Using Global Bag of Features Models in Random Fields for Joint Categorization and Segmentation of Objects

Dheeraj Singaraju René Vidal
Center for Imaging Science, Johns Hopkins University, Baltimore MD 21218, USA
{dheeraj,rvidal}@cis.jhu.edu http://www.vision.jhu.edu

Abstract

We propose to bridge the gap between Random Field (RF) formulations for joint categorization and segmentation (JCaS), which model local interactions among pixels and superpixels, and Bag of Features categorization algorithms, which use global descriptors. For this purpose, we introduce new higher order potentials that encode the classification cost of a histogram extracted from all the objects in an image that belong to a particular category, where the cost is given as the output of a classifier when applied to the histogram. The potentials efficiently encode the classification costs of several histograms resulting from the different possible segmentations of an image. They can be integrated with existing potentials, hence providing a natural unification of global and local interactions. The potentials' parameters can be treated as parameters of the RF and hence be jointly learnt along with the other parameters of the RF. Experiments show that our framework can be used to improve the performance of existing JCaS algorithms.

1. Introduction

JCaS corresponds to the problem of assigning an object category label to each pixel in a given image. Several solutions to this problem have been proposed using RF formulations [22, 16, 21, 20, 10, 3, 15, 5, 4, 18, 8, 2, 9]. These algorithms define a RF whose sites represent pixels in the image or superpixels obtained by oversegmentation of the image. In general, a unary potential is defined for each site, which models the cost of that site being assigned a particular category label. Algorithms also define pairwise potentials between neighboring sites to enforce spatial smoothness of the labels or to encode contextual information. In some cases, pairwise potentials may not be sufficient to describe the statistics of an object. To resolve this issue, recent algorithms have used higher order potentials that model the interactions among several sites [8, 17, 9].

While these methods have been fairly successful in practice, they have a few limitations. The unary potentials in most cases are obtained by integrating information across pre-defined local neighborhoods, e.g., [18, 8, 2]. Algorithms do not integrate information across arbitrarily large neighborhoods since they might cross the true boundaries of an object. Some methods such as [22, 20] do consider long range interactions between interest regions, but then encode them as pairwise potentials. The higher order interactions among different sites in the RF are typically restricted to fairly local neighborhoods such as neighboring pixels or superpixels. [17] is a notable exception which considers long-range higher order interactions among several sites to model co-occurrence statistics of object categories.

In general, most existing JCaS algorithms are primarily bottom-up and use fairly local interactions among neighboring sites to solve the problem. Moreover, most methods do not build a rich object model that considers long-range interactions among several sites to capture higher order statistics of the object. We argue that one can improve performance by analyzing global interactions across larger neighborhoods such as all the regions covered by an object.

Our concern regarding the lack of a global model also leads us to pose the question: *if we were given the true boundaries of the objects in the image, would we follow the same approach discussed above to categorize each segment?* In such a case, it seems more natural to use ideas from the genre of joint localization and categorization algorithms, e.g., [7]. These algorithms aim to locate each object in the image by placing a bounding box around it and also identify its category. One popular approach to this problem is to use the BoF approach. These algorithms extract interest points (visual features) from the region of interest (area enclosed by the bounding box) in the image. The descriptors for the interest points are then quantized using a previously learnt dictionary of *visual codewords*. The histograms of these quantized descriptors are assumed to follow characteristic models for each object category. Hence, the algorithms perform categorization based on the extracted histogram. This strategy can be useful for JCaS for the purpose of globally analyzing each object present in a given segmentation of an image. Unfortunately, it cannot be applied directly to JCaS since the interest regions given by the objects' segmentation are unknown beforehand.

Paper contributions. In this work, we address the aforementioned concerns and present a framework that integrates a global BoF based analysis into RF formulations for JCaS.

1) Our key contribution is a top-down cost function based

on the BoF approach. This cost function depends on the unknown segmentation of the image and is given as the output of a classifier applied to the histogram of all the interest points in the image that are assigned a particular category label. We term this output as the *classification cost* of the histogram. Although the number of possible segmentations of an image is exponentially large, the cost function efficiently encodes for each segmentation, the classification costs of the histograms extracted for each of the categories.

2) We show that the parameters of the classifiers may be treated as parameters of the cost function. Hence, we can learn the top-down parameters (the classifier parameters) and the bottom-up parameters (trade off between unary and pairwise terms) in a joint fashion. To this effect, we propose a max-margin framework for learning all the parameters.

3) The problem of computing the segmentation is a discrete optimization problem, which can be NP-hard in general. We show that our proposed top-down cost function is amenable to efficient discrete optimization schemes, i.e., a local optimum to the proposed optimization problem can be computed efficiently using existing inference algorithms, namely the graph cut based α-expansion method [1].

Related work. Recent work on JCaS has addressed the issue of using top-down object models that generalize to the case of multiple categories [10, 21, 5, 4, 23, 9]. Our framework differs from these works in that the object category information is encoded using the global BoF model. The works most closely related to our work are those of [21] and [9]. [21] also considers a global BoF approach by inspecting histograms of features extracted from the image. However, the framework in [21] considers the histogram extracted from the entire image rather than considering the histograms extracted from each segmented object region. Very recently, [9] proposed the use of object detectors to localize the objects in the estimated segmentation. We will show later that [9] is a particular case of our framework.

2. Review

2.1. RF formulations for JCaS

We define a RF whose sites correspond to the pixels or superpixels of an image I. The set of the sites in the RF is denoted as $\mathcal{V}$. A discrete valued random variable X_i is defined at each site $i \in \mathcal{V}$ and can take any value x_i in the set of possible labels $\mathcal{L} = \{1, \ldots, L\}$. These labels denote the different categories. Any assignment of labels to the random variables is referred to as a *labeling* and is denoted as $\boldsymbol{x} \in \mathcal{L}^{|\mathcal{V}|}$. We denote the restriction of the random variables and labeling to a set of sites $A \subseteq \mathcal{V}$ as $\boldsymbol{X}_A$ and $\boldsymbol{x}_A$, respectively. Note that x_i is the restriction of $\boldsymbol{x}$ to the site i.

