Multimedia analysis and retrieval

Vasileios Mezaris

(with the contribution of several colleagues)

Information Technologies Institute (ITI)

Centre for Research and Technology Hellas (CERTH)

ESSIR'15, Sept. 2015, Thessaloniki, Greece

Information Technologies Institute

Overview

- Introduction & motivation
- Technologies for multimedia analysis and retrieval
  - Video temporal segmentation to shots
  - Video temporal segmentation to scenes
  - Concept-based image/video indexing
  - Complex event detection
- A brief view on a few other relevant technologies and applications
- Concluding remarks

For each class of technologies, we go through:

- Problem statement
- Brief overview of the literature
- A closer look at one or more approaches
- Indicative experiments and results, demos
- Conclusions and future work
- Additional reading (references)

Introduction & motivation

- Why multimedia retrieval?
  - We are all multimedia consumers and multimedia producers
  - Multimedia is part of almost everything we do on the Web and Social Web…
    - Posting / watching videos in e.g. YouTube
    - Posting / looking at images in e.g. facebook, flickr, pinterest
    - Posting / watching lectures or slides in e.g. VideoLectures, slideshare
    - Posting / consuming tweets, watching or reading the news, etc.

  …regardless of what our goal is…
    - Entertainment, Information (news) or Education
    - Social activities & communication with friends
    - Work, search for product information & reviews, DIY instructions, etc.

  …and regardless of where & how we produce or consume the content
    - Using a PC or laptop, at home or office
    - Using a smart TV, at home
    - Using a mobile phone or tablet or smart watch or camera or…, anywhere and anytime.

Introduction & motivation

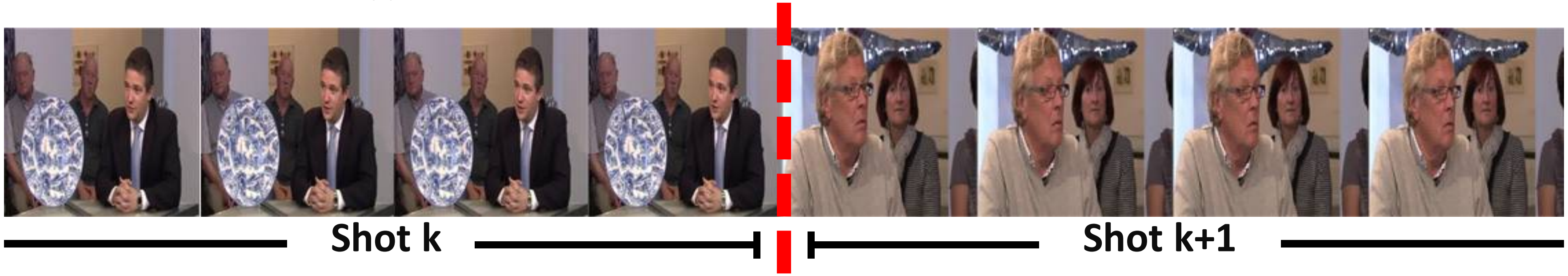
- Why focus on video retrieval?
  - Video is among the most expressive forms of multimedia
  - Video is predominant on the Web and will be even more so in the near future
  - A few Cisco predictions
    - Consumer internet video traffic will be 80 percent of all consumer Internet traffic in 2019
    - Every second, nearly a million minutes of video content will cross the network by 2019
  - A few YouTube statistics (August 2015)
    - YouTube has more than 1 billion users
    - 300 hours of video are uploaded to YouTube every minute
    - Every day people watch hundreds of millions of hours on YouTube and generate billions of views

Introduction & motivation

- Why do we need video analysis?
  - A human viewing a video sees objects and actors, can identify actors' actions, interactions between objects and actors, can recognize a temporal succession of events, possibly sees a story that has a meaningful structure, experiences feelings,…
  - A computer playing a video file experiences (after decoding the compressed video) a long sequence of pixel values, ordered in a 2D image grid and in time…not so exciting or helpful for meeting the information needs of the human video consumer
  - We need techniques that allow for automatically
    - Identifying a video's structure and constituent elements
    - Express in a human-understandable way the contents of each video element

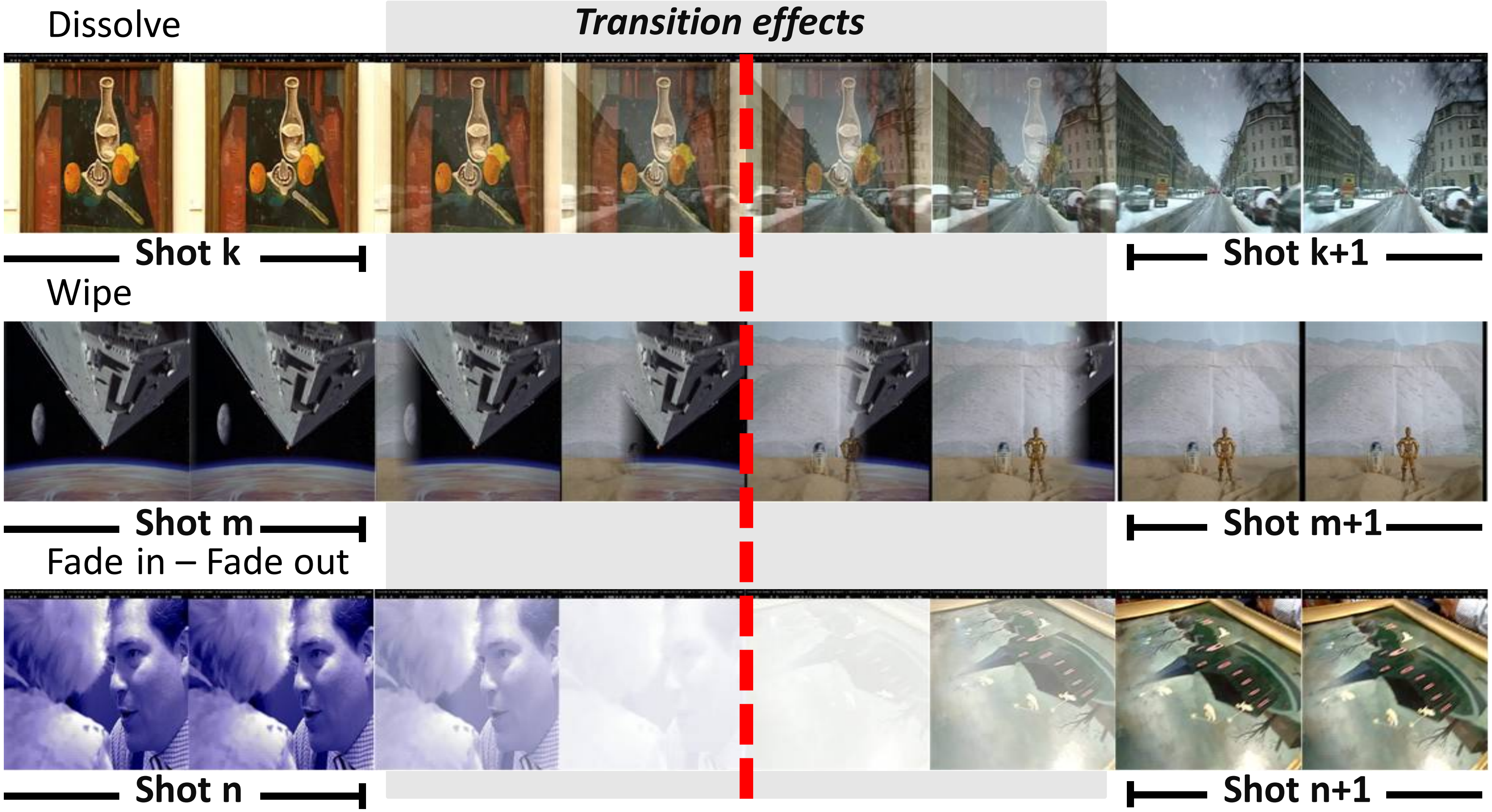
Video temporal segmentation to shots

- What is a shot: a sequence of consecutive frames taken without interruption by a single camera
- Elementary video unit for indexing; important for traditional and new retrieval paradigms (e.g. origin of video hyperlinks)
- Shot segmentation
  - Detecting the boundaries or changes between the shots
  - Important pre-processing step for further analysis (scene segmentation, concept/event detection, object re-detection)
- Shot change is manifested by a shift in visual content
  - Two basic types of transitions: **ABRUPT** transitions



