Multi-Instance Methods for Partially Supervised Image Segmentation

Andreas Müller* and Sven Behnke
amueller@ais.uni-bonn.de, behnke@cs.uni-bonn.de

Autonomous Intelligent Systems
Department of Computer Science
University of Bonn
53113 Bonn, Germany

Abstract. In this paper, we propose a new partially supervised multi-class image segmentation algorithm. We focus on the multi-class, single-label setup, where each image is assigned one of multiple classes. We formulate the problem of image segmentation as a multi-instance task on a given set of overlapping candidate segments. Using these candidate segments, we solve the multi-instance, multi-class problem using multi-instance kernels with an SVM. This computationally advantageous approach, which requires only convex optimization, yields encouraging results on the challenging problem of partially supervised image segmentation.

1 Introduction

The task of multi-class image segmentation is to create a pixel-wise labeling of an input image into regions belonging to one of several semantic classes. Most algorithms for this setting work with strong supervision: a pixel-wise labeling of training images. Methods that are used in this setting include random forests [24] and support vector machines (SVM). Usually the output of such algorithms is further processed by a conditional random field (CRF [14, 11, 13]). While these methods reach high accuracy, it is very time consuming to create pixel-level ground truth for real-world applications. This problem can be addressed in several ways: the LabelMe effort [23] tries to use the "wisdom of crowds" to obtain human labelings. Another possibility is to use only weak supervision, which is the approach we follow here.

In the weakly supervised setting, the ground truth for a given image is a list of semantic classes that occur in this image, instead of a pixel-level labeling as in the strongly supervised setting. Image-level labels are much easier to obtain, e. g. through online image libraries such as flickr and facebook.

The task of multi-class segmentation is often split up in a segmentation and a recognition part. Random forest methods often classify each pixel separately and segment using predicted classes [24] while SVM-based methods often work

* This work was funded by the B-IT research school.

on an over-segmentation of the image, called superpixels [14, 11]. Superpixels avoid the computational burden of classifying each pixel separately, but have two drawbacks:

1. A single superpixel does not provide enough context for classification [9].
2. Segment boundaries are decided on the lowest level by generating superpixels. This decision cannot be corrected afterwards [12].

In our approach, we work with a set of candidate segments, generated using constrained parametric min-cuts [2]. For each image, these segments are a set of overlapping, object-like regions, which serve as candidates for object locations.

We formulate weakly supervised multi-class image segmentation as a multi-instance problem, based upon candidate segments. In multi-instance learning [6], each training example is given as a multi-set of instances, called a bag. Each instance is represented as a feature vector x and a label y. A bag is labeled positive if it contains at least one positive example, and negative otherwise. During training, only the labels of the training bags, not of the instances inside the bags, are known. The goal is to learn a classifier for unseen bags. Formally, let $\mathcal{X}$ be the set of instances. To simply notation, we assume that bags are simply sets, not multi-sets. Then a bag is an element of the power set $2^{\mathcal{X}}$ and the task is to learn a function

$$f_{MI} \colon 2^{\mathcal{X}} \to \{-1, +1\} \tag{1}$$

from a set of training examples of the form (X_i, y_i) with bags $X_i \subset \mathcal{X}$ and labels $y_i \in \{-1, +1\}$. The f_{MI} function stems from the so-called underlying concept, given by an (unknown) function $f_I \colon \mathcal{X} \to \{-1, +1\}$, with

$$f_{MI}(X) = \max_{x \in X} f_I(x). \tag{2}$$

Sometimes, the goal of finding f_{MI} is extended to finding labels not only on bag-level but also for all the instances within a bag [17, 31], i. e. finding f_I. Even though finding f_I is sometimes included in the task statement, there has been very little work that actually reported accuracy on instance label prediction. Part of the reason for this might be that for many of the datasets used in multi-instance learning no ground truth exists.

