Unsupervised Semantic Scene Labeling for Streaming Data

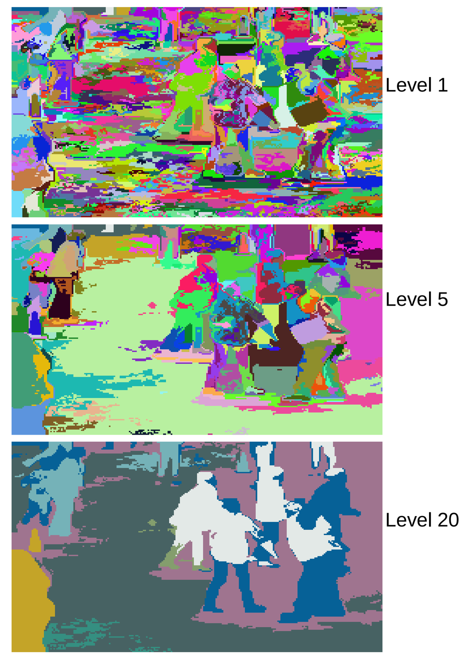
Maggie Wigness and John G. Rogers III
US Army Research Laboratory
maggie.b.wigness.civ@mail.mil, john.g.rogers59.civ@mail.mil

Motivation & Objectives

- Visual perception and semantic segmentation provide intelligent systems with information necessary to accomplish higher level tasks
- Shortcomings of state-of-the-art deep learning semantic labeling [1,2,3]
  - Large training sets requires significant human effort
  - Unable to discover novel concepts in streaming data
  - Often a domain mismatch between test environment and training data
- **Develop an unsupervised semantic scene labeling (USSL) approach that can learn from small sets of data on-line without human oversight to continuously model and discover novel concepts in a data stream**

Unsupervised Learning Challenges

- Parameter selection is difficult if number/types of concepts are unknown
- Changing visual properties in long data streams, e.g., illumination, weather
- Existing unsupervised video segmentation [4,5] side step these issues with hierarchical output and coherent region modeling, not semantics

