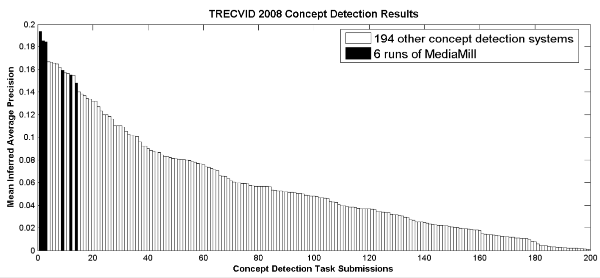
Multi-Frame, Multi-Modal, and Multi-Kernel Concept Detection in Video

Cees G.M. Snoek[1], Koen E.A. van de Sande[1], Jasper R.R. Uijlings[1], Miguel Bugalho[2], Isabel Trancoso[2], Fei Yan[3], Muhammed A. Tahir[3], Krystian Mikolajczyk[3], Josef Kittler[3], Theo Gevers[1], Dennis C. Koelma[1], Arnold W.M. Smeulders[1]

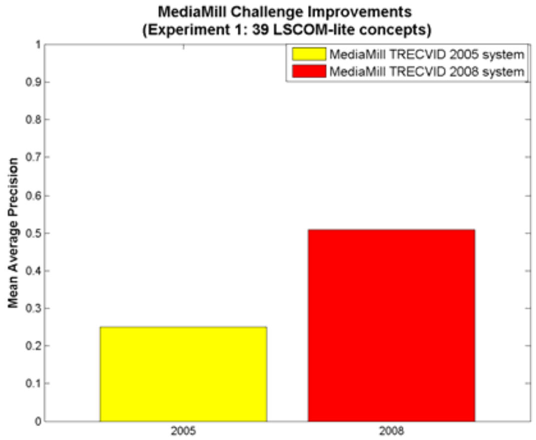
1 University of Amsterdam
2 inesc id lisboa
3 University of Surrey

Conclusions TRECVID 2008

- Good settings for Bag-of-Words
  - SIFT + colorSIFT improves ~8%
  - Soft codebook assignment improves ~7%
  - Multi-frame analysis improves ~20%



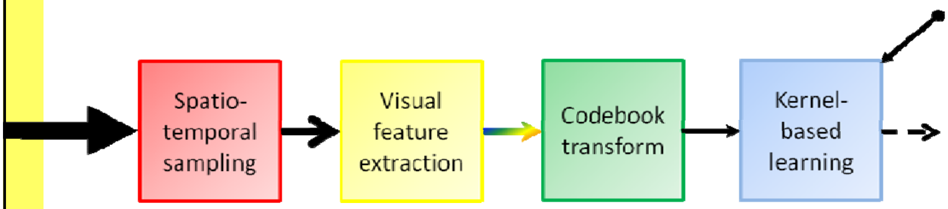
Myth: TRECVID incremental only



>100% improvement in just 3 years

Snoek et al, TRECVID 2008
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010

