Any Hope for Cross-Domain Concept Detection in Internet Video?

Cees G.M. Snoek, Koen E.A. van de Sande,
Dennis C. Koelma, & Arnold W.M. Smeulders

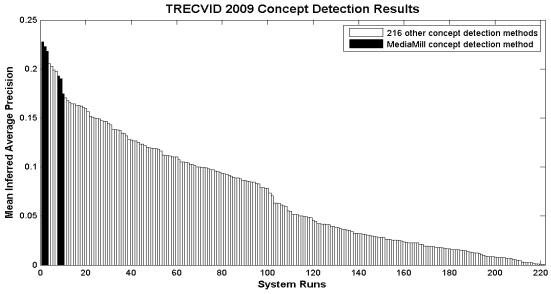
Intelligent Systems Lab Amsterdam
University of Amsterdam, The Netherlands

UNIVERSITY OF AMSTERDAM

stw FULBRIGHT

Conclusions TRECVID 2009

- Multi-frame is true performance booster
  - 30% improvement over single-frame baseline
  - Time for the community to move on to video analysis



Community myths or facts?

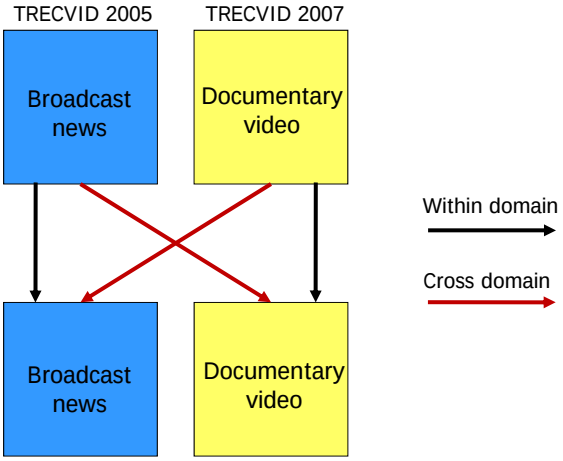
- Chua et al., ACM Multimedia 2007
  - Video search is practically solved and progress has only been incremental
- Yang and Hauptmann, ACM CIVR 2008
  - Current solutions are weak and generalize poorly

We have done an experiment

- Two video search engines from 2006 and 2009
  - MediaMill Challenge 2006 system
  - MediaMill TRECVID 2009 system
- How well do they detect 36 LSCOM concepts?