Video temporal segmentation to shots

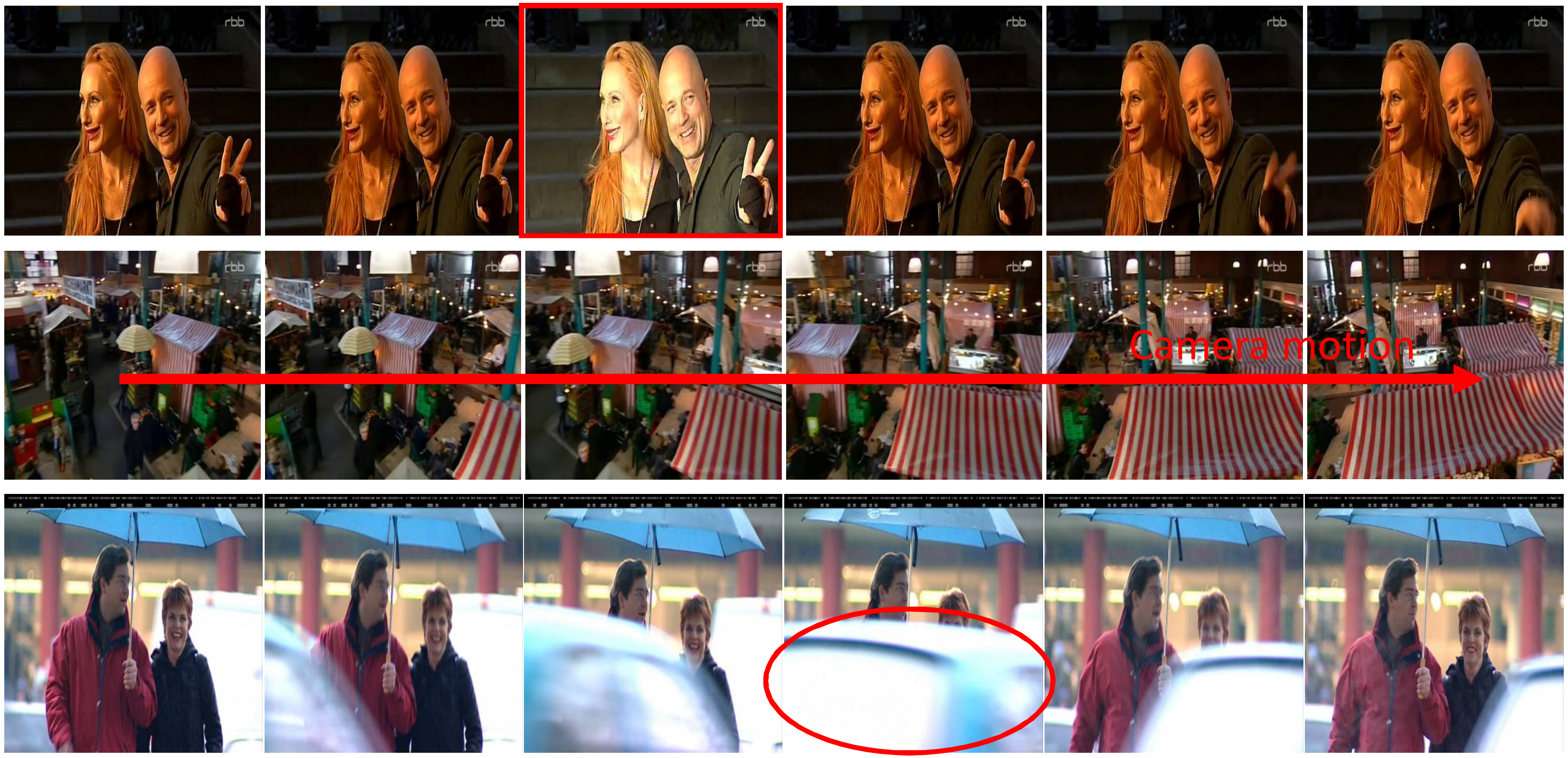
GRADUAL transitions



Video temporal segmentation to shots

- The analysis method should be robust to artifacts such as illumination changes, fast camera movement, rapid local (object) motion

Sudden illumination change



Partial occlusion due to object motion (car)

Related work

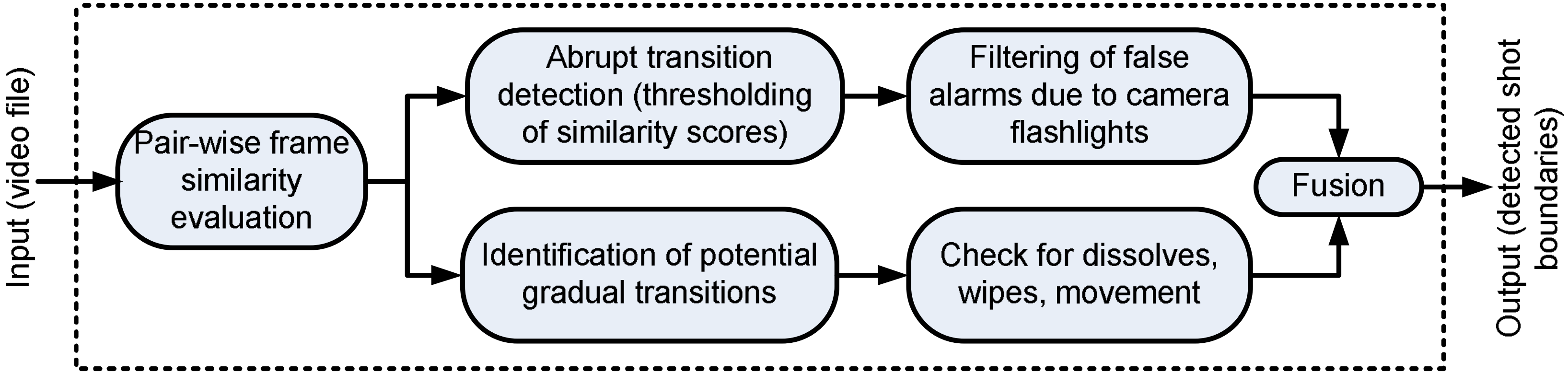
- Can generally be organized according to
  - Data to work with: uncompressed vs. compressed video
  - Features to use (also depends on the data)
  - Threshold-based vs. learning-based methods
- Compressed video methods
  - Reduce computational complexity by avoiding decoding, exploiting encoder results
    - Macroblock information of specific frames (e.g. intra-coded, skipped)
    - DC coefficients of the compressed images
    - Motion vectors included in the compressed data stream
  - Generally, very fast but not as accurate as uncompressed video methods

Related work

- Uncompressed video methods
  - Pair-wise pixel comparisons
  - Global visual feature comparisons (e.g. color histogram comparison)
  - Edge-based approaches, e.g. evaluating an edge change ratio
  - Motion-based approaches
  - Local visual features / Bag of Visual Words
    - Some features more computationally expensive than others
  - Deciding using experimentally-defined thresholds: often hard to tune → the alternative is to use machine learning techniques (often Support Vector Machines (SVMs)) for learning from different features
- General remark: high detection accuracy and relatively low computational complexity are possible when working with uncompressed data

An indicative approach

- Uncompressed video approach (based on [1])
- Detects both abrupt and gradual shot transitions by:
  - Extracting global and local visual features (HSV histograms, ORB descriptors) for every frame and evaluating the differences between every pair of consecutives frames (abrupt transitions)
  - Analyzing the sequence of similarity scores, to identify patterns that correspond to progressive changes of the visual content and then applying dissolve, wipe and movement detectors (gradual transitions)


Input (video file)
Pair-wise frame similarity evaluation
Abrupt transition detection (thresholding of similarity scores)
Filtering of false alarms due to camera flashlights
Identification of potential gradual transitions
Check for dissolves, wipes, movement
Fusion
Output (detected shot boundaries)

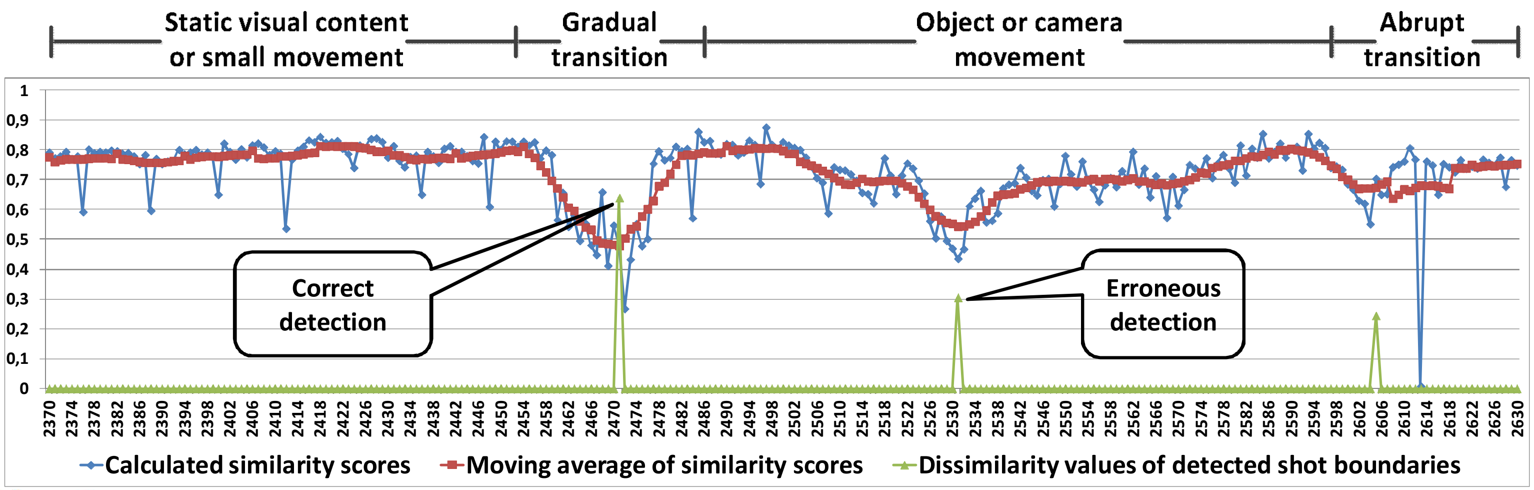
[1] E. Apostolidis, V. Mezaris, "Fast Shot Segmentation Combining Global and Local Visual Descriptors", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014.