The neighborhood structure of the RF is defined using the set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ and an edge that spans two sites i and j is denoted by $\boldsymbol{e}_{ij}$. A clique defines a set of sites $c \subset \mathcal{V}$ whose random variables $\boldsymbol{X}_c$ are conditionally dependent on each other, e.g., the set of pixels in a superpixel. We denote the set of all cliques in the RF as $\mathcal{C}$. One then defines *potential functions* (or potentials) to model the interactions among the sites in the RF as a function of their assigned labels. The following are a few commonly used potentials.

A *unary potential* $\psi_i^U(x_i; I)$ is defined for each site $i \in \mathcal{V}$, such that $\psi_i^U(l; I)$ defines the cost of assigning the label $l \in \mathcal{L}$ to the site i. Algorithms typically extract appearance and/or location descriptors for each site $i \in \mathcal{V}$ in an image. The descriptors extracted from the training images are used to train classifiers for each of the object categories. These classifiers are applied to the descriptor of each site i in the image I to construct $\psi_i^U(x_i; I)$. This is equivalent to using *local BoF models* for constructing the unary potentials.

A *pairwise potential* $\psi_{ij}^P(x_i, x_j; I)$ is defined for every pair of neighboring sites $i, j \in \mathcal{V}$, where $\boldsymbol{e}_{ij} \in \mathcal{E}$, such that $\psi_{ij}^P(l_i, l_j; I)$ defines the cost of assigning labels l_i and l_j to the sites i and j, respectively. These potentials are used to enforce the spatial smoothness of $\boldsymbol{x}$.They are also used to encode contextual information. In general, the potential $\psi_{ij}^P(x_i, x_j; I)$ for an edge $\boldsymbol{e}_{ij}$ is computed as a function of the appearance and/or location descriptors of sites i and j.

Notice that these potentials model interactions of at most two sites. One can also define a *higher order potential* $\psi_c^C(\boldsymbol{x}_c; I)$ on the clique $c \in \mathcal{C}$, such that $\psi_c^C(\boldsymbol{l}_c; I)$ defines the cost of assigning the labels $\boldsymbol{l}_c \in \mathcal{L}^{|c|}$ to the clique c. For example, the potential $\psi_c^C(\boldsymbol{x}_c; I)$ can be defined over the the clique of pixels that belong to a superpixel obtained from oversegmentation of an image. $\psi_c^C(\boldsymbol{x}_c; I)$ can then be used to enforce the consistency of the labels within a superpixel, while being tolerant towards some pixels taking labels that are different from the label of majority of the pixels in the superpixel [6, 8, 9].

As described earlier in §1, these potentials typically define bottom-up models for fairly local interactions. They can be used to construct a bottom-up energy function, $E_{bu}(\boldsymbol{x}; I)$, for solving the JCaS problem for an image I as

$$\begin{aligned} E_{bu}(\boldsymbol{x}; I) = \lambda_U \sum_{i \in \mathcal{V}} \psi_i^U(x_i; I) + \lambda_P \sum_{\boldsymbol{e}_{ij} \in \mathcal{E}} \psi_{ij}^P(x_i, x_j; I) \\ + \lambda_C \sum_{c \in \mathcal{C}} \psi_c^C(\boldsymbol{x}_c; I) = \boldsymbol{w}_{bu}^\top \Psi_{bu}(\boldsymbol{x}; I), \end{aligned} \tag{1}$$

where $\boldsymbol{w}_{bu}^\top = \begin{bmatrix} \lambda_U & \lambda_P & \lambda_C \end{bmatrix} \in \mathbb{R}^3$ denotes the relative contributions of the different genres of potentials and

$$\Psi_{bu}(\boldsymbol{x}; I) = \begin{bmatrix} \sum_{i \in \mathcal{V}} \psi_i^U(x_i; I) \\ \sum_{\boldsymbol{e}_{ij} \in \mathcal{E}} \psi_{ij}^P(x_i, x_j; I) \\ \sum_{c \in \mathcal{C}} \psi_c^C(\boldsymbol{x}_c; I) \end{bmatrix} \in \mathbb{R}^3. \tag{2}$$

2.2. BoF algorithms for categorization

BoF algorithms, e.g., [7], assume that the category of an object in an image can be inferred by analyzing the relative frequencies of certain visual keywords extracted from the image. Given a training set of N images $\{I^i\}_{i=1}^N$, BoF algo-

rithms detect interest points in each image such as corners, junctions or SIFT interest points [11]. A feature descriptor is associated with each of these interest points. These descriptors are quantized using clustering schemes such as K-means, to construct a dictionary of visual codewords. We denote the size of the constructed dictionary, i.e., the number of quantized codewords as K. BoF algorithms assume that the interest points in the image belonging to a particular object category follow a characteristic distribution. This distribution is represented as a histogram of the quantized descriptors of the interest points. We denote the histogram extracted from the s^{th} training image as $\boldsymbol{h}_t^s \in \mathbb{R}_+^K$. One can repeat this process for all the training images to get a total of S histograms $\{\boldsymbol{h}_t^1, \ldots, \boldsymbol{h}_t^S\}$, where $S = N$. We refer to these histograms as the *training histograms*.