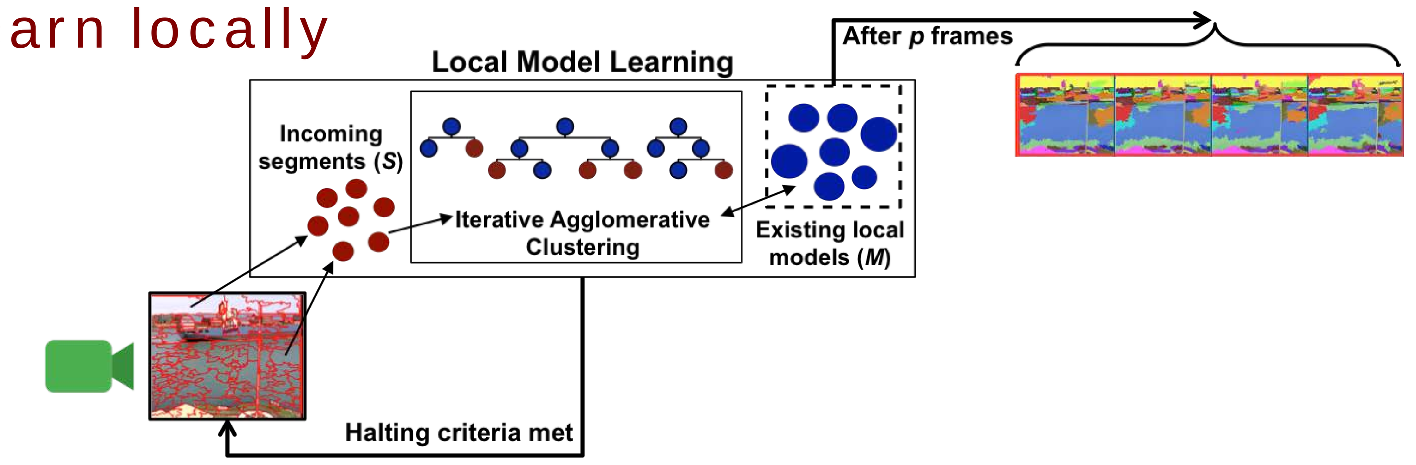


Above: Example segmentation output from our approach (USSL) and GBH [5]. USSL has semantic consistency in non-adjacent pixels, e.g., traffic cones. Right: Hierarchical output from GBH, requiring the user to select the best level, i.e., parameters, for the task.

Approach

- ***Over-learn locally*** in the data stream to minimize unsupervised noise
- Create an ***ensemble of local learners*** to create a better global output

Over-learn locally



- Iteratively cluster over-segmented superpixels from data stream frames
- Given superpixels, assume $\exists\, s_i, s_j \in I_i \rightarrow label(s_i) = label(s_j)$
  - Learn a merging threshold, α, from observed similarities *for each feature type* r, e.g., LAB, LBP, SIFT, HOG, etc.

Similarity history

$$H_r = \{S_r(s_i, NN(s_i)), \forall s_i \in \{I_1, \cdots, I_t\}\}$$

Merging Threshold

$$\alpha_r = \mu_{H_r} - \sigma_{H_r}$$

- Compare every m_i with its adjacent regions (build model locally) and k random non-adjacent regions (allow semantic modeling to expand)

Next Merge

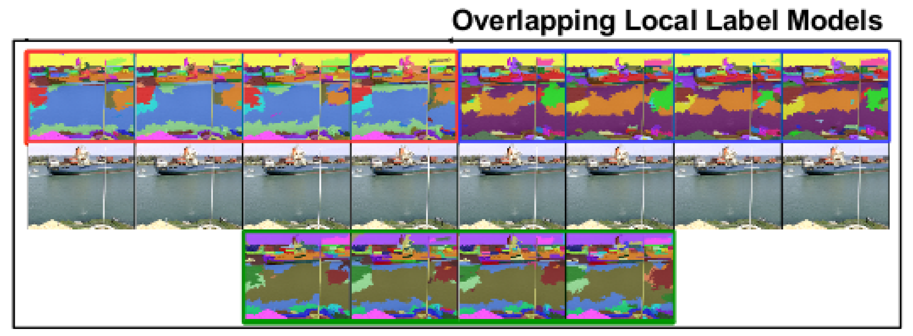
$$\text{Any } m_i, m_j \text{ such that } S_r(m_i, m_j) > \alpha_r, \forall f_r \in f$$

Setting high threshold to merge so local models are still over-learning

Approach

Ensemble of Local Learners

- Overlapping local models
- Graph-based encoding of label overlap
  - V: all m_i in from every W_i
  - E exist for any m_i, m_j that label at least one common pixel
- Cut edges with small weights to reconcile label noise and remaining connected components represent the global label set



