Detection Bank: An Object Detection Based Video Representation for Multimedia Event Recognition

Tim Althoff, Hyun Oh Song, Trevor Darrell
UC Berkeley EECS/ICSI

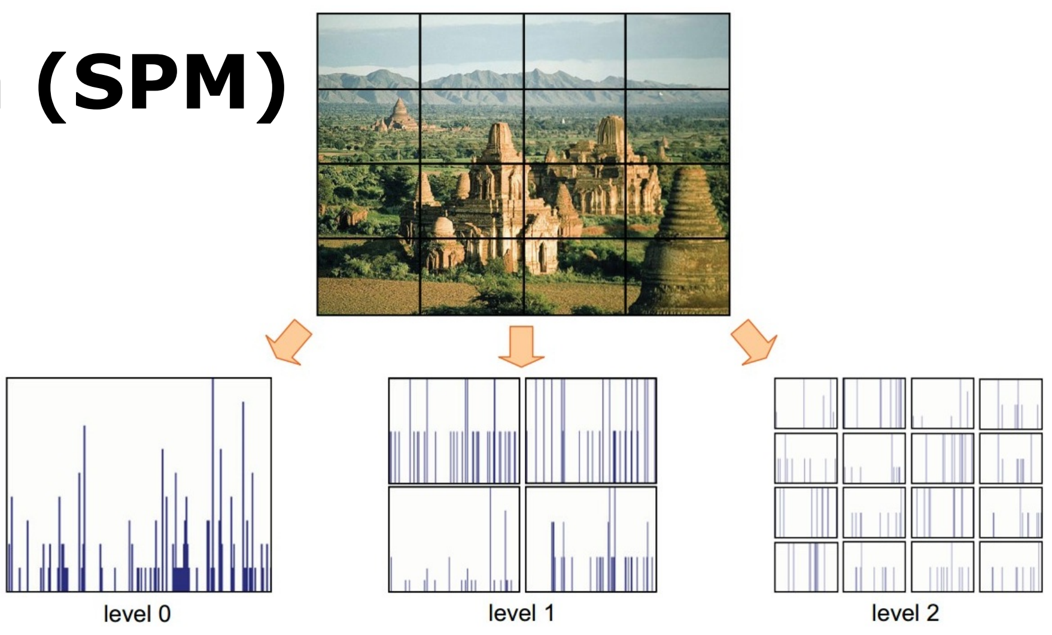
Multimedia Event Detection

Birthday Party vs **Wedding Ceremony**

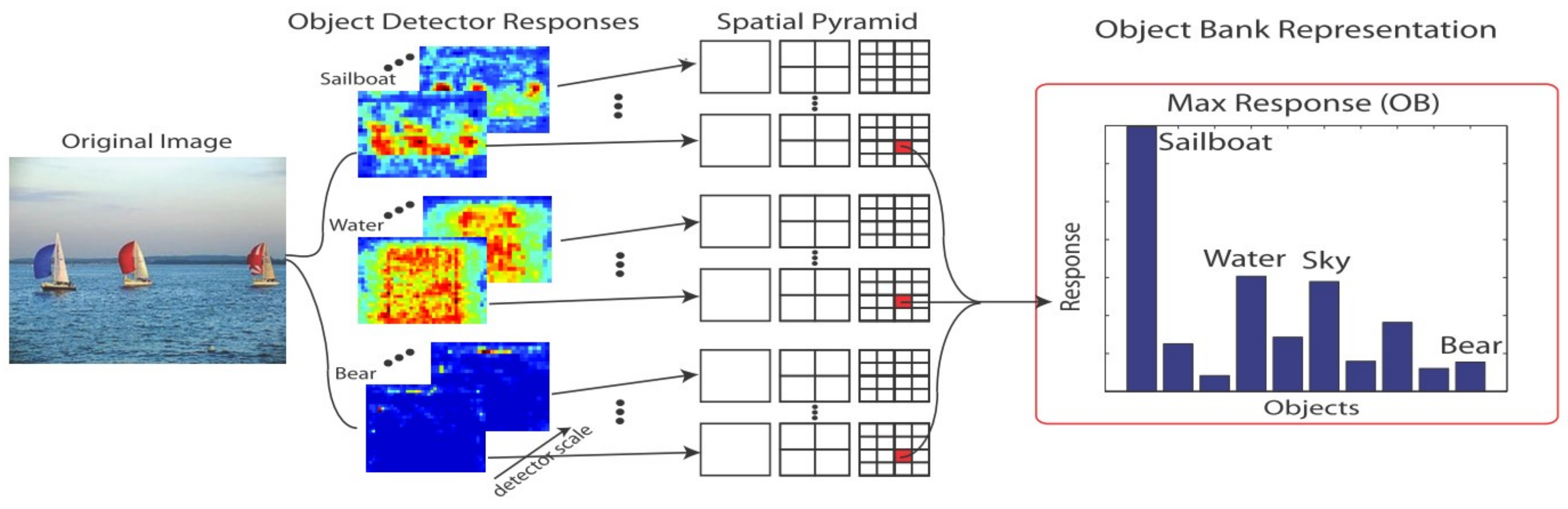
Look for: Balloon, Candle, Birthday Cake vs. Bride, Groom, Wedding Gown, Wedding Cake

