Using Inference Machines for Perception Tasks

Daniel Munoz DMUNOZ@RI.CMU.EDU
Alexander Grubb AGRUBB@CMU.EDU
J. Andrew Bagnell DBAGNELL@RI.CMU.EDU
Martial Hebert HEBERT@RI.CMU.EDU
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA USA

Abstract

Over the past few years we have developed the *inference machine* framework for addressing structured prediction problems arising in computer vision applications. This workshop paper serves to summarize our recent and ongoing work for scene parsing and human pose estimation.

1. Introduction

Many applications in computer vision can be framed as a form of structure prediction. An important problem is scene parsing, *i.e.*, semantic classification of all objects in an observed scene, as illustrated in Fig. 1. A prevalent method to encode structure/relations in the prediction is with a joint probabilistic or energy-based model which enables one to naturally write down these interactions (Lafferty et al., 2001; Taskar et al., 2003). Unfortunately performing inference over these expressive models leads to an NP-hard optimization problem which must be approximated and, consequently, poses theoretical and empirical difficulties when learning the model (Kulesza & Pereira, 2007; Finley & Joachims, 2008). Furthermore, using approximate inference on any learned model often leads to suboptimal predictions due to the approximations. As we ultimately care about predicting the correct labeling of an environment, and not necessarily learning a joint model of the data, we instead view the approximate inference process as a modular procedure that is directly trained in order to produce correct labelings, inspired by work in natural language processing (Cohen & Carvalho, 2005; Daume III et al., 2009). That is, we can view an iterative inference algorithm, such as message passing in a factor graph, as a network of computational modules taking in observations and other local computations on the graph (messages). We can then iteratively train each of these modules to output ideal intermediate messages, culminating in a holistic interpretation of the scene. In the following, we demonstrate that this iterative decoding approach achieves state-of-the-art performance on a variety of image and 3-D point cloud datasets while also being extremely computationally efficient in practice.

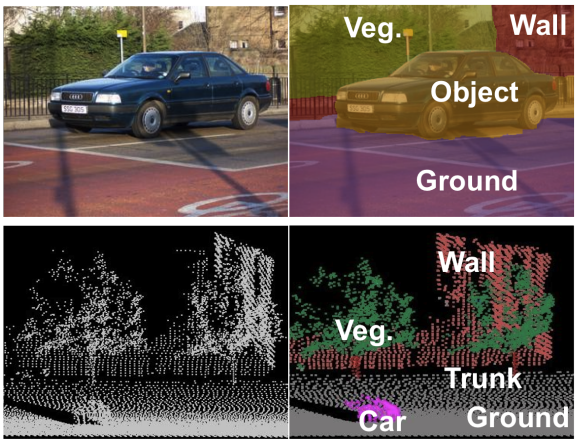

Figure 1. Examples inferred scene parsings in images (top) and 3-D point clouds (bottom).

2. Scene Parsing

2.1. Images

The following summarizes work originally presented at ECCV 2010 (Munoz et al., 2010) for parsing images, with using a more efficient and accurate implementation.

A synthetic illustration of the inference procedure is illustrated in Fig. 2. Given an image, we first create a hierarchy of regions that range from very large regions

Presented at the International Conference on Machine Learning (ICML) workshop on *Inferning: Interactions between Inference and Learning*, Atlanta, Georgia, USA, 2013.

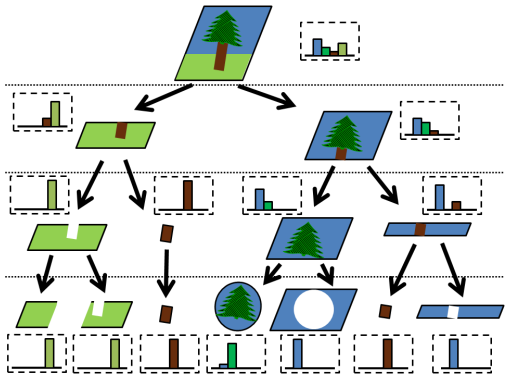

Figure 2. Hierarchical representation

	Segmentation	Features	Inference
Time (s)	0.095	0.462	0.037

Table 1. Average computation timings on SBD for each component of the entire inference procedure.

