Weakly Supervised Top-down Image Segmentation

Manuela Vasconcelos° Gustavo Carneiro⋆ Nuno Vasconcelos°
Statistical Visual Computing Lab° Integrated Data Systems⋆
University of California, San Diego Siemens Corporate Research

Abstract

There has recently been significant interest in top-down image segmentation methods, which incorporate the recognition of visual concepts as an intermediate step of segmentation. This work addresses the problem of top-down segmentation with weak supervision. Under this framework, learning does not require a set of manually segmented examples for each concept of interest, but simply a weakly labeled training set. This is a training set where images are annotated with a set of keywords describing their contents, but visual concepts are not explicitly segmented and no correspondence is specified between keywords and image regions. We demonstrate, both analytically and empirically, that weakly supervised segmentation is feasible when certain conditions hold. We also propose a simple weakly supervised segmentation algorithm that extends state-of-the-art bottom-up segmentation methods in the direction of perceptually meaningful segmentation[1].

1 Introduction

Image segmentation has been a subject of research in computer vision for many decades. Traditionally, it has been formulated as a problem of *bottom-up* processing, i.e. whose solution does not require (or assume) high-level knowledge about the scene under analysis. Instead, classical segmentation algorithms identify *image segments* or *regions* solely on the basis of low-level visual attributes. Examples include the definition of segments as regions enclosed by closed contours, or where the statistics of certain features (color, texture, etc.) are homogeneous, or both. While the low-level emphasis of these algorithms has some advantages, e.g. computational efficiency, the resulting segmentations usually have little resemblance to those produced by humans.

One of the main sources of difficulty seems to be that humans rely on different definitions of homogeneity in different image areas, depending on the scene content and higher levels goals that drive segmentation. This is exemplified by Figure 1, where we compare a human segmentation of an outdoor scene with the segmentation produced by a state-of-the-art bottom-up algorithm (N-cuts [11]). Note how the human segmentation consists of regions that are homogeneous with respect to different features, e.g. a *water* region of uniform *color*, a *sky* region of uniform *smoothness* (note that sky color varies across the image), a region of tree leaves of uniform *texture*, and so forth. On the other hand, the automatic segmentation algorithm applies a universal definition of uniformity (subjacent to the cost under which it is optimal) throughout the image, and cannot cope with the diversity of visual concepts that compose it.

To overcome this limitation, there has recently been interest in the direction of *top-down* segmentation. For example, because humans seem to be biased in favor of segmentations with some characteristics, e.g. a certain range of region sizes or a certain distribution of contrast along segment boundaries, it seems natural to model these biases. One possibility is to assemble a database of human-segmented images and use the examples in this database to learn a distribution on the space of image segmentations, namely the distribution of *perceptually plausible segmentations*. This type of effort is currently popular, and a number of techniques have been proposed to learn such distributions from hand-segmented imagery (e.g., see [10]). While it is likely that the resulting distributions will find wide application as priors for bottom-up image segmentation, they tend to be *universal statistical laws* that provide little help in terms of identifying the statistical homogeneities that are most relevant for the segmentation of a *particular scene*.

In general, this identification *cannot be successful in the absence of truly top-down processing*, i.e. processing that receives guidance and feedback from the higher levels of perception. For example, the mountain region of Figure 1 consists of 1) a brownish, approximately textureless, rocky formation on the left side of the image, 2) a vertically striped combination of rocks and vegetation in the center, and 3) a greenish randomly textured area of vegetation on the right. In the absence of explicit knowledge of 1) a (high-level) *mountain concept*, and 2) the fact that mountains exhibit all these different types of statistical homogeneity, it is virtually impossible to avoid oversegmenting the mountain into the sub-areas where each type of homogeneity dominates. This is exactly what N-cuts does, and also happens for the *tree* and *sky* concepts.

Top-down segmentation overcomes these difficulties by tying the *segmentation* and *recognition* problems, i.e. by *making (high-level) recognition an intermediate step of seg-*

[1] This work was performed while Gustavo Carneiro was with the University of British Columbia.