Four video data set mixtures

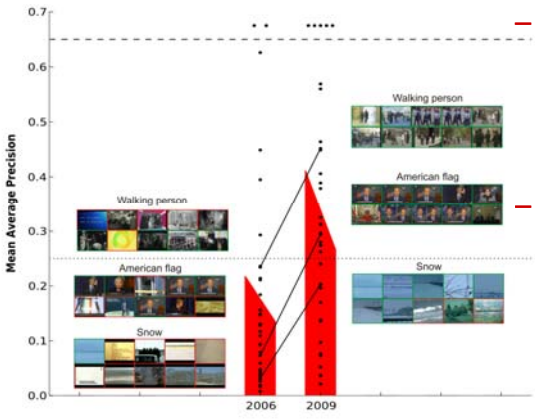
- Training
- Testing



Snoek & Smeulders, IEEE Computer 2010

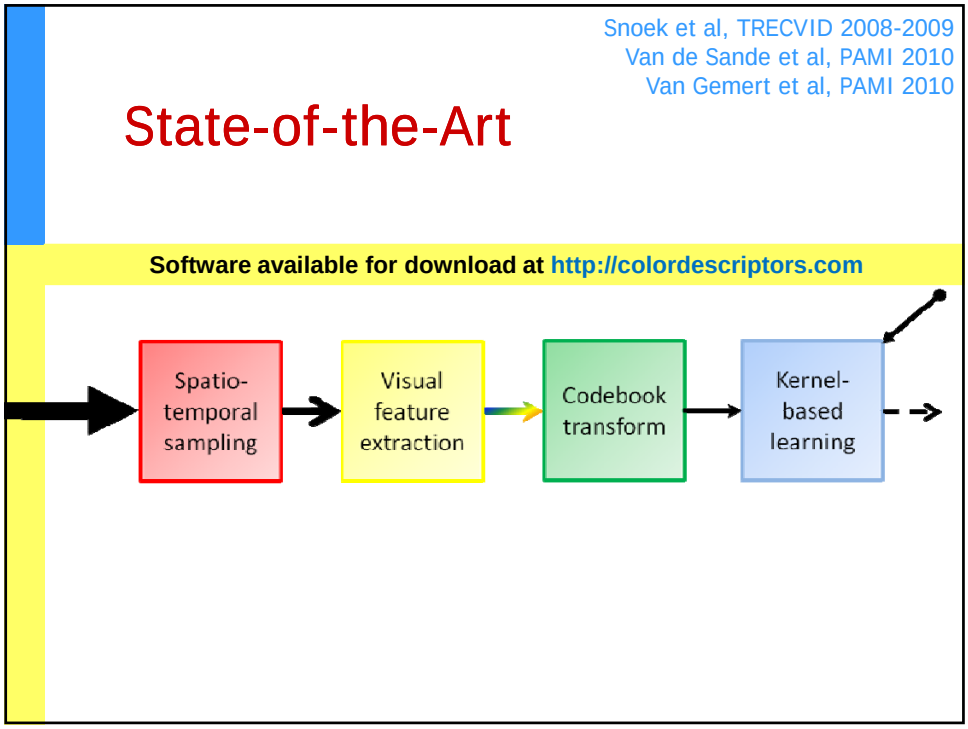
Performance doubled in just 3 years

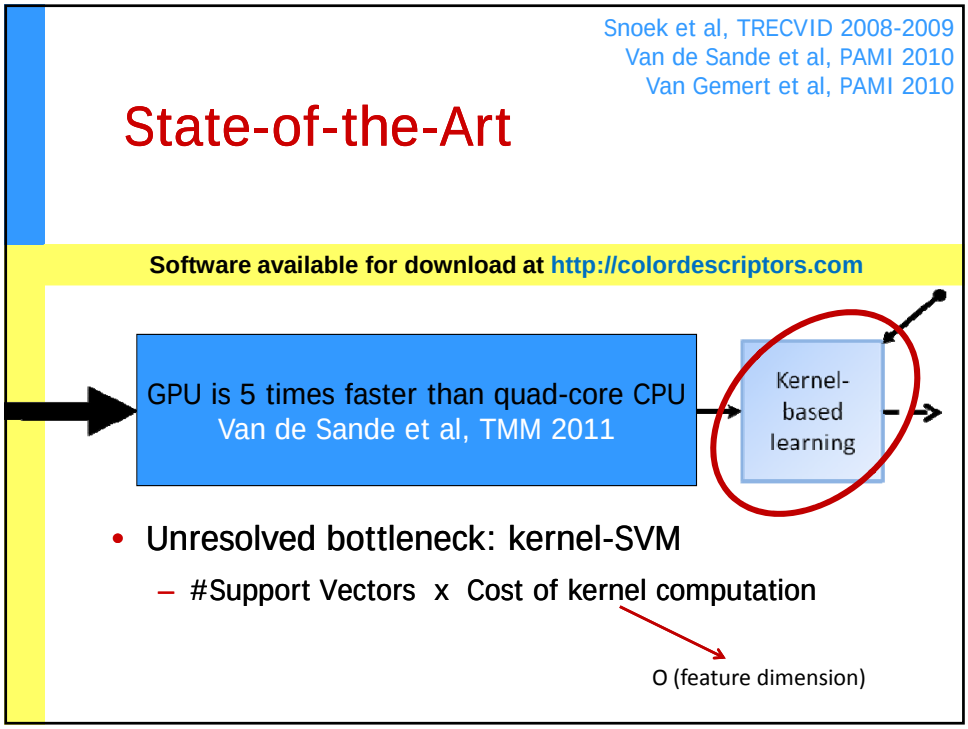
- 36 concept detectors



– Even when using training data of different origin

– Vocabulary still limited


Snoek et al, TRECVID 2008-2009
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010
State-of-the-Art
Software available for download at http://colordescriptors.com
Spatio-temporal sampling
Visual feature extraction
Codebook transform
Kernel-based learning


Snoek et al, TRECVID 2008-2009
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010
State-of-the-Art
Software available for download at http://colordescriptors.com
GPU is 5 times faster than quad-core CPU
Van de Sande et al, TMM 2011
Kernel-based learning
• Unresolved bottleneck: kernel-SVM
– #Support Vectors x Cost of kernel computation
O (feature dimension)

