# Using Models of Objects with Deformable Parts for Joint Categorization and Segmentation of Objects

Nikhil Naikal, Dheeraj Singaraju and S. Shankar Sastry

University of California, Berkeley

**Abstract.** Several formulations based on Random Fields (RFs) have been proposed for joint categorization and segmentation (JCaS) of objects in images. The RF's sites correspond to pixels or superpixels of an image and one defines potential functions (typically over local neighborhoods) which define costs for the different possible assignments of labels to several different sites. Since the segmentation is unknown a priori, one cannot define potential functions over arbitrarily large neighborhoods as that may cross object boundaries. Categorization algorithms extract a set of interest points from the entire image and solve the categorization problem by optimizing cost functions that depend on the feature descriptors extracted from these interest points. There is some disconnect between segmentation algorithms which consider local neighborhoods and categorization algorithms which consider non-local neighborhoods. In this work, we propose to bridge this gap by introducing a novel formulation which uses models of objects with deformable parts, classically used for object categorization, to solve the JCaS problem. We use these models to introduce two new classes of potential functions for JCaS; (a) the first class of potential functions encodes the model score for detecting an object as a function of its visible parts only, and (b) the second class of potential functions encodes *shape priors* for each visible part and is used to bias the segmentation of the pixels in the support region of the part, towards the foreground object label. We show that most existing deformable parts formulations can be used to define these potential functions and that the resulting potential functions can be optimized exactly using min-cut. As a result, these new potential functions can be integrated with most existing RF-based formulations for JCaS.

## 1 Introduction

The goal of JCaS is to assign an object category label to each pixel in the image. Several solutions to JCaS use RF-based formulations, wherein algorithms define a RF whose sites correspond to pixels in the image and/or superpixels of the image [1–3, 7–9, 13–16, 19–21, 23, 25, 26]. To solve the JCaS problem, one defines potential functions (or potentials) which define costs for the different assignments of category labels to the sites. These potentials aggregated over local neighborhoods are then used to define an energy function over the different labelings, the minimizer of which is used to obtain a labeling for the image.

The potential functions used by most of the existing algorithms are local in nature. The unary potential for a site, which depends on the label of that single site only, is typically defined by using feature descriptors extracted from a local neighborhood of the site, e.g., [21, 14]. The features cannot be extracted from arbitrarily large neighborhoods since they might cross the objects' boundaries. Some methods consider non-local interest regions [26, 23] and use them to define pairwise potentials, which depend on labels of just two sites. Unary and pairwise potentials are typically not sufficient to describe all relationships amongst the sites. Hence, some algorithms use higher order potentials that depend on several sites [14, 20, 13]. While these potentials are also defined over local neighborhoods such as neighboring pixels or superpixels, there are a few exceptions [20, 22].

We argue that one can improve performance by using potentials that are defined over larger non-local neighborhoods, preferably all the regions covered by an object. However, such potentials can lead to a computational bottleneck. Therefore, it is preferable to define potentials over some representative subset region of the object. In this work, we propose to use models of objects with deformable parts [4, 6, 28], which have traditionally been used for object categorization, to define higher order potential functions over non-trivial non-local neighborhoods. These models assume that each object has a set of parts and the problem of detection corresponds to finding the locations of these parts in the image. Our work is motivated by the fact that the locations of the object's parts help define the non-local neighborhoods for our proposed potentials.

**Paper contributions.** We propose to address the aforementioned issues by integrating deformable parts models with RF formulations for JCaS. We assume that we are given a set of hypotheses as the output of detectors based on deformable parts models. Each hypothesis specifies for the object, a size, a pose and the locations for the object's parts. Given this, we propose a new energy function for JCaS with the following properties.

1) The energy function solves for detection and segmentation in a unified framework. The solution obtained by minimizing this function provides (i) a segmentation of the image, (ii) a list of the hypotheses that are accepted from the given ones, and (iii) a list of the visible parts for each of the accepted hypothesis.

2) Our key contribution is the design of two new higher order potential functions for defining the above energy function. The first family of potentials models the detection score for the deformable parts model. The binary-valued variables of this family of potentials indicate whether a part is detected/occluded at a certain location and the potential encodes the object detection score as a function of the visible parts only. The second family of potentials is used to model the *shape prior* of a part. Specifically, a part's shape prior provides for each pixel in the support region of that part, the probability that it belongs to the foreground object. Our proposed potentials use these probabilities to bias the segmentations of the pixels towards the foreground object label.

3) The problem of computing the minimizer of our proposed function is a discrete optimization problem, which can be NP-hard in general. We show that a global optimum to our optimization problem can be computed using min-cut.

**Related work.** The following are a few examples that have used object models for non-local potentials for JCaS. [15] modeled the object using multiple blobs. [8] and [9] used the output of object detectors to localize the objects in images. [22] and [25] used Bag of Features as the object model, while [1] and [2] used Poselets for their model. Our work, in contrast, uses the deformable parts model.

The works most closely related to our work are those of [12], [13] and [27]. [12] was perhaps the first work to use deformable parts models for object segmentation. The solution is obtained via an iterative process where the algorithm alternates between sampling from the space of possible hypothesis and computing the segmentation given the hypothesis. [27] extends [12] to deal with multiple object categories. [27] takes as input a set of hypothesis, all of which are used for segmentation. Our proposed framework has a few differences with these. First, we model the detection score as a function of the visible parts, while the above do not. Second, [27] computes the solution using EM, while we compute the segmentation in a single step using min-cut. Finally, in contrast to [27], our algorithm allows for rejection of some of the hypotheses provided as input.

[13] takes as input a set of hypotheses giving the locations of different parts of the image. Given these hypotheses, [13] defines a potential function which penalizes the number of pixels in the support region for each object part, which deviate from the foreground label. There is no shape prior used to bias the pixels differently based on their location in an object part's support region. Moreover, they do not model the detection score as a function of the visible parts.

