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 require significant human effort
  - Unable to discover novel concepts in streaming 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

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

Over-learn locally
- Iteratively cluster over-segmented superpixels from data stream frames
- Given superpixels, assume $S_k, S_j \in I \rightarrow label(S_k) = label(S_j)$
- Learn a merging threshold, $a$, from observed similarities for each feature type, e.g., LAB, LBP, SIFT, HOG, etc.

Similarity history
$$ H_t = \{S_k(S_i, N_S(S_i)), \forall S_i \in \{I_1, ..., I_t\} \} $$

Merging Threshold
$$ a_t = \mu _{I_t} - \sigma _{I_t} $$

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

Next Merge
$$ a_n, \forall f \in f $$

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

Ensemble of Local Learners
- Overlapping local models
  - Graph-based encoding of label overlap
    - $V$: all $m_i$ in from every $W_f$
    - $E$ 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

Approach
- Ensemble of Local Learners
  - Overlapping local models
  - Graph-based encoding of label overlap
    - $V$: all $m_i$ in from every $W_f$
    - $E$ 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

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

Comparison 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

References