Semantic Part Segmentation using Compositional Model combining Shape and Appearance

Jianyu Wang, Alan Yuille
University of California, Los Angeles

The past few years have witnessed significant progress on various object-level visual recognition tasks, such as object detection [3], object segmentation [2], etc. Understanding how different parts of an object are related and where the parts are located have been an increasingly important topic in computer vision. There is extensive study on some part-level visual recognition tasks, such as human pose estimation (predicting joints) [4] and landmark localization (predicting keypoints) [1]. But there are only a few pieces of works on semantic part segmentation.

In this paper, we study the problem of semantic part segmentation for animals. This is more challenging because semantic parts of animals often have similar appearance and highly varying shapes. To tackle these challenges, we build a mixture of compositional models to represent the object boundary and the boundaries of semantic parts. Each mixture component is able to handle local deformation of shapes and different mixtures deal with global variations due to viewpoints and poses. Figure 1 (a) shows the visualization of one compositional tree. We formulate the compositional part-subpart relation by a probabilistic graphical model. The MAP inference performs the following energy minimization

$$E(\mathbf{I}) = \min_{S_{\mathcal{V}}} E(S_{\mathcal{V}}, \mathbf{I}) = \sum \phi(S_v, \mathbf{I}) + \sum \psi(S_{ch(v)}), \quad (1)$$

where $\mathbf{I}$ denotes the image, $v \in \mathcal{V}$ denotes a node (part/subpart) in the graph, S_v denotes the location of node v, and $ch(v)$ denotes the children of node v. $\phi(\cdot)$ models the interaction with the image, and $\psi(\cdot)$ captures the part-subpart spatial relations.

For the leaf node v (oriented edgelet), the unary potential function is

$$\phi(S_v, \mathbf{I}) = \phi^{\text{edge}}(S_v, \mathbf{I}) + \phi^{\text{app}}(S_v, \mathbf{I}). \quad (2)$$

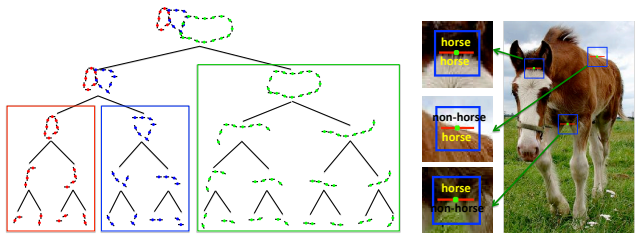
The first term $\phi^{\text{edge}}(S_v, \mathbf{I})$ characterizes how well the orientation at location S_v in the image matches the model orientation. The second term $\phi^{\text{app}}(S_v, \mathbf{I})$ captures the local appearance information at location S_v, i.e., which side of the leaf node is object side, and which side is non-object (background) side, as shown in Figure 1 (b). For the non-leaf node v, the unary term $\phi(S_v, \mathbf{I})$ indicates the confidence of part v being at location S_v. The confidence score can be from some part detection algorithm for animals.

Given an image, the goal of inference is to find the best mixture (i.e. the best viewpoint and pose) and specify locations of all the part/subparts. Specifically, for each mixture, we solve the minimization problem (1) by standard dynamic programming. And then we select the mixture with the minimal energy as the best mixture. The standard dynamic programming requires quadratic complexity $O(|\mathcal{D}|^2)$, where $\mathcal{D} = \{1,..,H\} \times \{1,..,W\}$ is the image grid. In this paper, we give the constrained generalized distance transform (CGDT) algorithm which achieves linear complexity with little accuracy sacrifice. Detailed description is in Section 4 of the paper.

Structure learning refers to learning the hierarchical graph to represent the animal and part shapes under various poses and viewpoints. As for the leaf nodes, we consider eight orientations which are equally distributed from 0 to π, and three polarity values for each orientation which represent object region on one side, object region on the other side, and object region on both sides respectively, as shown in Figure 1 (b). Note that leaf nodes are shared across different mixtures.

We use compositional models to represent big semantic parts such as head, neck and torso. The structure learning algorithm proceeds in the following four steps.

1. Clustering: Given part-level annotations, we extract the masks for head, neck and torso. Then we apply the K-medoids clustering algorithm to find K representative shapes from the training data. And we build K compositional mixtures based on the K representative shapes.



(a) (b)

Figure 1: (a) Illustration of compositional model for a particular horse shape. Red for head, blue for neck and green for torso. Due to space limitation, the leaf nodes (oriented edgelet of eight orientations) are not shown. (b) Three types of polarity value for a leaf node with a horizontal orientation. Green dot represents center location and red line segment represents orientation. Best viewed in color.

2. Sampling: We evenly sample fixed number of landmarks along the boundary of each semantic part.

3. Matching: We match each landmark to one of the 24 leaf nodes.

4. Composing: Starting from the landmarks (leaf nodes), we compose each two adjacent nodes (children) into a higher-level node (parent) and record the spatial relation between the two children nodes. The parent location is the average of two children locations. We run this procedure level-by-level up to the top level.

As for the parameter learning, we adopt latent SVM for learning the model parameters. These parameters strike a balance between the prior shape, appearance cues, orientation confidence and part confidence. Considering the extremely high variability of animal legs, we take a coarse-to-fine approach to segment legs. Specifically, after segmenting the animal body (head, neck, torso), we can narrow down the search region for legs since we know that most of the time the legs appear underneath the torso.

We use a newly annotated dataset on Pascal VOC 2010 to evaluate our part segmentation algorithm. Table 1 shows the quantitative results of our method against several comparison baselines. More experimental results are in Section 7 of the paper.

Method	head	neck+torso	leg
Our model	41.55	60.98	30.98
PD+OS	26.77	53.79	11.18
Mask+OS	33.19	56.69	11.31
PD+GT	38.66	60.63	19.36

Table 1: Part segmentation result for cows. The performance measure is IOU (%). PD+OS refers to the method that combines part detection bounding box and object segmentation (first baseline). Mask+OS refers to the method that uses oracle mask selection and object segmentation (second baseline). PD+GT refers to the oracle method that combines part detection bounding box and groundtruth segmentation.

[1] Peter N Belhumeur, David W Jacobs, D Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[2] J Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

[3] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[4] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.