Our TRECVID 2010 focus

- Baseline: TRECVID 2009 system
  - 6 extra i-frames per shot ~ 600K frames in test set
- Revisit multi-frame for Internet video
- Training from multiple domains
  - Add 50K labels from TRECVID05-09 ~ 170K frames train set
  - Requires efficient prediction

Maji et al., CVPR 2008

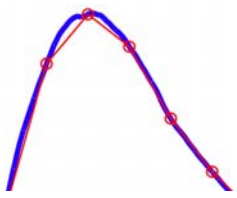
$$K(a,b) = \sum_{i=1}^{n} min(a_i, b_i)$$ is efficient

$$h(x) = \sum_{i=1}^{\#dim} \left(\sum_{j=1}^{\#sv} \alpha^j \min(x_i, x_i^j) \right) + b$$

$$= \sum_{i=1}^{\#dim} h_i(x_i)$$

$$h_i(x_i) = \sum_{j=1}^{\#sv} \alpha^j \min(x_i, x_i^j) + b$$

$$= \sum_{x_i^j < x_i} \alpha^j x_i^j + \left(\sum_{x_i^j \geq x_i} \alpha^j \right) x_i$$

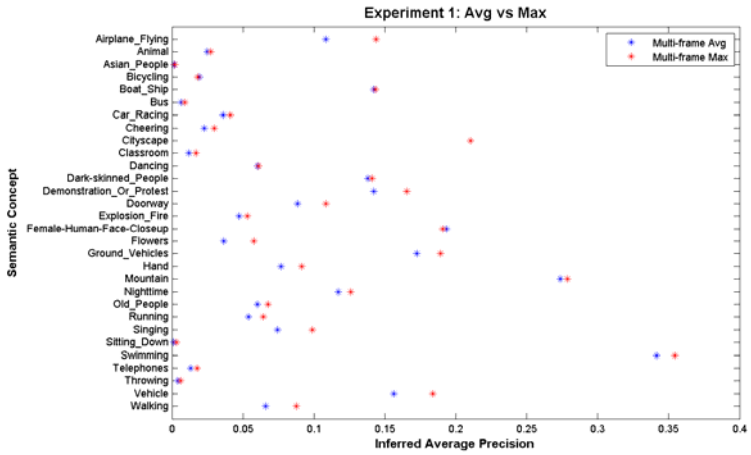


For the Intersection Kernel h_i is piecewise linear, and quite smooth, blue plot. We can *approximate* with fewer uniformly spaced segments, red plot. Saves time & space!

Slide credit: Subhransu Maji

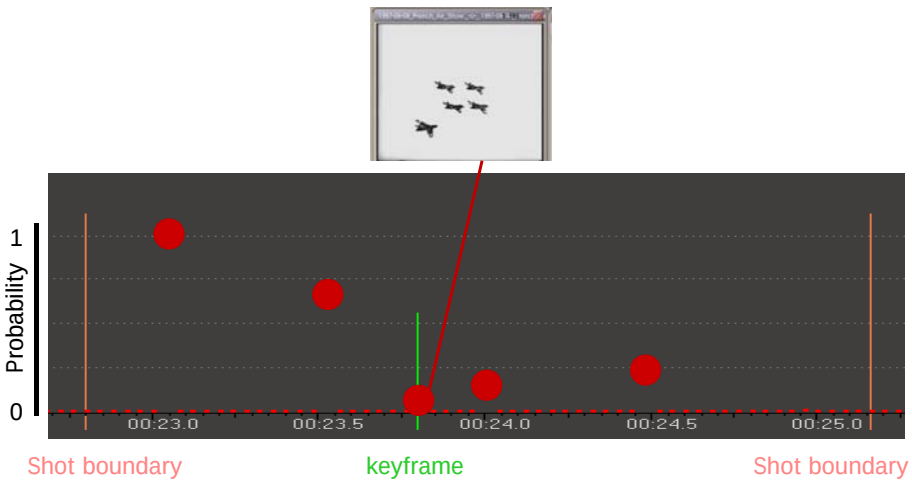
Experiment 1: Avg vs Max (χ^2)