Previous Work

Spatial Pyramid Match (SPM)



Object Bank (OB)



Problem

Scene-level descriptors cannot capture *fine-grained phenoma* that discriminate between events. **Object Bank** lacks immediate sense of whether or not there are *objects present in the image* and if so how many.

Idea

- ObjectBank omits the following steps that are standard in a detection pipeline:
  - *Thresholding of score maps*
  - *Non-maximum suppression*
  - *Pooling across all scales*
- We compute different *detection count statistics* to capture e.g. max number of detections, sum of detection scores, probablity of detection based on the detection images from a large number of windowed object detectors.

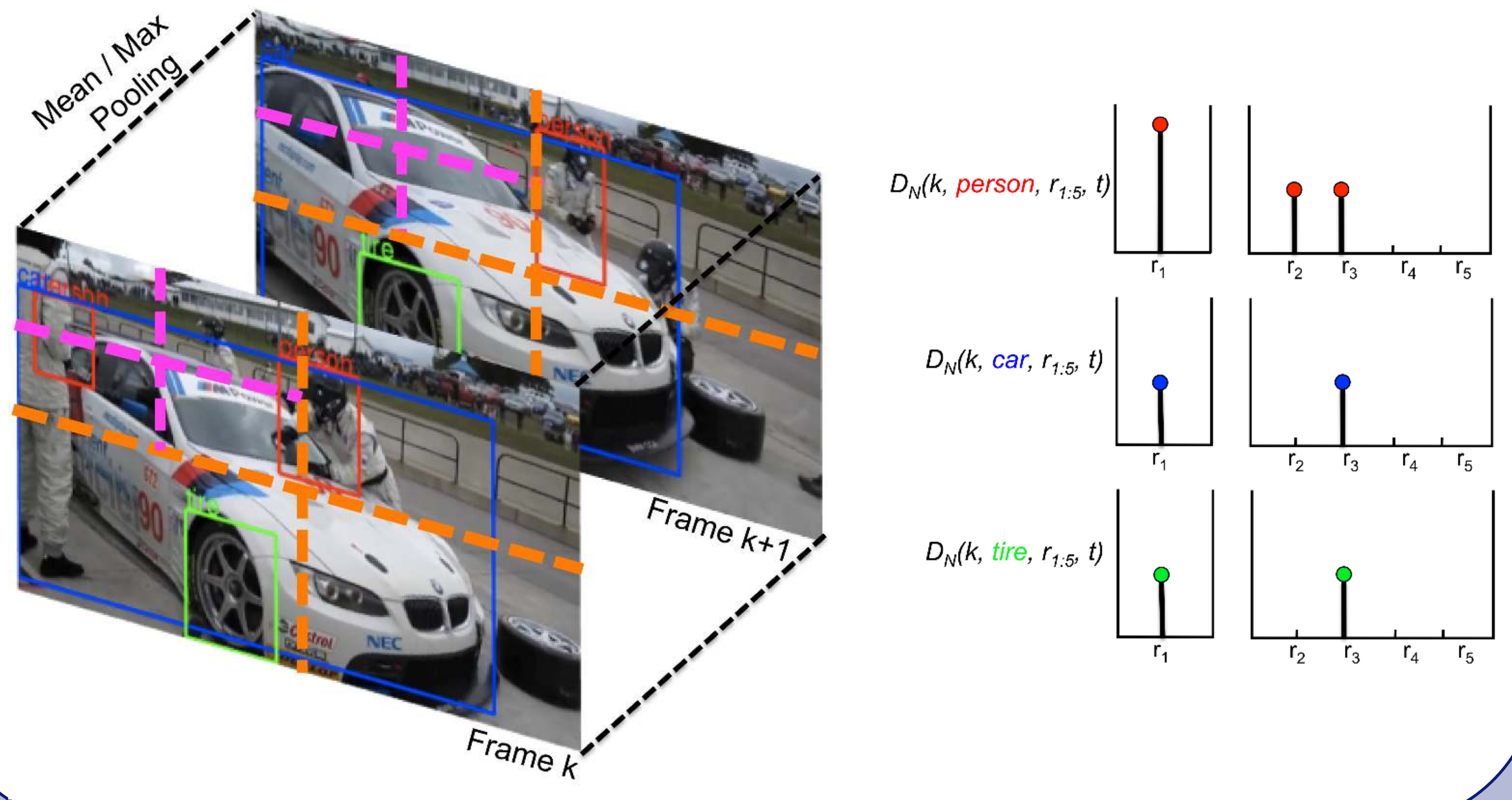
Detection Count Statistics

$$D_S\left(k,c,r,t\right)=\sum_{i=1}^{P}\mathbb{I}\left[\ \overline{\mathbf{b}_{c,i}}\in\mathcal{I}\left(r\right)\ \right]\mathbb{I}\left[s\left(\mathbf{b}_{c,i}\right)\geq t\right]s\left(\mathbf{b}_{c,i}\right)$$

$$D_N\left(k,c,r,t\right)=\sum_{i=1}^{P}\mathbb{I}\left[\ \overline{\mathbf{b}_{c,i}}\in\mathcal{I}\left(r\right)\ \right]\mathbb{I}\left[s\left(\mathbf{b}_{c,i}\right)\geq t\right]$$

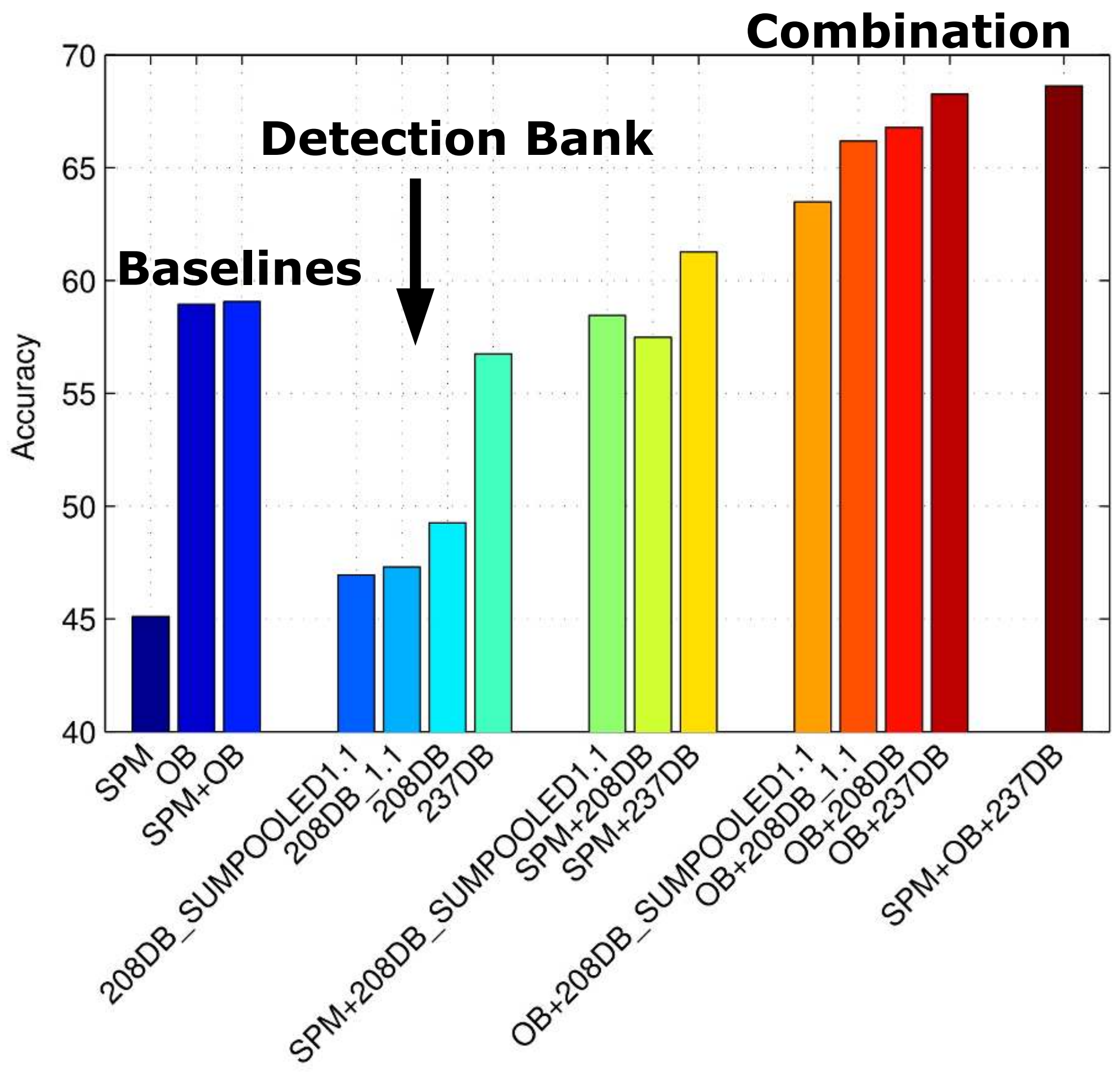
$$D_0\left(k,c,r,t\right)=\mathbb{I}\left[\sum_{i=1}^{P}\left(\mathbb{I}\left[\ \overline{\mathbf{b}_{c,i}}\in\mathcal{I}\left(r\right)\ \right]\mathbb{I}\left[s\left(\mathbf{b}_{c,i}\right)\geq t\right]\right)>0\right]$$