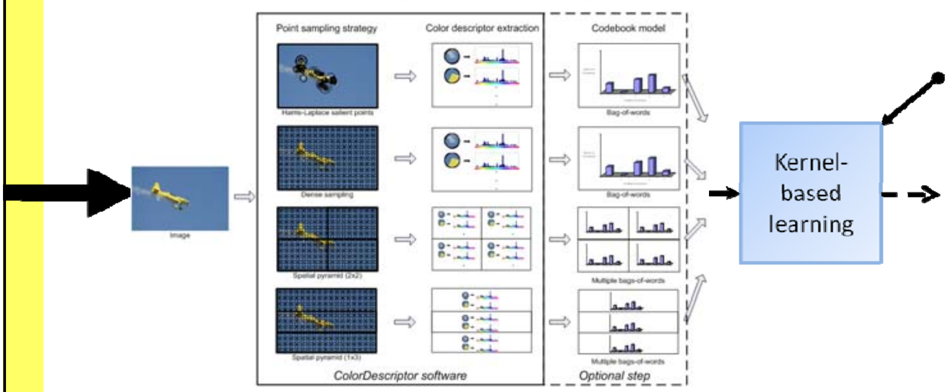
State-of-the-Art



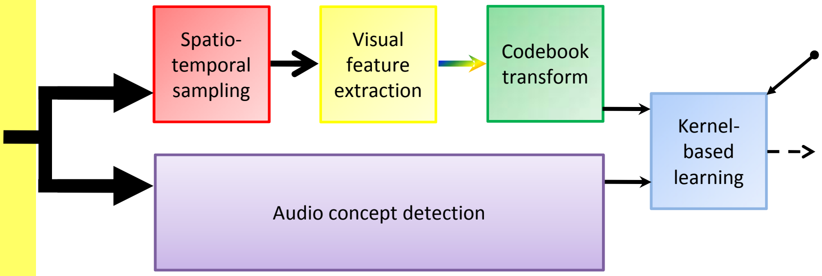
Snoek et al, TRECVID 2008
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010

State-of-the-Art

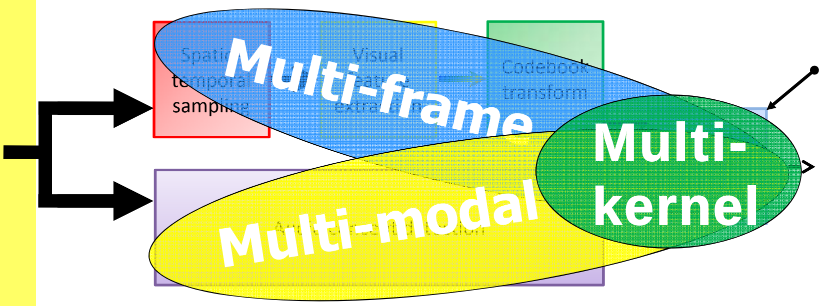
Software available for download at http://colordescriptors.com