	VMR-Oakland	Freiburg
Number of 3-D points	44,198	452,330
Total time (s) w/ F-H	0.794	3.15
Total time (s) w/ Grids	**0.215**	**0.597**

Table 2. Average processing timings for processing 3-D point clouds in two datasets when using a region hierarchy formed using F-H segmentation and gridded partitioning (Hu et al., 2013).

in the image (potentially including the image itself as one region at the top) down to small regions (*e.g.*, superpixels) at the bottom; we use a standard graph-based segmentation algorithm (Felzenszwalb & Huttenlocher, 2004) to create this region hierarchy. We represent the inference process as a series of predictions that traverses over the regions in the hierarchy. It is important to note that we do not expect each region to contain one class/label. Instead, we explicitly model the distribution of object classes within in each region. That is, given the region's computed feature descriptor, we train a predictor to match the empirical ground truth distribution of object categories contained within each region. In practice, we train a multi-class MaxEnt model to predict the per-region probability distributions.

We initialize the inference procedure by making an initial prediction at the bottom level of this hierarchy. These predictions are passed in an bottom-up manner to the subsequent coarser levels and used as additional features in the MaxEnt predictors. Similarly, predictions are refined as the procedure traverses top-down to the child level in the hierarchy. Since we model label proportions over regions: we are robust to imperfect segmentation, we can use features defined over large regions, and we do not make hard commitments during inference. Furthermore, since we are no longer attempting to model a join probability distribution, we can encode arbitrary interactions/predictions among the regions in the scene, and the entire inference procedure is a deterministic sequence of efficient MaxEnt predictions.

We evaluate our approach on the popular Stanford Background Dataset (SBD) (Gould et al., 2009), which contains 8 classes, and the Cambridge Video Dataset (CamVid) (Brostow et al., 2008), which contains 11 classes; we follow the same training/testing evaluation procedures as originally described in the respective works.

Table 1 breaks down the computation timings for SBD of the three main components of classification: 1) hierarchical image segmentation, 2) feature computation time for all regions, 3) inference over the hierarchy via MaxEnt predictions. All times were computed on a 4-core Intel i7-2960XM processor. The table demonstrates the feature extraction dominates the current pipeline with the actual MaxEnt predictions constituting a small portion of the time. In Table 3, we demonstrate that our approach achieves state-of-the-art classification performance.

2.2. 3-D Point Clouds

As regions are the atomic elements of how the data is represented, our inference procedure is invariant to the underlying form of the data, *e.g.*, it does not assume an organized lattice structure of the data. Our only requirements are that regions can be extracted from the data and that discriminative feature representations can be computed. Hence, parsing scenes in 3-D data can be analogously performed under this framework by extracting and operating over 3-D regions.

In work presented at ICRA 2011 (Xiong et al., 2011), we constructed this 3-D hierarchical segmentation again using the F-H graph-based segmentation technique (Felzenszwalb & Huttenlocher, 2004). Classification performances on the VMR-Oakland (Xiong et al., 2011) and GML-PCV (Shapovalov et al., 2010) are shown in Table 4.

Recently we have demonstrated improved efficiency in 3-D classification by using a much simpler representation of the scene (Hu et al., 2013). Instead of using a precise segmentation algorithm, such as F-H, that attempts to obey borders/discontinuities, we observed that using multiple, inexact partitions/grids of the en-

	Sky	Tree	Road	Grass	Water	Bldg.	Mtn.	Object	Avg.	Pixel
(Farabet et al., 2013)	95.7	78.7	88.1	89.7	68.7	79.9	44.6	62.3	76.0	81.4
(Socher et al., 2011)	-	-	-	-	-	-	-	-	-	78.1
Inference Machine	92.4	76.6	90.5	81.7	68.2	82.8	13.2	68.8	71.8	81.6

	Bldg.	Tree	Sky	Car	Sign	Road	Ped.	Fence	Pole	Sdwlk.	Bike	Avg.	Pixel
(de Nijs et al., 2012)	59	75	93	84	45	90	53	27	0	55	21	54.7	75.0
(Ladicky et al., 2010)†	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8
Inference Machine	83.5	85.1	94.5	78.3	41.7	95.5	38.7	18.0	15.2	78.3	36.2	60.5	85.7

Table 3. Accuracies on 2-D datasets `SBD` (top) and `CamVid` (bottom) where *Avg.* is the average class accuracy and *Pixel* is the per-pixel accuracy. †Uses additional training data not leveraged by other techniques.

vironment is able achieve much faster classifications without any loss in accuracy. Table 2 demonstrates the improvement in efficiency of the entire inference pipeline on the `VMR-Oakland` and `Freiburg` (Behley et al., 2012) datasets.

