Semantic Scene Segmentation using Random Multinomial Logit

Ananth Ranganathan
http://www.ananth.in

Honda Research Institute, USA
Cambridge, MA, USA

Semantic segmentation or scene analysis involves labeling all the components in an image, usually at the pixel level. While recent work in this area has led to exciting results [1, 3], existing techniques face difficulties in dealing with object categories with wide intra-class variation. In addition, real-time operation is also a distant dream.

We are interested in the task of classifying street scenes for use in intelligent transportation systems. This involves detecting the road, other vehicles, and pedestrians to alert the user in potentially dangerous situations. For instance, the detection of fast moving vehicles, such as motorbikes, would help decrease the number of accidents significantly.

In this paper, we address these challenges by introducing Random Multinomial Logit (RML), a general purpose classifier, and applying it to the task of texture-based scene segmentation. RML consists of an ensemble of multinomial logistic regression models, each of which operates on a randomly selected subset of features and outputs a probability distribution on the label of the pixel corresponding to the features. The use of a bagging framework overcomes the inability of multinomial logistic regression to operate in large feature spaces.

Random Multinomial Logit Classifiers

An RML classifier consists of N multinomial logistic regression models, each of which models the probability distribution of the label y given the input vector x as

$$\pi_{il} = p_i(y=l|\mathbf{x},\beta_i) = \begin{cases} \exp\left(\beta_{il0} + \sum_{f=1:M} \beta_{ilf}\phi_f(\mathbf{x})\right)/Z, & l = 1:L-1 \\ 1/Z, & l = L \end{cases} \quad (1)$$

where i and l are indices into the model and label set respectively, and Z is the normalizing constant that makes the distribution sum to unity. The $\phi(.)$ are feature functions computed on the input vector $\mathbf{x}$, and β_{il} is the vector of coefficients of length $(L-1)$ that define the detection function for object category l.

Training for the RML classifier, which involves learning the β coefficients from data, proceeds in a manner similar to many other ensemble methods. The labeled training set is sampled with replacement to get N smaller sets using which the individual regression models are learned. The features for the individual models are also selected randomly, with M features per model. The final output label distribution of the RML is computed by averaging over the output of the individual models

$$\hat{\pi}_l = \sum_{i=1:N} \pi_{il} \quad (2)$$

The coefficients β for the individual regression models are learned in a maximum a posteriori (MAP) framework with a L2 regularizing prior on the model parameters β.

Our image representation uses texture-layout features [2], which consist of a rectangle r and a texton word t. For every pixel p, the feature computes the proportion of the texton word t inside the rectangle r, where r has been translated to be in the coordinate system with p as the origin. Texture-layout features capture local textural context in the image, for instance the relationship that a boat is usually surrounded by water. In addition, this contextual relationship, expressed as a linear combination of multiple texture-layout feature values, is sufficient to do pixel-wise scene labeling [2].

An illustration of the RML training process is given in Figure 1.

Feature Selection

As a secondary contribution, we describe an algorithm for feature selection that, in essence, performs a random search through feature space to find a statistically significant set of features. Feature selection is performed to improve the quality of each regression model and reduce the total number of models required.

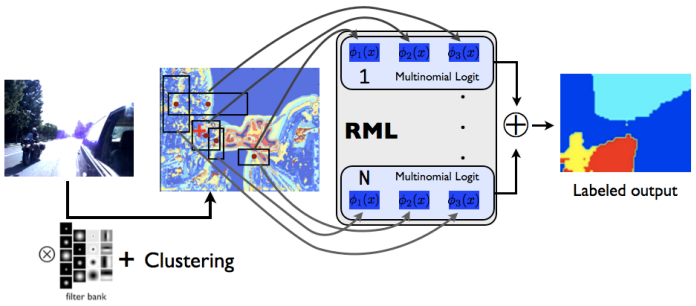

Figure 1: An overview of our scene segmentation system using RML: training images are convolved with a filter bank and clustered to yield texton images, on which texture-layout features operate. Five texture-layout features are shown in the second box from the left. Subsets of such features are randomly selected and passed into an ensemble of multinomial logistic regression models (here having three features each), the outputs from which are averaged to yield the final semantic segmentation.

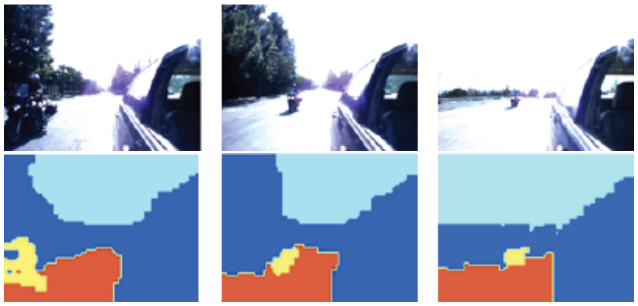

Figure 2: Representative results on a few frames of the traffic scene dataset with sky labeled in light blue, road in red, bike in yellow, and background in dark blue.

Feature selection improves performance since many of the features picked randomly will not have a bearing on the labeling decision. For example, the shape of a texture-layout filter may be too big or too small to provide any useful information. These features can be weeded out using a simple test based on the significance of the corresponding β coefficients.

A feature does not contribute in the regression model (1) if the column of coefficients corresponding to it are all extremely small. This can be determined in a scale-independent manner by comparing the coefficients with their standard deviation. The statistically insignificant features are then swapped out and the model is re-learnt.

Experimental Verification

We evaluate RML on two datasets - the 20-class VOC 2008 dataset and an extremely challenging real-world video dataset of traffic scenes with large illumination, perspective, and intra-class variation. Comparisons with recent techniques such as TextonBoost[2] on the latter dataset demonstrate RML as being state of the art, and advancing it in many cases. A sample result on the traffic scene dataset is shown in Figure 2.

[1] G. Csurka and F. Perronnin. A simple high performance approach to semantic segmentation. In *British Machine Vision Conf. (BMVC)*, 2008.

[2] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *Intl. J. of Computer Vision*, 81(1):2 –, 2009.

[3] E. Sudderth, A. Torralba, W. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *Intl. J. of Computer Vision*, March 2008.