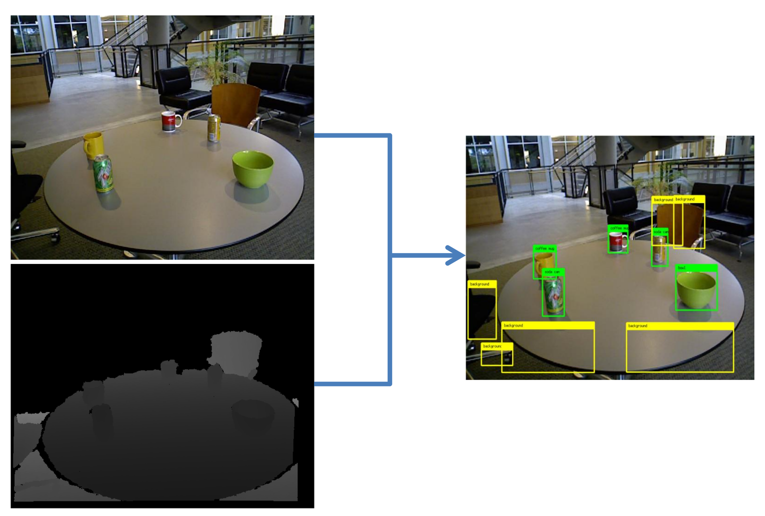
RGB-D Multi-View Object Detection with Object Proposals and Shape Contexts

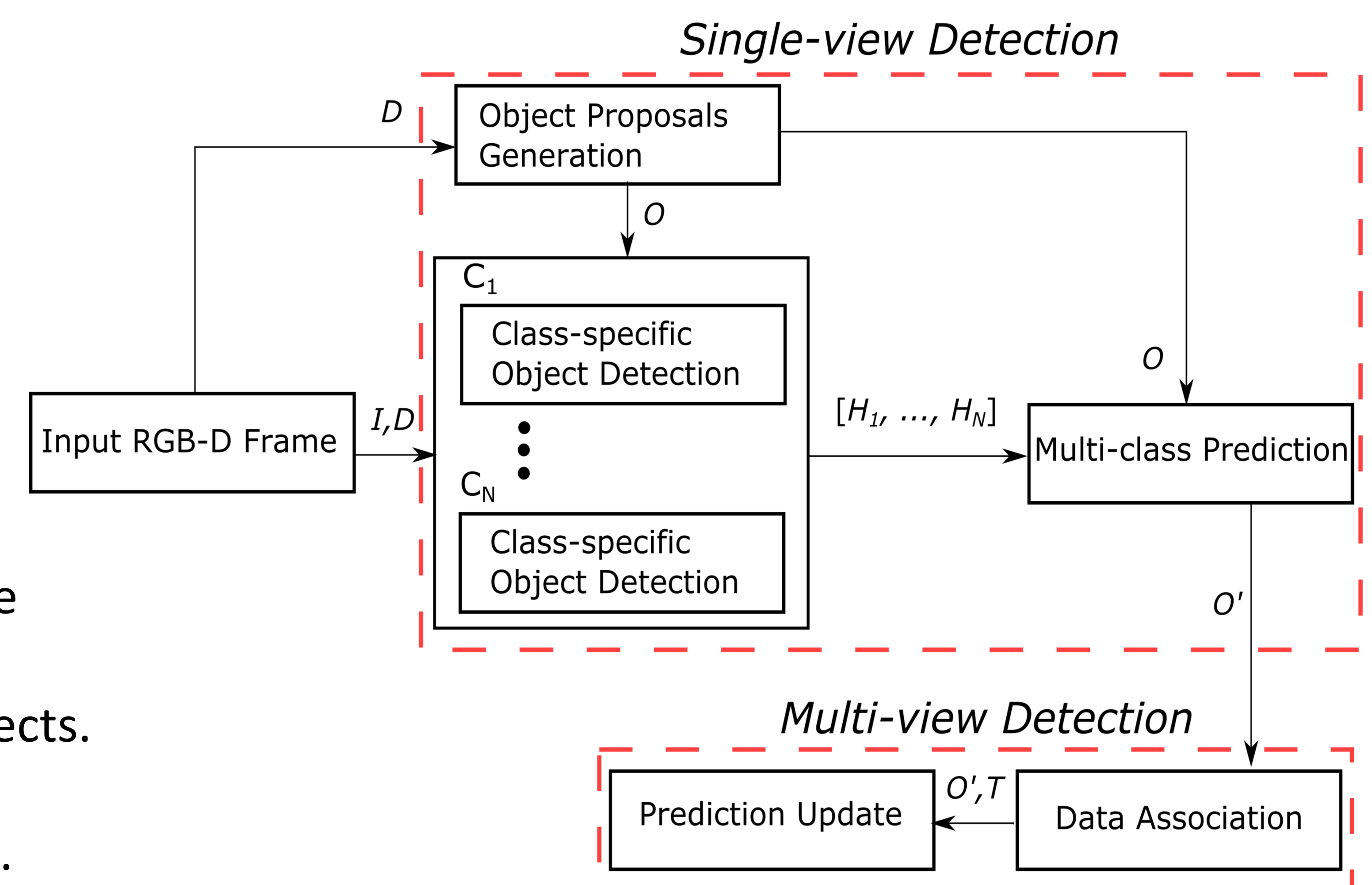
Georgios Georgakis, Md. Alimoor Reza, and Jana Kosecka
George Mason University

