Deep Nets for Detecting, Combining, and Localizing Concepts in Video

Cees Snoek, Daniel Fontijne,
Zhenyang Li, Koen van de Sande, Arnold Smeulders
University of Amsterdam

Acknowledgement

This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Inspiration

- Deep nets prevail in many AI benchmarks,
  - except in TRECVID.
- The aim of our TRECVID experiments is to understand the value of deep nets for
  - video concept detection.

Related work: Concept detection

Bag of codes

- Few examples
- Human encodings
- Quantization loss
- Rigid pyramids
- One vs All
- Forward-propagation

Sande et al., TPAMI 2010
Sanchez et al., IJCV 2013
…

Net of convolutions

- Many examples
- Machine encoding
- Loss minimized
- Learned pyramids
- Multi-class
- Back-propagation

Krizhevsky et al., NIPS 2012
Coates et al., ICML 2013
…

We consider mutual benefit of bag of codes and deep nets

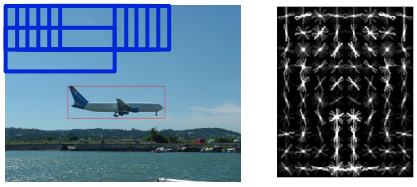
Related work: Object localization

Exhaustive search

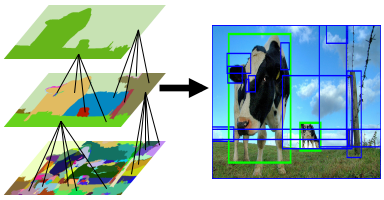
- Part-based
- Cheap encoding mandatory

Selective search

- Segment-based
- Expensive encoding feasible



Felzenswalb et al., TPAMI 2010



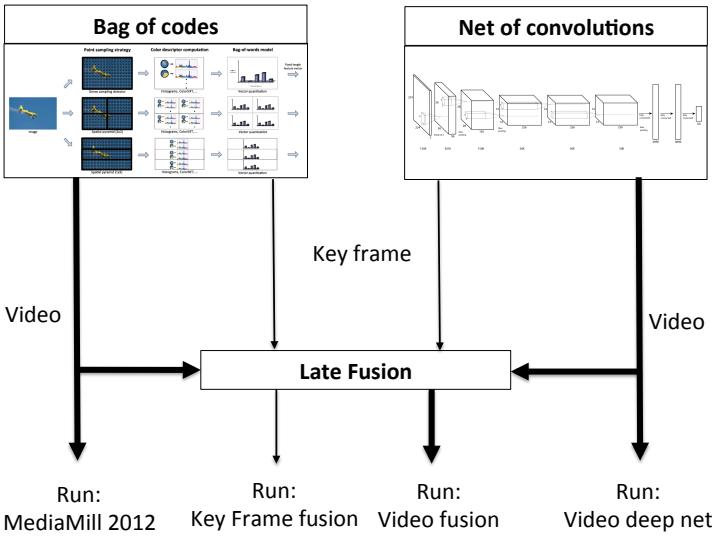
Uijlings et al., IJCV 2013

Selective search with precise encoding is better

TASK I

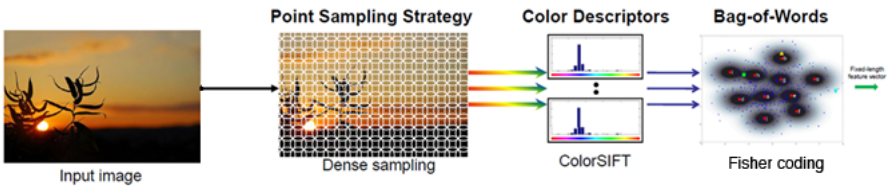
DETECTING CONCEPTS

MediaMill TRECVID 2013 runs



MediaMill: Color difference coding

- Densely sampled points
- SIFT, RGB-SIFT and T-SIFT descriptors
- PCA reduction to 80D
- Fisher vector coding with codebook size 256
- Spatial pyramid 1x1+1x3
- Linear classifier



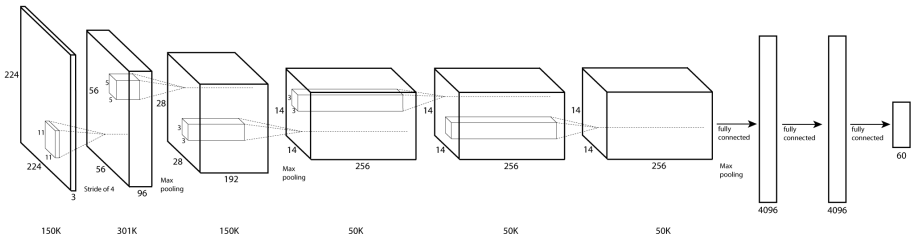
Color Descriptor software available for download at http://colordescriptors.com