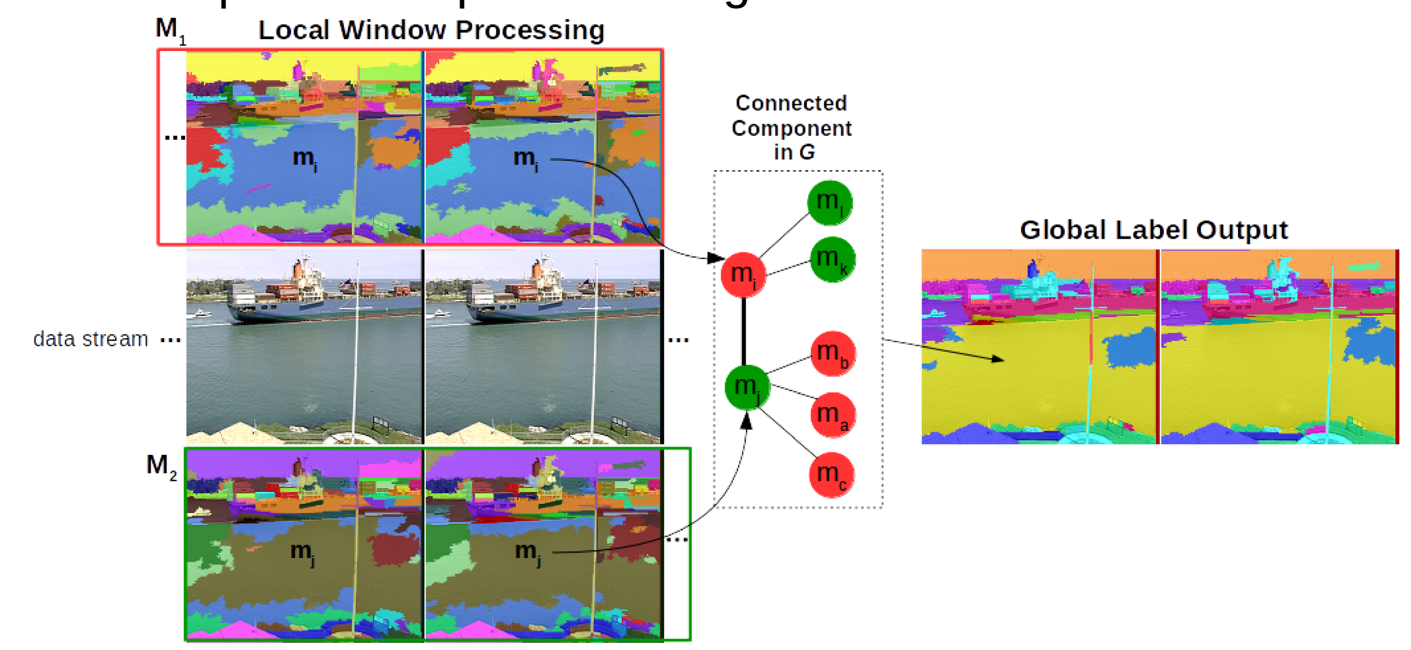
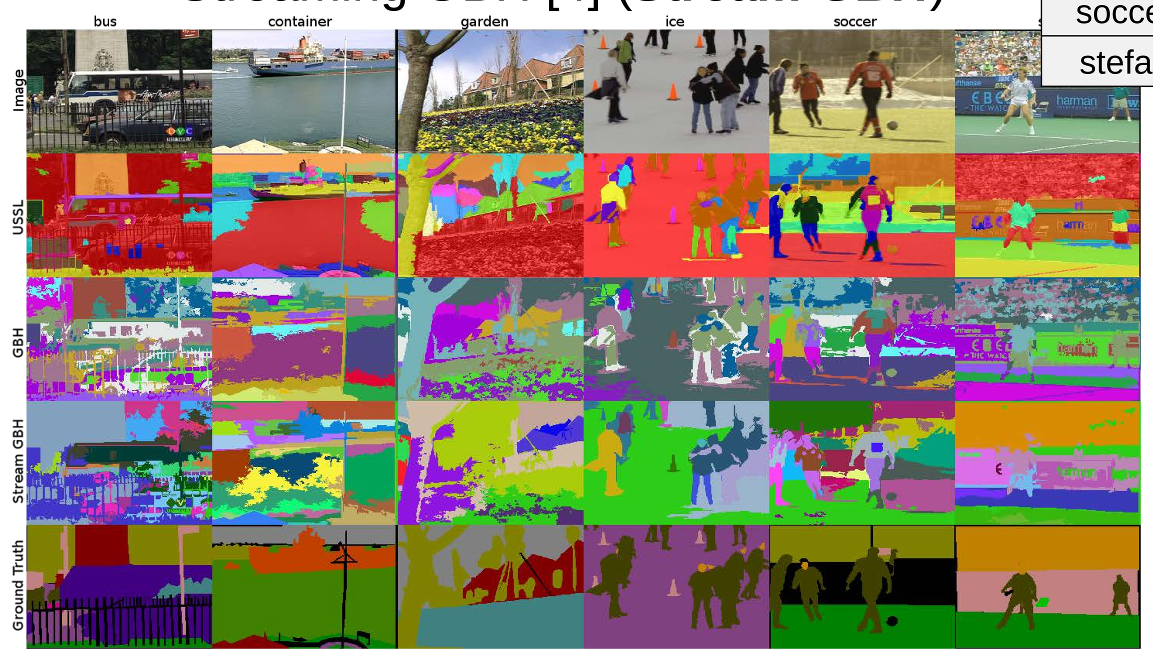