**Paper outline.** In §2, we review some definitions that are relevant to our proposed formulation. In §3, we propose a new cost function for JCaS. We introduce two new higher order potentials for this cost function and discuss the constraints on these potentials that make them amenable to efficient inference using min-cut. We outline how the parameters of our cost function can be learned using max-margin methods. In §4, we evaluate the performance of our formulation on the PARSE dataset [17] and highlight our framework's advantages/limitations.

## 2 Review

In this section, we briefly review some concepts relevant to our formulation.

### 2.1 Random fields (RFs) formulations for JCaS

Given an image $I$, we define a RF, the set of whose sites is denoted as $\mathcal{V}$. These sites correspond to pixels or superpixels of the image. A binary-valued random variable $X(v_i)$ is defined at each site $v_i \in \mathcal{V}$ and can take any value $x(v_i)$ in the set of possible labels $\mathcal{B} = \{0, 1\}$. Any assignment of labels to the random variables is referred to as a *labeling* and is denoted as $\mathbf{x} \in \mathcal{B}^{|\mathcal{V}|}$. We denote the restriction of the random variables and labeling to a set of sites $A \subseteq \mathcal{V}$ as $\mathbf{X}(A)$ and $\mathbf{x}(A)$, respectively. Note that $x(v_i)$ is the restriction of $\mathbf{x}$ to the site $v_i$. Though the set of possible labels can contain several values for multiple categories, we restrict our analysis to the case of two labels for the ease of exposition.

The neighborhood of the RF is defined using the set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. An edge that spans two sites $v_i$ and $v_j$ is denoted by $\mathbf{e}_{ij}$. Larger neighborhoods are defined using cliques, where a clique $c \subset \mathcal{V}$ defines a set of sites, e.g., the set of pixels in a superpixel. We denote the set of all cliques in the RF as $\mathcal{C}$. One defines *potential functions* for each clique to model the scores for different assignments of labels to the clique. The following are a few commonly used potentials.

A *unary potential* $\psi_i(x(v_i); I)$ is defined for each site $i \in \mathcal{V}$, such that $\psi_i(b; I)$ defines the cost of assigning the label $b \in \mathcal{B}$ to the site $i$. This cost is typically computed using appearance-based or location-based feature descriptors. A *pairwise potential* $\psi_{ij}(x(v_i), x(v_j); I)$ is defined for each pair of neighboring sites $v_i$ and $v_j \in \mathcal{V}$, where $\mathbf{e}_{ij} \in \mathcal{E}$, such that $\psi_{ij}(b_i, b_j; I)$ defines the cost of assigning labels $b_i$ and $b_j$ to the sites $v_i$ and $v_j$, respectively. These potentials help enforce the spatial smoothness of $\mathbf{x}$ and align the edges across which the labeling changes with the edges in the image. They are also used to encode context.

Recent work has addressed the use of higher order potentials defined on larger cliques [14, 20, 13, 22]. A *higher order potential* $\psi_c(\mathbf{x}(c); I)$ is defined on the clique $c \in \mathcal{C}$, such that $\psi_c(\mathbf{b}_c; I)$ is the cost of assigning the labels $\mathbf{b}_c \in \mathcal{B}^{|c|}$ to the clique c. The potential $\psi_c(\mathbf{x}(c); I)$ can be defined over the the clique of pixels that belong to a superpixel. It can also be used to encode higher order contextual information about co-occurence of different categories [20] or to encode bin counts of histograms of quantized descriptors of interest points [22].

Most algorithms solve JCaS by minimizing an energy function of the form

$$E_1(\mathbf{x}; I) = \sum_{c \in \mathcal{C}} \lambda_c \psi_c(\mathbf{x}(c); I), \tag{1}$$

where $\forall c \in C, \lambda_c \in \mathbb{R}$. Note that (1) includes unary and pairwise potentials as special cases when $|c| = 1$ and 2, respectively. $E_1(\mathbf{x}; I)$ is typically designed such that min-cut based solvers provide the global minimum for the 2-label case and a local minimum (with optimality bounds) for the multi-label case.

As described in §1,it is preferable to have global object models that consider larger non-local neighborhoods, preferably all the sites with the same label. Such neighborhoods cannot be imposed apriori because the labeling is unknown.

### 2.2 Detection of objects with deformable parts

The algorithms in this genre assume that an object consists of $P \in \mathbb{Z}_+$ parts [4–6, 28]. Given an image $I$, a hypothesis $\theta$ specifies the object's pose $\pi(\theta)$, object's scale (size) $s(\theta)$ and a set of locations $l(\theta) = [l_1(\theta), \ldots, l_P(\theta)]^\top \in \Omega(I)^P$ for the different parts, where $\Omega(I) \in \mathbb{R}^2_+$ denotes the pixel domain of image $I$. The algorithms compute a detection score for the different hypotheses. The hypotheses with scores better than a threshold (say $\kappa$) are treated as accepted.

To define a detection score for a hypothesis $\theta$, the algorithms consider two different kinds of cost functions. The first type of cost function is an appearance-based cost for each of the different parts. For the $p^{\text{th}}$ part ($p = 1, \ldots, P$), one extracts feature descriptors from a support region (say $R_p(\theta)$) around $l_p(\theta)$,

where the size of the support region depends on the object's pose, scale and the part. The appearance-based cost $\phi_p^{\text{app}}(\theta; I)$ for the $p^{\text{th}}$ part is then computed as the output of a linear filter applied to these features descriptors, where the filter's coefficients depend on the pose, scale and part.