MediaMill: Video deep learning

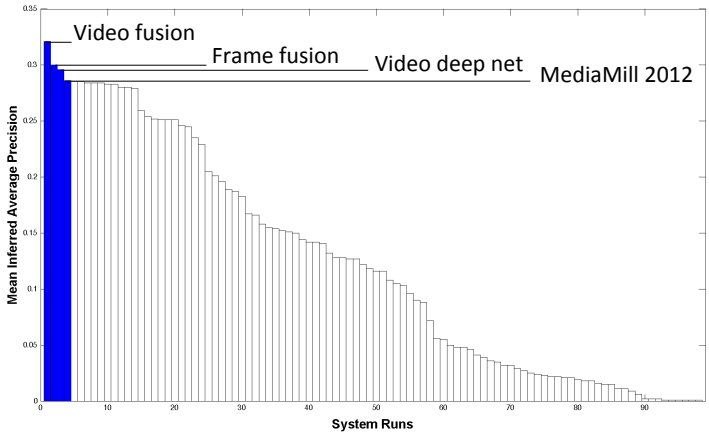
Convolutional neural network with 8 layers with weights

Trained using error back propagation

- ImageNet for pre-training

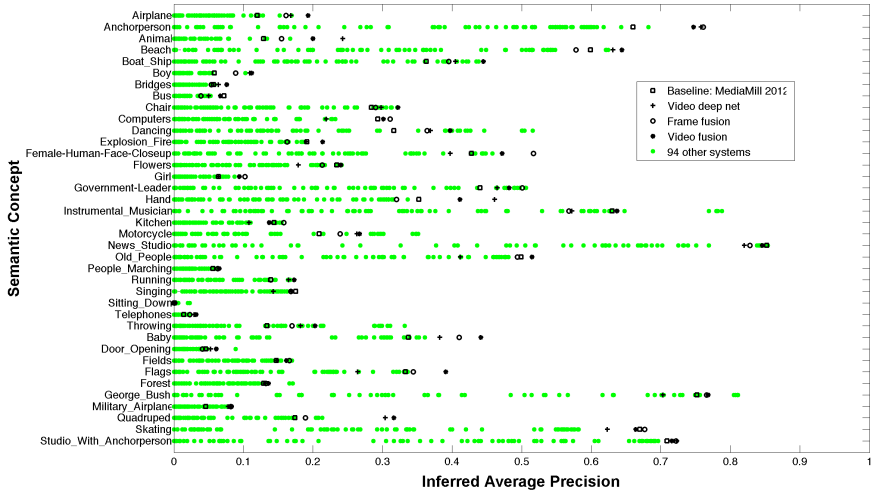


Results



Bag of codes and deep net profit from each other

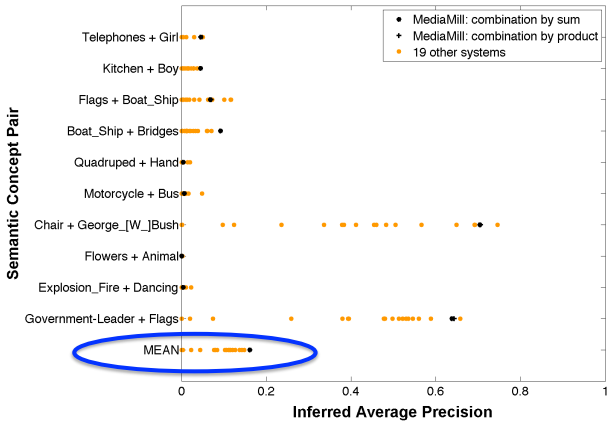
Results per concept



TASK II

COMBINING CONCEPTS

Results



Simple combination of robust detectors is good baseline

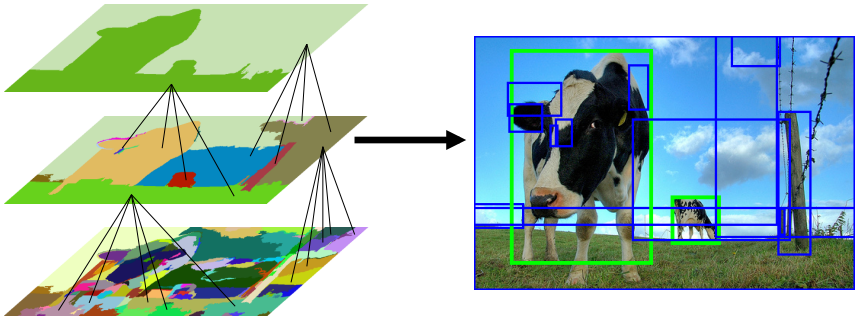
TASK III

LOCALIZING CONCEPTS

Uijlings et al., IJCV 2013