We look explicitly at accuracy on instance-level since we are interested in actually segmenting images, not just classifying them. For multi-class image segmentation, there are some hand-labeled datasets that provide ground truth on pixel level. We use this ground truth to evaluate the performance of our method. This approach does not exactly correspond to instance-level ground truth – since the instances are segments, not pixels – but relates to it closely.

In this work, we explore the application of multi-instance learning algorithms to the task of partially supervised image segmentation. Multi-instance learning is a natural formulation for image classification and has been successfully applied in this task [35]. We propose to go a step further and apply multi-instance learning to the task of object-class segmentation in natural images. To our knowledge, all previous methods in the field use strong supervision, meaning manual pixel-wise annotation of training images. This approach does not scale to larger datasets, especially if one expects consistency and quality in the segmentations.

2 Related Work

2.1 Proposal Object Segments

Most work on multi-class segmentation focuses on strong supervision on superpixel level. There is still little work on using candidate segments. The method we use for generating candidate segments is Constraint Parametric Min-Cuts (CPMC) from Carreira and Sminchisescu [2]. This method creates a wide variety of overlapping segments. Support vector regression (SVR) is trained on these segments to estimate the overlap of segments with ground truth object-class labeling from the Pascal VOC dataset [8]. This provides a ranking of candidate segments, according to how "object-like" they are, which allows for selecting only a limited number of very object-like segments. The method performed well on a variety of datasets. A similar approach was investigated by Endres and Hoiem [7].

2.2 Multi-Instance Methods

Multi-instance learning was formally introduced in Dietterich et al. [6]. Since then, many algorithms were proposed to solve the multi-instance learning problem using many different approaches [1, 10, 34, 18, 33, 21, 15, 4]. We will discuss only those that are relevant to this work.

Gärtner et al. [10] introduced the concept of a multi-instance kernel on bags, defined in terms of a kernel on instances. The basic principle of multi-instance kernel is similar to a soft-max over instances in each bag. This can be viewed as approximating the kernel value of the "closest pair" given by two bags. They show that the multi-instance kernel is able to separate bags if and only if the original kernel on instances is able to separate the underlying concepts. The method of Gärtner et al. [10] has a particular appeal in that it neatly transforms a multi-instance problem into a standard classification problem by changing the kernel. The downside of this approach is that it does not directly label instances, only bags.

Zhou et al. [34] explicitly address non-i.i.d. labels, leading to an algorithm that can take advantage of correlations inside bags. Computational costs of their algorithm does not scale well with the number of instances, although a heuristic algorithm is proposed to overcome this restriction. Zhou et al. [34] demonstrated only a slight advantage of their algorithm over the MI-kernel of Gärtner et al. [10], so we use the MI-kernel for better scalability.

Li and Sminchisescu [17] compute likelihood ratios for instances, giving a new convex formulation of the multi-instance problem. Using these likelihood ratios, classification can be performed directly on the instances, provided an appropriate threshold for classifying instances as positive is known. We circumvent this problem by applying the same classifier to instances and bags, thereby obtaining hard class decisions for each instance.

2.3 Semantic Scene Segmentation via Multi-Instance Learning

Recently, several methods have been proposed to obtain semantic segmentations of images using only image-level supervision [29, 27, 28]. Vezhnevets et al. [29], for example, report impressive results on the MSRC dataset.

While semantic segmentation is closely related to multi-class image segmentation, there are important distinctions: In semantic segmentation, each pixel has a semantic annotation, also containing "background" classes like "sky", "grass" and "water". In multi-class image segmentation, the focus is on objects, with possibly large parts of the image being labeled as unspecific "background". The unspecific background class contains much more clutter than for example "grass" and is therefore much harder to model. This makes disseminating the interesting part in multi-class object recognition challenging, since it is not necessary possible to identify non-object regions easily.