An indicative approach

- Gradual transition detection
  - **Step 1:** pair-wise frame similarity scores *F(x)* (blue curve) → moving average *G(k)* (red)
  - **Step 2:** compute first order deriv. of *G(k)* and store local minima, maxima in *Gmin*, *Gmax*
  - **Step 3:** If frame k is the p-th element in *Gmin*, compute

$$D(k)=\left|F(\mathrm{Gmin}(p))-F(\mathrm{Gmax}(p-1))\right|+\left|F(\mathrm{Gmin}(p))-F(\mathrm{Gmax}(p))\right|$$

  - **Step 4:** Compute C(k) (green) as the sum of the values of *D(k)* within a sliding window
  - **Step 5:** Potential gradual transitions are identified by thresholding *C(k)*
  - **Detection:** Gradual transitions are defined by evaluating each potential transition using (a) dissolve detector, (b) wipe detection and (c) detector for object/camera movement



Indicative experiments and results

- Dataset
  - About 7 hours of video
    - 150 min. of news shows
    - 140 min. of cultural heritage shows
    - 140 min. of various other genres
- Ground-truth (generated via manual annotation)
  - 3647 shot changes
    - 3216 abrupt transitions
    - 431 gradual transitions
- System specifications
  - Intel Core i7 processor at 3.4GHz
  - 8GB RAM memory

Indicative experiments and results

- Detection accuracy expressed in terms of:
  - Precision (P): the fraction of detected shots that correspond to actual shots of the videos
  - Recall (R): the fraction of actual shots of the videos, that have been successfully detected
  - F-Score: 2(PR)/(P+R)
- Time performance: runs ~ 7-8 times faster than real-time (utilizing multi-threading)

Experimental Results	
Precision	**94.3 %**
Recall	**94.1 %**
F-Score	**0.942**

- Demo video http://www.youtube.com/watch?v=0IeVkXRTYu8
- Software available at http://mklab.iti.gr/project/video-shot-segm

Shot detection conclusions

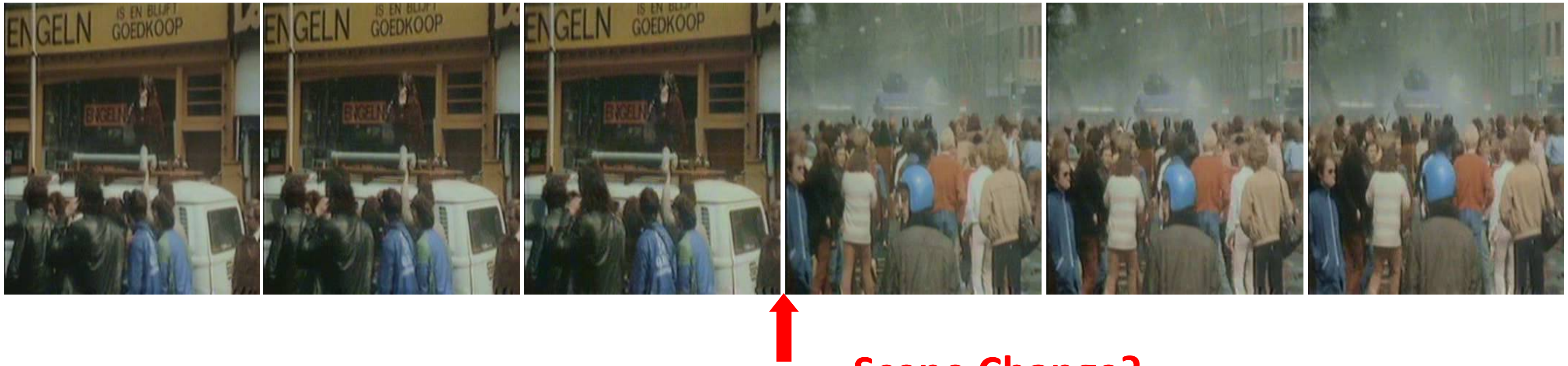
- Overall accuracy of shot detection methods is high, sufficient for any application
- Detection of abrupt transitions is very easy; gradual transitions & handling of intense motion still a bit more challenging
- Several times faster-than-real-time processing is feasible
- More elaborate motion-based analysis is still possible (but would increase the computational load)

Shot detection: additional reading

- E. Apostolidis, V. Mezaris, "Fast Shot Segmentation Combining Global and Local Visual Descriptors", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014.
- E. Tsamoura, V. Mezaris, and I. Kompatsiaris, “Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework”, Proc. 15th IEEE Int. Conf. on Image Processing, 2008.
- V. Chasanis, A. Likas, and N. Galatsanos, “Simultaneous detection of abrupt cuts and dissolves in videos using support vector machines,” Pattern Recogn. Lett., vol. 30, no. 1, pp. 55–65, Jan. 2009.
- Z. Qu, Y. Liu, L. Ren, Y. Chen, and R. Zheng, “A method of shot detection based on color and edge features,” Proc. 1st IEEE Symposium on Web Society, 2009. SWS ’09. 2009, pp. 1–4.
- J. Lankinen and J.-K. Kamarainen, “Video shot boundary detection using visual bag-of-words,” Proc. Int. Conf. on Computer Vision Theory and Applications (VISAPP), Barcelona, Spain, 2013.
- J. Li, Y. Ding, Y. Shi, and W. Li, “A divide-and-rule scheme for shot boundary detection based on sift,” JDCTA, pp. 202–214, 2010.
- C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan, L.-H. Chen, “A motion-tolerant dissolve detection algorithm”, IEEE Transactions on Multimedia, vol. 7, no. 6, pp. 1106–1113, 2005.
- K. D. Seo, S. Park, S. H. Jung, “Wipe scene-change detector based on visual rhythm spectrum”, IEEE Transactions on Consumer Electronics, vol. 55, no. 2, pp. 831–838, May 2009.
- D. Lelescu and D. Schonfeld, “Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream”, IEEE Transactions on Multimedia,,vol. 5, no. 1, pp. 106–117, 2003.
- J. H. Nam and A.H. Tewfik, “Detection of gradual transitions in video sequences using b-spline interpolation”, IEEE Transactions on Multimedia, vol. 7, no. 4, pp. 667–679, 2005.
- C. Grana and R. Cucchiara, “Linear transition detection as a unified shot detection approach”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 4, pp. 483–489, 2007.
- Z. Cernekova, N. Nikolaidis, and I. Pitas, “Temporal video segmentation by graph partitioning”, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2006.

Video temporal segmentation to scenes

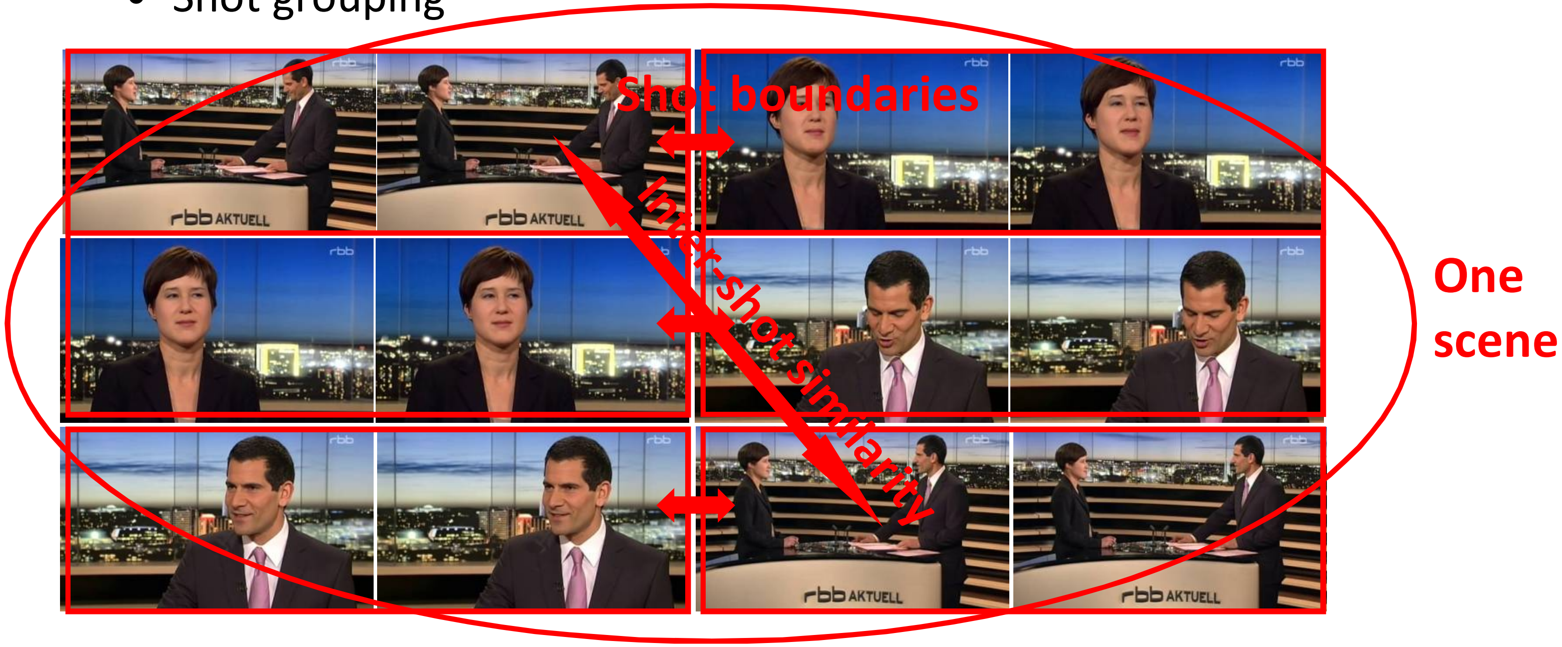
- What is a scene: a higher-level temporal video segment that is elementary in terms of semantic content, covering either a single event or several related events taking place in parallel
- Scene segmentation: temporal decomposition of videos into such basic story-telling units
- Scene change is not manifested by just a change in visual content



Scene Change?