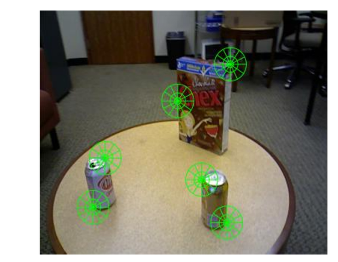
Problem and Approach

• Multi-View Object Detection in RGB-D indoor table-top scenes.

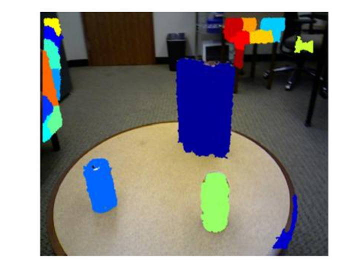


Contributions:
• Resolve limitations of single-view detections such as occlusion or view-dependent ambiguities by integrating evidence from multiple views.
• Extract Shape Contexts on depth discontinuities to capture objects shape properties.
• Improve accuracy for texture-less objects.
• Unsupervised 3D object proposal generation that supports the detection.

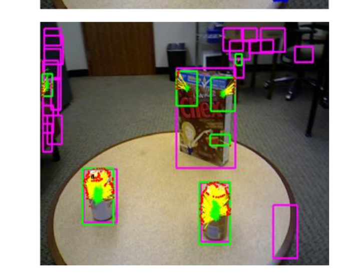




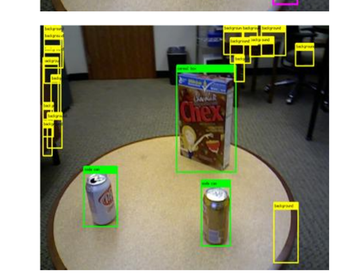
1. Scaled Shape Contexts extracted on depth discontinuities. Scaling depends on object class and sampled depth from test image.