Illustration

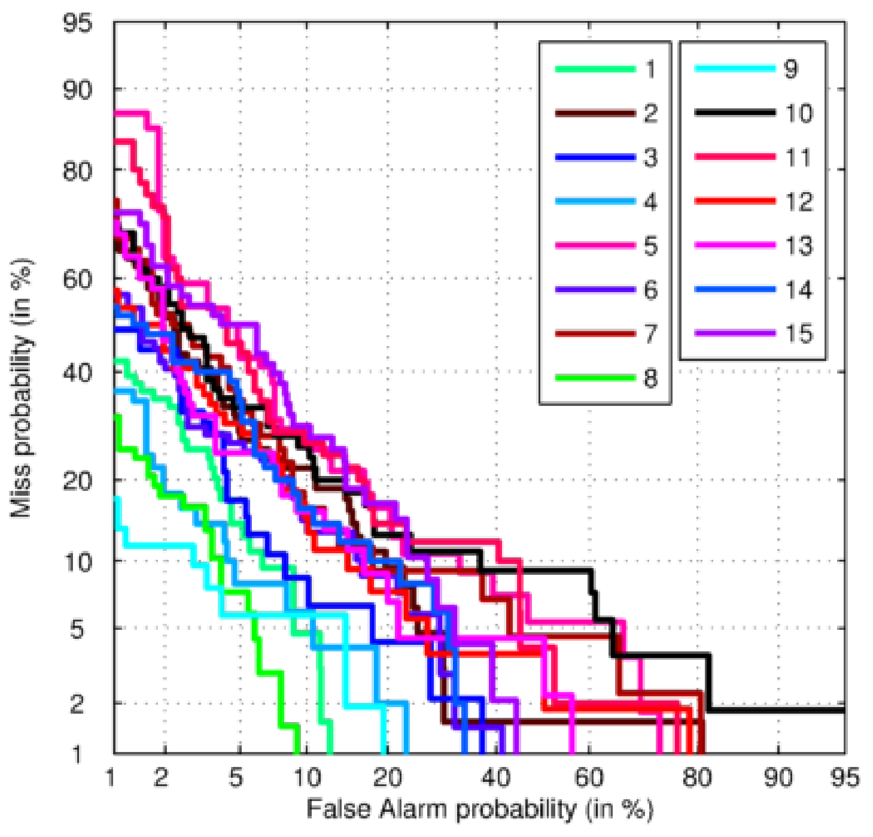

Mean / Max Pooling
Frame k+1
Frame k
$D_N(k, person, r_{1:5}, t)$
$D_N(k, car, r_{1:5}, t)$
$D_N(k, tire, r_{1:5}, t)$
r_1
r_2
r_3
r_4
r_5

Experiments

Classification Accuracy on TRECVID MED



DET Curves for all 15 Events



Conclusion

- Significant performance increase in Multimedia Event Classification Task
- Provides complementary discriminative information to current state-of-the-art image representations such as Spatial Pyramid Matching and Object Bank

References

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR, 2006.

L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. NIPS, 2010.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.
Disclaimer: The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.
H. Song was supported by Samsung Scholarship Foundation.