Illustration of the USSL graph-based encoding of overlapping local models

Results

- Evaluation on xiph.org video subset [6]
- Comparisons
  - Hierarchical Graph Based [5] (***GBH***)
  - Streaming GBH [4] (***Stream GBH***)

	Number of Segments			
Video	Ground Truth	**USSL**	GBH	Stream GBH
bus	10	17	19	25
container	7	24	24	31
garden	4	17	17	18
ice	4	11	15	11
soccer	6	19	20	24
stefan	5	13	15	12



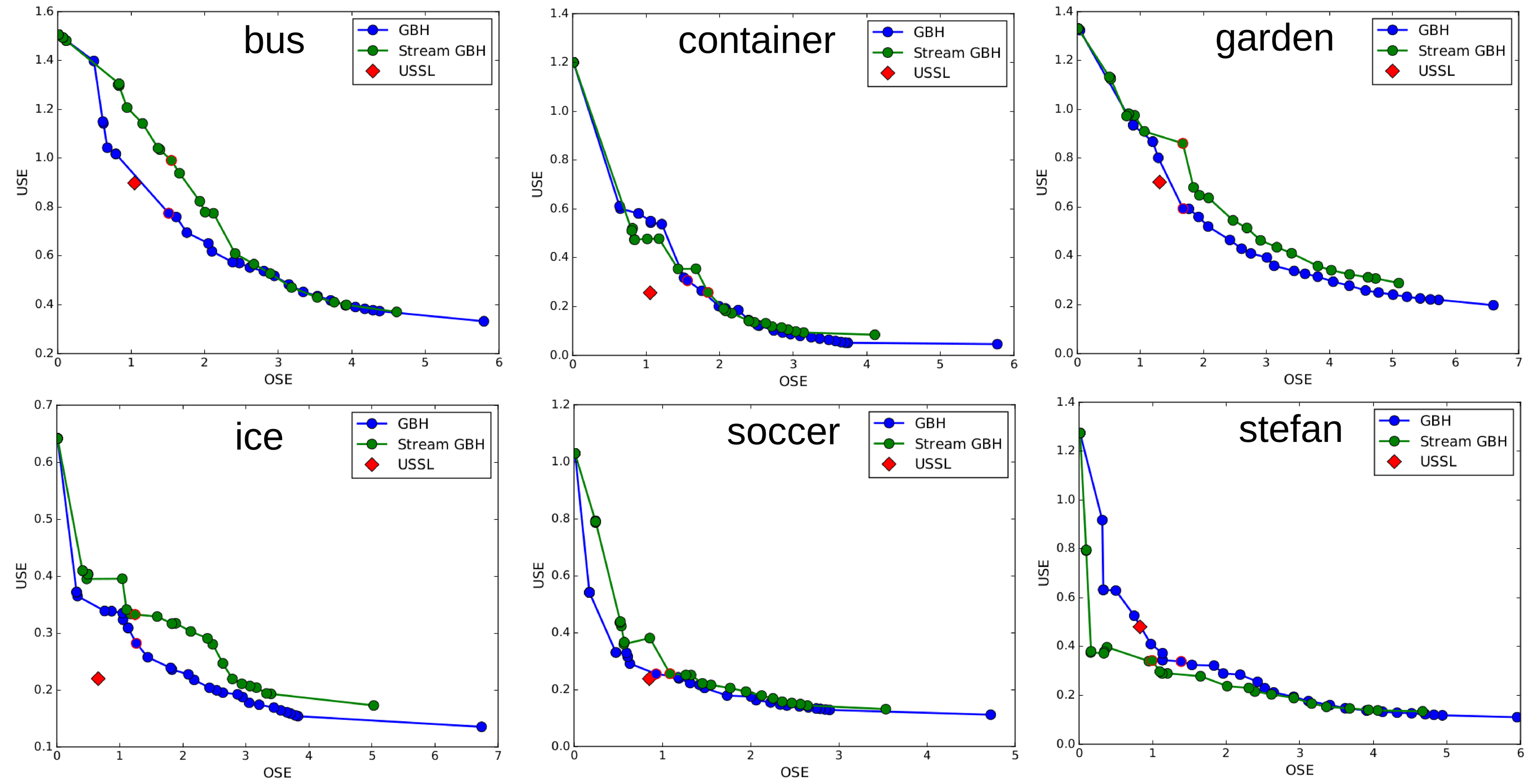
Above: Summary of segmentation output of USSL and the hierarchical graph-based approaches from the level that is most similar to the USSL output. Left: Qualitative comparison of output of each segmentation technique