3 Multi-Instance Kernels for Image Segmentation

3.1 Constraint Parametric Min Cuts (CPMC)

To generate proposal segments, we use the CPCM framework from Carreira and Sminchisescu [2]. We construct initial segments using graph cuts, on the image graph. The energy function for these cuts uses pixel color and the response of the global probability of boundary (gPb) detector [20]. As much as ten thousand initial segments are generated from foreground and background seeds. A fast rejection based on segment size and ratio cut [30] reduced these to about 2000 overlapping segments per image. Then, the segments are ranked according to a measure of object-likeness that is based on region and Gestalt properties. This ranking is computed using an SVR model [2], which is available online. For computing the global probability of boundary (gPb), we used the CUDA implementation of Catanzaro et al. [3], instead of the original one, for speed.

3.2 Multi-Instance Learning using MI-Kernels

Since scalability is very important in real-world computer vision applications, and natural images might need hundreds of segments to account for all possible object boundaries, we use the efficient multi-instance kernels [10]. Multi-instance kernels are a form of set kernels that transform a kernel on instance level to a kernel on bag level. We reduce the multi-instance multi-class problem to a multi-instance problem by using the one-vs-all approach.

With k_I denoting a kernel on instances $x, x' \in \mathcal{X}$, we define the corresponding multi-instance kernel k_{MI} on bags $X, X' \in 2^{\mathcal{X}}$ as

$$k_{MI}(X, X') := \sum_{x \in X, x' \in X'} k^p(x, x'), \tag{3}$$

where $p \in \mathbb{N}$ is a parameter [10]. As we use the RBF-kernel k_{rbf} as kernel on $\mathcal{X}$ and powers of RBF-kernels are again RBF-kernels, we will not consider p in the following.

We normalize the kernel k_{MI} [10] using

$$k(X, X') := \frac{k_{MI}(X, X')}{\sqrt{k_{MI}(X, X) k_{MI}(X', X')}}. \tag{4}$$

Training an SVM with this kernel produces a bag-level classifier for each class, which we will refer to as MIK. This procedure is very efficient since the resulting kernel matrix is of size number of bags, which is much smaller than a kernel matrix of size number of instances, as is commonly used in the literature [1, 22, 32]. Another advantage over other methods is, that it uses a single convex optimization, whereas other approaches often use iterative algorithms [1] or need to fit complex probabilistic models [31].

While using MIK has many advantages, it produces only an instance-level classifier. We propose to transform a bag-level classifier f_{MI} as given by the SVM and Equation (3) into an instance-level classifier by setting $f_I(x) := f_{MI}(\{x\})$, in other words, by considering each instance as its own bag.

3.3 Segment Features

To describe single segments, we make use of densely computed SIFT [19] and ColorSIFT [26] features, from which we compute bag of visual word histograms. Additionally, we use histograms of oriented gradients [5] on the segments. We use RBF-kernels for all of the features, constructing one MI-kernel per feature. These are then combined using multiple kernel learning to produce a single kernel matrix. This kernel matrix can then be used for all classes, making classification particularly efficient.

3.4 Combining Segments

The framework described above yields an image-level and a segment-level classification. In our setup, each segment might be given multiple labels. To obtain a pixel-level object-class segmentation, we have to combine these. When building the segmentation for a given image, we only consider classes whose presence was predicted on image level. Since we do not make use of the ground truth segmentation during training, we cannot learn an optimal combination as in Li et al. [16] but perform a simple majority vote instead. We merge segments into pixel-level class labels by setting the label y_x of a pixel x according to

$$y_x = \text{argmax}_{y \in Y} \#\{S_i | p \in S_i \wedge y_{S_i} = y\}, \tag{5}$$

where $Y = \{\text{car, bike, person}\}$, S_i enumerates all segments within an image and y_{S_i} is the label of segment S_i. In words: each pixel is assigned the class with the highest number of class segments containing it. This simple heuristic yields good results in practice.