2. Generation of Object Proposals by removing the support surfaces and clustering the remaining 3D points.

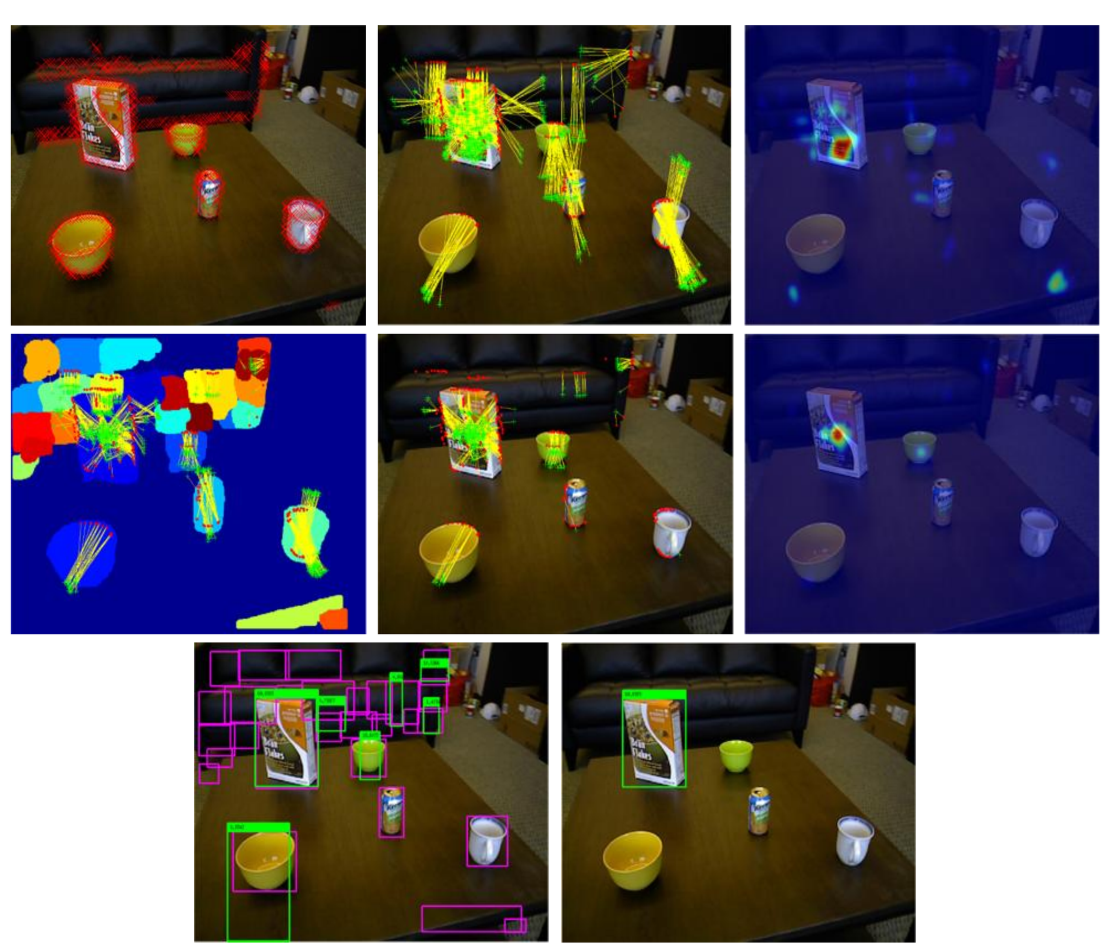


3. Detection Stage: Shape context matching and generation of class-specific hypotheses with verification from object proposals.



4. Result after all detectors are applied and Multi-View information is incorporated.

Class-specific Detection



Detection Example for Cereal Box:
• Matching of local descriptors and voting for object center following the implicit shape model (row 1).
• Votes are scaled based on depth ratio to avoid performing detection in several scales and pruned if they contribute outside a proposal's region (row 2).
• Keep hypotheses consistent with the object proposals based on their IOU overlap (row 3).

Multi-Class Prediction

• Detectors applied sequentially for all object categories.
• For each class the score depends on number and concentration of votes.
• Scores across classes are normalized based on samples of the number of edge points.

$$\bar{s}_j^c = \frac{s_j^c - \mu^c}{\sigma^c}$$

• Hypotheses from detection provide a score distribution over the classes for each proposal.

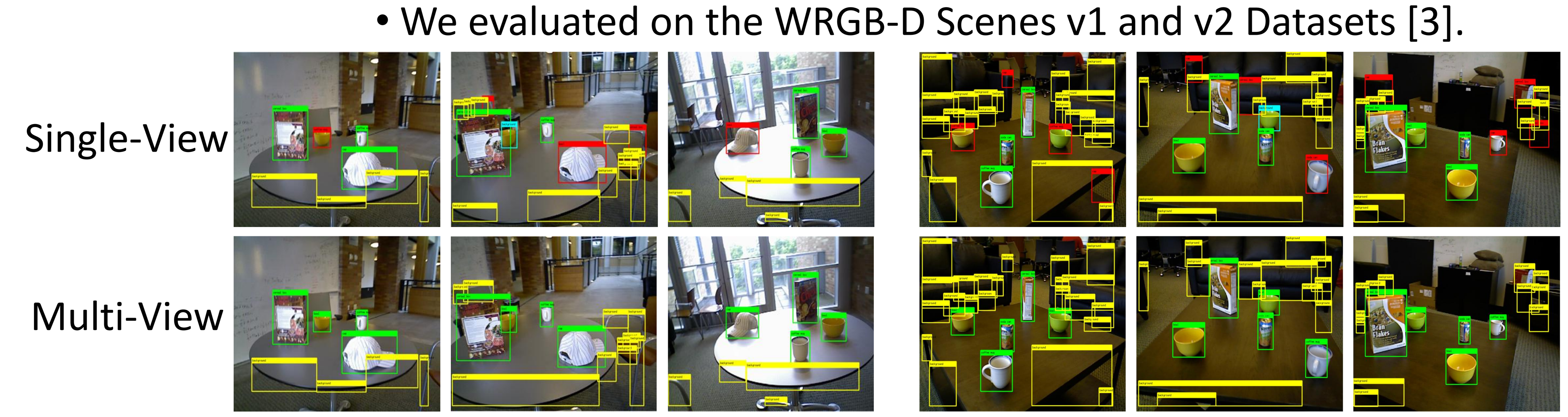
Multi-View Detection

• Create tracks of object proposals based on their 3D centroid proximity in the scene.
• Update the class probabilities using Bayes rule every time a new proposal is added to a track.

