to obtain a set of segmentation hypotheses $B_i, i = 1, 2, ..., K$ for K classes. The categorization task is to find the best segmentation

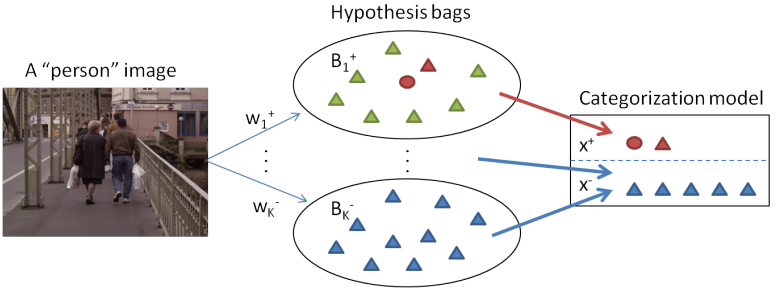

Figure 3: Training the proposed model for object categorization. Given an image labeled with person, we can generate positive and negative hypotheses bags, B_i^+ and B_i^-. To train the classification model, only red ones in B_i^+ are selected as positive samples x^+, where the circle one is the ground truth and the triangle one is the best segmentation hypothesis. For negative samples x^-, we use all the samples from all B_i^-, denoted as blue triangles.

with the correct label among all the hypothesis sets B_i. To tackle this problem, we first divide our hypothesis set into positive and negative bags, denoted by B_i^+ and B_i^- respectively. The positive bag consists of the hypotheses generated with the positive segmentation classifier w_i^+, and likewise negative bags contain examples from negative segmentation classifiers w_i^- (See Figure 3).

During the training process, we solve the categorization problem with a SVM model by collecting all the samples. To reduce the uncertainty in each positive bag, we only choose the best segmentation among a positive bag and the ground truth segmentation as positive samples x^+. In the meanwhile, we use all negative samples x^- from all the negative bags to reduce the chances of false positives. We train a categorization model v by solving the standard SVM optimization problem:

$$\min_{\mathbf{v}} \frac{1}{2}\|\mathbf{v}\|^2 + \frac{C}{N}\sum_n \ell(y_i \cdot v^T x_i), \tag{5}$$

where $\ell(t) = \max(0, 1-t)$ is the hinge loss function, x_i is the feature vector and $y_i \in (1,-1)$ denotes the label for positive or negative samples. We use the stochastic gradient descent based SVM solver [34] to train object categorization models.

Inference. Given a test image, we generate a bag of segmentation hypotheses from each segmentation model as the process in the training stage. We choose the best hypothesis for each bag by measuring the classifier scores, and find the highest one as the target bag, thereby determining the image categorization label. To produce the final segmentation result, we re-rank all the hypotheses in the predicted target bag. The ranking process can be carried out by the class-wise SVR in a way similar to [24]. We train a SVR for each class using all the positive samples and their scores measuring the overlapping ratio between the segment and the ground truth. Figure 4 illustrates the inference process.

Feature representation. In addition to the features we use in the segmentation stage, we extract one more bag-of-words histogram of local shape context descriptors [3] to encode object shape information. For SIFT histograms, we use the spatial pyramid max pooling on the foreground mask, and the global max pooling on the background mask to better represent the object structure [29]. We introduce parameter settings for feature representation in Section 4.

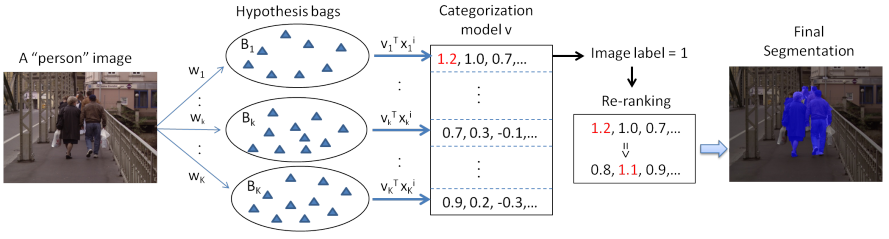

Figure 4: Inference process of the proposed algorithm. Given an image, we can generate a hypothesis bag B_i from each model w_i. From the learned categorization model, each hypothesis is assigned a classifier score. By sorting all the scores from each bag, the class label is first decided with the highest score, and then we re-rank all the hypotheses within the target bag to find the final segmentation result.