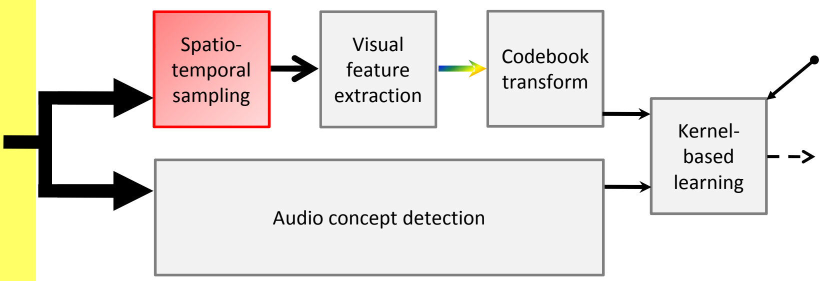
Our TRECVID 2009 focus



Our TRECVID 2009 focus



Roadmap

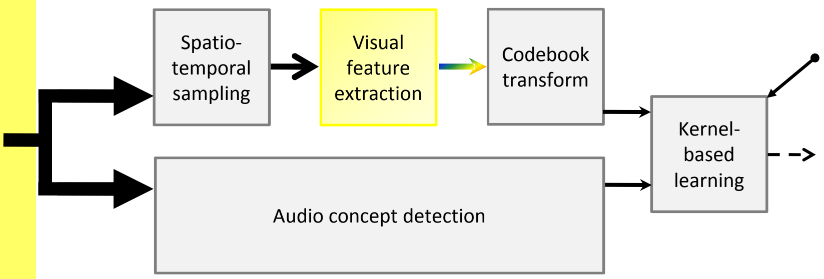


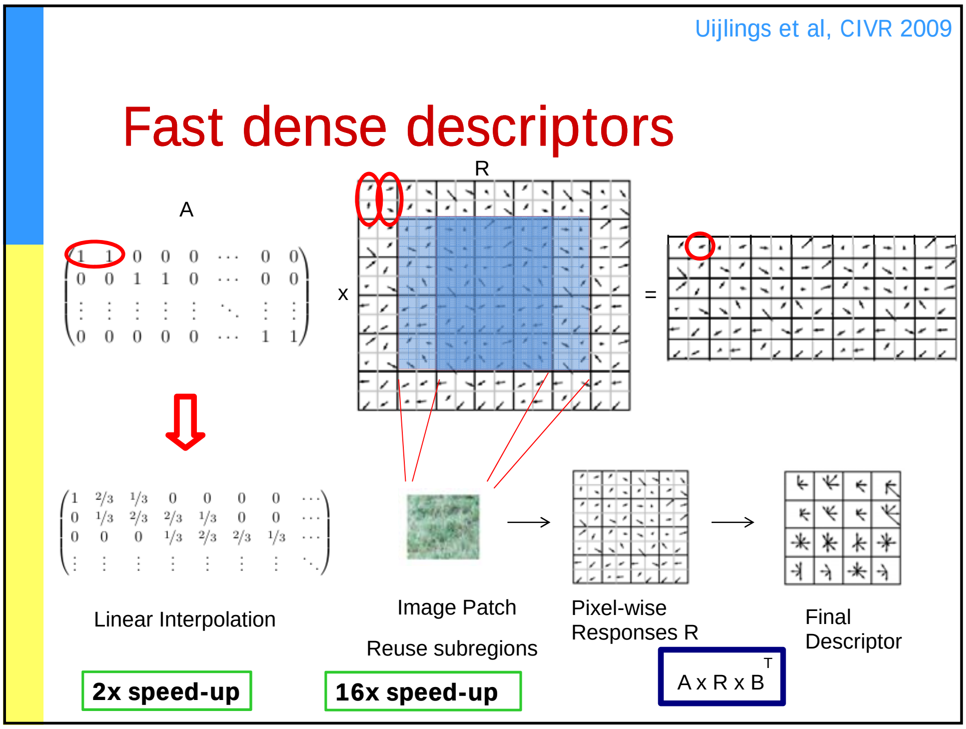
Snoek et al, ICME 2005

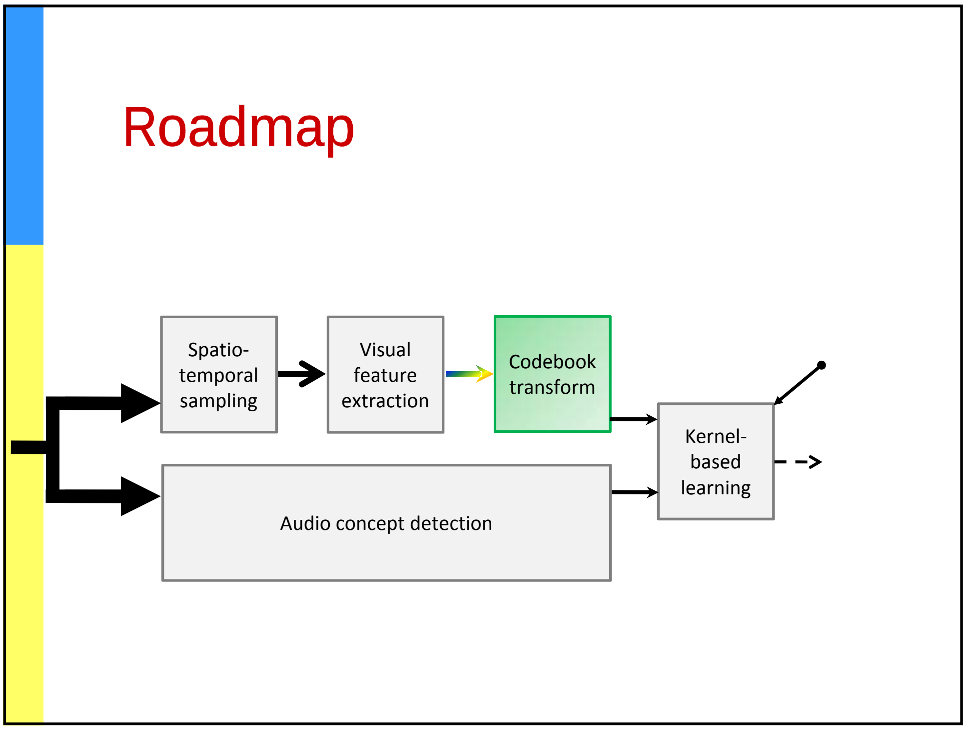
1,000,000 frames analyzed

- Multi-frame biggest improvement in 2008
  - Extend further by analyzing up to 10 extra i-frames/shot
  - Yields 1M frames to analyze for the test set collection
- Need to speed-up by being "smart and strong"
  - Speed-up feature extraction
  - Speed-up quantization
  - Speed-up kernel-based learning
  - Speed-up by computing