BoF algorithms train classifiers $\phi_l(\boldsymbol{h}) : \mathbb{R}_+^K \rightarrow \mathbb{R}$ for each category $l \in \mathcal{L}$. We define the classification cost of a histogram $\boldsymbol{h}$ with respect to the classifier for a category l as the value given by $\phi_l(\boldsymbol{h})$. In general, the classifiers are trained such that they satisfy the constraint $\phi_l(\boldsymbol{h}) \geq 0$ if $\boldsymbol{h}$ is extracted from an object of category l and $\phi_l(\boldsymbol{h}) < 0$ otherwise. While one may choose from several classifiers for these histograms, we focus on classifiers based on the histogram intersection kernel [12]. Consider a test histogram $\boldsymbol{h}$ extracted from an object whose category we want to infer. Define the intersection kernel $\text{int}(\boldsymbol{h}, \boldsymbol{h}_t^s)$ operating on the test histogram $\boldsymbol{h}$ and the s^{th} training histogram $\boldsymbol{h}_t^s$, as

$$\text{int}(\boldsymbol{h}, \boldsymbol{h}_t^s) = \sum_{k=1}^{K} \min\{h_k, h_t^{s,k}\}, \tag{3}$$

where $s = 1, \ldots, S$, $h_t^{s,k}$ denotes the k^{th} bin count of the histogram $\boldsymbol{h}_t^s$ and h_k denotes the k^{th} bin count of the histogram $\boldsymbol{h}$. The general form of a classifier for category l using the histogram intersection kernel is then given as

$$\phi_l(\boldsymbol{h}) = \sum_{s=1}^{S} a_{l,s} \text{int}(\boldsymbol{h}, \boldsymbol{h}_t^s) + b_l. \tag{4}$$

Notice that the classifier $\phi_l(\boldsymbol{h})$ is linear in its parameters $\boldsymbol{a}_l = \begin{bmatrix} a_{l,1} & \ldots & a_{l,S} \end{bmatrix} \in \mathbb{R}^S$ and $b_l \in \mathbb{R}$. However, the kernel $\text{int}(\boldsymbol{h}, \boldsymbol{h}_t^s)$ is a non-linear function of the entries of the histograms $\boldsymbol{h}$ and $\boldsymbol{h}_t^s$.

Remark 1. *(Linear classifiers as a special case) The family of linear classifiers, whose general form is given as*

$$\phi_l(\boldsymbol{h}) = \sum_{k=1}^{K} a_{l,k} h_k + b_l, \tag{5}$$

is a special case of the family of the classifiers defined in (4)*. Specifically, this special case is obtained bysetting all the entries of each training histogram $\boldsymbol{h}_t^s$ to ∞ or to any large finite number M which is greater than the maximum number of interest points possible in an image.*

We note that one may use normalized histograms that satisfy the constraint $|\boldsymbol{h}|_1 = \sum_{k=1}^{K} h_k = 1$. This normalization introduces invariance to changes in scale of the object. However, as shown in [7], unnormalized histograms may also be effectively used for categorization.

3. A top-down energy function for JCaS

In this section, we will define a new top-down energy function by using concepts from BoF categorization algorithms. Specifically, we introduce a new family of top-down higher order potentials that model long-range interactions among all the interest points detected in a given image.

We first introduce some additional notation. We denote the subset of sites in an image where interest points have been detected as $\mathcal{I} \subset \mathcal{V}$. We divide $\mathcal{I}$ into K disjoint subsets $\mathcal{I}_k, k = 1, \ldots, K$, where $\mathcal{I}_k$ is the set of interest points whose descriptors are quantized as the k^{th} codeword in the learnt dictionary. Given a labeling $\boldsymbol{x}$ for an image I, we use $\boldsymbol{h}_l(\boldsymbol{x})$ to denote the histogram of the quantized descriptors for all those interest points $i \in \mathcal{I}$ that have been assigned the label $l \in \mathcal{L}$. In this work, we use unnormalized histograms, the reason for which will be made clearer later.

Since the ground truth labeling $\{\boldsymbol{y}^i\}_{i=1}^N$ is provided for the images $\{I^i\}_{i=1}^N$ in the training set, we can extract the training histograms $\{\boldsymbol{h}_t^s\}_{s=1}^S$ from the training images as follows. Given a training image I^i, we denote the set of distinct category labels present in its ground truth segmentation $\boldsymbol{y}^i$ as $\mathcal{L}^+(\boldsymbol{y}^i)$. We extract the interest points in the image I^i and then use the labeling $\boldsymbol{y}^i$ to extract for each category $l \in \mathcal{L}^+(\boldsymbol{y}^i)$, the interest points in the image that belong to that category. In this manner, we can construct $|\mathcal{L}^+(\boldsymbol{y}^i)|$ training histograms for the image I^i, where $|\mathcal{L}^+(\boldsymbol{y}^i)|$ denotes the number of elements in the set $\mathcal{L}^+(\boldsymbol{y}^i)$. We repeat this process for all the training images to get a total of S histograms $\{\boldsymbol{h}_t^1, \ldots, \boldsymbol{h}_t^S\}$, where $S = \sum_{i=1}^N |\mathcal{L}^+(\boldsymbol{y}^i)| \leq LN$. We can then use these training histograms to train classifiers $\{\phi_l\}_{l \in \mathcal{L}}$ for the histograms. Given a new test image I, we propose to use these classifiers to define a new top-down energy function, $E_{td}(\boldsymbol{x}; I)$, that depends on the labels of the interest points detected in the image, as

$$E_{td}(\boldsymbol{x}; I) = \sum_{l \in \mathcal{L}^+(\boldsymbol{x})} \phi_l(\boldsymbol{h}_l(\boldsymbol{x})). \tag{6}$$

Notice that the term $\phi_l(\boldsymbol{h}_l(\boldsymbol{x}))$, which is used to compute a top-down score for the object category l, is derived from the BoF framework discussed in the previous section. The main difference is that the histogram is now a function of the unknown labeling $\boldsymbol{x}$. Also, recall that we mentioned that BoF algorithms train the classifiers to satisfy the constraint $\phi_l(\boldsymbol{h}) \geq 0$ if $\boldsymbol{h}$ is extracted from an object of category l and $\phi_l(\boldsymbol{h}) < 0$ otherwise. Without loss of generality, we modify the constraints that the classifiers must satisfy. Specifically, for any $\boldsymbol{h}$ extracted from an object belonging to category $l \in \mathcal{L}$, the classifiers are trained to satisfy the constraint $\forall l' \in \mathcal{L} \setminus l : \phi_{l'}(\boldsymbol{h}) > \phi_l(\boldsymbol{h})$. This is done to ensure that the energy $E_{td}(\boldsymbol{x}; I)$ is lower when computed for accurate segmentations as compared to erroneous segmentations.