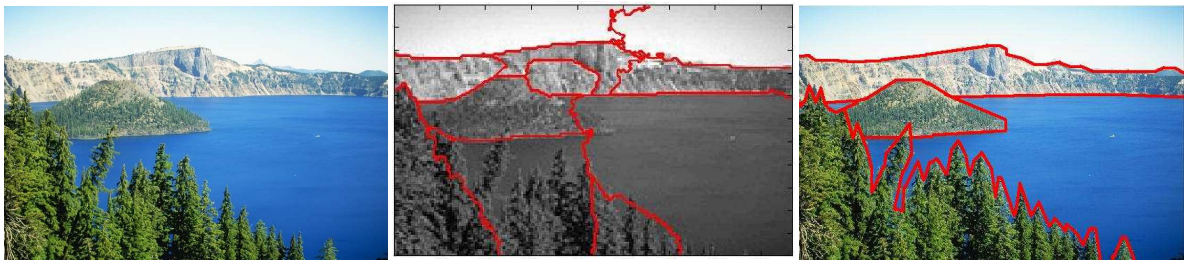

Figure 1. Left: an image, center: segmentation by the N-cuts algorithm, right: segmentation by a human.

mentation. It has roots on the observation that, given a large vocabulary of visual concepts, and a library of statistical appearance models for these concepts, segmentation reduces to the simple assignment of each image pixel to the model that best explains it. The main difficulty is that, as is common in segmentation problems, this introduces a "chicken-and-egg" type of roadblock: in the absence of a set of segmented images it is not feasible to learn concept models, and in the absence of concept models it is not possible to perform the segmentation. One possibility to overcome this problem is to rely on a set of manually segmented images to bootstrap the process [1, 6, 7]. This approach, which we refer to as *strongly supervised* segmentation, is quite non-scalable in the size of the target concept vocabulary, and therefore unlikely to be a suitable replacement for existing general-purpose bottom-up algorithms.

In this work we study the alternative problem of top-down segmentation with *weak supervision*. The basic goal is to relax the supervision requirements from *image segmentation* to *image annotation*. That is, to require a training set where each image is complemented with a *caption* that describes the visual concepts depicted in it, rather than a training set of *manually segmented* examples for each concept in the vocabulary. The motivation is that *annotating images is significantly easier than segmenting them* (as shown by the existence on the web of a number of databases of the former type - *flicker*, *ESP*, *corbis*, etc. - and virtually none of the latter[2]) and weakly supervised segmentation is therefore significantly more scalable than its strongly supervised counterpart. The main contribution of this work is the *demonstration*, both analytical and experimentally, that *weakly supervised segmentation is possible*, when certain conditions hold. We also propose a simple *weakly supervised segmentation algorithm* that extends state-of-the-art bottom-up segmentation methods in the direction of perceptual segmentations.

2 Weakly Supervised Top-down Segmentation

The inspiration for weakly supervised segmentation comes from three areas of vision and learning: multiple instance learning [9], semantic image labeling and retrieval [13], and recognition from cluttered scenes [5]. In all these areas it has been observed that *the empirical distribution of a collection of feature vectors extracted from images containing a common visual concept tends to approximate the distribution of this concept*. This appears to happen even when the images are from *scenes that include various other concepts*, as long as no other concept is common to the entire image set. Although the convergence to the concept distribution has only been demonstrated experimentally, the experimental evidence is substantial. For example, [9] has shown that the peak of the empirical distribution tends to occur in the region of support of the concept, [13] has shown that the empirical distribution performs well when used as the concept's class conditional distribution for image classification, and [5] has shown that clustering the collection of feature vectors produces a codebook of concept parts (e.g. eyes, mouth, or nose, for face concept).

Under the assumption that the convergence indeed holds, the design of a weakly supervised segmentation algorithm is relatively straightforward. It consists of two stages: *training* and *segmentation*. Training can be implemented as follows:

1. define a concept vocabulary $\mathcal{L} = \{c_1, \ldots, c_C\}$.
2. for each concept c assemble a collection of images $\mathcal{D}^c = \{I_1^c, \ldots, I_N^c\}$ of scenes that contain the concept (and possibly other concepts as well).
3. for each c, extract a set of feature vectors $\mathcal{X}^c = \{\mathbf{x}_1^c, \ldots, \mathbf{x}_F^c\}$ from $\mathcal{D}^c$ and obtain an estimate of the concept distribution $\hat{P}_c(\mathbf{x})$ by applying a standard density estimation procedure (e.g. a kernel density estimator [12], a mixture model [2], etc.) to $\mathcal{X}^c$.