Roadmap

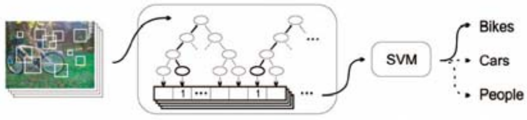



Uijlings et al, CIVR 2009
Fast dense descriptors
A
R
x
=
Linear Interpolation
Image Patch
Reuse subregions
Pixel-wise
Responses R
Final
Descriptor
A x R x B T
2x speed-up
16x speed-up


Roadmap
Spatio-
temporal
sampling
Visual
feature
extraction
Codebook
transform
Kernel-
based
learning
Audio concept detection

Moosman, PAMI 2008
Uijlings et al, CIVR 2009

Fast quantization

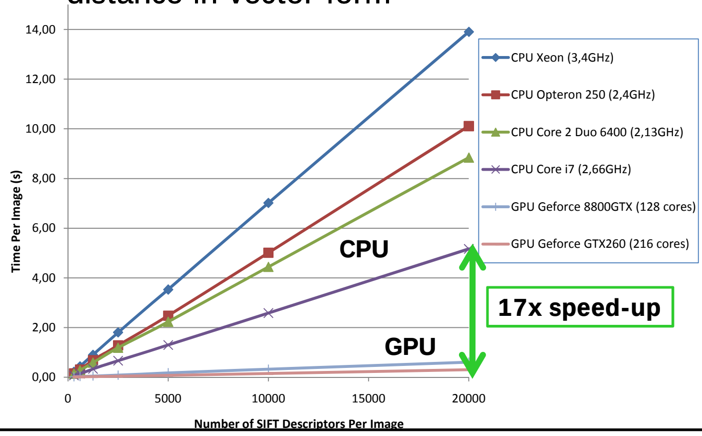


- Random forests
  - Randomized process makes it very fast to build
  - Tree structure allows fast vector quantization
  - Logarithmic rather than linear projection time
- Real-time BoW
  - When used with fast dense sampling
  - SURF 2x2 descriptor instead of 4x4
  - RBF kernel