Problem statement

- Basic assumptions
  - A shot cannot belong to more than one scenes
  - Scene boundaries are a subset of the visual shot boundaries of the video
- Then, scene segmentation can (and typically is) performed by
  - Shot segmentation, and
  - Shot grouping

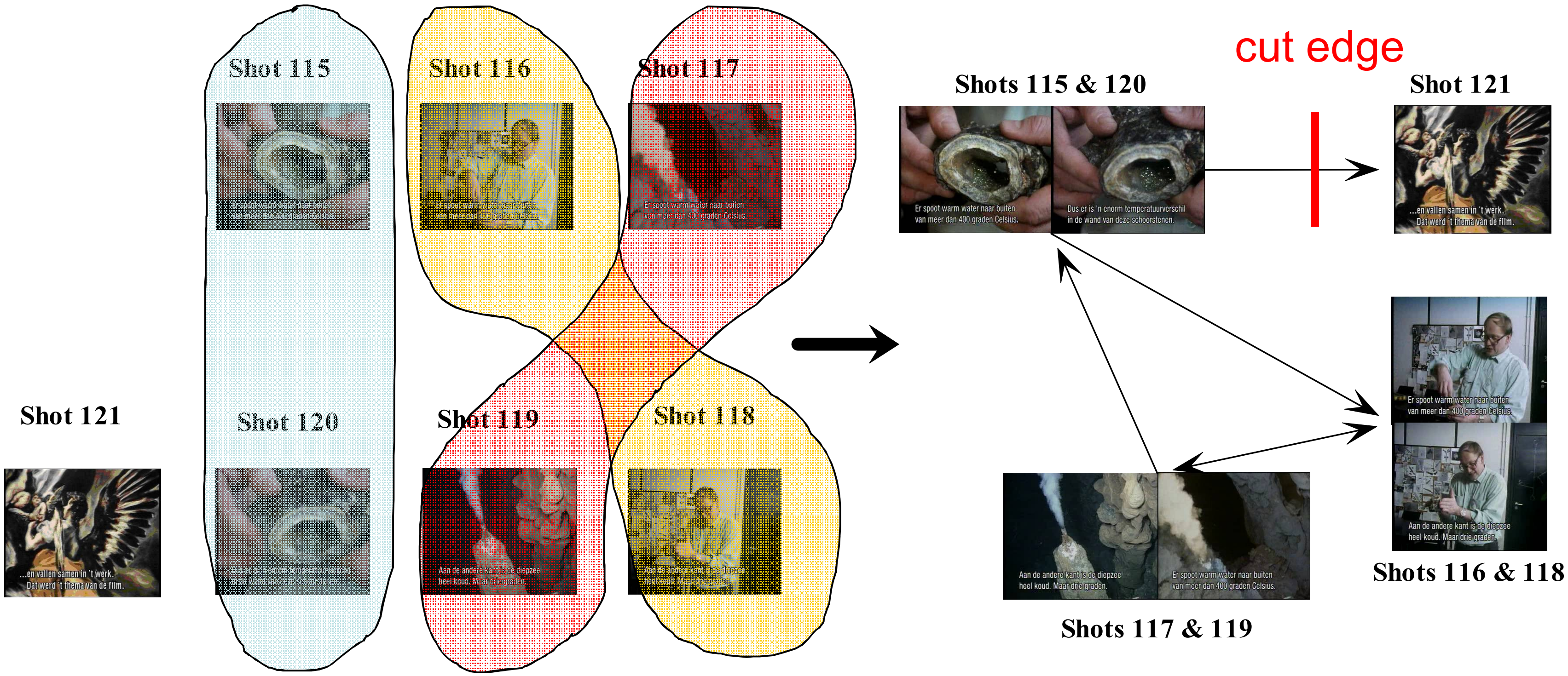


Related work

- Can generally be organized according to
  - Data to work with: uni-modal vs. multi-modal
  - Dependence or not on specific-domain knowledge; domain of choice
  - Algorithms used
- Uni-modal vs. multi-modal
  - Uni-modal methods use one type of information, typically visual cues
  - Multi-modal ones may combine visual cues, audio, speech transcripts, …
- Domain-specific vs. domain-independent
  - Domain-independent methods are generally applicable
  - News-domain (e.g. using knowledge of news structure), TV broadcast domain (e.g. based on advertisement detection), etc.
- Algorithms
  - Graph-based, e.g. the Scene Transition Graph
  - Clustering-based, e.g. using hierarchical clustering
  - Based on statistical methods, e.g. on Markov Chain Monte Carlo (MCMC)

An indicative approach

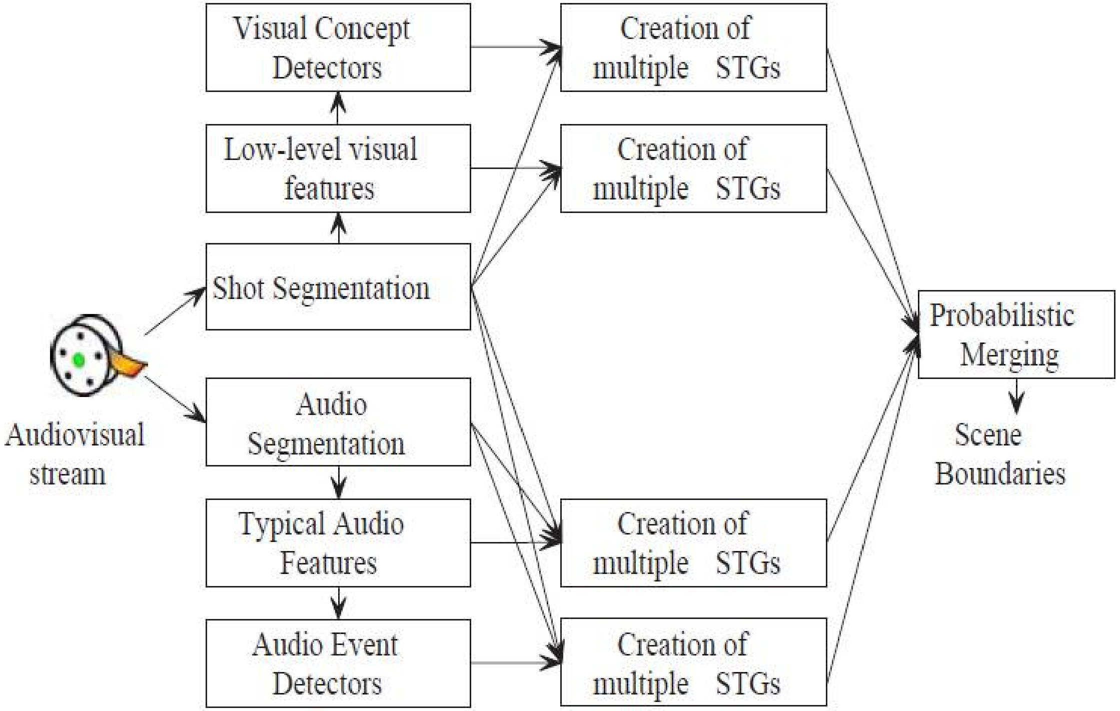
- Based on the Scene Transition Graph (STG) algorithm [2]


Shot 115
Shot 116
Shot 117
Shot 121
Shot 120
Shot 119
Shot 118
...en vallen samen in 't werk. Dat werd 't thema van de film.
cut edge
Shots 115 & 120
Shot 121
Er spoot warm water naar buiten van meer dan 400 graden Celsius.
Dus er is 'n enorm temperatuurverschil in de wand van deze schoorstenen.
...en vallen samen in 't werk. Dat werd 't thema van de film.
Er spoot warm water naar buiten van meer dan 400 graden Celsius.
Aan de andere kant is de diepzee heel koud. Maar drie graden.
Aan de andere kant is de diepzee heel koud. Maar drie graden.
Er spoot warm water naar buiten van meer dan 400 graden Celsius.
Shots 117 & 119
Shots 116 & 118

[2] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. Computer Vision Image Understanding, 71(1):94–109, July 1998.

An indicative approach

- Introduces two extensions of the STG [3]
  - Fast STG approximation (scenes as convex sets of shots; linking transitivity rules)
  - Generalized STG (probabilistic merging of multiple STGs created with different parameter values, different features)



[3] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. IEEE Transactions on Circuits and Systems for Video Technology, 21(8):1163 –1177, August 2011.

Indicative experiments and results

- Dataset
  - 513 min. of documentaries (A)
  - 643 min. of movies (B)
- Ground-truth (generated via manual annotation)
  - 3459 (in A) + 6665 (in B) = 10125 shot changes
  - 525 (in A) + 357 (in B) = 882 scene changes
- System specifications
  - Intel Core i7 processor at 3.4GHz
  - 8GB RAM memory