We now introduce a new family of top-down potentials that we refer to as the *classification potentials*, to define the top-down energy $E_{td}(\boldsymbol{x}; I)$. We will show that these potentials can be modeled using the robust higher order Potts potential functions introduced in [6]. This is of particular importance since it was shown in [6] that the latter potentials can be efficiently optimized using α-expansion.

3.1. Construction of the classification potentials

Our goal is to define for each category $l \in \mathcal{L}$, a higher order potential function $\psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I)$ on the clique of interest points $\mathcal{I}$ detected in an image I, such that it encodes the classification cost $\phi_l(\boldsymbol{h}_l(\boldsymbol{x}))$ for all the possible labelings $\boldsymbol{x} \in \mathcal{L}^{|\mathcal{V}|}$. Specifically, this potential must satisfy

$$\forall \boldsymbol{x} \in \mathcal{L}^{|\mathcal{V}|}, \psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I) = \begin{cases} \phi_l(\boldsymbol{h}_l(\boldsymbol{x})) & \text{if } \|\boldsymbol{h}_l(\boldsymbol{x})\| > 0 \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

In particular, the classification cost for an object category $l \in \mathcal{L}$ is accounted for, only when at least one interest point has been assigned label l, i.e., if $\|\boldsymbol{h}_l(\boldsymbol{x})\| > 0$. Now, we see that the potential $\psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I)$ in (7) can be rewritten as

$$\begin{aligned} \psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I) &= \sum_{s=1}^{S} a_{l,s} \text{int}(\boldsymbol{h}_l(\boldsymbol{x}), \boldsymbol{h}^s_t) + b_l \delta(\|\boldsymbol{h}_l(\boldsymbol{x})\| > 0) \\ &= \sum_{s=1}^{S} a_{l,s} \sum_{k=1}^{K} \min\left\{h_{l,k}(\boldsymbol{x}), h^{s,k}_t\right\} + b_l \min\left\{\sum_{k=1}^{K} h_{l,k}(\boldsymbol{x}), 1\right\}, \end{aligned} \tag{8}$$

where $h_{l,k}(\boldsymbol{x})$ is the k^{th} bin count of $\boldsymbol{h}_l(\boldsymbol{x})$. Notice that when $\|\boldsymbol{h}_l(\boldsymbol{x})\| = 0$, we have $\forall k = 1, \ldots, K$, $h_{l,k}(\boldsymbol{x}) = 0$. This can be used to verify that $\psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I)$ satisfies the property $\psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}) = 0$ when $\|\boldsymbol{h}_l(\boldsymbol{x})\| = 0$. Now, we note that due to the use of unnormalized histograms, we have

$$h_{l,k}(\boldsymbol{x}) = \sum_{i \in \mathcal{I}_k} \delta(x_i = l). \tag{9}$$

This can be used to rewrite (8) as

$$\begin{aligned} \psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I) = \sum_{s=1}^{S} a_{l,s} \sum_{k=1}^{K} \min\left\{\sum_{i \in \mathcal{I}_k} \delta(x_i = l), h^{s,k}_t\right\} \\ + b_l \min\left\{\sum_{i \in \mathcal{I}} \delta(x_i = l), 1\right\}, \end{aligned} \tag{10}$$

where $\delta(A) = 1$ if the event "A" is true and $\delta(A) = 0$ if "A" is false. The last expression in (10) is obtained as a result of the fact that $\sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \delta(x_i = l) = \sum_{i \in \mathcal{I}} \delta(x_i = l)$, since $\cup_{k=1}^{K} \mathcal{I}_k = \mathcal{I}$. To further simplify the representation of $\psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I)$, we define two kinds of potentials as follows

$$\begin{aligned} \psi^{H}_{\mathcal{I}_k,l,s}(\boldsymbol{x}_{\mathcal{I}_k}; I) &= \min\left\{\sum_{i \in \mathcal{I}_k} \delta(x_i = l), h^{s,k}_t\right\} \text{ and} \\ \psi^{\delta}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I) &= \min\left\{\sum_{i \in \mathcal{I}} \delta(x_i = l), 1\right\}. \end{aligned} \tag{11}$$

Here, the potential $\psi^{H}_{\mathcal{I}_k,l,s}(\boldsymbol{x}_{\mathcal{I}_k}; I)$ encodes the "minimum" function over the k^{th} bin counts of the histogram $\boldsymbol{h}_l(\boldsymbol{x})$ and the s^{th} training histogram $\boldsymbol{h}^s_t$. The potential $\psi^{\delta}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I)$ encodes the indicator function $\delta(\|\boldsymbol{h}_l(\boldsymbol{x})\| > 0)$. Notice that all the potentials defined in (11) belong to the family of higher order potentials defined on a set $A \subset \mathcal{V}$ as

$$\psi^{\text{RHOP}}_{A,l}(\boldsymbol{x}_A; I) = \min\left\{\sum_{i \in A} \delta(x_i = l), c\right\}, \text{ where } c \in \mathbb{R}. \tag{12}$$

This is precisely the family of robust higher order Potts potential introduced in [6].