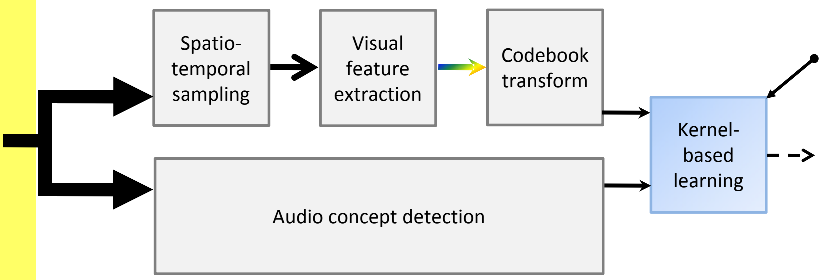
Van de Sande et al, ASCI 2009

GPU-empowered quantization

- Achieve data-parallelism by writing Euclidean distance in vector form



Roadmap



SVM pre-computed kernel trick

- Use distance between feature vectors
  - Feature length easily > 100,000

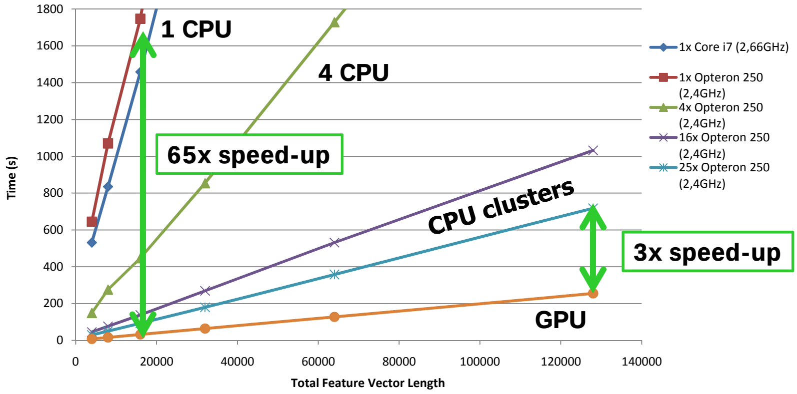
$$k(\vec{F}, \vec{F}') = e^{-\frac{1}{A} dist(\vec{F}, \vec{F}')}$$

- Increase efficiency significantly
  - Pre-compute the SVM kernel matrix
  - Long vectors possible as we only need 2 in memory
  - Parameter optimization re-uses pre-computed matrix

Van de Sande et al, ASCI 2009

GPU-empowered pre-computed kernel

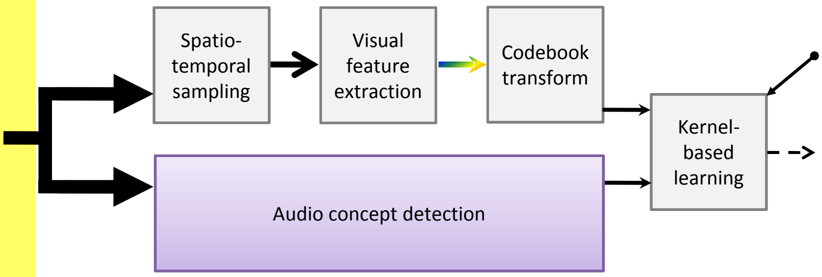
1. Compute average distances per N^2 kernel sub-block
2. Compute kernel function values



Computing

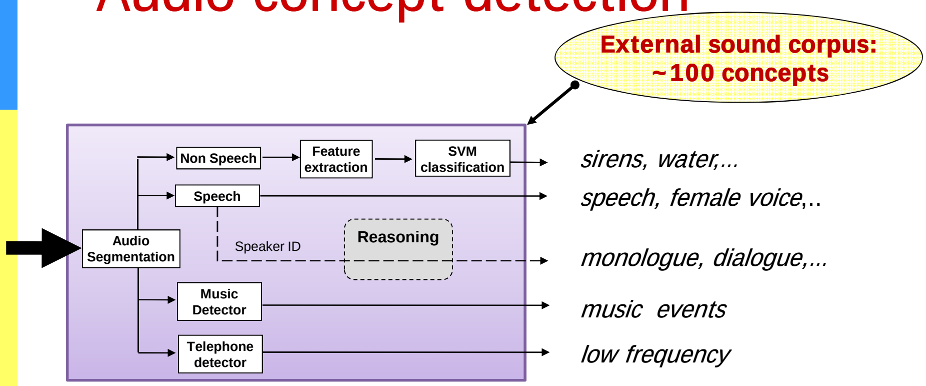
- 2009 system much more efficient than 2008 system
  - 6x more visual data analyzed using less compute power
- Some best estimates:
  - Visual feature extraction: 8400 Processor-Node-Hours
  - Training concept detectors: 4000 PNH
  - Applying concept detectors: ~1 week GPU

Roadmap



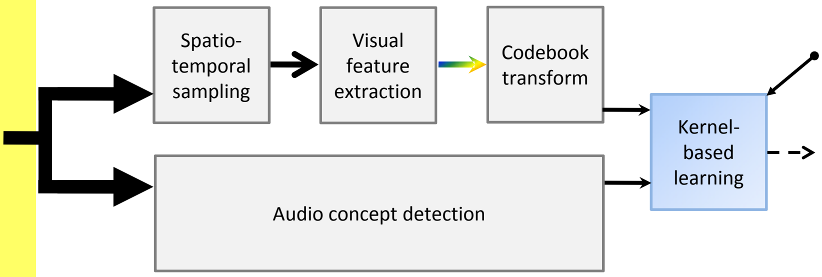
Bugalho et al, InterSpeech 2009
Trancoso et al, ICME 2009