Max multi-frame appears best choice for online video



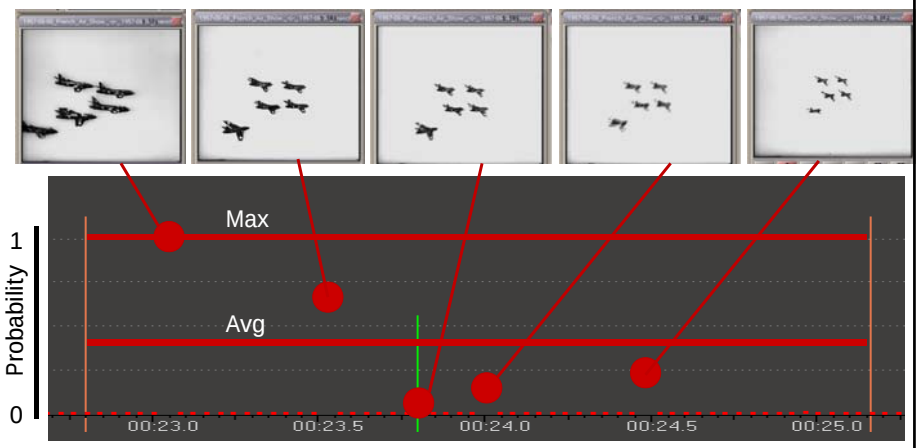
= Emphasis added

Moving object appearance



= Emphasis added

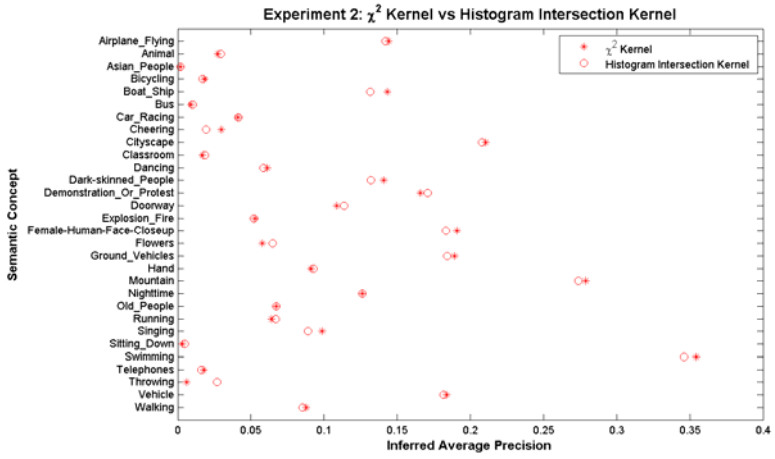
Moving object appearance



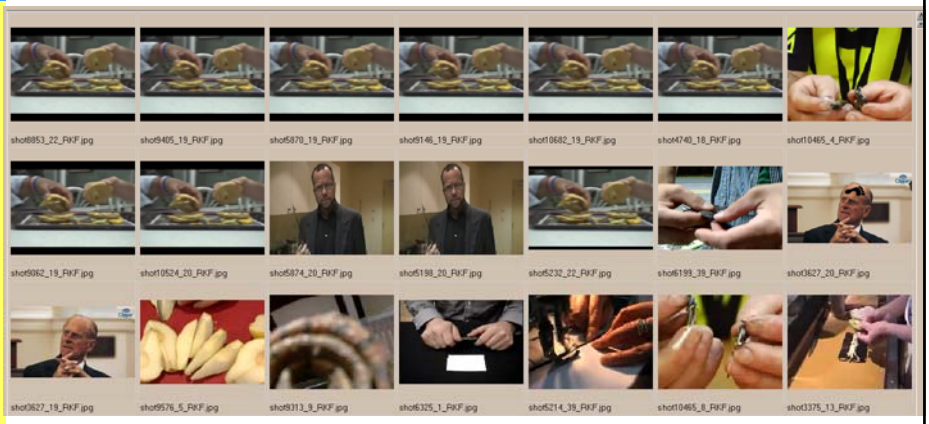
Note: we submitted avg...

Experiment 2: χ^2 vs HIK (max)

HIK 75 times faster, negligible loss in average precision



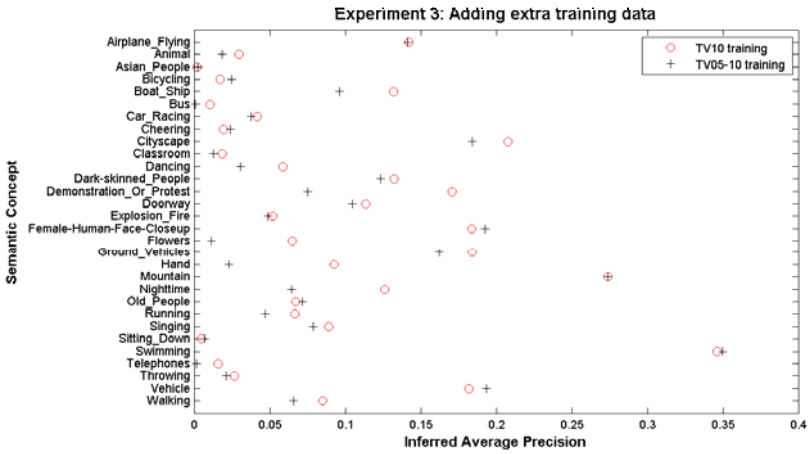
Top 21 results for “hand”