Indicative experiments and results

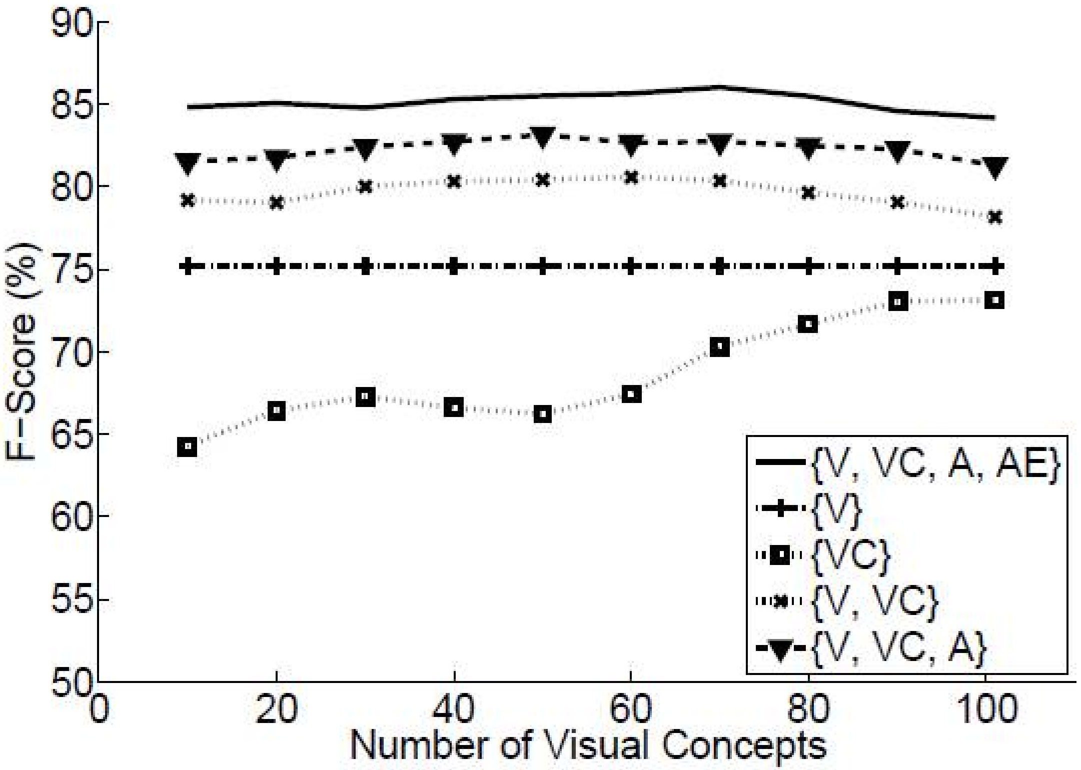
- Detection accuracy expressed in terms of:
  - Coverage (C): to what extent frames belonging to the same scene are correctly grouped together (optimal value 100%)
  - Overflow (O): the quantity of frames that, although not belonging to the same scene, are erroneously grouped together (optimal value 0%)
  - F-Score = 2C(1−O)/(C+(1−O))

	Coverage (%)	Overflow (%)	F-Score (%)
Documentaries	76.96	20.80	78.06
Movies	73.55	26.11	73.72

- Time performance: the algorithm runs in 0,015x real time (i.e., ~67 times faster than real-time), as long as the features have been extracted
- Software available at http://mklab.iti.gr/project/video-shot-segm (using low-level visual features only)

Indicative experiments and results

- Contribution of different modalities (on a different dataset)



V: low-level visual features

VC: visual concept detection results

A: low-level audio features

AE: audio event detection results

Scene detection conclusions

- Automatic scene segmentation less accurate than shot segmentation…
- …but the results are good enough for improving access to meaningful fragments in various applications (e.g. retrieval, video hyperlinking)
- Using more than just low-level visual features helps a lot
- The choice of domain-specific vs. domain-independent method should be taken seriously

Scene detection: additional reading

- M. Yeung, B.-L. Yeo, and B. Liu, “Segmentation of video by clustering and graph analysis”, Computer Vision Image Understanding, vol. 71, no. 1, pp. 94–109, July 1998.
- P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, “Temporal video segmentation to scenes using high-level audiovisual features”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 8, pp. 1163 –1177, August 2011.
- P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, J. Kittler, “Differential Edit Distance: A metric for scene segmentation evaluation”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 6, pp. 904-914, June 2012.
- Z. Rasheed and M. Shah, “Detection and representation of scenes in videos”, IEEE Transactions on Multimedia, vol. 7, no. 6, pp. 1097–1105, December 2005.
- C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, “Video summarization and scene detection by graph modeling”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 2, pp. 296–305, February 2005.
- Y. Zhao, T. Wang, P. Wang, W. Hu, Y. Du, Y. Zhang, G. Xu, “Scene segmentation and categorization using N-cuts”, IEEE Conf. on Computer Vision and Pattern Recognition, 2007.
- X. Zhu, A.K. Elmagarmid, X. Xue, L. Wu, and A.C. Catlin, “Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval”, IEEE Transactions on Multimedia, vol. 7, no. 4, pp. 648– 666, August 2005.
- B. T. Truong, S. Venkatesh, and C. Dorai, “Scene extraction in motion pictures”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 1, pp. 5–15, January 2003.
- X.-S. Hua, L. Lu, and H.-J. Zhang, “Optimization-based automated home video editing system”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 5, pp. 572–583, May 2004.
- D. Gatica-Perez, A. Loui, and M.-T. Sun, “Finding structure in consumer videos by probabilistic hierarchical clustering”, IEEE Transactions on Circuits and Systems for Video Technology, 2002.

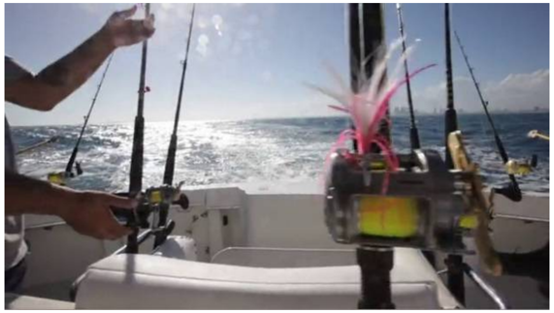
Scene detection: additional reading

- J. Liao and B. Zhang, “A robust clustering algorithm for video shots using haar wavelet transformation”, Proc. SIGMOD2007 Workshop on Innovative Database Research (IDAR2007), Beijing, China, June 2007.
- Y. Zhai and M. Shah, “Video scene segmentation using markov chain monte carlo”, IEEE Transactions on Multimedia, vol. 8, no. 4, pp. 686–697, August 2006.
- M. Sugano, K. Hoashi, K. Matsumoto, and Y. Nakajima, “Shot boundary determination on MPEG compressed domain and story segmentation experiments for trecvid 2004, in trec video retrieval evaluation forum”, Proc. TREC Video Retrieval Evaluation (TRECVID), Washington, DC, USA, pp. 109–120, 2004.
- L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, “Structure analysis of soccer video with domain knowledge and hidden markov models”, Pattern Recogn. Lett., vol. 25, no. 7, pp. 767–775, May 2004.
- H. Lu, Z. Li, and Y.-P. Tan, “Model-based video scene clustering with noise analysis”, Proc. Int. Symposium on Circuits and Systems, ISCAS ’04, vol. 2, pp. 105–108, May 2004.
- Y. Ariki, M. Kumano, and K. Tsukada, “Highlight scene extraction in real time from baseball live video”, Proc. 5th ACM SIGMM international workshop on Multimedia information retrieval, MIR ’03, pp. 209–214, New York, NY, USA, 2003.
- U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll, “New approaches to audio-visual segmentation of tv news for automatic topic retrieval”, Proc. IEEE Int. Conf. on the Acoustics, Speech, and Signal Processing (ICASSP ’01), vol. 3, pp. 1397–1400, Washington, DC, USA, 2001.
- Y. Cao, W. Tavanapong, K. Kim, and J.H. Oh, “Audio-assisted scene segmentation for story browsing”, Proc. 2nd Int. Conf. on Image and Video Retrieval, CIVR’03, pp. 446–455, Springer-Verlag.
- A. Velivelli, C.-W. Ngo, and T. S. Huang, “Detection of documentary scene changes by audio-visual fusion”, Proc. 2nd Int. Conf. on Image and Video Retrieval, CIVR’03, pp. 227–238, Springer-Verlag.

Concept-based image/video indexing

- Goal: assign one or more semantic concepts to images or temporal video fragments (typically, shots), from a predefined concept list
  - Input: image, video fragment or representative information (e.g. keyframes)
  - Output: concept labels and associated confidence scores (DoC)
- Application: find similar media items (concept-based search and retrieval; also clustering, summarization, further analysis e.g. event detection, etc.)
- Concept detection is challenging: semantic gap, annotation effort, computational requirements,…

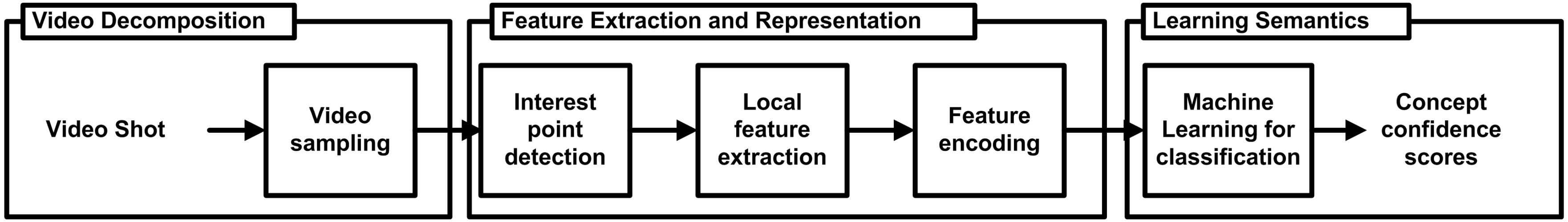
Sample keyframe



hand: 0.97,
sky: 0.93,
sea: 0.91,
boat: 0.91, …

Related work – handcrafted features

- Typical concept detection system : made of independent concept detectors
  - Feature extraction (typically local features; choices of IP detectors/ descriptors)
  - Feature encoding (Bag of Words, Fisher vectors, VLAD,…)
  - Training/classification (supervised learning; need for annotated training data)
  - Late fusion and output score post-processing



- How to build a competitive system
  - Use color-, rotation- , scale- invariant descriptors; SoA encoding of them
  - Fuse multiple descriptors and machine learning-based detectors
  - Exploit concept correlations (e.g., sun & sky often appearing together)
  - Exploit temporal information (videos)

Related work – handcrafted features

- Feature extraction
  - Visual features
    - Global vs. local
    - Popular local descriptors SIFT, Color SIFT , SURF; also look at binary local descriptors
    - Interest point detection: Harris-Laplace, Hessian, dense sampling
  - Motion features (STIP, MoSIFT, feature trajectories,…)
  - Others modalities (text, audio): of limited use
- Feature encoding
  - Bag-of-words (BoW): codebook construction (K-means); hard/soft assignment
  - Fisher vectors: characterize each keyframe by a gradient vector
  - Others: VLAD, Super Vector (SV),…
  - Capture spatial information: Pyramidal decomposition,…
  - Capture information at multiple scales: Pyramid Histogram Of visual Words (PHOW)