2.3. Image + 3-D Data

Inexpensive sensors returning both image and depth data, *e.g.*, the Microsoft Kinect, have re-spurred interest in multi-modal data processing. Some sensors may observe readings that have close to one-to-one correspondence between modalities. In such cases, it is natural to restrict the representation to a single modality and incorporate feature statistics computed over the other. However, in general (and is typically the case in mobile robotics), there is a severe discrepancy and often many-to-one correspondences between the data sources. In work presented at ECCV 2012 (Munoz et al., 2012), we demonstrate that it is favorable to not restrict the representation to one modalities and to instead treat both modalities as first-class objects by iterating predictions over both hierarchies simultaneously, *i.e.*, co-inference, as illustrated in Fig. 3.

3. Pose Estimation

Another challenging vision problem is estimating human body pose, *i.e.*, predicting individual limb/joint locations, from a single image. As the locations of joint locations are highly correlated this is another natural structured prediction problem. In joint work with Varun Ramakrishna, we have applied the inference machine framework to this problem to jointly decode part locations. An example parsing is illustrated in Fig. 4, and we plan to present more thorough results at the workshop.

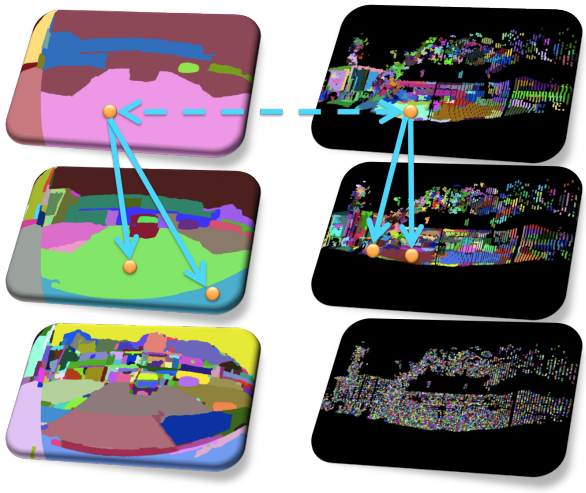

Figure 3. Illustration of the co-inference procedure. The region hierarchies are constructed separately in the image (left) and 3-D point cloud (right) domains with predictions propagated across and within domains.

References

Behley, J., Steinhage, V., and Cremers, A.B. Performance of histogram descriptors for the classification of 3d laser range data in urban environments. In *ICRA*, 2012.

Brostow, Gabriel J., Shotton, Jamie, Fauqueur, Julien, and Cipolla, Roberto. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008.

Cohen, William W. and Carvalho, Vitor R. Stacked sequential learning. In *IJCAI*, 2005.

Daume III, Hal, Langford, John, and Marcu, Daniel.

	Wire	Pole	Ground	Veg.	Trunk	Bldg.	Vehicle	Avg.
(Munoz et al., 2009)	72	63	99	94	30	92	43	70.4
Inference Machine	75	67	98	93	41	93	74	77.3

	Ground	Bldg.	Tree	Bush	Car	Avg.
(Shapovalov et al., 2010)	96	58	99	9	16	55.6
Inference Machine	98	77	98	36	10	63.8

	Ground	Bldg.	Tree	Bush	Avg.
(Shapovalov et al., 2010)	98	81	89	57	81.3
Inference Machine	99	92	97	52	85.0

Table 4. Accuracies on 3-D datasets `VMR-Oakland` (top), `GML-PCV-A` (bottom-left), and `GML-PCV-B` (bottom-right).