	Average Per-Class Accuracy			Overall Pixel Accuracy		
Video	**USSL**	GBH	Stream GBH	**USSL**	GBH	Stream GBH
bus	0.294	**0.314**	0.137	0.401	**0.647**	0.370
container	0.613	0.491	**0.641**	**0.907**	0.786	0.855
garden	**0.638**	0.627	0.418	0.686	**0.689**	0.438
ice	**0.628**	0.524	0.534	**0.941**	0.898	0.870
soccer	**0.446**	0.426	0.438	**0.910**	0.876	0.892
stefan	0.544	**0.571**	0.541	0.841	**0.878**	0.837
Average	**0.527**	0.492	0.452	0.781	**0.796**	0.710

Comparison of average per-class accuracy and overall pixel-wise accuracy achieved by USSL and graph-based video segmentation variants.

Results

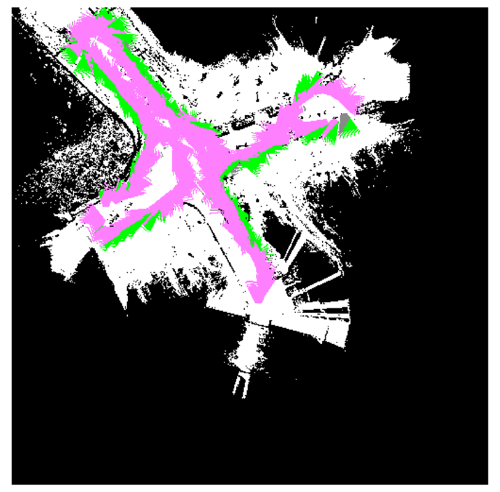
- Under segmentation entropy (USE) vs over-segmentation entropy (OSE)



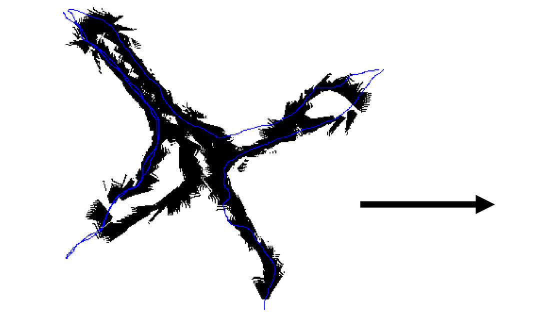
Comparison of over-segmentation and under-segmentation entropy achieved by our USSL approach, which produces a single segmentation output, and hierarchical graph-based approaches, which produce many levels of output (seen by the curve) using changing parameters.