Audio concept detection



- Early fusion of features
  - MFCCs (+ deltas), PLPs (+ deltas), Brightness, Bandwidth, ZCR, Pitch, Harmonicity, Shifted delta cepstra, Audio spectrum envelope and flatness
  - 0.50s window length, with 0.25s spacing

Roadmap



Tahir et al, ICCV-Subspace 2009
Yan et al, ICDM 2009

Multi-kernel learning

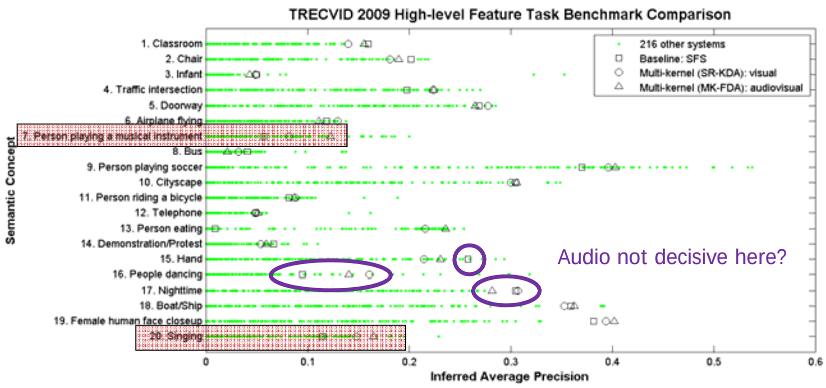
- Kernel Discriminant Analysis combined with spectral regression [Tahir09]
  - We use SR-KDA with 6 visual kernels
  - Weighted output combined using SUM rule
- Multi-Kernel Fisher Discriminant Analysis
  - We use non-sparse *L2* MK-FDA [Yan09]
  - Fusion of 1 audio and 6 visual kernels
    - 20 audio concept detector scores used as input for RBF kernel

Experiments (all type A)

- Baseline: single-frame SFS on all visual kernels
- Experiment 1: multi-modal & multi-kernel
  - SR-KDA (visual only)
  - MK-FDA (audiovisual fusion)
- Experiment 2: multi-frame
  - Visual fusion: 5 extra i-frames + MAX fusion [donated]
  - Best-of: 1 to 10 extra i-frames + MAX/AVG fusion
  - SFS: all multi-frame visual kernel combinations

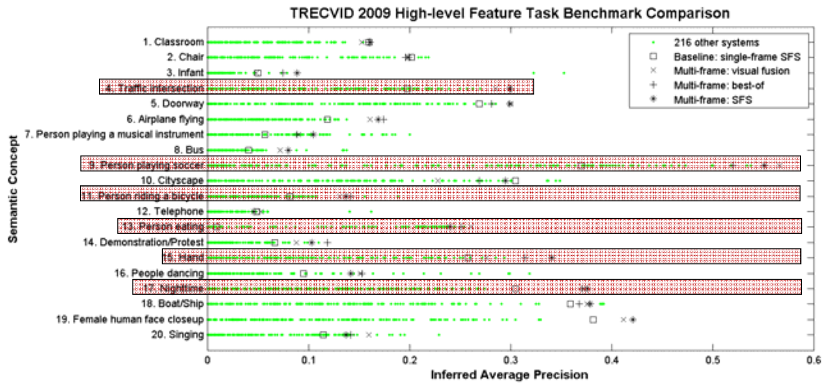
Results: experiment 1

- Multi-kernel improves upon baseline: ~9%
- Multi-modal kernel outperforms uni-modal kernel only slightly: ~2%
  - ...but for specific (audiovisual) concepts more impressive improvement, up to 50%