Related work – handcrafted features

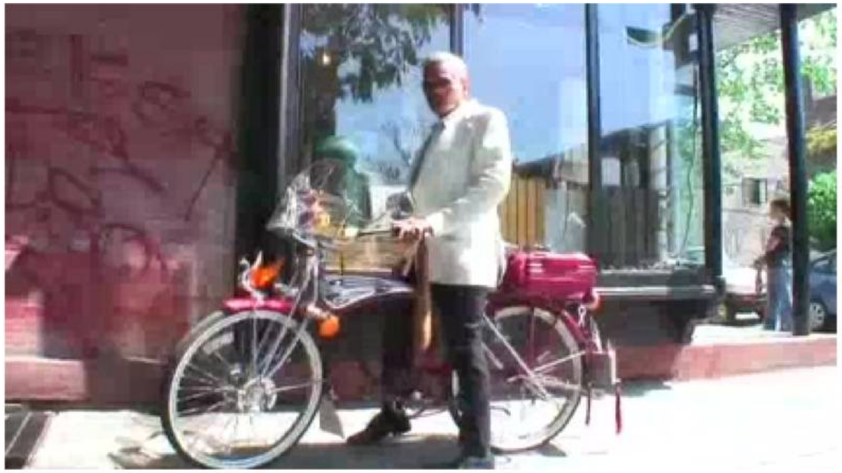
- Machine learning
  - Binary classification: Support Vector Machines (SVMs), Logistic Regression,…
    - Linear vs. kernel methods
  - Multi-label learning approaches; Stacking
- Fusion
  - Of features (early / late)
  - Of detectors
- Post-processing: temporal re-ranking,…

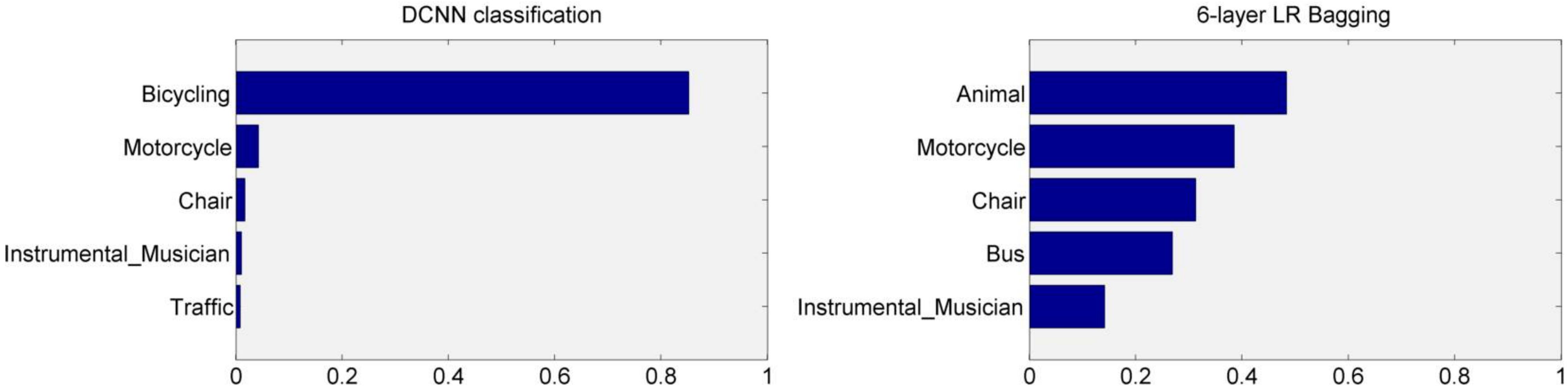
Related work – DCNNs

- Deep Convolutional Neural Networks
  - Neural networks that are made of multiple layers
  - Input to a DCNN is the visual information (the images pixel values)
  - Output is a vector of responses, each corresponding to one of the different concepts that the DCNN was trained for
- For concept-based indexing, a DCNN can be used
  - As an end-to-end concept detector
  - As a generator of (learned) features, which can be used as input to a machine learning approach similar to those used for handcrafted features
    - As features, the output of the last as well as other DCNN layers can be used
- Several DCNN software libraries available, e.g. Caffe, MatConvNet
- Different DCNN architectures have been proposed for image annotation, e.g. CaffeNet, GoogLeNet, VGG ConvNet

DCNNs as end-to-end concept detectors vs. as generators of features

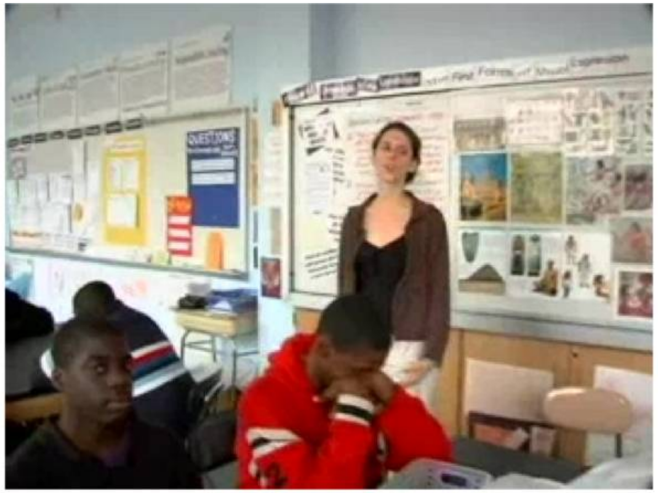
- Example results for the video annotation problem

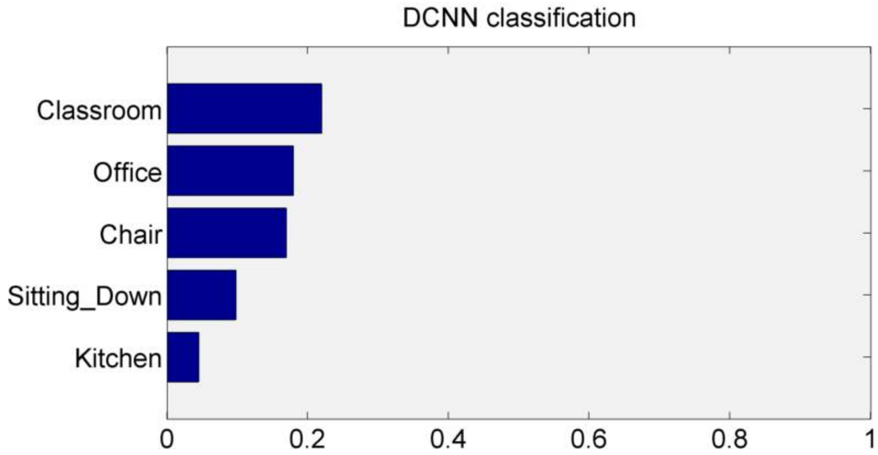




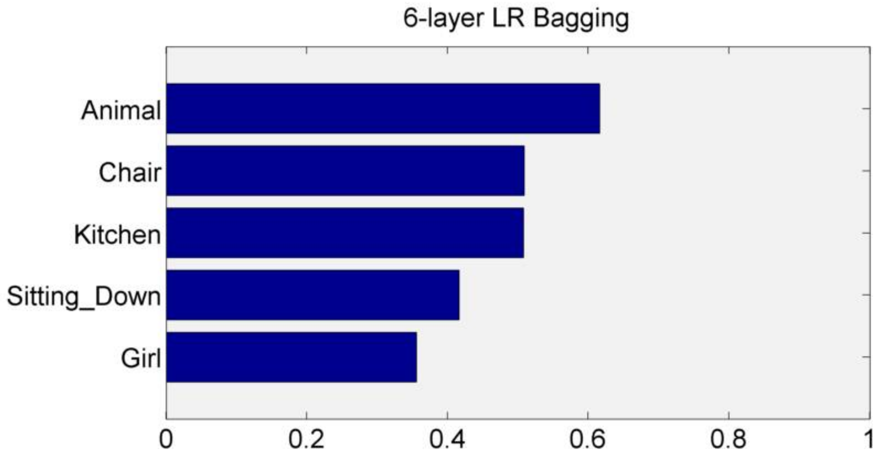
DCNNs as end-to-end concept detectors vs. as generators of features