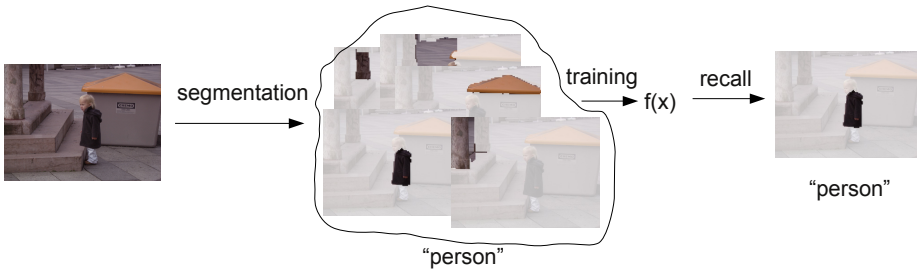

Fig. 1. Overview of our method. See text for details.

4 Experiments

4.1 Instance-Level Predictions using Multi-Instance Kernels

To assess the validity of instance-level predictions using multi-instance kernels, we transform f_I back to an instance-level classifier, using the multi-instance learning assumption (Equation (2)). We refer to these instance-based MIK predictions as MIK-instance. In all experiments, the parameters of the MI-Kernel and SVM are adjusted using MIK and then used with both MIK and MIK-instance. This facilitates very fast parameter scans since MIK is very efficient to compute. Note that we cannot adjust parameters using instance prediction error, as we assume no instance labels to be known.

Table 1. Bag level performance of various MIL algorithms on the standard Musk datasets. All but MIK provide instance-level labeling.

	SVM-SVR	EMDD	mi-SVM	MI-SVM	MICA	MIK	MIK-instance
Musk1	87.9	84.9	87.4	77.9	84.3	88.0	88.0
Musk2	85.4	84.8	83.6	84.3	90.5	89.3	85.2

We compared the performance of MIK, MIK-instance and state-of-the-art MI methods on the Musk benchmark datasets [6], see Table 1. Results were obtained using 10-fold cross-validation. While the computational complexity of MIK-instance is very low compared to the other methods, it achieves competitive results. Using instance-level labels results in a slight loss of accuracy of MIK-instances, compared to MIK. This small degradation of performance is quite surprising, since the model was not trained to provide any instance-level labels.

For multi-class image segmentation, it is beneficial to have a low witness rate, i. e. only a few instances are assumed to be positive in a positive bag. Since an object might not be very prominent in an image, only a fraction of segments might correspond to the object. Table 2 compares the witness rates of MIK-instance,

miSVM [1] and SVR-SVM [17] on the Musk datasets. MIK-instance is able to achieve similar accuracy with much less witnesses than the other methods. Note that Musk1 consists of very small bags while Musk2 contains significantly larger bags, more similar to the image/segment setup.

Table 2. MIL algorithms and the empirical witness rates of the classifiers.

	Musk1		Musk2	
	accuracy	witness-rate	accuracy	witness-rate
mi-SVM	87.4	100%	83.6	83.9%
SVM-SVR	87.9	100%	85.4	89.5%
MIK-instance	88.0	99%	85.2	62.3%

4.2 Partially Supervised Image Segmentation on Graz 02

We evaluate the performance of the proposed algorithm for object-class segmentation on the challenging Graz-02 dataset. This dataset contains 1096 images of three object classes, bike, car and person. Each image may contain multiple instances of the same class.

We adjusted parameters on a hold-out validation set using bag-level information and used the training and test sets as given by the dataset. It is straight-forward to extend the binary MIK method to the multi-class setting using a one-vs-all strategy. We train one MKL-SVM per class using MIK and predict class labels on segment level using MIK-instance. If at least one SVM classifies a segment as positive, it is associated with the most confident class. Otherwise, it is assigned "background" or no class. This yields a classification of each segment into one of four classes: car, bike, person, or background. We merge segments into pixel-level class labels as described in Section 3.4.