Note that the images are not segmented and a subset of the features in $\mathcal{X}^c$ can be unrelated to concept c. Given the learned sequence of concept distributions $\hat{P}_c(\mathbf{x}), c \in \{1, \ldots, C\}$, and a new image I, segmentation consists of:

1. determine the set of concepts $\mathcal{L}' \subset \mathcal{L}$ present in the image. This can be user-specified, or done automatically as discussed below.
2. extract a feature vector $\mathbf{x}$ at each location (i, j) of I and assign it to one of the concepts using a standard

[2] Ignoring, of course, those produced by the vision community.

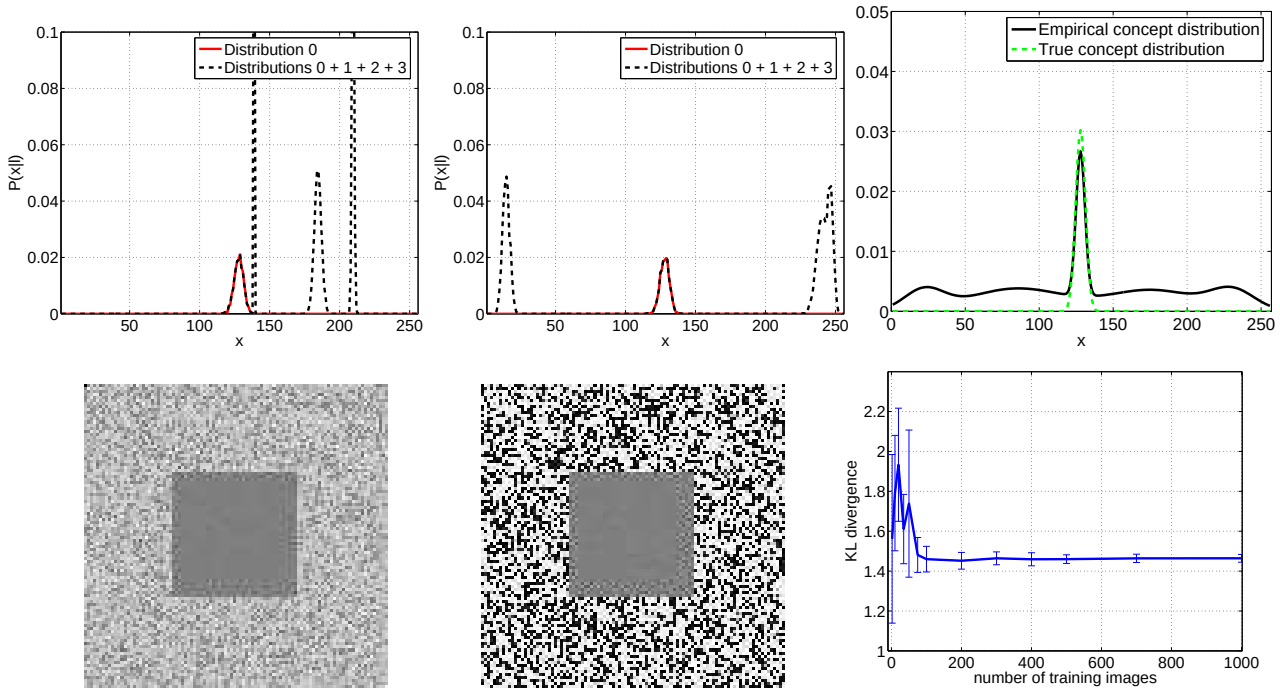

Figure 2. Convergence to the density of concept c, shown in red in the top left images and in green on the top right.

minimum probability of error (MPE) decision rule

$$l(i,j) = \arg\max_{c \in \mathcal{L}'} \hat{P}_c(\mathbf{x})\pi_c \quad (1)$$

where π_k is a set of prior concept probabilities, which in the absence of reasons to favor some concepts over others can be set to a a uniform distribution $\pi_k = 1/C$.

The automatic determination of the concepts present in the image can be achieved with a procedure similar to (1) but applied to all feature vectors $\mathbf{x}_i$ extracted from the image. Assuming that the vectors are sampled independently, concepts can be ordered by posterior probability, by computing

$$\lambda_c = \prod_i \hat{P}_c(\mathbf{x}_i)\pi_c \quad (2)$$

for all $c \in \mathcal{L}$ and ordering the λ_c by decreasing magnitude.