The second cost function takes into account the constraints on the relative locations of the different parts. Given locations $l_{p_1}(\theta)$ and $l_{p_2}(\theta)$ for parts $p_1$ and $p_2$, the cost $\phi_{p_1,p_2}^{\text{def}}(\theta)$ is a quadratic function of the entries of the vector $l_{p_1}(\theta) - l_{p_2}(\theta)$, where the coefficients of the quadratic function depend on the object's pose and scale. While one may construct a deformation cost for each of the $\frac{P(P-1)}{2}$ possible pairs of parts, most algorithms assume, for computational ease, that only a subset of these pairs are relevant for the purpose of detection. In fact, it is assumed that this subset of pairs can be represented using a tree. These connections between the parts are defined by the set of edges $\mathcal{E}_{\text{obj}} = (p_1, p_2)$.

Given an image $I$, one then defines the detection score for a candidate $\theta$ as

$$E_2(\theta; I) = \sum_{p=1}^{P} \phi_p^{\text{app}}(\theta; I) + \sum_{(p_1,p_2)\in\mathcal{E}_{\text{obj}}} \phi_{p_1,p_2}^{\text{def}}(\theta) \tag{2}$$

Due to the small number (say $M$) of possible poses and number (say $S$) of scales considered by detection algorithms, it is possible to do a brute force computation of the energy for the different poses and scales. Since the number of possible locations of the parts is high, i.e., $|\Omega(I)|^P$, it is not possible to compute the energy for all possible locations. To address this, given partial information for a hypothesis $\theta$ in terms of the pose $\pi(\theta)$ and scale $s(\theta)$, the part locations $l(\theta)$ that minimize $E_2(\theta; I)$ can be found using dynamic programming with time complexity $O(|\Omega(I)|P)$ [5]. We note that in the literature, the best hypothesis is typically obtained by solving a maximization problem. We can always reformulate the problem to get an equivalent minimization problem. In our work, we assume, without loss of generality, that the best hypotheses are obtained by solving a minimization problem, i.e., lower hypothesis scores are considered better.

In this formulation, the algorithm assumes that for a given pose, each part is assumed to be detected/visible in the image. There has been work to deal with occlusions, but an occluded pose is modeled as a pose different from the original pose [4]. We argue that it is of interest to explicitly model occlusion of parts within a certain pose rather than modeling occlusions using different poses.

## 3 A Novel Energy Function for JCaS

In this section, we define a new energy function to model non-local interactions amongst the sites of RFs for JCaS. For expositional ease, we make two simplifying assumptions. First, the number of parts (say $P$) is the same for all the poses. This is not necessary in practice. Second, we assume that the image is segmented into two groups only – an object of a particular category vs. background. Our analysis can be extended to deal with multiple semantic categories too.

We now introduce some notation. Given an image $I$, we denote the set of sites representing the pixels as $\mathcal{V}_{\text{pixels}} = \{v_1, \ldots, v_N\}$, where $N = |\Omega(I)|$. We do not introduce any additional sites for superpixels since potentials defined over superpixels can be redefined as potentials defined over the pixels [14]. We assume that we are given a set of $H$ hypotheses, $\Theta = \{\theta_1, \ldots, \theta_H\}$. For each hypothesis $\theta_h$ (where $h = 1, \ldots, H$), we define a set of $P+1$ sites $\mathcal{V}_{\text{obj}}(\theta_h) = \{v_0^h, v_1^h, \ldots, v_P^h\}$. The site $v_0^h$ is used to represent the $h^{\text{th}}$ hypothesis $\theta_h$ and for $p = 1, \ldots, P$, the site $v_p^h$ is used to represent the $p^{\text{th}}$ object part for hypothesis $\theta_h$.

We introduce binary-valued variables for the sites. For each site $v_i \in \mathcal{V}_{\text{pixels}}$, the variable $x(v_i)$ takes value 0 or 1 and represents segmentation as background or object, respectively. For each site $v_0^h$, the variable $x(v_0^h)$ takes value 0 or 1 and represents whether the hypothesis $\theta_h$ is rejected or accepted, respectively. For each site $v_p^h \in \mathcal{V}_{\text{obj}}(\theta_h)$, where $p > 0$, $x(v_p^h)$ takes value 0 or 1 and represents whether the $p^{\text{th}}$ part for the $h^{\text{th}}$ hypothesis is occluded or visible, respectively.

In this work, we propose to solve the JCaS problem by computing the values for the variables $\mathbf{x}$ that minimize an energy function of the form

$$\begin{aligned} E^{\text{JCaS}}(\mathbf{x}; I, \Theta) = \lambda^{\text{seg}} E^{\text{seg}}(\mathbf{x}(\mathcal{V}_{\text{pixels}}); I) + \lambda^{\text{hyp}} \sum_{h=1}^{H} E^{\text{det}}(\mathbf{x}(\mathcal{V}_{\text{obj}}(h)); I, \theta_h) \\ + \sum_{h=1}^{H} \sum_{p=1}^{P} \lambda_p^{\text{shape}}(\pi(\theta_h)) E_p^{\text{shape}}(x(v_p^h), \mathbf{x}(R_p(\theta_h)); I, \theta_h), \end{aligned} \tag{3}$$

where $\lambda^{\text{seg}}, \lambda^{\text{hyp}}$ and $\lambda_p^{\text{shape}}(\cdot)$ are all non-negative scalars, and $R_p(\theta_h)$ is the support region for the $p$'th part in hypothesis $\theta_h$. The term $E^{\text{seg}}(\cdot)$ is a segmentation-based energy function. It encodes the cost of assigning segmentation labels to the pixels, where the cost is computed using feature descriptors such as color, texture, etc. This energy can also be thought of in more general terms and can be replaced by energy functions used by existing JCaS algorithms.