- Example results for the video annotation problem





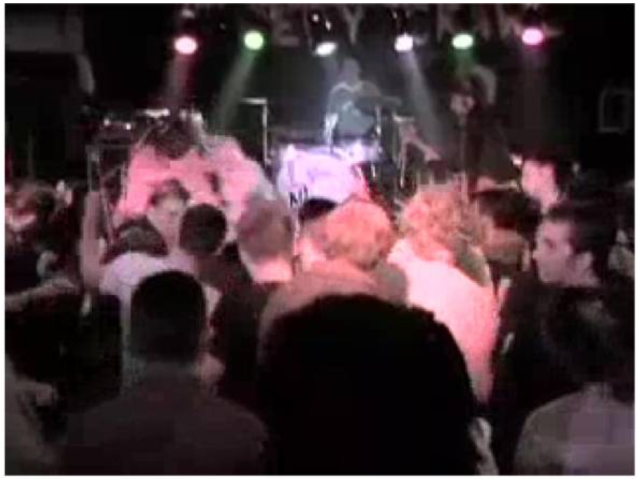
DCNN classification

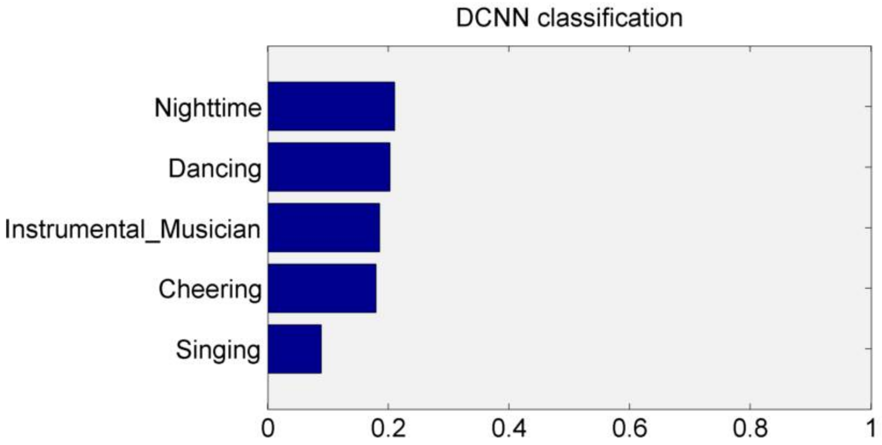


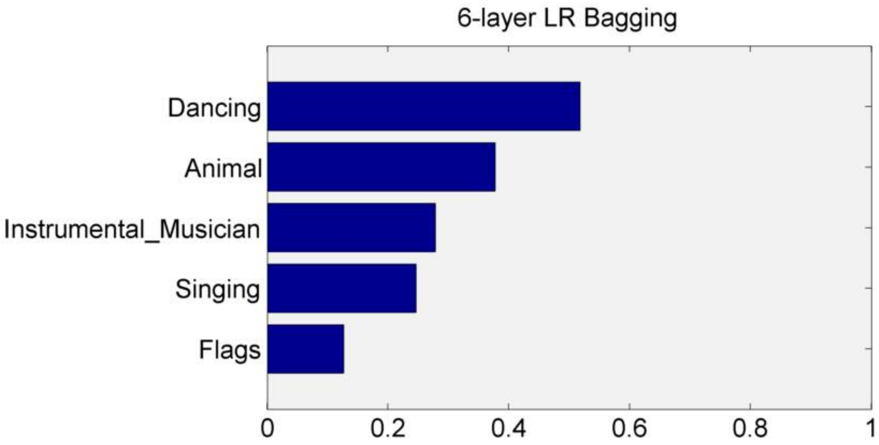
6-layer LR Bagging

DCNNs as end-to-end concept detectors vs. as generators of features

- Example results for the video annotation problem

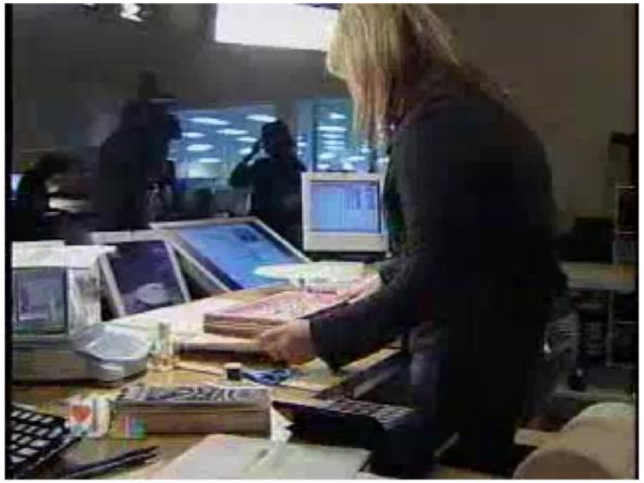


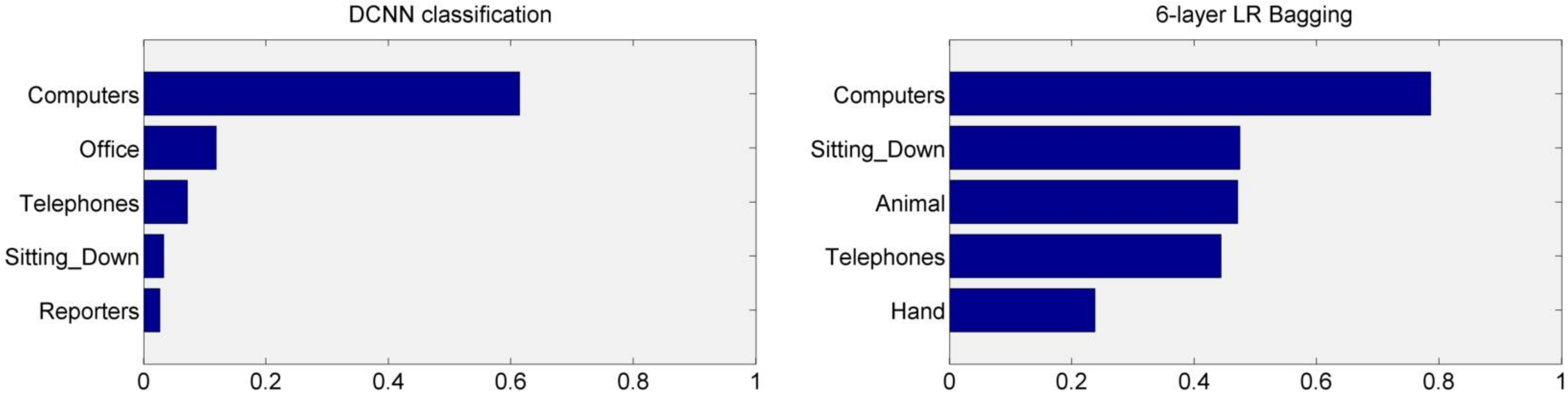




DCNNs as end-to-end concept detectors vs. as generators of features

- Example results for the video annotation problem



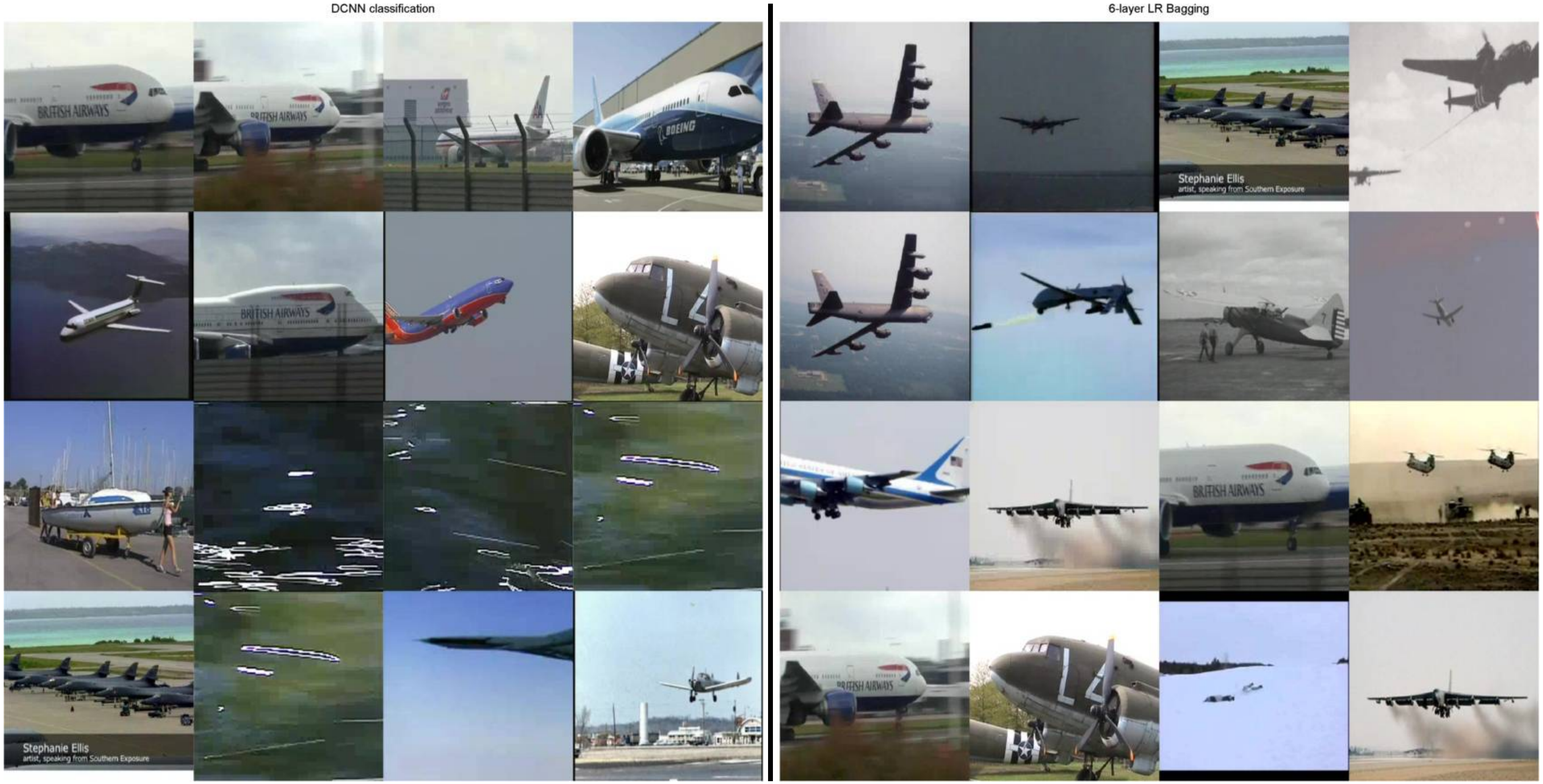


DCNN classification