3 Motivation

In this section we motivate weakly supervised segmentation by analyzing a simple synthetic example. In this example, concepts are squares textured with independent Gaussian noise, $P_c(\mathbf{x}) = \mathcal{G}(x, \mu_c, \sigma_c)$, where $\mu_c = 127$ and $\sigma_c = 10$, and $\mathcal{G}(.)$ is used throughout the text to represent the Gaussian probability density function

$$\mathcal{G}(\mathbf{x}, \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}. \quad (3)$$

In each image, a concept is presented against a background of pixels randomly drawn from a mixture of three Gaussians

$$P_B(x) = \sum_{i=1}^{3} \gamma_i \mathcal{G}(x, \mu_i, \sigma_i) \quad (4)$$

whose means μ_i and variances σ_i are sampled independently from uniform distributions of range [0,255] and [0.1,25], respectively.

Figure 2 illustrates the convergence of the empirical estimate $\hat{P}_c(\mathbf{x})$ to the density of the concept. The top row shows the mixture distributions associated with two images in the concept's training set (in each case the concept density is shown in red). The images themselves are shown immediately below, in the second row of the figure. The top right plot shows the empirical distribution $\hat{P}_c(x)$ estimated from the entire training set, and a scaled replica of the true concept distribution $P_c(x)$. Note that the empirical estimate converges to a mixture of the true concept density and an almost uniform component. The bottom right plot shows the Kullback-Leibler (KL) divergence between the true concept distribution and the estimate, as a function of the number of training images used to learn the concept. Note that the convergence to the asymptotic distance is quite fast.

4 Theoretical Analysis

In this section we study the convergence of the empirical estimate to the concept distribution.

4.1 Definitions

Consider a feature space $\mathcal{X} \subset \mathbb{R}^d$. Images are represented as collections of feature vectors, i.e. $I_i = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_n^i\}$ for the i^{th} image, drawn from a random variable $\mathbf{X}$ defined on $\mathcal{X}$. Visual concepts define probability distributions on $\mathcal{X}$. For example, if the features are the average values of the three color channels over a localized neighborhood then the "face" concept will assign a large probability to the region of skin tones, and small probability mass to other regions. Concepts are drawn from a random variable Y that assigns a probability distribution to a concept vocabulary $\mathcal{L} = \{c_1, \ldots, c_C\}$. The goal is to learn the probability distribution associated with a certain concept c, $P_{\mathbf{X}|Y}(\mathbf{x}|c)$, which we will refer to as $P_c(\mathbf{x})$ for simplicity.

Learning is weakly supervised because the *learner is not provided with cleanly cropped image regions of the concept*. Although a collection of training images containing regions that depict c is available, *regions are not identified*. Hence, in a training image set $\mathcal{D} = \{I_1, \ldots, I_N\}$, each image I_i is a sample of a feature distribution

$$P_{\mathbf{X}}(\mathbf{x}) = \pi P_c(\mathbf{x}) + (1 - \pi) P_B(\mathbf{x}) \quad (5)$$

where π is the percent of the image area which is covered by c and $P_B(\mathbf{x})$ a background distribution that accounts for everything else. Since any probability distribution can be approximated arbitrarily well by a (potentially infinite) mixture of Gaussians, we assume that the background density is of this form. We further assume that it is a mixture of $K-1$ equal probability $(1/K)$ components[3] and that $\pi = 1/K$.

Definition 1 *Image I_i in the training set $\mathcal{D}$ is a sample from a random variable of probability density function*

$$P_{\mathbf{X}}^i(\mathbf{x}) = \frac{1}{K}\left(P_c(\mathbf{x}) + \sum_{j=1}^{K-1} \mathcal{G}(\mathbf{x}, \mu_j^i, \mathbf{\Sigma}_j^i)\right). \quad (6)$$

The training set $\mathcal{D}$ is denoted as *diverse* if the background distributions are themselves a diverse set. This can be formalized by making the Gaussian parameters $\mu_j^i, \mathbf{\Sigma}_j^i$ samples from some random variable.