Our main contribution is the design of the energies $E^{\text{det}}(\cdot)$ and $E_p^{\text{shape}}(\cdot)$. The term $E^{\text{det}}(\cdot)$ is a detection-based energy function and computes the detection score for each of the $H$ hypotheses, as a function of the visible parts only. The third term $E_p^{\text{shape}}(\cdot)$ is an energy function that connects the segmentation and detection terms. It helps encode how the $p^{\text{th}}$ part, if visible, affects the segmentation of the image region where the part is detected. We will discuss later, that it also helps in the use of the segmentation of an image region to verify whether a part is visible or not. In what follows, we define these energy functions and discuss how we can obtain $\mathbf{x}$ as the minimizer of $E^{\text{JCaS}}(\cdot)$.

### 3.1 Definition of the energy terms

**Detection.** We first extend the detection score defined in (2) by introducing the binary-valued variables $x(v_p^h)$ that model the visibility/occlusion of the parts, as

$$\phi(\mathbf{x}; I, \theta_h) = \sum_{p=1}^{P} \phi_p^{\text{app}}(\theta_h; I) x(v_p^h) + \sum_{(p_1, p_2) \in \mathcal{E}_{\text{obj}}} \phi_{p_1, p_2}^{\text{def}}(\theta_h) x(v_{p_1}^h) x(v_{p_2}^h). \tag{4}$$

The appearance score for the $p^{\text{th}}$ part is accounted for only if it is visible ($x(v_p^h) = 1$). The deformation score for a pair of parts is accounted for only when both parts are visible. Hence, the detection score depends on the visible parts only.

Given this definition of the score, we define the hypothesis score as follows

$$E^{\text{det}}(\mathbf{x}(\mathcal{V}_{\text{obj}}(\theta_h)); I, \theta_h) = \begin{cases} \phi(\mathbf{x}; I, \theta_h) - \kappa & \text{if } \phi(\mathbf{x}; I, \theta_h) \leq \kappa \\ 0 & \text{if } \phi(\mathbf{x}; I, \theta_h) \geq \kappa \text{ and } \mathbf{x}(\mathcal{V}_{\text{obj}}(\theta_h)) = \mathbf{0}, \\ \infty & \text{if } \phi(\mathbf{x}; I, \theta_h) \geq \kappa \text{ and } \mathbf{x}(\mathcal{V}_{\text{obj}}(\theta_h)) \neq \mathbf{0} \end{cases} \tag{5}$$

where $\kappa$ is a pre-defined threshold applied to the detection score, to accept a hypothesis (recall from §2.2). We have considered three different cases in defining $E^{\text{det}}(\cdot)$. In the first case, the detection score is below the threshold and the hypothesis is accepted. The cost paid in this case is a negative value and is precisely equal to $\phi(\mathbf{x}; I, \theta_h) - \kappa$. When the detection score is above the threshold, we want to reject the hypothesis and we don't want any of the parts to be detected. The third case in (5) ensures that none of the parts are detected when the detection score is more than $\kappa$, by assigning a very high cost, i.e., $\infty$, to this undesirable case. The second case in (5) corresponds to the case when we reject the hypothesis and no part is detected. In this case, we pay a constant cost 0.

**Shape prior.** $E^{\text{seg}}(\cdot)$ and $E^{\text{det}}(\cdot)$ are defined on disjoint sets of vertices, i.e., $\mathcal{V}_{\text{pixels}}$ and $\mathcal{V}_{\text{obj}}(\cdot)$, respectively. $E_p^{\text{shape}}(\cdot)$ serves to connect the segmentation variables with the hypotheses variables. Given a hypothesis $\theta_h$, we define for each part $p$, a shape prior (see Figure 1(a)) over its support region $R_p(\theta_h)$, as

$$\forall v_i \in R_p(\theta_h) : \xi(v_i, p) = prob(x(v_i) = 1 | x(v_p^h) = 1) \tag{6}$$

More specifically, the shape prior specifies for each pixel in the support region of a visible part, the probability that it will be assigned to the foreground object.

Now, note that it is straightforward to define an energy function for the $p^{\text{th}}$ part, as $\sum_{v_i \in R_p(\theta_h)} \left( -\xi(v_i, p) x(v_i) x(v_p^h) \right)$. Specifically, when the $p^{\text{th}}$ part is detected ($x(v_p^h) = 1$) and a pixel $v_i$ in its support region is assigned to the foreground, a negative cost $-\xi(v_i, p)$ is paid. This implies that all the pixels which have a high probability (as given by the shape prior) of belonging to the foreground, will have a greater bias towards being segmented as foreground. In this manner, we see how the detection of parts can help improve the segmentation. However, there must be a symbiotic interplay between segmentation and detection, and we argue that segmentation must also help improve the detection. We propose a constraint that a part must be treated as being detected/visible, only if a sufficient number of pixels in its support region are segmented as belonging to the foreground. To this effect, we define the energy function as

$$E_p^{\text{shape}}(x(v_p^h), \mathbf{x}(R_p(\theta_h); I, \theta_h), = \begin{cases} \sum_{v_i \in R_p(\theta_h)} -\xi(v_i, p) \beta_p(\pi(\theta_h)) & \text{if } x(v_p^h) = 0 \\ \sum_{v_i \in R_p(\theta_h)} -\xi(v_i, p) x(v_i) & \text{if } x(v_p^h) = 1 \end{cases}, \tag{7}$$

where $\beta_p(\pi(\theta_h)) > 0$. When the $p^{\text{th}}$ part is not detected, a constant cost (which does not depend on the segmentation) is paid. When the part is detected, the cost depends on the segmentation in the support region. Moreover, notice from (7) that when a sufficient number of pixels in $R_p(\theta_h)$ are assigned to the foreground, i.e., when $\sum_{v_i \in R_p(\theta_h)} -\xi(v_i,p)x(v_i) \leq \beta_p(\pi(\theta_h)) \sum_{v_i \in R_p(\theta_h)} -\xi(v_i,p)$, this energy function biases the $p^{\text{th}}$ part towards being detected.

