**Paper-1:**

**Region-based Segmentation and Object Detection**

**Stephen Gould**1 **Tianshi Gao**1 **Daphne Koller**2

1 Department of Electrical Engineering, Stanford University

2 Department of Computer Science, Stanford University

**1 Introduction**

Object detection and multi-class image segmentation are two closely related tasks

that can be greatly improved when solved jointly by feeding information from

one task to the other. However, current state-of-the-art models use a

separate representation for each task making joint inference clumsy and leaving

the classification of many parts of the scene ambiguous.

This work propose a hierarchical region-based approach to joint object

detection and image segmentation. This approach simultaneously reasons about

pixels, regions and objects in a coherent probabilistic model. Pixel appearance

features allow us to perform well on classifying amorphous background classes,

while the explicit representation of regions facilitate the computation of more sophisticated

features necessary for object detection. Importantly, this model gives

a single unified description of the scene—we explain *every* pixel in the image and

enforce global consistency between all random variables in our model.

We can run experiments on the challenging Street Scene dataset and show significant

improvement over state-of-the-art results for object detection accuracy.

This work proposes a more integrated region-based approach that combines multi-class image

segmentation with object detection. Specifically, it proposes a hierarchical model that reasons

simultaneously about pixels, regions and objects in the image, rather than scanning arbitrary windows.

At the region level the pixels are labelled as belonging to one of a number of background classes

(currently *sky*, *tree*, *road*, *grass*, *water*, *building*, *mountain*) or a single foreground class. The foreground

class is then further classified, at the object level, into one of the known object classes

(currently *car* and *pedestrian*) or *unknown*.

**2 Background and Related Work**

This method inherits features from the sliding-window object detector works, such as Torralba et al. and Dalal and Triggs , and the multi-class image segmentation work of Shotton et al. .

The model also has many novel ideas for improving object detection via scene

context. The innovative works that inspire us include predicting camera viewpoint for estimating

the real world size of object candidates , relating “things” (objects) to nearby “stuff” (regions)

, co-occurrence of object classes , and general scene “gist” .

Current models are not tightly coupled and may result in incoherent outputs (e.g., the pixels in

a bounding box identified as “car” by the object detector, may be labeled as “sky” by an image

segmentation task). In this method, all tasks use the same region-based representation which forces

consistency between variables. Intuitively this leads to more robust predictions.

The decomposition of a scene into regions to provide the basis for vision tasks

**3 Region-based Model for Object Detection**

This model presents an overview of our joint object detection and scene segmentation model. This model

combines scene structure and semantics in a coherent energy function.

This work will allow an object to be composed of many regions (rather than trying to force dissimilar

regions to merge). The object to which a region belongs is denoted by its object-correspondence

variable Or ∈ {ᵠ, 1, . . . ,N}. Some regions, such as background, do not belong to any object

which we denote by Or = ᵠ. Like regions, the set of pixels that comprise the o-th object is denoted

by Po = Ur Or=o Pr.

Random variables are associated with the various entities (pixels, regions and objects) in this

model. Each pixel has a local appearance feature vector αp ∈ Rn (see [7]). Each region has an

appearance variable Ar that summarizes the appearance of the region as a whole, a semantic class

label Sr (such as “road” or “foreground object”), and an object-correspondence variable Or. Each

object, in turn, has an associated object class label Co (such as “car” or “pedestrian”). The final

component in our model is the horizon which captures global geometry information. Assume

that the image was taken by a camera with horizontal axis parallel to the ground and model the

horizon Vvhz ∈ [0, 1] as the normalized row in the image corresponding to its location. Quantize

vvhz into the same number of rows as the image.

The variables are combined in this model into a single coherent energy function that captures the

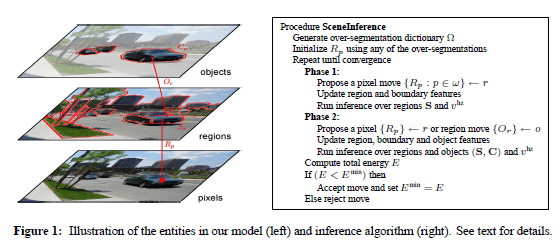
structure and semantics of the scene. The energy function includes terms for modelling the location

of the horizon, region label preferences, region boundary quality, object labels, and contextual relationships

between objects and regions. These terms are described in detail below. The combined

energy function E(R, S,O,C, vhz | I, θ) has the form:





**4 Object Detectors**

Here in addition to raw appearances feature vector φo feature derived from object detection models is appended to the object

In this approach, they treat the object detector as a black-box that returns a score per

candidate window. In the black-box approach, they naively place a bounding

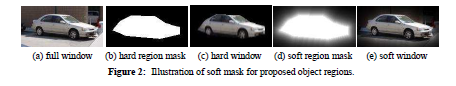
box (at the correct aspect ratio) around these pixels and classify the entire contents of the box. To

make classification more robust they search candidate windows in a small neighborhood (defined over

scale and position) around this bounding box, and take as the feature the output of highest scoring

window.

They also mask out all pixels not belonging to the object. In the implementation, they use a soft mask that attenuates the intensity of pixels outside the object based on their distance to the object boundary. This has the dual advantage of preventing hard edge artifacts and being less sensitive to segmentation errors.



**5 Experiments**

We can conduct experiments on the challenging Street Scene dataset .A dataset consisting of

3547 high-resolution images of urban environments .The image can be rescaled to 320 × 240 before

running our algorithm. The dataset comes with hand-annotated region labels and object boundaries.

However, the annotations use rough overlapping polygons, so we use Amazon’s Mechanical Turk

to improve the labeling of the background classes only. We keep the original object polygons to be

consistent with other results on this dataset.

The dataset are divided into five folds—the first fold (710 images) can be used for testing and the

remaining four can be used for training. The multi-class image segmentation component of this model

is said to achieves an overall pixel-level accuracy of 84.2% across the eight semantic classes.

Desired output



**6 Limitation**

One of the difficulties in this model is learning the trade-off between energy terms—too strong a

boundary penalty and all regions will be merged together, while too weak a penalty and the scene

will be split into too many segments