Now, note that $E_{td}(\boldsymbol{x}; I)$ defined in (6) can be written as

$$E_{td}(\boldsymbol{x}; I) = \sum_{l \in \mathcal{L}^+(\boldsymbol{x})} \phi_l(\boldsymbol{h}_l(\boldsymbol{x})) = \sum_{l \in \mathcal{L}} \psi^{\text{BoF}}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I). \tag{13}$$

Notice that we can use (10) and (11) to further simplify the expression of $E_{td}(\boldsymbol{x}; I)$ as $\boldsymbol{w}^{\top}_{td} \Psi_{td}(\boldsymbol{x}; I)$, where

$$\boldsymbol{w}^{\top}_{td} = \begin{bmatrix} \cdots & \boldsymbol{a}^{\top}_l & b_l & \cdots \end{bmatrix} \in \mathbb{R}^{L(S+1)} \text{ and}$$

$$\Psi_{td}(\boldsymbol{x}; I) = \begin{bmatrix} \vdots \\ \sum_{k=1}^{K} \psi^{H}_{\mathcal{I}_k,l,s}(\boldsymbol{x}_{\mathcal{I}_k}; I) \\ \psi^{\delta}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I) \\ \vdots \end{bmatrix} \in \mathbb{R}^{L(S+1)}. \tag{14}$$

Remark 2. *[9] uses the output of object detectors to first localize objects with bounding boxes in the image. Each pixel in a bounding box pays a constant cost if it is assigned a category label that is different from the category prevalent in the box. [9] also ensures that no additional cost is payed when the number of pixels deviating from the prevalent category exceeds a pre-decided number. It can be verified that this is equivalent to using our framework by choosing all the points inside the box as the interest points and using the linear classifiers of (5) for defining the classification costs.*

3.2. Constraints on the top-down parameters $\boldsymbol{w}_{td}$

We previously mentioned that the potentials in (11) can be minimized using α-expansion. However, $E_{td}(\boldsymbol{x}; I)$ is constructed using scaled versions of these potentials, i.e., $\left\{a_{l,s} \psi^{H}_{\mathcal{I}_k,l,s}(\boldsymbol{x}_{\mathcal{I}_k}; I)\right\}^{s=1,\ldots,S}_{l \in \mathcal{L}}$ and $\left\{b_l \psi^{\delta}_{\mathcal{I},l}(\boldsymbol{x}_{\mathcal{I}}; I)\right\}_{l \in \mathcal{L}}$. In order to optimize these potentials using α-expansion, they need to belong to the family of the robust higher order Potts potentials. It can be verified that this holds true when the classifier parameters for each category $l \in \mathcal{L}$ satisfy the constraints $\forall s = 1, \ldots, S$, $a_{l,s} \geq 0$ and $b_l \geq 0$. Theorem 1 shows that these constraints are not restrictive. Specifically, given classifiers with real valued parameters, Theorem 1 describes the construction of new classifiers with non-negative parameters that give identical classification results.

Theorem 1. *Given a set of classifiers $\{\phi_l\}_{l \in \mathcal{L}}$ with parameters $(a_{l,1}, \ldots, a_{l,S}) \in \mathbb{R}^S$ and $b_l \in \mathbb{R}$, define a new set of classifiers $\{\tilde{\phi}_l\}_{l \in \mathcal{L}}$ with parameters $\forall s = 1, \ldots, S$*

$$\tilde{a}_{l,s} = a_{l,s} - \min_{l' \in \mathcal{L}} a_{l',s} \;\; \textit{and} \;\; \tilde{b}_l = b_l - \min_{l' \in \mathcal{L}} b_{l'} \qquad (15)$$

for all $l \in \mathcal{L}$. Then, the parameters of the new classifiers $\tilde{\boldsymbol{a}}_l$ and $\tilde{b}_l$ are non-negative. Moreover, the classification results obtained with the classifiers $\{\tilde{\phi}_l\}_{l \in \mathcal{L}}$ are identical to those obtained with the classifiers $\{\phi_l\}_{l \in \mathcal{L}}$.

While such constraints do not affect the classification results, they do affect the JCaS problem. Notice that b_l can be thought of as a category-level prior, i.e., a cost paid when one of the interest points in the image is assigned label l. The non-negativity of b_l biases our algorithm towards assigning a lower number of distinct category labels to the interest points. Also, notice that the potential $a_{l,s}\psi^H_{\mathcal{I}_k,l,s}(\boldsymbol{x}_{\mathcal{I}_k}; I)$ encodes the fact that a non-negative cost $a_{l,s}$ is paid when an interest point of type k is assigned category label $l \in \mathcal{L}$. However, these potentials incorporate *robustness* by encoding the fact that beyond a certain number of interest points (given by $h^{s,k}_t$) being assigned label l, the potential does not pay any additional cost. Now, since $\boldsymbol{a}_l$ is non-negative, this would imply that fewer sites should be assigned category l. However, the classifier parameters are non-negative for all the categories. Hence, we expect the assignment of the labels to be balanced across the different categories. The relative values of the parameters $\boldsymbol{a}_l$ for the different categories $l \in \mathcal{L}$ determine how the interest points' features are distributed across the different categories.

4. A new energy function for JCaS

The energy $E_{td}(\boldsymbol{x}; I)$ proposed in §3 is defined only over the labels of the interest points detected in an image I. Hence, $E_{td}(\boldsymbol{x}; I)$ alone cannot be used to solve the JCaS problem for an image since it would not give a label for each pixel. To this effect, we propose to solve the JCaS problem by minimizing a new energy $E(\boldsymbol{x}; I)$ that we define by combining the bottom-up energy $E_{bu}(\boldsymbol{x}; I)$ discussed in §2.1, with our proposed top-down energy $E_{td}(\boldsymbol{x}; I)$, as

$$\begin{aligned} E(\boldsymbol{x}; I) &= E_{bu}(\boldsymbol{x}; I) + E_{td}(\boldsymbol{x}; I) \\ &= \begin{bmatrix} \boldsymbol{w}_{bu}^\top & \boldsymbol{w}_{td}^\top \end{bmatrix} \begin{bmatrix} \Psi_{bu}(\boldsymbol{x}; I) \\ \Psi_{td}(\boldsymbol{x}; I) \end{bmatrix} = \boldsymbol{w}^\top \Psi(\boldsymbol{x}; I). \end{aligned} \qquad (16)$$

As a result, the segmentation for the regions that do not contain interest points is driven by the bottom-up potentials, which we believe is a fair strategy if the regions themselves do not offer any informative cue for BoF classification.