Definition 2 *$\mathcal{D}$ is a diverse training set if μ_j^i, and $\mathbf{\Sigma}_j^i$ are independent samples from two independent random variables with probability density functions*

$$P_\mu(\mu) = \mathcal{G}(\mu, \mu_0, \mathbf{\Sigma}_0)$$

and $P_{\mathbf{\Sigma}}(\mathbf{\Sigma})$, such that $E_{\mathbf{\Sigma}}[\Sigma] = \mathbf{S}$, and (for $\epsilon \geq 0$)

$$|E_{\mathbf{\Sigma}}[\mathcal{G}(\mathbf{x}, \mu_0, \mathbf{\Sigma} + \mathbf{\Sigma}_0)] - \mathcal{G}(\mathbf{x}, \mu_0, \mathbf{S} + \mathbf{\Sigma}_0)| \leq \epsilon. \quad (7)$$

[3]This is mostly to simplify notation, all results that follow could be extended to the case where each component has an individual weight.

The assumption of a Gaussian distribution for μ is not crucial for the discussion that follows. In particular, all results could be generalized to the case of a Gaussian mixture and, therefore, any $P_\mu(\mu)$ of practical interest. The Gaussian assumption is adopted because it makes the notation much simpler. We refer to $\mathbf{\Sigma}_0$ as the *diversity parameter* of $\mathcal{D}$.

(7) is a technical condition, required by the proofs of the subsequent sections. We note, however, that for most practical purposes it is a very mild restriction on $P_{\mathbf{\Sigma}}(\mathbf{\Sigma})$. If, for example, the Gaussian components of (6) are produced by a kernel density estimator, it is common practice for all covariances to be identical, i.e. $\mathbf{\Sigma}_j^i = \mathbf{S}$. In this case $P_{\mathbf{\Sigma}}(\mathbf{\Sigma})$ is a delta function centered at $\mathbf{S}$, and (7) holds with $\epsilon = 0$. In general, the condition will hold if $\mathbf{\Sigma} + \mathbf{\Sigma}_0 \approx \mathbf{S} + \mathbf{\Sigma}_0$ for all $\mathbf{\Sigma}$ such that $P_{\mathbf{\Sigma}}(\mathbf{\Sigma}) > 0$, i.e. if the spread of $P_{\mathbf{\Sigma}}(\mathbf{\Sigma})$ around the mean value $\mathbf{S}$ is small compared to $\mathbf{S} + \mathbf{\Sigma}_0$. This is true whenever $\mathbf{\Sigma}_0$ is large, which (as we will see below) is a necessary condition for the concept distribution to be learnable. Note that, as long as the support of $P_{\mathbf{\Sigma}}(\mathbf{\Sigma})$ is bounded, it is possible, by making $\mathbf{\Sigma}_0$ arbitrarily large, to make (7) hold with arbitrarily small ϵ.

4.2 Concept Learnability

The following theorem shows that the distribution of a diverse set of images of concept c converges to a mixture of the concept distribution and a background component of spread determined by the diversity parameter $\mathbf{\Sigma}_0$.

Theorem 1 *If $\mathcal{D}$ is a diverse training set, according to Definitions 1 and 2, then*

$$P_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} P_{\mathbf{X}}^i(\mathbf{x}) \quad (8)$$

satisfies

$$\lim_{N\to\infty} |P_N(\mathbf{x}) - f(\mathbf{x})| \leq \delta \quad (9)$$

with

$$f(\mathbf{x}) = \frac{1}{K} P_{\mathbf{C}}(\mathbf{x}) + \left(1 - \frac{1}{K}\right)\mathcal{G}(\mathbf{x}, \mu_0, \mathbf{S} + \mathbf{\Sigma}_0) \quad (10)$$

and

$$\delta = (1 - 1/K)\,\epsilon. \quad (11)$$

Proof: Available from the authors.

Note that, as long as the diversity of $\mathcal{D}$ is large (large $\mathbf{\Sigma}_0$), the background component will have small amplitude and the limit distribution of $P_N(\mathbf{x})$ is dominated by the concept distribution.