Selective Search: Approach

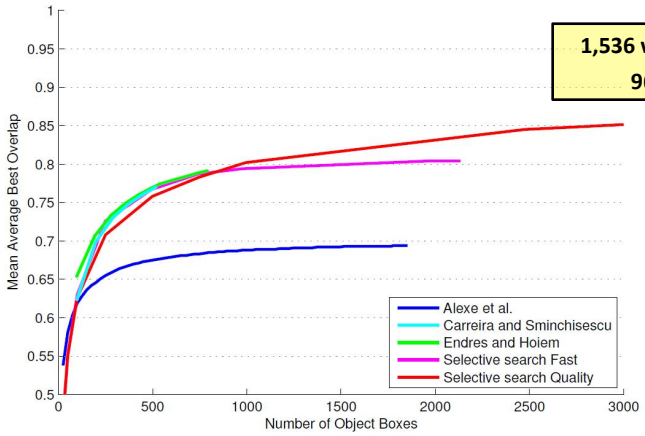
Hypotheses based on hierarchical grouping



Group adjacent regions on color/texture cues

Uijlings et al., IJCV 2013

Selective Search



1,536 windows/image
96.7% recall

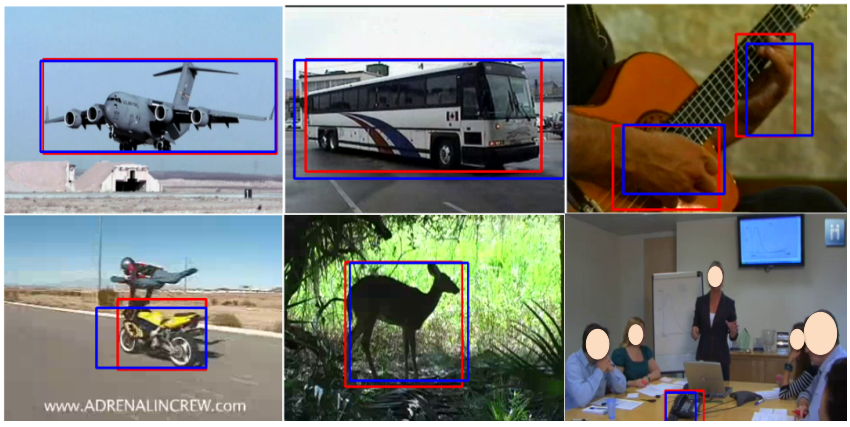
Location hypotheses are class-independent

Software available for download at http://koen.me/research/selectivesearch/

Provided i-frame JPG export is too lossy

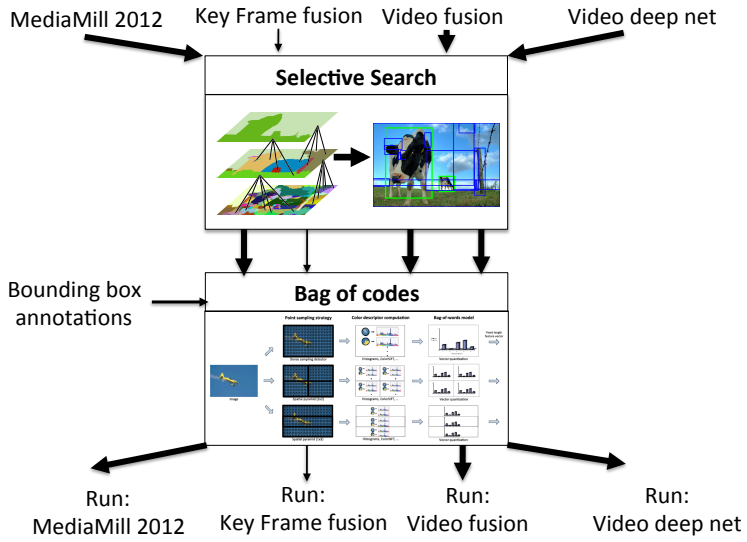
Re-extracted as PNG for consistency with training data