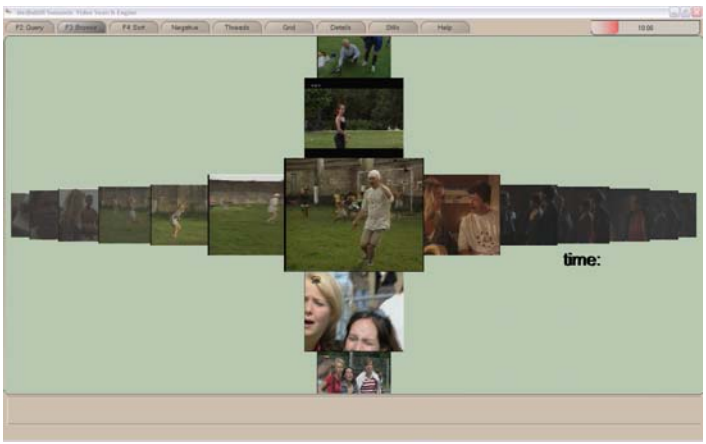
Results: experiment 2

- Multi-frame is true performance booster, improvement over baseline: ~30%
- Best to select optimal number of extra frames, per kernel, per concept,
  - On average 6 additional i-frames with MAX or AVG fusion is a solid choice



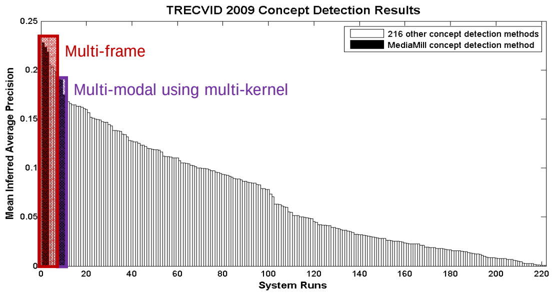
http://www.MediaMill.nl

Visualizing multi-frame impact



Conclusions TRECVID 2009

- Multi-modal using multi-kernel seems promising
  - More experiments needed to be conclusive
- Multi-frame is true performance booster
  - 30% improvement over single-frame baseline
  - Time for the community to move on to video analysis



http://www.vidivideo.eu

References I

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

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 (in press), 2010.

Visual Word Ambiguity. Jan C. van Gemert, Cor J. Veenman, Arnold W. M. Smeulders, Jan-Mark Geusebroek. IEEE Trans. Pattern Analysis and Machine Intelligence (in press), 2009.

On the Surplus Value of Semantic Video Analysis Beyond the Key Frame. Cees G. M. Snoek, Marcel Worring, Jan-Mark Geusebroek, Dennis C. Koelma, and Frank J. Seinstra. Proc. IEEE Int'l Conference on Multimedia & Expo, 2005.

Real-Time Bag of Words, Approximately. Jasper R. R. Uijlings, Arnold W. M. Smeulders, R. J. H. Scha. ACM Int'l Conference on Image and Video Retrieval, 2009.

Empowering Visual Categorization with the GPU. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. In Proc. Annual Conference of the Advanced School for Computing and Imaging, 2009.

http://www.vidivideo.eu

References II

Detecting Audio Events for Semantic Video Search. M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad. In InterSpeech, 2009.
Audio Contributions to Semantic Video Search. I. Trancoso, T. Pellegrini, J. Portelo, H. Meinedo, M. Bugalho, A. Abad, and J. Neto. Proc. IEEE Int'l Conference on Multimedia & Expo, 2009.
Visual Category Recognition using Spectral Regression and Kernel Discriminant Analysis. M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. van de Sande, and T. Gevers. In Proc. 2nd Int'l Workshop on Subspace, In Conjunction with ICCV, 2009.
Nonsparse Multiple Kernel Learning for Fisher Discriminant Analysis. F. Yan, J. Kittler, K. Mikolajczyk, and A. Tahir. In IEEE Int'l conf. Data Mining, 2009.
The MediaMill TRECVID 2009 Semantic Video Search Engine. C.G.M. Snoek et al. Proceedings of the TRECVID Workshop, 2009.
Concept-Based Video Retrieval. C.G.M. Snoek, M. Worring. Foundations and Trends in Information Retrieval, Vol. 4 (2), page 215-322, 2009.

Contact info

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