Applications

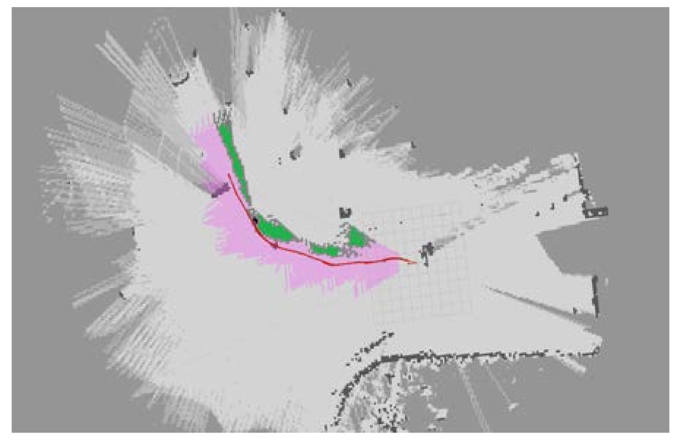
- Autonomous unmanned ground vehicles
  - Provide adaptable visual perception
  - Use modeling uncertainty to guide exploration in a new environment
  - Learn terrain traversability cost of unsupervised concepts



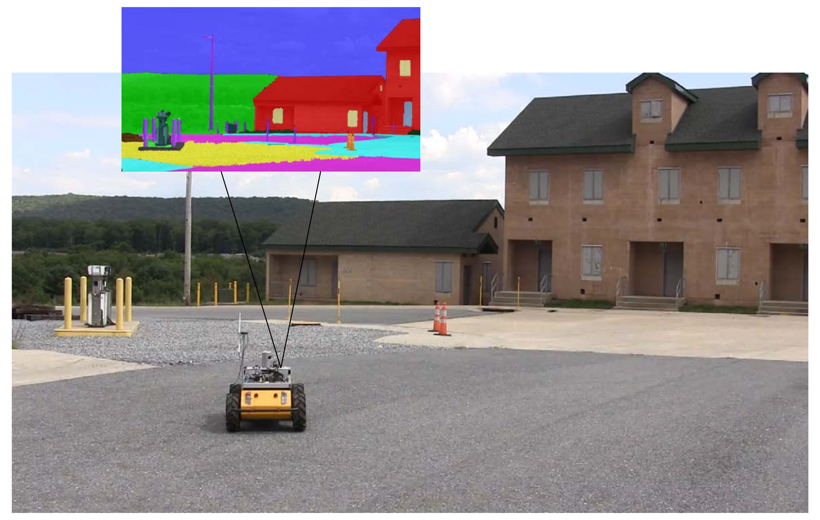
USSL labeling output



Example desired trajectories



Visual perception navigation



References

[1] *Convolutional nets and watershed cuts for real-time semantic labeling of rgbd videos.* C. Couprie, C. Farabet, L. Najman, and Y. Lecun. The Journal of Machine Learning Research, 2014

[2] *Learning hierarchical features for scene labeling.* C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Transactions on Pattern Analysis and Machine Intelligence, 2013.

[3] *Fully convolutional networks for semantic segmentation*. J. Long, E. Shelhamer, and T. Darrell. Computer Vision and Pattern Recognition, 2015

[4] *Streaming hierarchical video segmentation*. C. Xu, C. Xiong and J.J. Corso. European Conference on Computer Vision, 2012.

[5] *Efficient hierarchical graph-based video segmentation*. M. Grundmann, V. Kwatra, M. Han and I. Essa. Computer Vision and Pattern Recognition, 2010.

[6] *Propagating multi-class pixel labels throughout video frames.* A.Y. Chen and J.J. Corso. Western New York Image Processing Workshops, 2010.