Notice that $\boldsymbol{w}$ contains the bottom-up parameters λ_U, λ_P and λ_C that are used to regulate the relative contributions of the potential functions modeling the local interactions, as well as the top-down parameters of the classifiers $\{\boldsymbol{a}_l, b_l\}_{l \in \mathcal{L}}$ that are used for modeling the global interactions. Hence, the process of learning $\boldsymbol{w}$ is equivalent to the joint learning of all the parameters needed for JCaS.

Note that as in [9], one may learn the top-down parameters separately from the bottom-up parameters. Specifically, one may extract histograms from the ground truth labelings of the training images and use them to train one-vs.-rest SVM classifiers $\{\boldsymbol{a}_l, b_l\}$ for each of the categories $l \in \mathcal{L}$. These classifiers can be used to construct the higher order potentials discussed in the previous section. However, one can construct several classifiers that partition the space of histograms extracted from the objects and achieve good classification. Not all the classifiers trained in this fashion might give good potentials for the purpose of JCaS. Hence, we need to *learn classifiers specifically for the JCaS task*.

4.1. A max-margin formulation for learning $\boldsymbol{w}$

Recall that we propose to segment an image I by minimizing the energy $E(\boldsymbol{x}; I)$. Hence, we would want that for any image I, the ground truth segmentation $\boldsymbol{y}$ minimizes the energy $E(\boldsymbol{x}; I)$ as $\forall \boldsymbol{x} \in \mathcal{L}^{|\mathcal{V}|} \setminus \boldsymbol{y}, E(\boldsymbol{x}; I) > E(\boldsymbol{y}; I)$, i.e., $\boldsymbol{w}^\top \Psi(\boldsymbol{x}; I) > \boldsymbol{w}^\top \Psi(\boldsymbol{y}; I)$. We will now describe a strategy to learn $\boldsymbol{w}$, motivated by this desired property.

Assume that we are given a training set of N images $\{I^i\}_{i=1}^N$ with ground truth labelings $\{\boldsymbol{y}^i\}_{i=1}^N$. We refer to any labeling of an image that is different from $\boldsymbol{y}^i$ as a negative example of segmentation for that image. We denote the set of negative examples of segmentations for an image I^i as $\mathcal{S}_i^-$. Now, note that all negative segmentation examples should not be treated equally. For example, a labeling which has a few errors is not the same as a labeling with 50% errors. Hence, we propose to enforce the constraint

$$\forall \boldsymbol{x} \in \mathcal{S}_i^- : \boldsymbol{w}^\top \big(\Psi(\boldsymbol{x}; I^i) - \Psi(\boldsymbol{y}^i; I^i)\big) > \ell(\boldsymbol{x}, \boldsymbol{y}^i). \qquad (17)$$

Here $\ell(\boldsymbol{x}, \boldsymbol{y}^i)$ measures the error in the labeling $\boldsymbol{x}$ as the average fraction of misclassified sites per category, as

$$\ell(\boldsymbol{x}, \boldsymbol{y}^i) = \sum_{l \in \mathcal{L}^+(\boldsymbol{y}^i)} \frac{\Delta_l(\boldsymbol{x}, \boldsymbol{y}^i)}{|\boldsymbol{y}^i_l|}, \qquad (18)$$

where $|\boldsymbol{y}^i_l|$ is the number of sites whose true category label is l and $\Delta_l(\boldsymbol{x}, \boldsymbol{y}^i)$ is the number of sites whose true label is l, but are assigned a different label as per $\boldsymbol{x}$. Given $\mu > 0$, we propose to learn the parameters $\boldsymbol{w}$ as the solution of

$$\begin{aligned} &\{\boldsymbol{w}^*, \{\xi_i^*\}_{i=1}^N\} = \operatorname*{argmin}_{\boldsymbol{w}, \{\xi_i\}_{i=1}^N} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\mu}{N}\sum_{i=1}^N \xi_i, \text{ subject to} \\ &\text{(a) } \forall i = 1, \ldots, N\text{: } \forall \boldsymbol{x} \in \mathcal{S}_i^- : \\ &\qquad \boldsymbol{w}^\top \big(\Psi(\boldsymbol{x}; I^i) - \Psi(\boldsymbol{y}^i; I^i)\big) \geq \ell(\boldsymbol{x}, \boldsymbol{y}^i) - \xi_i, \\ &\text{(b) } \forall i = 1, \ldots, N\text{: } \xi_i \geq 0 \quad \text{and} \quad \text{(c) } \boldsymbol{w} \geq \boldsymbol{0}. \end{aligned} \qquad (19)$$

While we refer the readers to [19] for a detailed explanation of this max-margin formulation, we now provide an intuition for (19). The constraint (a) is similar to (17) and the loss function $\ell(\boldsymbol{x}, \boldsymbol{y}^i)$ represents the margin between the positive and negative examples of segmentation. Notice that the margin has been *rescaled* as a function of the errors in the segmentation. The difference between constraint (a) and (17) is the non-negative valued slack variable ξ_i. Now, note that it might not be possible to find weights $\boldsymbol{w}$ that ensure that the ground truth segmentations have the minimum energy for each training image. The slack variables are introduced precisely to account for the violation of (17).

Finally, we enforce the constraint (b) in order to ensure that the classifier parameters are non-negative, so that the resulting energy can be optimized using α-expansion.