### 3.2 Inference

The potentials defined in (5) and (7) are higher order potentials that depend on the labels of more than two sites. We will now show how these potentials can be expressed using unary and pairwise potentials. Since energy functions with unary and pairwise potentials can be minimized using min-cut, our proposed potentials can be integrated into existing JCaS algorithms that use min-cut based solvers.

To find the global optimum using min-cut, there are no constraints on the unary potentials. The pairwise potentials, however, do need to satisfy the *submodularity* constraint [11]. If one considers pairwise potentials that are defined over two binary-valued variables, say $y_1$ and $y_2$, the pairwise potentials $\gamma_1 y_1 y_2$ and $\gamma_2 \bar{y}_1 y_2$ (where $\bar{y}_1 = 1 - y_1$) are submodular only if $\gamma_1 \leq 0$ and $\gamma_2 \geq 0$ [11].

We first see that the energy $E_p^{\text{shape}}(\cdot)$ defined in (7) can be rewritten as

$$E_p^{\text{shape}}(x(v_p^h), \mathbf{x}(R_p(\theta_h); I, \theta_h) = \sum_{v_i \in R_p(\theta_h)} -\xi(v_i,p) \left(\beta_p(\pi(\theta_h))\bar{x}(v_p^h) + x(v_i)x(v_p^h)\right). \tag{8}$$

It is easy to verify that for $x(v_p^h) = 0$ and $x(v_p^h) = 1$, the score in (8) is exactly the same as that in (7). The first term in the summation in (8) is a unary term that depends only on $x(v_p^h)$. The second term is a pairwise potential that depends on $x(v_i)$ and $x(v_p^h)$. In this case, we note that by its definition in (6), $\xi(v_i,p) \geq 0$. Therefore, the potential $-\xi(v_i,p)x(v_i)x(v_p^h)$ is submodular by construction.

We now use the variable $x(v_0^h)$ to rewrite $E^{\text{det}}(\cdot)$ defined in (5), as

$$E^{\text{det}}(\mathbf{x}(\mathcal{V}_{\text{obj}}(\theta_h)); I, \theta_h) = (\phi(\mathbf{x}; I, \theta_h) - \kappa)\, x(v_0^h) + \infty \sum_{p=1}^{P} (\bar{x}(v_0^h)x(v_p^h)). \tag{9}$$

When $x(v_0^h) = 1$, the hypothesis is accepted and the cost is equal to $\phi(\mathbf{x}; I, \theta_h) - \kappa$. When $x(v_0^h) = 0$, the hypothesis is rejected and the third term $\infty(\bar{x}(v_0^h)x(v_p^h))$ ensures that all the $x(v_p^h) = 0$ when $x(v_0^h) = 0$. The cost paid when $\mathbf{x}(\mathcal{V}_{\text{obj}}(\theta_h)) = \mathbf{0}$ is equal to 0. Therefore, this energy represents the detection energy in (5). Now, given the expression in (4), we can rewrite the right hand side of (9) as

$$\begin{aligned}(\phi(\mathbf{x}; I, \theta_h) - \kappa)x(v_0^h) + \infty \sum_{p=1}^{P} (\bar{x}(v_0^h)x(v_p^h)) &= \sum_{p=1}^{P} \phi_p^{\text{app}}(\theta_h; I)x(v_p^h)x(v_0^h) \\ + \sum_{(p_1,p_2) \in \mathcal{E}_{\text{obj}}} \phi_{p_1,p_2}^{\text{def}}(\theta_h)x(v_{p_1}^h)x(v_{p_2}^h)x(v_0^h) - \kappa x(v_0^h) &+ \infty \sum_{p=1}^{P} (\bar{x}(v_0^h)x(v_p^h)).\end{aligned} \tag{10}$$

Notice that the first and second expressions are potentials defined over two variables $\left(x(v_p^h)x(v_0^h)\right)$ and three variables $\left(x(v_{p_1}^h)x(v_{p_2}^h)x(v_0^h)\right)$, respectively. However any solution $\mathbf{x}^*$ that minimizes the energy satisfies the constraint that $\forall p = 1,\ldots,P, x^*(v_p^h) = 1$, only if $x^*(v_0^h) = 1$. To this effect, $x^*(v_p^h)x^*(v_0^h) = 1$, only if $x^*(v_p^h) = 1$. Similarly, $x^*(v_{p_1}^h)x^*(v_{p_2}^h)x^*(v_0^h) = 1$, only if $x^*(v_{p_1}^h)x^*(v_{p_2}^h) = 1$. Hence, we can drop $x(v_0^h)$ in the first and second terms and rewrite $E^{\text{det}}(\cdot)$ as

$$\begin{aligned} E^{\text{det}}(\mathbf{x}(\mathcal{V}_{\text{obj}}(\theta_h)); I, \theta_h) = \sum_{p=1}^{P} \phi_p^{\text{app}}(\theta_h; I)x(v_p^h) + \sum_{(p_1,p_2)\in\mathcal{E}_{\text{obj}}} \phi_{p_1,p_2}^{\text{def}}(\theta_h)x(v_{p_1}^h)x(v_{p_2}^h) \\ - \kappa x(v_0^h) + \infty \sum_{p=1}^{P} (\bar{x}(v_0^h)x(v_p^h)), \end{aligned} \tag{11}$$

such that the minimizers of (9) and (11) are the same. The first and third terms in (11) are unary potentials. The fourth term $\infty\bar{x}(v_0^h)x(v_p^h)$ is submodular by construction. The second term $\phi_{p_1,p_2}^{\text{def}}(\theta_h)x(v_{p_1}^h)x(v_{p_2}^h)$ is submodular if and only if $\phi_{p_1,p_2}^{\text{def}}(\theta_h) \leq 0$. This score $\phi_{p_1,p_2}^{\text{def}}(\theta_h)$ is a quadratic function computed as