3.4 Discussions

From the proposed algorithm, we can summarize that our learning pipeline consists of class-wise segmentation models, the multi-class categorization classifier and class-wise regression models for ranking. For each stage, we deal with specific problems separately, but still sharing the information between each sub-task. This is, segmentation models give useful object figure-ground cues for categorization, which also relaxes the multi-class labeling problem. Likewise, regression models only need to take care of selecting the best segments from a class-wise subset due to the decided class label from the categorization models.

Another potential of our algorithm is that since we decompose the process into sub-problems, for each stage we can use any proper model for specific tasks. For example, we can use other methods to produce more diverse multiple hypotheses, such as [2, 15]. For the categorization, we only use the best positive samples for training, but during the inference, the segmentation results from test images are usually not as good as training ones. To make up this gap, we can train a multi-class regression model considering all the positive samples. To re-rank segments, instead of simply using regression models, a structural SVM can be learned so that the loss function is defined in a relative way, saying that the score of the best segment should be always larger than all the others [35].

4 Experimental Results

4.1 Graz-02

Experimental Settings. The Graz-02 [28] dataset, consisting of 3 object classes (bike, car and person) in different views and background images, is challenging for object segmentation and recognition. There are 300 images for each class and the odd-number images from each object class are used for training, and the others for evaluation. We use all the training images to train a codebook with 512 codewords for SIFT histograms, a codebook of 512 codewords for shape context histograms, and a codebook of 128 codewords for color histograms. Considering large shape variations, we vary the value of λ (Equation 4) from -2 to 2 with the increment 0.1 to generate multiple segmentation hypotheses. After filtering our duplicated hypotheses, the average number of hypotheses in each class is 10.

Experimental Results. To evaluate the quality of segmentation, we compute the commonly

Figure 5: Segmentation results on the Graz-02 test set. Bikes, cars and people are highlighted with red, green and blue masks, respectively. Best viewed in color.

used union-over-intersection accuracy. Table 1 shows the results compared to state-of-the-art methods using the same experimental settings. The proposed algorithm consistently performs better than the other methods in all categories. With the re-ranking strategy, our results achieve 69.06% and outperform the state-of-art methods by 16%. We also report the pixel level accuracy in Table 2. Figure 5 shows some qualitative results and our algorithm performs well in segmenting multiple occluded objects in different views with large scale variation. More results can be found in the supplementary document.

To evaluate the performance for categorization, we compute the classification accuracy in the second stage of the proposed algorithm. The image based object categorization accuracy rates for bike, car and person are 94%, 92%, and 90.7%, respectively.

Table 1: Graz-02 segmentation results using intersection/union overlap metric.

Method	Background	Bicycle	Car	Person	mean
[31]	82.32	46.18	36.49	38.99	50.99
[17]	77.97	55.60	41.51	37.26	53.08
Proposed	91.20	64.95	59.60	60.49	69.06

Table 2: Graz-02 segmentation results using pixel accuracy metric.

Method	Background	Bicycle	Car	Person	mean
[31]	86.44	73.01	68.71	71.32	74.87
[17]	75.90	84.91	76.74	79.78	79.33
Proposed	95.72	75.27	80.19	76.52	81.93

4.2 Caltech-101

Experimental Settings. The Caltech-101 [25] dataset is a commonly used benchmark for object categorization. For each class, we randomly select 30 images from each class for training and all the others for tests, and repeat the experiments three times. For feature extraction, we train codebooks of size 1024 for SIFT and shape contexts histograms, and a

Table 3: Caltech-101 classification results. SFea denotes single feature while MFea denotes multiple features. Geo denotes the geometric information (segmentation, saliency or deformable matching).

	Method	30 training
SFea + Geo	Yang et al. [36]	76.1 ± 1.3
	Feng et al. [13]	82.6
	Duchenne et al. [10]	80.3 ± 1.2
MFea	NBNN [5]	73.0
	Gehler et al. [13]	73.1
MFea + Geo	Gu et al. [14]	77.7
	SvrSegm [24]	82.3
	Proposed	84.2 ± 0.3

Table 4: Segmentation influence on categorization with the Caltech-101 dataset.

Method	SvrSegm [24]	Proposed
Predicted segmentation	82.3	84.2
Upper bound segmentation	82.5	84.7
Ground truth segmentation	89.3	89.8

codebook of size 128 for color histograms. Since the within-class appearance variations of the Caltech-101 dataset is smaller than that of the Graz-02 dataset, we vary the value of λ from -1 to 1 with the increment 0.2 to generate multiple segmentation hypotheses. For each class, the maximum number of hypotheses is 11 per image.