4.2. An iterative algorithm for learning $\boldsymbol{w}$

Note that the number of negative examples of segmentation for each image I^i is exponentially large (i.e., $L^{|I^i|}$). Hence, it is infeasible to solve (19) by considering all the negative examples for all the training images. For this purpose, we use an iterative algorithm to learn $\boldsymbol{w}$. The algorithm proceeds by sampling the space of negative examples of segmentation rather than considering the entire space. In particular, notice that for any image I^i, the slack variable ξ_i in (19)(a), depends only on the segmentation $\boldsymbol{x}^i_-$ that violates the constraint in (17) the most, which satisfies

$$\begin{aligned}&\forall \boldsymbol{x} \in \mathcal{S}_i^- \; : \; \boldsymbol{w}^\top \big(\Psi(\boldsymbol{x}; I^i) - \Psi(\boldsymbol{y}^i; I^i)\big) - \ell(\boldsymbol{x}, \boldsymbol{y}^i) \\ &\geq \boldsymbol{w}^\top \big(\Psi(\boldsymbol{x}^i_-; I^i) - \Psi(\boldsymbol{y}^i; I^i)\big) - \ell(\boldsymbol{x}^i_-, \boldsymbol{y}^i) = -\xi_i.\end{aligned} \tag{20}$$

Hence, we propose to learn $\boldsymbol{w}$ using an iterative procedure that aims to find $\boldsymbol{x}^i_-$ for each image I^i, by alternating between the following two steps.

First step. Given an estimate of $\boldsymbol{w}$, the first step of the algorithm finds the segmentation $\boldsymbol{x}^i_-$ that satisfies (20). It can be verified that given $\boldsymbol{w}$, $\boldsymbol{x}^i_-$ can be computed as

$$\boldsymbol{x}^i_- = \arg\min_{\boldsymbol{x}} \; \left[\boldsymbol{w}^\top \Psi(\boldsymbol{x}; I^i) - \ell(\boldsymbol{x}, \boldsymbol{y}^i)\right], \tag{21}$$

using α-expansion. Specifically, it can be shown that the loss function $\ell(\boldsymbol{x}, \boldsymbol{y}^i)$ can be represented using unary potentials. Moreover, we have shown in the previous section that $E(\boldsymbol{x}; I) = \boldsymbol{w}^\top \Psi(\boldsymbol{x}; I)$ can be minimized using α-expansion. Finally, notice that the term $\Psi(\boldsymbol{y}^i; I^i)$ can be ignored since it doesn't affect the optimization problem. Hence, the objective function in (21), i.e., $E(\boldsymbol{x}; I^i) - \ell(\boldsymbol{x}, \boldsymbol{y}^i)$, can also be minimized using α-expansion.

Second step. Once we have $\boldsymbol{x}^i_-$, we add this labeling to the set of negative examples $\mathcal{S}^i_-$ for the training image I^i. Having updated the set of negative examples of segmentation for each training image, we re-estimate $\boldsymbol{w}$ using (19) to ensure that the desired constraints are satisfied.

The problem in (19) has a unique solution for a given set of negative examples for the segmentations. Since there are at most a finite number of such examples for each image, there are a finite number of constraints that can be imposed in (19). The parameters estimated by the described iterative algorithm will converge to the solution that would be given by (19) after including the constraints for all the possible negative examples of segmentations (see [19]). In our experiments, we restricted the number of iterations to 15 since the algorithm typically converges in so many iterations.

5. Experiments

We evaluate our framework on the Graz-02 dataset [13], which contains 3 object categories (bicycles, cars and people) and the background category. For each of the three categories, there are 150 training images and 150 testing images of a single or multiple objects against the background. Hence, we have 450 training images and 450 test images.

We will compare the JCaS results obtained by minimizing the bottom-up energy $E_{bu}(\boldsymbol{x}; I)$ in (1), with those obtained by minimizing the energy $E(\boldsymbol{x}; I)$ in (16), which contains the bottom-up as well as top-down potentials.

Constructing the bottom-up potentials. In general, we can choose any algorithm to construct the bottom-up potentials. We adopt the strategy of [2] which was shown to produce good results. [2] oversegments a given image and works under the assumption that all the pixels in a given superpixel have the same label. This is equivalent to using a higher order potential for the superpixel which constrains all its constituent pixels to have the same label. For a fair comparison, we used the same constraint. Due to this constraint, we treat each superpixel as a single node and define unary and pairwise potentials over these nodes, as in [2].

A SIFT descriptor is extracted at each pixel and these descriptors are quantized using K-means to create a dictionary with 400 codewords. For each superpixel, the algorithm constructs a histogram of the quantized SIFT descriptors of the pixels in that superpixel. In order to account for spatial context, the histogram for each superpixel is updated by aggregating all the histograms over a neighborhood of size 4. A one-vs.-rest multi-class SVM with an RBF χ^2 kernel is trained on these histograms (after normalization), for each category. These classifiers are then used to analyze the histogram of a query superpixel $i \in \mathcal{V}$ and construct $\psi^U_i(x_i; I)$.

The pairwise potential $\psi^P_{i,j}(x_i, x_j; I)$ for neighboring superpixels i and j is defined as $\psi^P_{i,j}(x_i, x_j; I) = \frac{L(i,j)}{1+\|f(i)-f(j)\|}\delta(\boldsymbol{x}_i \neq \boldsymbol{x}_j)$, where $f(i)$ is the mean LUV color of the superpixel i and $L(i, j)$ is the length of the common boundary between superpixels i and j.

Constructing the top-down classification potentials. Notice that one could use all the pixels in an image as interest points for defining these potentials. However, the optimization of potentials defined on such a large clique would become intractable. Hence, we generate a sparse set of interest points $\mathcal{I}$ by using the SIFT interest point detector [11].