$$\phi_{p_1,p_2}^{\text{def}}(\theta_h) = \begin{bmatrix} dl_1 \\ dl_2 \\ 1 \end{bmatrix}^\top \begin{bmatrix} a_1(s(\theta_h),\pi(\theta_h)) & 0 & b_1(s(\theta_h),\pi(\theta_h)) \\ 0 & a_2(s(\theta_h),\pi(\theta_h)) & b_2(s(\theta_h),\pi(\theta_h)) \\ b_1(s(\theta_h),\pi(\theta_h)) & b_2(s(\theta_h),\pi(\theta_h)) & c(s(\theta_h),\pi(\theta_h)) \end{bmatrix} \begin{bmatrix} dl_1 \\ dl_2 \\ 1 \end{bmatrix}, \tag{12}$$

where $[dl_1, dl_2]^\top = l_{p_1}(\theta_h) - l_{p_2}(\theta_h)$ [5]. Note that by definition, $a_1(s(\theta_h),\pi(\theta_h)) > 0$ and $a_2(s(\theta_h),\pi(\theta_h)) > 0$ [5]. We now describe how the parameters of $\phi_{p_1,p_2}^{\text{def}}(\cdot)$ can be updated for a given image, such that the classification results are not affected and $\phi_{p_1,p_2}^{\text{def}}(\theta_h) \leq 0$ for all possible $(l_{p_1}(\theta_h), l_{p_2}(\theta_h))$.

If we update $\phi_{p_1,p_2}^{\text{def}}(\cdot)$ to $\tilde{\phi}_{p_1,p_2}^{\text{def}}(\cdot)$, such that all the parameters are kept constant but $c(\cdot)$ is updated as $\tilde{c}(s(\theta_h),\pi(\theta_h)) = c(s(\theta_h),\pi(\theta_h)) + \Delta c(s(\theta_h),\pi(\theta_h))$, we have for all $\theta_h$, $\tilde{\phi}_{p_1,p_2}^{\text{def}}(\theta_h) = \phi_{p_1,p_2}^{\text{def}}(\theta_h) + \Delta c(s(\theta_h),\pi(\theta_h))$. This does not alter the relative ordering of the scores of the different hypotheses. The detection results are the same if one updates the threshold $\kappa$ as $\tilde{\kappa} = \kappa + \Delta c(s(\theta_h),\pi(\theta_h))$.

Given an image $I$, since there are only a finite number of locations used to compute the expression in (12), we can always find a $\Delta c(s(\theta_h),\pi(\theta_h))$ for that image, such that $\tilde{\phi}_{p_1,p_2}^{\text{def}}(\theta_h) \leq 0$ for all possible $(l_{p_1}, l_{p_2})$. This implies that we can always update the parameters to construct submodular pairwise potentials.

### 3.3 Parameter learning

Notice that the energy $E^{\text{JCaS}}(\mathbf{x}; I, \Theta)$ defined in (3), can be rewritten as

$$\begin{aligned} E^{\text{JCaS}}(\mathbf{x}; I, \Theta) &= \mathbf{w}^\top \Psi(\mathbf{x}; I, \Theta) \\ &= \begin{bmatrix} \lambda^{\text{seg}} \\ \lambda^{\text{hyp}} \\ \vdots \\ \lambda_p^{\text{shape}}(\pi_m) \\ \vdots \end{bmatrix}^\top \begin{bmatrix} E^{\text{seg}}\left(\mathbf{x}(\mathcal{V}_{\text{pixels}}); I\right) \\ \sum_{h=1}^{H} E^{\text{det}}\left(\mathbf{x}(\mathcal{V}_{\text{obj}}(h)); I, \Theta\right) \\ \vdots \\ \sum_{h=1}^{H} \delta(\pi(\theta_h) = \pi_m) E_p^{\text{shape}}(x(v_p^h), \mathbf{x}\left(R_p(\theta_h)\right); I, \theta_h) \\ \vdots \end{bmatrix}, \end{aligned} \tag{13}$$

where $\delta(\cdot)$ is the 0-1 indicator function and $\mathbf{w} \in \mathbb{R}^{2+(P\times M)}$ contains the parameters that regulate the relative contributions of the different potentials.

Recall that we segment an image $I$ by minimizing $E(\mathbf{x}; I, \Theta)$. Hence, we want that the true segmentation $\mathbf{y}$ of image $I$ minimize the energy $E(\mathbf{x}; I)$ as $\forall \mathbf{x} \in \mathcal{B}^{N+(H\times(P+1))} \setminus \mathbf{y}, E(\mathbf{x}; I, \Theta) > E(\mathbf{y}; I, \Theta)$, i.e., $\mathbf{w}^\top \Psi(\mathbf{x}; I, \Theta) > \mathbf{w}^\top \Psi(\mathbf{y}; I, \Theta)$. We now describe an optimization problem to learn $\mathbf{w}$, motivated by this property.

Assume that we are given a training set of $T$ images $\{I_t\}_{t=1}^T$ with ground truth labelings $\{\mathbf{y}_t\}_{t=1}^T$. We refer to any labeling of an image that is different from $\mathbf{y}_t$ as a negative example of segmentation. We denote the set of negative examples of segmentations for an image $I_t$ as $\mathcal{U}_t^-$. Since all negative segmentation examples should not be treated equally, we propose to enforce the constraint

$$\forall \mathbf{x} \in \mathcal{U}_t^- : \mathbf{w}^\top \big(\Psi(\mathbf{x}; I_t, \Theta_t) - \Psi(\mathbf{y}_t; I_t, \Theta_t)\big) > \ell(\mathbf{x}, \mathbf{y}_t), \tag{14}$$

where $\ell(\mathbf{x}, \mathbf{y}_t)$ is a loss function that quantifies errors in the segmentation, as

$$\ell(\mathbf{x}, \mathbf{y}_t) = \frac{\sum_{v_i \in \mathcal{V}_{\text{pixels}}} y_t(v_i)\bar{x}(v_i)}{\sum_{v_i \in \mathcal{V}_{\text{pixels}}} y_t(v_i)} + \frac{\sum_{v_i \in \mathcal{V}_{\text{pixels}}} \bar{y}_t(v_i)x(v_i)}{\sum_{v_i \in \mathcal{V}_{\text{pixels}}} \bar{y}_t(v_i)}. \tag{15}$$

$\ell(\mathbf{x}, \mathbf{y}_t)$ computes the sum of fractions of misclassified sites per category.