5 Connections to bottom-up segmentation

It is well known that, in the absence of a prior that favors spatially smooth segmentations, these tend to be quite noisy. While the theoretical analysis above is valid for any concept

probability model, the complexity of learning, in a weakly supervised manner, both the observation model $P_{\mathbf{X}|Y}(\mathbf{x}|c)$ and the prior $P_Y(c)$, appears non-trivial when the latter is smoothness enforcing, e.g. a Markov random field or equivalent. One possibility, that we explore in this work, is to rely on weakly supervised learning to estimate the observation component $P_{\mathbf{X}|Y}(\mathbf{x}|c)$ and a standard bottom-up segmentation algorithm to learn the prior. This enables an interpretation of weakly supervised segmentation as a direct extension of various existing bottom-up segmentation algorithms which support a supervised mode, where the observation component is known [8, 14]. The extension consists of learning the observation component in a weakly supervised manner, and then learning the prior in the standard bottom-up manner. We have implemented our weakly supervised learning algorithm using this strategy, and tested two state-of-the-art bottom-up methods, that of wavelet-based priors [8], and the min-cut algorithm [14].

6 Experimental Results

In this section, we report on weakly-supervised top-down segmentation experiments on the Corel data set of [4]. This is a dataset of $5,000$ images from 50 Corel Stock Photo CDs, divided into a training set of $4,500$ images, and a test set of 500 images. Each image has been manually labeled with a caption of $1-5$ keywords, and there are a total of 371 keywords (concepts) in the data set. All images consist of three color channels (YBR color space) which were decomposed into a set of overlapping $8 \times 8 \times 3$ windows. The discrete cosine transform (DCT) was applied to each color channel of each window, and each image represented as a bag of independent feature vectors containing the first 21 DCT coefficients of each color channel.

The observation model for each concept was a mixture of Gaussians learned from all the images labeled with the associated keyword, using the method of [13], while the prior was learned as discussed above. In Figures 3-5, 'LKL SEG' indicates an independent prior (no spatial smoothness), 'WAV SEG' a wavelet prior, and 'MINCUT SEG' the mincut prior.

The first experiment was designed to evaluate the feasibility of *weakly unsupervised concept detection*. For this, we trained the classifier of (1) in a one-vs-all manner, where one class contained the training images labeled with the concept of interest, and the other the remaining training images. Figures 3 and 4 show the detection results for water, and sky, respectively. Note that although the segmentations produced by the independent prior tends to be noisy, the introduction of smoothness enforcing priors makes them reasonably precise. For brevity, the segmentations are only shown for the mincut prior, but the two priors produced similar results. Given the large intra-class variability of these classes and the relatively small number of training examples (883 for sky and 1004 for water) these results can be considered very promising.

Figure 5 presents results for a multi-class segmentation problem based on the the automatic labeling scheme of [3]. This method extracts the top 5 labels for each test image, from which we manually selected 2-4 according to five constraints: 1) uniqueness among the 5 concepts (e.g., if an image is labeled with both 'tiger' and 'cat', we rejected one of the two); 2) ability to localize the concept in the image (e.g., we rejected abstract concepts like 'city' or 'outdoor'); 3) variability of the concept training set (e.g., we rejected the concept 'horses' because, on Corel, horses always appear with 'grass' and it is impossible to distinguish the two concepts); 4) training set size (we rejected concepts with unreasonably small training sets); and 5) actual presence of the concept in the test image (to avoid labeling errors).

These results show that weakly-supervised top-down segmentation can produce quite stable segmentation results. Once again, we note that the number examples available for each concept is small (in the figure, 59 for the concept with fewest examples (petals) and 267 for that with the most (snow)). Also, on Corel, most images of various classes are presented against similar backgrounds (e.g., the horse discussion above). Interestingly, while this is a property that simplifies problems such as image retrieval or semantic labeling, it significantly increases the difficulty of weakly supervised segmentation, by reducing the covariance of background distributions. In this sense, the Corel set is close to a worst-case scenario. Overall, while the segmentations are clearly not perfect, we believe that these results indicate great promise for weakly supervised top-down segmentation, when combined with more sophisticated probabilistic representations than the simple mixture of Gaussians adopted here.

Acknowledgements This work was partially supported by NSF Career award IIS-0448609, and NSF IIS-0534985. Gustavo Carneiro also wishes to acknowledge funding received from NSERC (Canada).

References

[1] E. Borenstein and S. Ullman. Class specific top down-segmentation. *ECCV*. 2002.

[2] R. Duda, P. Hart, and D. Stork. Pattern Classification (2nd Ed.). *Wiley, New York*. 2001.

[3] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. *IEEE CVPR*. 2005.