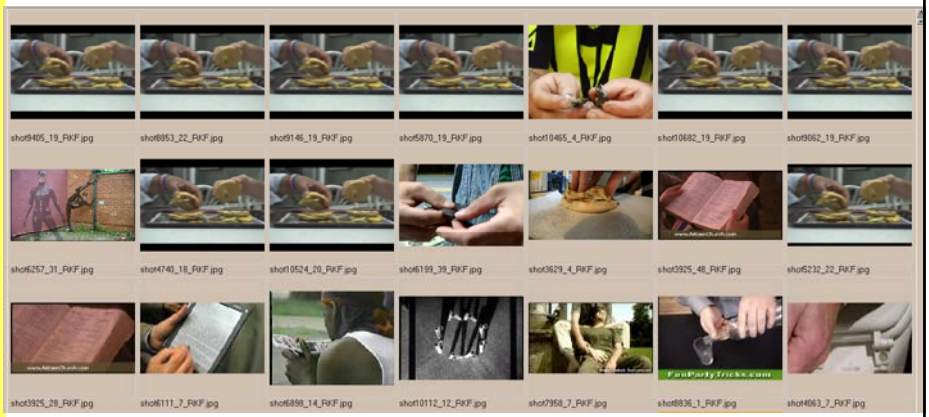
Top 21 results for “protest”

Experiment 3: adding labels

At best on par, often worse.

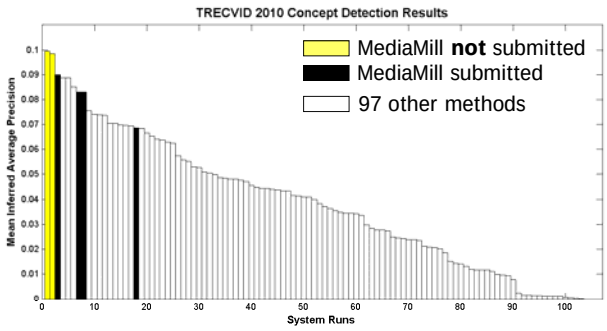


Top 21 results for “hand”



Top 21 results for “protest”

TRECVID 2010 results



- When considering submitted runs only
  - Best performer for 6 concepts
  - Best overall

Conclusions TRECVID 2010

- Internet video concept detection is feasible
  - Use max for effective multi-frame fusion
  - Use histogram intersection kernel for fast prediction
- We do not know how to exploit extra labeled training samples from other domains
  - A good challenge!

Contact info

- Cees Snoek
  http://staff.science.uva.nl/~cgmsnoek

- **We are hiring!**
  - PhD's and Postoc on video event retrieval

http://www.mediamill.nl

References

The MediaMill TRECVID 2008-2010 Semantic Video Search Engine. C.G.M. Snoek et al. Proceedings of the TRECVID Workshop.

Evaluating Color Descriptors for Object and Scene Recognition. K.E.A. van de Sande, Th. Gevers, C.G.M. Snoek. IEEE Trans. Pattern Analysis and Machine Intelligence, 2010.

On the Surplus Value of Semantic Video Analysis Beyond the Key Frame. C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra. Proc. IEEE Int'l Conference on Multimedia & Expo, 2005.

Empowering Visual Categorization with the GPU. K. E. A. van de Sande, T. Gevers, and C.G.M. Snoek. IEEE Trans. Multimedia, 2011.

Classification using Intersection Kernel Support Vector Machines is Efficient. S. Maji, A.C. Berg and J. Malik. Proc. IEEE CVPR, 2008.

Concept-Based Video Retrieval. C.G.M. Snoek, M. Worring. Foundations and Trends in Information Retrieval, Vol. 4 (2), page 215-322, 2009.

Visual-Concept Search Solved? C.G.M. Snoek, A.W.M. Smeulders. IEEE Computer, vol. 43(6), page. 76-78, 2010.