Given a regularization parameter $C > 0$, we propose to learn $\mathbf{w}$ by solving

$$\begin{aligned} &\{\mathbf{w}^*, \{\eta_t^*\}_{t=1}^T\} = \underset{\mathbf{w}, \{\xi_t\}_{i=1}^N}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{T}\sum_{t=1}^T \eta_t, \text{ subject to } \forall t = 1, \ldots, T \\ &\text{(a) } \forall \mathbf{x} \in \mathcal{U}_t^- : \mathbf{w}^\top \big(\Psi(\mathbf{x}; I_t, \Theta_t) - \Psi(\mathbf{y}_t; I_t, \Theta_t)\big) \geq \ell(\mathbf{x}, \mathbf{y}_t, \Theta_t) - \eta_t, \\ &\text{(b) } \eta_t \geq 0 \quad \text{and} \quad \text{(c) } \mathbf{w} \geq \mathbf{0}. \end{aligned} \tag{16}$$

This formulation is mostly based on [24] and we solve (16) using the cutting-plane algorithm described in [24]. While we refer the readers to [24] for the details, we now provide some intuition for (16). The constraint (a) is similar to (14) except for the non-negative valued slack variable $\eta_t$ which allow for the violation of (14). Constraint (c) ensures that the resulting energy is submodular.

## 4 Experiments

*Description of dataset.* For the evaluation, we use the Image Parse Dataset [17] which consists of 305 articulated full-body images of people. The first 100 images are used as training data and the remaining 205 as test data. We have manually segmented the images in the dataset for our quantitative evaluation.

*Algorithms compared in the evaluation.* We first describe the construction of the energy $E^{\text{JCaS}}(\cdot)$. We define $E^{\text{det}}(\cdot)$ using the outputs of the detector of [28]. Although our method can handle multiple detection hypotheses, we use only the highest scoring hypothesis for each image in our experiments. We now describe how we construct the shape priors for $E_p^{\text{shape}}(\cdot)$. We run the detection algorithm of [28] on the training images. For each part type detected in an image, we find


(a) (b)

**Fig. 1.** (a) Shape priors generated for 4 part types using the parts model of [28]. (b) Two examples of shape priors being superimposed to generate foreground hypothesis.

the associated patch in its ground truth segmentation enclosed by the detection box. Averaging the segmentation patches over all the training images provides shape priors similar to those shown in Figure 1(a). Figure 1(b) shows examples where the learned shape priors are placed at part detection sites. It is clear from this image how the shape priors influence the segmentation of the people.

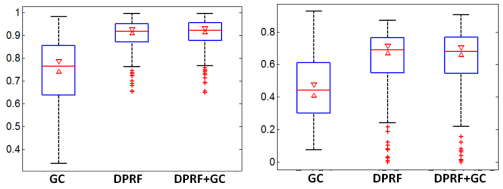
To construct $E^{\text{seg}}(\cdot)$, we use the given hypothesis to create a color-based unary potential. We fit a a Gaussian Mixture Model (GMM) with 5 components to the $RGB$-colors of all the image's pixels that lie outside the detection boxes for the parts. Given the color of a pixel $v_i$ in the image, we use this GMM to define the background unary potential $\psi_i^{clr}(0; I)$. We set the foreground unary potentials to zero, i.e., $\psi_i^{clr}(1; I) = 0$. This reduces dependency on color for segmenting the foreground while relying entirely on the detection and shape prior potentials. We also define a color-based pairwise potential as $\psi_{ij}^{clr}(x(v_i), x(v_j)) = \delta(x(v_i) \neq x(v_j))e^{-\beta\|\mathbf{z}(v_i)-\mathbf{z}(v_j)\|^2}$ where $\mathbf{z}(v_i)$ is the $RGB$ color at pixel $v_i$, and each $v_j$ is in the 4-neighborhood of pixel $v_i$. In all our experiments we set $\beta = 10$.

As a baseline for comparison, we consider the GrabCut algorithm [18], which considers only unary and pairwise potentials. It alternates between (a) fitting GMMs of color for the foreground/background, given the segmentation, and (b) computing the segmentation, given the potentials constructed with these GMMs. We initialize GrabCut with a segmentation, where we label all the pixels inside the detection boxes for the parts as the foreground, and the rest as background. We run 10 iterations of GrabCut. Unlike the traditional GrabCut, we cannot place any hard constraints on the pixels' labels, since the detection boxes contain pixels belonging to the background as well as foreground. We choose this baseline to show that even if one is given a good initial object detection, using low-level features such as color need not produce good JCaS results. This motivates our argument for object models defined over non-trivial non-local neighborhoods.

We also consider a third algorithm, where we combine our algorithm with GrabCut. We alternate between (a) computing the segmentation by minimizing $E^{\text{JCaS}}(\cdot)$, and (b) improving the color models given the segmentation.