[4] P. Duygulu, K. Barnard, and D. Forsyth, and N. Freitas. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. *ECCV*. 2002.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[6] M. Kumar, P. Torr, and A. Zisserman. Obj cut. *IEEE CVPR*. 2005.

[7] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. *ECCV'04 Workshop on Statistical Learning in Computer Vision*. 2004.

[8] M. Figueiredo. Bayesian image segmentation using wavelet-based priors. *IEEE CVPR*. 2005.

[9] O. Maron and A. Ratan. Multiple-Instance Learning for Natural Scene Classification. ICML. 1998.

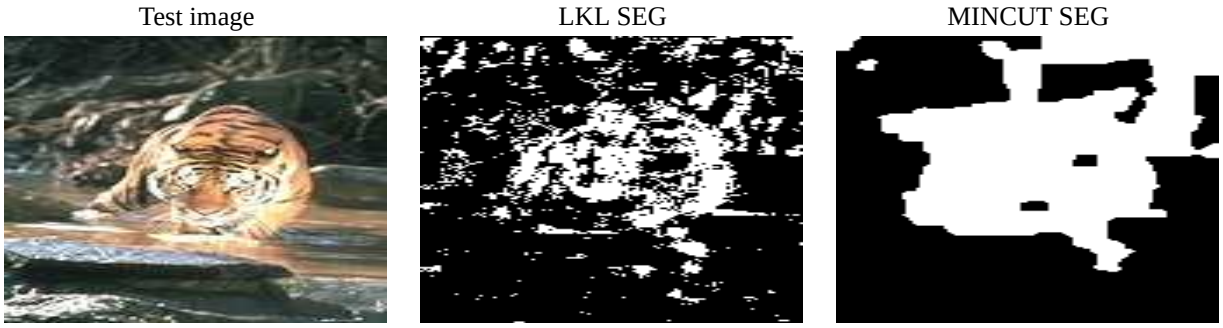

Figure 3. Water detection results. Dark area represents regions of the image containing the concept water.

Figure 4. Sky detection results. Dark area represents regions of the image containing the concept sky.

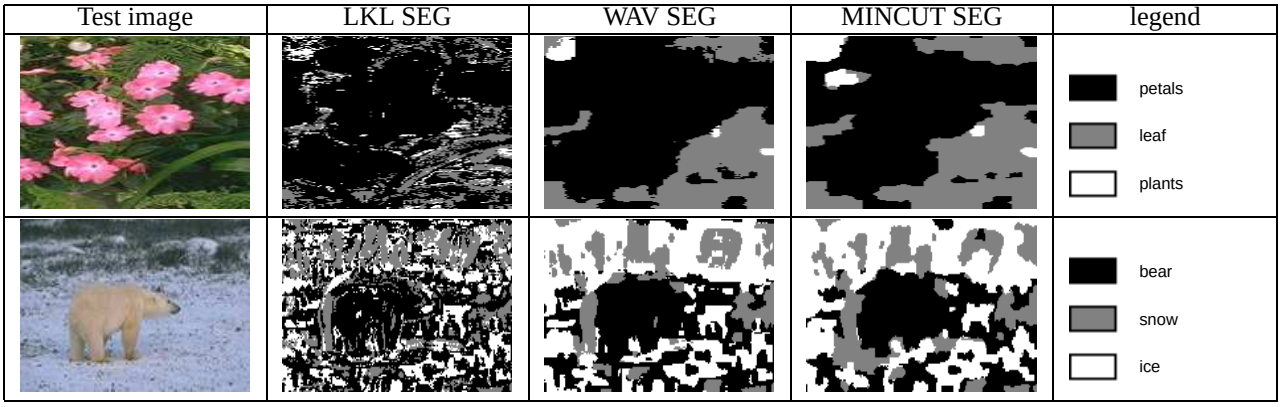

Figure 5. Weakly supervised multi-class segmentation.

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*. 2001.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE PAMI*, 22(8):888–905, 2000.

[12] J. Simonoff Smoothing methods in statistics *Springer-Verlag New York*. 1996.

[13] N. Vasconcelos. Exploiting group structure to improve retrieval accuracy and speed in image databases. *ICIP*. 2002.

[14] R. Zabih and V. Kolmogorov. Spatially coherent clustering using graph cuts. *IEEE CVPR*. 2004.