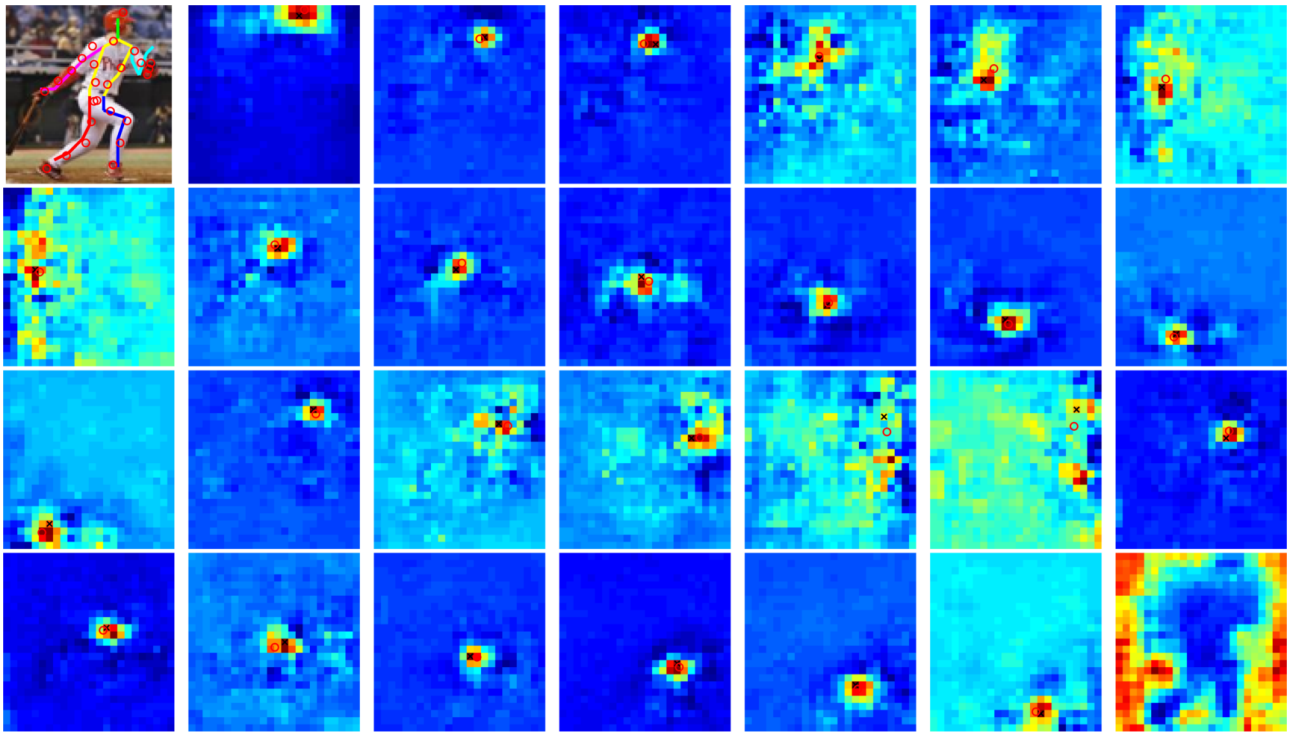

Figure 4. Human pose estimation with inference machines. In the top-left image, the colored lines trace the inferred joint locations, whose ground truth locations are indicated by the red circles. The remaining images are predicted score maps (red = high score, blue = low score) for each joint location at each position in the image, with the lower-right image denoting the background class.

Search-based structured prediction. *MLJ*, 75(3), 2009.

de Nijs, Roderick, Ramos, Sebastian, Roig, Gemma, Boix, Xavier, Van Gool, Luc, and Kuhnlenz, Kolja. On-line semantic perception using uncertainty. In *IROS*, 2012.

Farabet, Clement, Couprie, Camille, Najman, Laurent, and LeCun, Yann. Learning hierarchical features for scene labeling. In *T-PAMI*, 2013.

Felzenszwalb, Pedro F. and Huttenlocher, Daniel P. Efficient graph-based image segmentation. *IJCV*, 59(2), 2004.

Finley, Thomas and Joachims, Thorsten. Training structural svms when exact inference is intractable. In *ICML*, 2008.

Gould, Stephen, Fulton, Richard, and Koller, Daphne. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

Hu, Hanzhang, Munoz, Daniel, Bagnell, J. Andrew, and Hebert, Martial. Efficient 3-d scene analysis from streaming data. In *ICRA*, 2013.

Kulesza, Alex and Pereira, Fernando. Structured learning with approximate inference. In *NIPS*, 2007.

Ladicky, Lubor, Sturgess, Paul, Alahari, Karteek, Russell, Chris, and Torr, Philip H.S. What, where & how many? combining object detectors and crfs. In *ECCV*, 2010.

Lafferty, J., McCallum, A., and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

Munoz, Daniel, Bagnell, J. Andrew, Vandapel, Nicolas, and Hebert, Martial. Contextual classification with functional max-margin markov networks. In *CVPR*, 2009.

Munoz, Daniel, Bagnell, J. Andrew, and Hebert, Martial. Stacked hierarchical labeling. In *ECCV*, 2010.

Munoz, Daniel, Bagnell, J. Andrew, and Hebert, Martial. Co-inference for multi-modal scene analysis. In *ECCV*, 2012.

Shapovalov, Roman, Velizhev, Alexander, and Barinova, Olga. Non-associative markov networks for 3d point cloud classification. In *Photogrammetric Computer Vision and Image Analysis*, 2010.

Socher, Richard, Lin, Cliff, Ng, Andrew Y., and Manning, Christopher D. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011.

Taskar, B., Guestrin, C., and Koller, D. Max-margin markov networks. In *NIPS*, 2003.

Xiong, Xuehan, Munoz, Daniel, Bagnell, J. Andrew, and Hebert, Martial. 3-d scene analysis via sequenced predictions over points and regions. In *ICRA*, 2011.