The parameters for all the three algorithms are learnt as described in §3.3. In what follows, we refer to our method as DPRF (deformable parts + random fields) and to GrabCut as GC. The third algorithm is referred to as DPRF+GC.

| | Foreground | | Background | |
|---|---|---|---|---|
| Method | mean | std | mean | std |
| GC | 0.74 | 0.15 | 0.46 | 0.20 |
| DPRF | 0.90 | 0.07 | **0.63** | **0.19** |
| DPRF +GC | **0.91** | **0.06** | **0.63** | **0.19** |




(a) Foreground I/U (b) Background I/U

**Fig. 2.** Comparison of I/U for the segmentation results produced by the 3 methods.

*Evaluation.* We evaluate the segmentations using the Intersection/Union (I/U) metric given by $\frac{\#TP}{\#TP+\#FP+\#FN}$, where $TP$ = true positives, $FP$ = false positives and $FN$ = false negatives. Better segmentation corresponds to higher I/U.

The results are presented in the table and the boxplots in Fig. 2. The top/bottom edge of each boxplot for a set of values indicates the maximum/minimum of the values. The bottom/top extents of the box mark the 25/75 percentile. The red line in the box indicates the median and the red crosses outside the boxes show potential outliers. The 5% confidence intervals for determining statistical significance of difference between the medians are shown as red triangles.

The median I/U is notably lower for GC in comparison to DPRF and the 5% confidence intervals for these results do not overlap. However, the medians for DPRF and DPRF+GC are very similar and the 5% confidence intervals do overlap. This combined with the results in the table help us conclude that (a) DPRF produces better results than GC, and (b) the introduction of color information into DPRF, i.e., DPRF+GC does not produce any significant improvement.

Figure 3 presents a qualitative comparison of the results. The first column shows the hypothesis for the deformable parts (from [28]) overlaid on the image. The second column shows the result of pruning some of the detections using DPRF. The third, fourth and fifth columns of the figure show the segmentation produced by GC, DPRF and DPRF+GC, respectively. The first 3 rows show examples where the results of GC are inferior to those produced by DPRF and DPRF+GC. In these examples, the detection algorithm has fit the articulated models to the data reasonably well. The next two examples show scenarios where the detection algorithm errs and detects an extra limb (circled in white). As seen in the second column, DPRF prunes out these errors and provides a better segmentation than GC. The last two rows show examples where GC performs comparable to or outperforms DPRF. The last failure case is due to the poor part detection as seen in the first column in the last row.

## 5 Conclusion

We presented a JCaS framework where we proposed two new families of potentials that combine detection hypothesis with the segmentation of the image. These potentials can be integrated with existing RF-based JCaS algorithms. Results show that the detection hypothesis helps provide good segmentation results, and the segmentation can be used to prune some errors in the hypothesis.

**Fig. 3.** Column 1 shows the articulated model overlaid on the images. Column 2 shows the pruned model that has rejected some part detections using DPRF. Columns 3-5 show the segmentations given by GC, DPRF and DPRF+GC. See text for explanation.

**Acknowledgement**. This research was supported by ARL MAST-CTA W911NF-08-2-0004, DARPA FA8650-11-1-7153 and ONR N00014-12-1-0609. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute for Government purposes notwithstanding any copyright notation herein.

## References

1. Arbelaez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., Malik, J.: Semantic Segmentation using Regions and Parts. CVPR (2012)
2. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object Segmentation by Alignment of Poselet Activations to Image Contours. CVPR (2011)

3. Carreira, J., Sminchisescu, C.: CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. PAMI **34(7)** (2012)
4. Felzenszwalb, P.: Object Detection Grammars ICCV Workshops (2011)
5. Felzenszwalb, P., Huttenlocher, D.: Pictorial Structures for Object Recognition. IJCV **61(1)** (2005)
6. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. PAMI **32(9)** (2010)
7. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. CVPR (2008)
8. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. IJCV (2008)
9. Gould, S., Gao, T., Koller, D.: Region-based segmentation and object detection. NIPS (2009)
10. Kohli, P., Ladicky, L., Torr, P.: Robust higher order potentials for enforcing label consistency. CVPR (2008)
11. Kolmogorov, V., Zabih, R.: What Energy Functions Can Be Minimized via Graph Cuts? PAMI **26(2)**, (2004)
12. Kumar, M., Torr, P., Zisserman, A.: An object category specific MRF for segmentation. Toward Category-Level Object Recognition. (2006) 596–616
13. Ladicky, L., Sturgess, P., Alahari, K., Russell, C., Torr, P.: What, where & how many? Combining object detectors and CRFs. ECCV (2010)
14. Ladicky, L., Rusell, C., Kohli, P., Torr, P.: Associative hierarchical CRFs for object class image segmentation. ICCV (2009)
15. Larlus, D., Jurie, F.: Combining appearance models and MRFs for category level object segmentation. CVPR (2008)
16. Pantofaru, C., Schmid, C., Hebert, M.,: Object recognition by integrating multiple image segmentations. ECCV (2008)
17. Ramanan, D.: Learning to Parse Images of Articulated Bodies. NIPS (2006)
18. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. SIGGRAPH (2004)
19. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. ICCV (2007)
20. Rusell, C., Ladicky, L., Kohli, P., Torr, P.: Graph cut based inference with co-occurrence statistics. ECCV (2010)
21. Shotton, J., Winn, J., Rother, C., Criminisi, A., Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV (2009)
22. Singaraju, D., Vidal, R.: Using Global Bag of Features Models in Random Fields for Joint Categorization and Segmentation of Objects. CVPR (2011)
23. Torralba, A., Murphy, K., Freeman, W.: Contextual models for object detection using boosted random fields. NIPS (2004)
24. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. JMLR (2005)
25. Verbeek, J., Triggs, B.: Scene segmentation with CRFs learned from partially labeled images. NIPS (2008)
26. Winn, J., Shotton, J: The layout consistent random field for recognizing and segmenting partially occluded objects. CVPR (2006)
27. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.: Layered object detection for multi-class segmentation. CVPR (2010)
28. Yang, Y., Ramanan D.: Articulated pose estimation with flexible mixtures-of-parts. CVPR (2011)