Experimental Results. In Table 3, we present several state-of-art approaches that use multiple types of features and/or geometric information (segmentation, saliency or matching). Overall, the proposed algorithm outperforms the other methods by at least 1.6%. Note that most of the evaluated methods use up to 15 images per class as the test set, and do not provide the standard deviation of classification accuracy. Considering the underlying sample bias, we also carry out experiments with 15 randomly selected test images per class for 20 times in each trial. In this setting, the average accuracy our method is 84.4% with standard deviation 0.22.

Table 4 demonstrates the effects of segmentation results on categorization tasks. We compare the categorization results by using the ground truth segmentation, the upper bound segmentation and the predicted segmentation. The ground truth segmentation gives a high categorization accuracy 89.8% while the upper bound segmentation produces 84.7% accuracy on average. The results indicate the potential of figure-ground segmentation for object categorization. In addition, the categorization result produced by the predicted segmentation is very close to the one by upper bound segmentation (only 0.5% difference). This result shows that our categorization stage can actually find out good segmentations for classification purpose. Compared to [24], our categorization results are better and closer to the one using ground truth masks, which demonstrates our category-specific segmentations are more reliable.

We also evaluate the category-level segmentation results by measuring average overlap, recall, and precision rates over 101 classes. The results are 73.27%, 83.94%, and 86.05%, respectively.

5 Conclusion

In this paper, we proposed a decomposed learning approach for joint segmentation and categorization that takes the interaction of sub-tasks into account. The class label knowledge is first used by the segmentation model for better object representations, which in turn helps the categorization model for predicting the desired class. By recognizing the gap between the outputs from classification and segmentation, the predicted class label is used for re-ranking all the segmentation hypotheses and generating the final joint segmentation and categorization results. Experimental results on the Graz-02 and Caltech-101 datasets show that the proposed algorithm performs favorably against the state-of-the-arts methods in segmentation and classification.

Acknowledgements

This work is supported in part by the NSF CAREER Grant # 1149783 and NSF IIS Grant # 1152576.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011. ISSN 0162-8828.

[2] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in Markov random fields. In *ECCV*, 2012.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24:509–522, 2001.

[4] L. Bertelli, T. Yu, D. Vu, and B. Gokturk. Kernelized structural svm learning for supervised object segmentation. In *CVPR*, 2011.

[5] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[6] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[7] Y. Chai, V. Lempitsky, and A. Zisserman. Bicos : A bi-level co-segmentation method for image classification. In *ICCV*, 2011.

[8] Y. Chai, E. Rahtu, V. Lempitsky, L. Van Gool, and A. Zisserman. Tricos: a tri-level class-discriminative co-segmentation method for image classification. In *ECCV*, 2012.

[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artifical Intelligence*, 89(1-2):31–71, 1997.

[10] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *ICCV*, 2011.

[11] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010.

[12] J. Feng, B. Ni, Q. Tian, and S. Yan. Geometric lp-norm feature pooling for image classification. In *CVPR*, 2011.

[13] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[14] C. Gu, J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, 2009.

[15] A. Guzman-Rivera, D. Batra, and P. Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *NIPS*, 2012.

[16] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.

[17] A. Jain, L. Zappella, P. McClure, and R. Vidal. Visual dictionary learning for joint object categorization and segmentation. In *ECCV*, 2012.

[18] J. Kim and K. Grauman. Shape sharing for object segmentation. In *ECCV*, 2012.

[19] V. Kolmogorov, Y. Boykov, and C. Rother. Applications of parametric maxflow in computer vision. In *ICCV*, 2007.

[20] M. Pawan Kumar, P. Torr, and A. Zisserman. Obj cut. In *CVPR*, 2005.

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[22] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.

[23] V. Lempitsky, A. Vedaldi, and A. Zisserman. A pylon model for semantic segmentation. In *NIPS*, 2011.

[24] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010.

[25] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, April 2007.

[26] T. Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007.

[27] M. H. Nguyen, L. Torresani, L. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009.

[28] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *PAMI*, 28(3):416–431, 2006.

[29] O. Russakovsky, Y. Lin, K. Yu, and Li F.-F. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[30] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[31] D. Singaraju and R. Vidal. Using global bag of features models in random fields for joint categorization and segmentation of objects. In *CVPR*, 2011.

[32] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2010.

[33] S. Todorovic and N. Ahuja. Learning subcategory relevances for category recognition. In *CVPR*, 2008.

[34] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[35] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *CVPR*, 2013.

[36] J. Yang and M.-H. Yang. Learning hierarchical image representation with sparsity, saliency and locality. In *BMVC*, 2011.