Table 3. Pixel-level accuracy on the Graz-02 dataset.

	car	bike	person
Segment based MIK-instance (proposed method)	0.30	0.45	0.43
Best strongly supervised approaches [9, 25]	0.72	0.72	0.66

Per-class pixel accuracies are reported in Table 3; some qualitative results are shown in Figure 2. The overall accuracy on images labels, which is the task that was actually trained, is 90%. The performance of our multiple-instance based approach is far from current methods that use pixel-level annotations, whose pixel-level accuracy is around 70% [9, 25] on pixel-level. This is no surprise as our method has no access to the pixel labels. Rather, it is noteworthy that learning segmentation is possible without pixel labels at all.

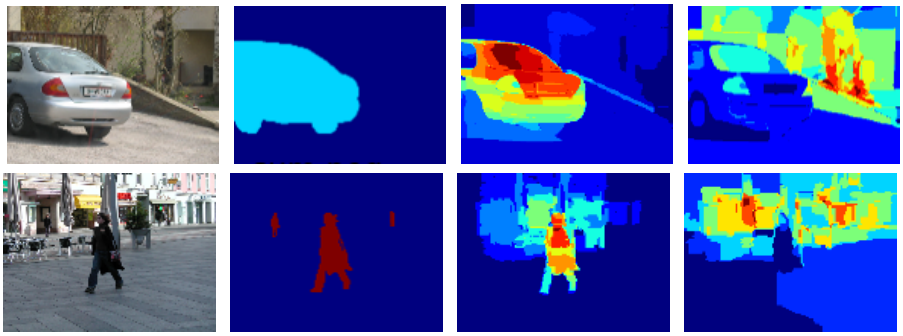

Fig. 2. Qualitative results on on the Graz-02 dataset. Top: Results on category "car". Bottom: Results on category "person". From left to right: original image, ground truth segmentation, segment votes for correct class, segment votes against correct class (red many, blue few votes).

5 Conclusions

We proposed an algorithm for object-class segmentation using only weak supervision based on multiple-instance learning. In our approach, each image is represented as a bag of object-like proposal segments.

We described a way to extend bag-level predictions made by the multi-instance kernel method to instance level while remaining competitive with the state-of-the-art in bag label prediction.

Finally, we evaluated the proposed object-class segmentation method on the challenging Graz02 dataset. While not reaching the performance of methods requiring strong supervision, our result can serve as a baseline for further research into weakly supervised object-class segmentation.

In future work, we plan to scale our approach to much larger image datasets. As much more images with weak annotations are available than with pixel-level segmentation, we hope that we can improve upon the state-of-the-art in object-class segmentation by making use of larger bodies of training images.

References

[1] Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. pp. 577–584 (2003)

[2] Carreira, J., Sminchisescu, C.: Constrained parametric min-cuts for automatic object segmentation. In: Conference on Computer Vision and Pattern Recognition. pp. 3241–3248 (2010)

[3] Catanzaro, B., Su, B., Sundaram, N., Lee, Y., Murphy, M., Keutzer, K.: Efficient, high-quality image contour detection. In: International Conference on Computer Vision. pp. 2381–2388 (2009)