Selective search for Internet video



Best hypothesis examples generated by selective search

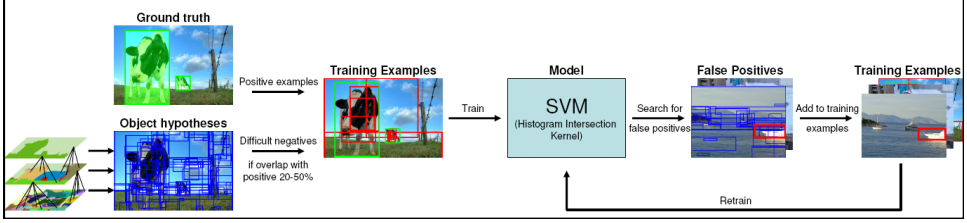
MediaMill TRECVID 2013 runs



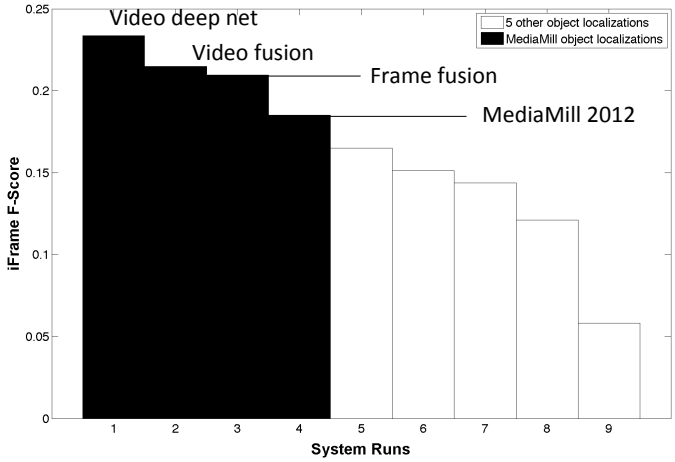
Uijlings et al., IJCV 2013

Implementation details

- PCA-reduced ColorSIFT descriptors to 80D
- Hard assignment Bag-of-Words
- Spatial pyramid
- Fast Intersection Kernel SVM Maji et al., PAMI 2013
- Hard negative mining