$$p(C_t|y^{1:n}) = \frac{p(y^n|C_t)p(C_t|y^{1:n-1})}{p(y^n|y^{1:n-1})}$$

Results

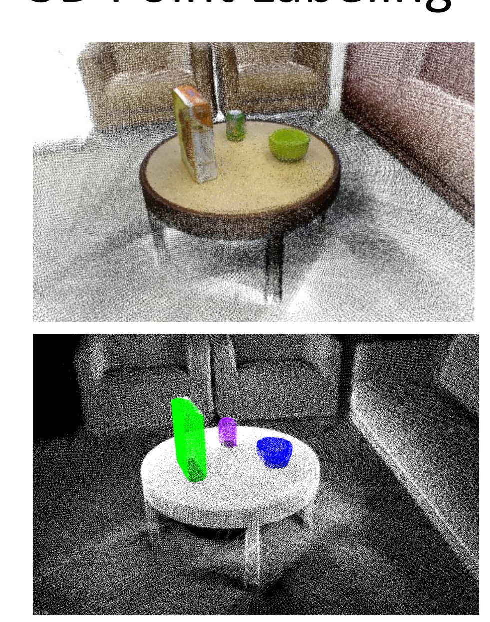
• We evaluated on the WRGB-D Scenes v1 and v2 Datasets [3].



	Bowl	Cap	Cereal Box	Coffee Mug	Soda Can	Flashlight	Average
Tang et al. [2] (HOG)	51.6	33.3	21.4	54.1	**71.0**	32.1	43.9
Tang et al. [2] (HH)	71.6	71.4	50.0	61.8	60.6	44.4	60.0
Ours	**75.1**	**74.5**	**61.2**	**62.8**	69.5	**73.6**	**69.5**

Table 2: Average Precision for class-specific object detection on the WRGB-D v1 scenes Dataset [4].

3D Point Labeling



	Bowl	Cap	Cereal Box	Coffee Mug	Soda Can	Background	Average
Single-View							
Pillai et al. [1]	**88.6/71.6**	85.2/**62.0**	83.8/75.4	70.8/**50.8**	78.3/42.0	**95.0**/90.0	81.5/59.4
Ours	70.7/56.8	**87.2**/49.0	**84.6/83.3**	**83.7**/34.3	**85.6/55.6**	89.0/**98.1**	**83.5/62.8**
Multi-View							
Pillai et al. [1]	88.7/70.2	89.4/72.0	**95.6**/84.3	80.1/64.1	**89.1**/75.6	96.6/96.8	89.8/72.0
Ours	**92.7/89.8**	**96.9/81.0**	87.4/**97.8**	**88.4/87.0**	86.7/**84.2**	**97.3/98.0**	**91.6/89.6**
3D Point Labeling							
HMP2D+3D [3]	**97.0**/89.1	**82.7/99.0**	**96.2/99.3**	81.0/92.6	**97.7/98.0**	95.8/95.0	**91.7**/95.5
Ours	88.5/**95.1**	79.3/95.6	91.0/98.6	**85.0/95.3**	85.8/93.4	**99.6/98.7**	88.2/**96.1**

Table 1: Precision/recall (%) results for the single-view, multi-view, and 3D point labeling experiments on the WRGB-D v2 scenes dataset [3].

Conclusions

• The 3D Class agnostic object proposals support the implicit shape model favorably by reducing the false positives.
• Integrating the evidence from multiple views can increase the performance considerably.

References
1. S. Pillai and J. Leonard. Monocular SLAM supported object recognition. (RSS). 2015.
2. S. Tang, X. Wang, X. LV, T. X. Han, and J. Keller. Histogram of oriented normal vectors for object recognition with depth sensor. (ACCV) 2012.
3. K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. (ICRA). 2014.
4. K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. (ICRA). 2011
5. L. Wang, J. Shi, G. Song, and I. Shen. Object Detection Combining recognition and segmentation. (ACCV). 2007.