[4] Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1931–1947 (2006)
[5] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Converence on Computer Vision and Pattern Recognition. vol. 1, pp. 886–893 (2005)
[6] Dietterich, T., Lathrop, R., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1-2), 31–71 (1997)
[7] Endres, I., Hoiem, D.: Category independent object proposals. pp. 575–588. Springer (2010)
[8] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenger. International Journal of Computer Vision 88(2), 303–338 (Jun 2010)
[9] Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: International Conference on Computer Vision. pp. 670–677 (2009)
[10] Gärtner, T., Flach, P., Kowalczyk, A., Smola, A.: Multi-instance kernels. In: International Conference on Machine Learning. pp. 179–186 (2002)
[11] Gonfaus, J., Boix, X., van de Weijer, J., Bagdanov, A., Serrat, J., Gonzalez, J.: Harmony potentials for joint classification and segmentation. In: Conference on Computer Vision and Pattern Recognition (2010)
[12] Hanbury, A.: How do superpixels affect image segmentation? Progress in Pattern Recognition, Image Analysis and Applications pp. 178–186 (2008)
[13] Jiang, J., Tu, Z.: Efficient scale space auto-context for image segmentation and labeling. In: Conference on Computer Vision and Pattern Recognition. pp. 1810–1817 (2009)
[14] Ladicky, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical CRFs for object class image segmentation. In: International Conference on Computer Vision. pp. 739–746 (2009)
[15] Leistner, C., Saffari, A., Bischof, H.: MIForests: Multiple-instance learning with randomized trees. European Conference on Computer Vision pp. 29–42 (2010)
[16] Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: Conference on Computer Vision and Pattern Recognition. pp. 1712–1719 (2010)
[17] Li, F., Sminchisescu, C.: Convex multiple-instance learning by estimating likelihood ratio. In: Advances in Neural Information Processing Systems. pp. 1360–1368 (2010)
[18] Li, Y., Kwok, J., Tsang, I., Zhou, Z.: A convex method for locating regions of interest with multi-instance learning. Machine Learning and Knowledge Discovery in Databases pp. 15–30 (2009)
[19] Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
[20] Maire, M., Arbeláez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008)

[21] Mangasarian, O., Wild, E.: Multiple instance classification via successive linear programming. Journal of Optimization Theory and Applications 137(3), 555–568 (2008)
[22] Nguyen, N.: A New SVM Approach to Multi-instance Multi-label Learning. In: International Conference on Data Mining. pp. 384–392 (2010)
[23] Russell, B., Torralba, A., Murphy, K., Freeman, W.: LabelMe: A database and web-based tool for image annotation. International Journal of Computer Vision 77(1), 157–173 (2008)
[24] Schroff, F., Criminisi, A., Zisserman, A.: Object class segmentation using random forests. In: British Machine Vision Conference (2008)
[25] Schulz, H., Behnke, S.: Object-class segmentation using deep convolutional neural networks. In: Hammer, B., Villmann, T. (eds.) Proceedings of the DAGM Workshop on New Challenges in Neural Computation 2011. Machine Learning Reports, vol. 5, pp. 58–61 (2011)
[26] Van De Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. Transactions on Pattern Analysis and Machine Intelligence pp. 1582–1596 (2009)
[27] Verbeek, J., Triggs, B.: Region classification with Markov field aspect models. In: Computer Vision and Pattern Recognition. pp. 1–8 (2007)
[28] Vezhnevets, A., Buhmann, J.: Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In: Computer Vision and Pattern Recognition. pp. 3249–3256 (2010)
[29] Vezhnevets, A., Ferrari, V., Buhmann, J.: Weakly supervised semantic segmentation with a multi-image model. In: International Conference on Computer Vision (2011)
[30] Wang, S., Siskind, J.: Image segmentation with Ratio Cut. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 675–690 (2003)
[31] Zha, Z., Hua, X., Mei, T., Wang, J., Qi, G., Wang, Z.: Joint multi-label multi-instance learning for image classification. In: Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008)
[32] Zhang, M., Zhou, Z.: M3miml: A maximum margin method for multi-instance multi-label learning. In: International Conference on Data Mining. pp. 688–697 (2008)
[33] Zhang, Q., Goldman, S.: Em-dd: An improved multiple-instance learning technique. Advances in Neural Information Processing Systems 2, 1073–1080 (2002)
[34] Zhou, Z., Sun, Y., Li, Y.: Multi-instance learning by treating instances as non-iid samples. In: International Conference on Machine Learning. pp. 1249–1256 (2009)
[35] Zhou, Z., Zhang, M.: Multi-instance multi-label learning with application to scene classification. In: Advances in Neural Information Processing Systems. pp. 1609–1616 (2006)