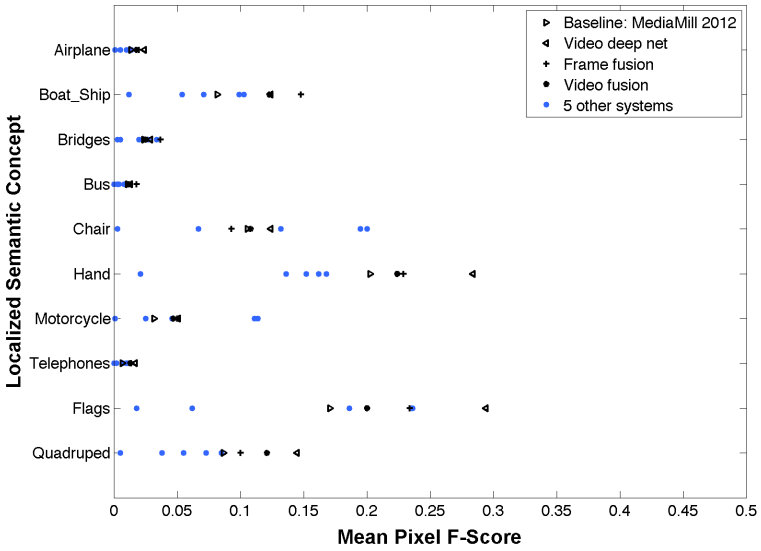


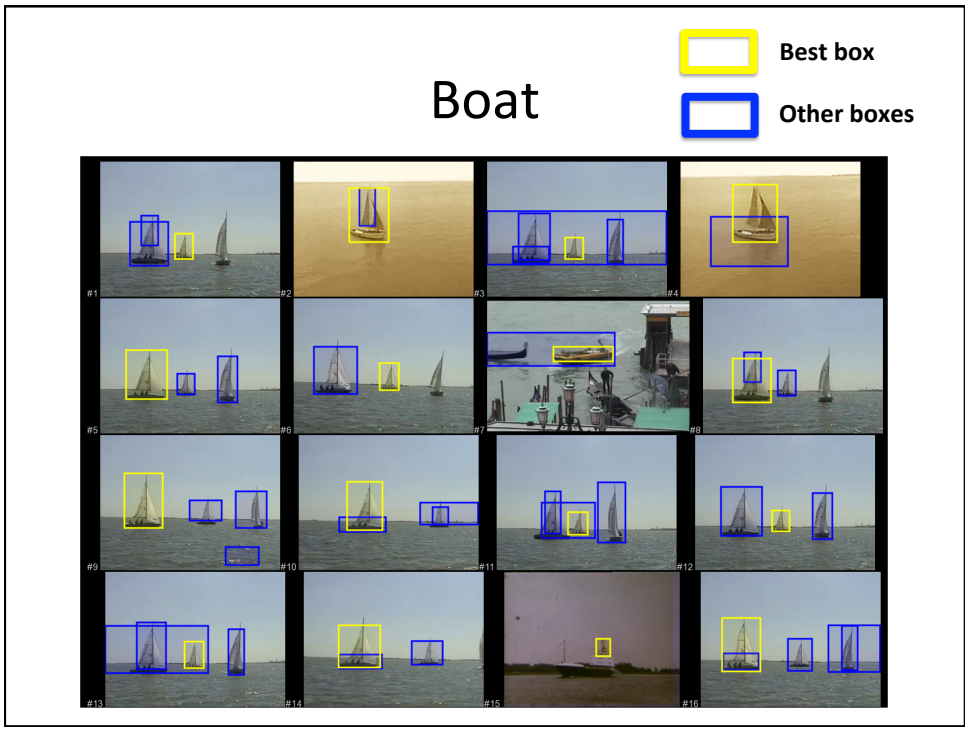
Results

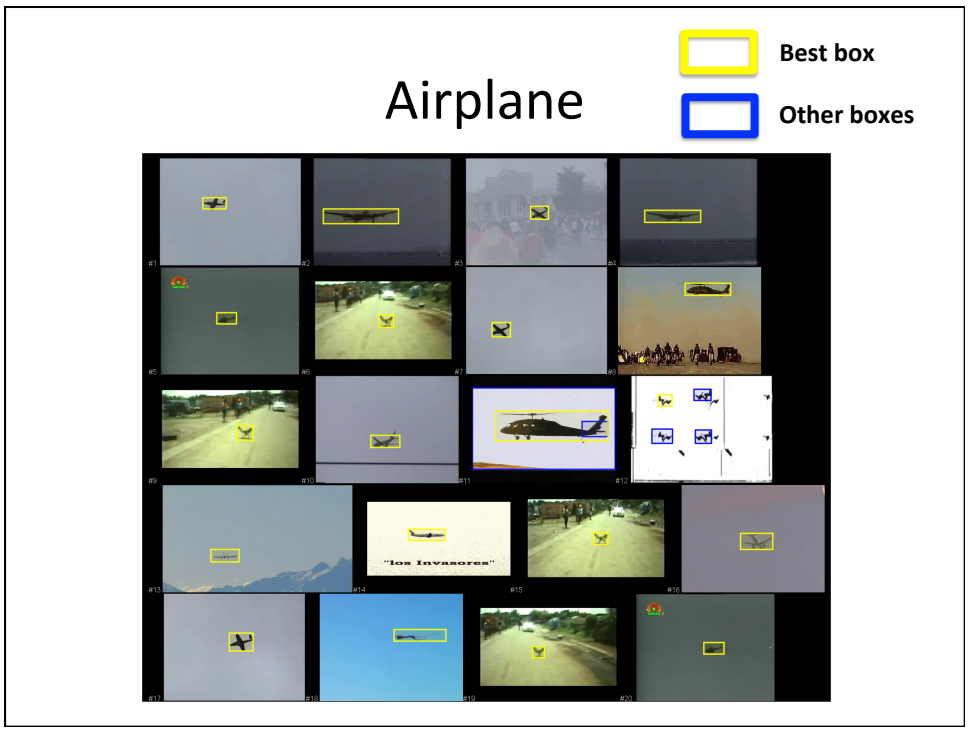


Selective search after deep nets is best

Results per object




Boat
Best box
Other boxes


Airplane
Best box
Other boxes

Conclusion

Deep Nets for Detecting, Combining, and Localizing Concepts in Video

www.mediamill.nl

Try it yourself: Impala iPhone App