We use K-means to quantize the SIFT descriptors of $\mathcal{I}$ into a dictionary of K clusters. Notice that we could have used the dictionary already learnt for the unary potentials, to quantize the interest points. However, the SIFT interest points form a subset of all the pixels in the image. Hence, we learn a new dictionary with $K = 20$ clusters, to get a quantization of the descriptors of the interest points only.

Learning the energies' parameters. The iterative algorithm described in §4.2 computes an optimal set of parameters for the energy $E(\boldsymbol{x}; I)$. We can also modify this algorithm to compute an optimal set of parameters for the energy $E_{bu}(\boldsymbol{x}; I)$, by hardcoding the parameters for the classifica-

tion potentials to be equal to 0. In both cases, we initialize the algorithm with $\boldsymbol{w} = \begin{bmatrix} 1 & \mathbf{0}_{2+L(S+1)} \end{bmatrix}$, to avoid any unwanted biases due to the difference in the initializations.

Results. We evaluate the JCaS results with the intersection/union metric which is given as $\frac{100 \times \#\text{TP}}{\#\text{TP}+\#\text{FP}+\#\text{FN}}$, where TP = true positives, FP = false positives and FN = false negatives. The evaluation is presented in Table 1. Notice that the JCaS results (say $\boldsymbol{x}^*_{bu+td}$) obtained by minimizing $E(\boldsymbol{x}; I)$ are quantitatively better than those (say $\boldsymbol{x}^*_{bu}$) obtained by minimizing $E_{bu}(\boldsymbol{x}; I)$, across all the categories. In general, we have noticed that our proposed potentials help reduce several false positives for the object categories, therefore leading to this improvement in performance.

Due to space constraints, we have provided in Fig. 1, a few examples where $\boldsymbol{x}^*_{bu+td}$ is better than $\boldsymbol{x}^*_{bu}$. We have provided more qualitative results in the additional material to provide examples of cases where $\boldsymbol{x}^*_{bu+td}$ is better than $\boldsymbol{x}^*_{bu}$ as well as cases where $\boldsymbol{x}^*_{bu+td}$ is worse than $\boldsymbol{x}^*_{bu}$.

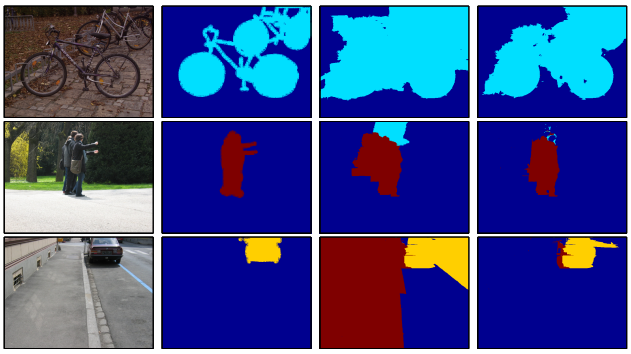
JCaS result	background	bicycles	cars	people	mean
$\boldsymbol{x}^*_{bu}$	76.25	40.58	34.66	37.17	47.16
$\boldsymbol{x}^*_{bu+td}$	**82.32**	**46.18**	**36.49**	**38.99**	**50.99**

Table 1. Analysis of JCaS results for the Graz-02 dataset, using the intersection/union metric.

6. Conclusions and future work

Experiments on the Graz-02 dataset show that our proposed potentials can be used to improve JCaS results. Although we used SIFT features in our evaluation, our method can be applied to any set of regions whose features can be classified with the intersection kernel. In fact, our framework can be used to go beyond traditional pairwise co-occurrence statistics by using the intersection kernel for higher order co-occurrence statistics. Now, recall from §3, that the number of training histograms S is roughly linear in the number of training images N. Also, recall that the number of parameters to be learnt for the classification potentials is equal to $L(S+1)$ which increases with the number of training images and categories. Hence, learning these parameters can be computationally expensive on large datasets with more categories, such as in the PASCAL VOC challenge. Future work entails an algorithm for learning these parameters efficiently. Finally, it is of interest to develop optimization schemes to deal with normalized histograms which provide some invariance to scale chances and to multiple instances of objects of the same category in an image.

Acknowledgements. This work was supported by grants NSF CAREER 0447739 and ONR YIP N00014-09-1-0839.



(a) Image (b) ground truth (c) $\boldsymbol{x}^*_{bu}$ (d) $\boldsymbol{x}^*_{bu+td}$

Figure 1. Examples of JCaS results for the Graz-02 dataset. The results are color coded as dark blue, light blue, yellow and red for the background, bicycles, cars and people, respectively.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE PAMI*, 23(11):1222–1239, 2001.

[2] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[3] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[4] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009.

[5] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 2008.

[6] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[7] C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE PAMI*, 2009.

[8] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, 2009.

[9] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What, where & how many? Combining object detectors and CRFs. In *ECCV*, 2010.

[10] D. Larlus and F. Jurie. Combining appearance models and MRFs for category level object segmentation. In *CVPR*, 2008.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, volume 20, pages 91–110, 2003.

[12] F. Odone, A. Barla, and A. Verri. Building kernels from binary strings for image matching. *IEEE TIP*, 14(2):169–180, 2005.

[13] A. Opelt and A. Pinz. The TU Graz-02 database. http://www.emt.tugraz.at/~pinz/data/GRAZ_02/.

[14] X. He, R. Zemel and M. Carreira-Perpiñán. Multiscale conditional random fields for image labelling. In *CVPR*, 2004.

[15] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *ECCV*, 2008.

[16] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[17] C. Russell, L. Ladicky, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[19] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 2005.

[20] A. B. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2004.

[21] J. Verbeek and B. Triggs. Scene segmentation with CRFs learned from partially labeled images. In *NIPS*, 2008.

[22] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.

[23] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010.

[24] L. Zhu, Y. Chen, Y. Lin, C. Lin and A. Yuille: A Hierarchical Image Model for Polynomial-Time 2D Parsing. In *NIPS*, 2008.