6-layer LR Bagging

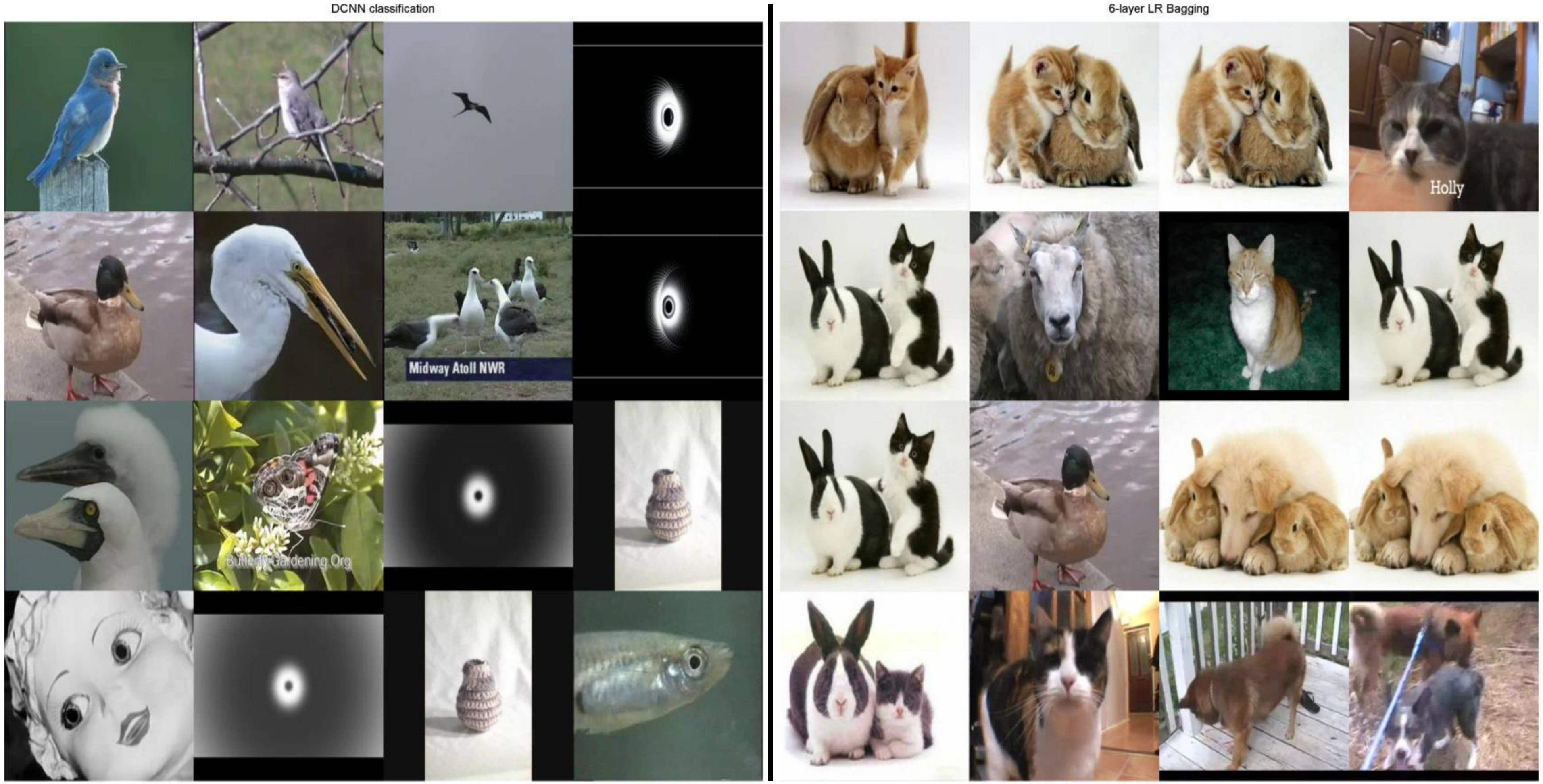
DCNNs as end-to-end concept detectors vs. as generators of features

- Example results for the video retrieval problem (concept: airplane)



DCNNs as end-to-end concept detectors vs. as generators of features

- Example results for the video retrieval problem (concept: animal)

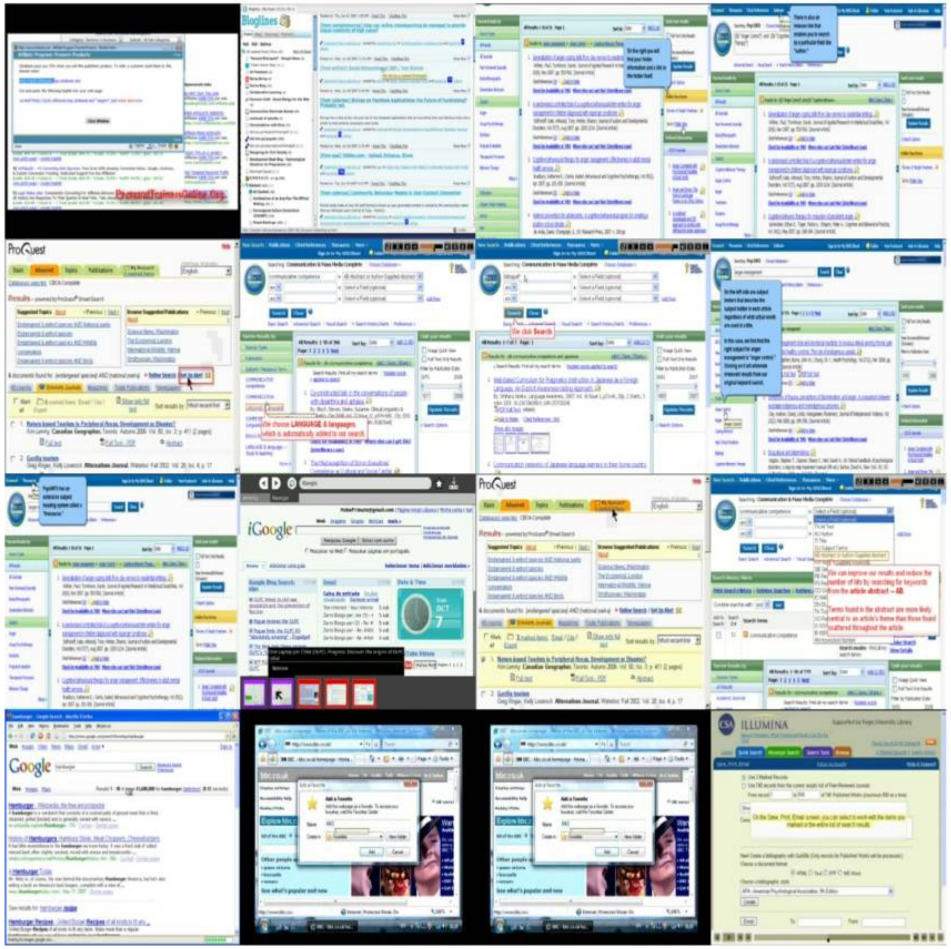

DCNN classification

6-layer LR Bagging

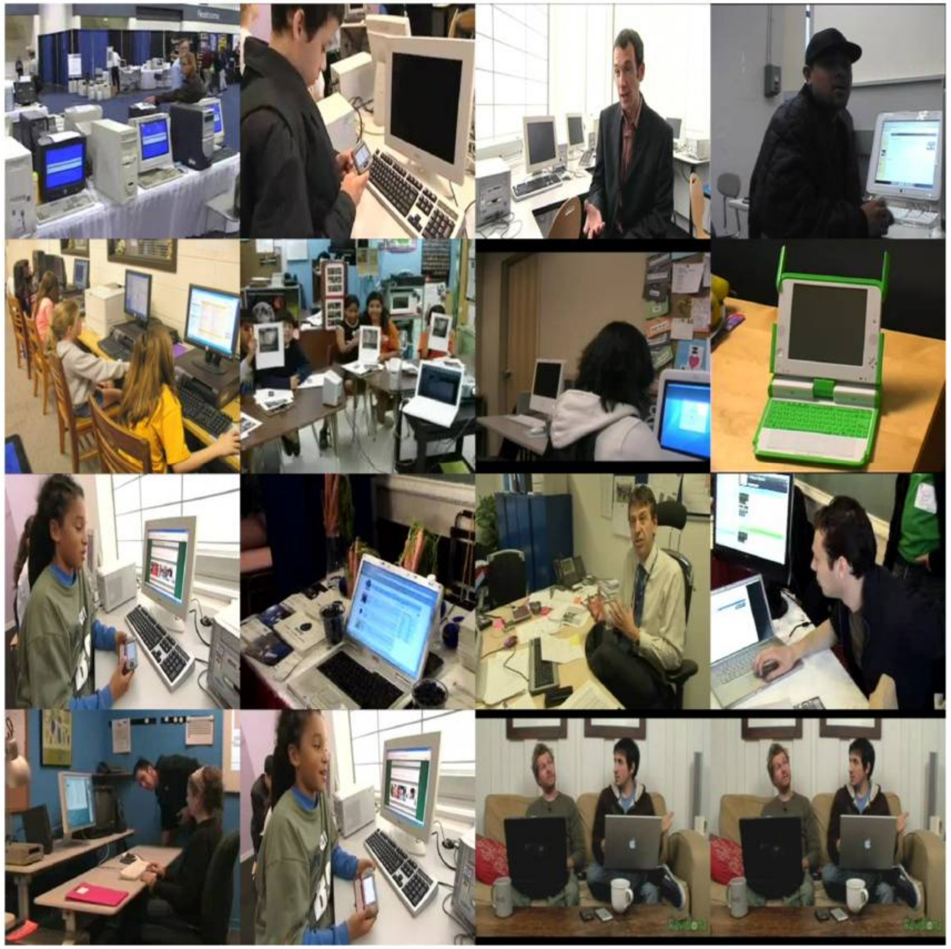
DCNNs as end-to-end concept detectors vs. as generators of features

- Example results for the video retrieval problem (concept: computers)

DCNN classification



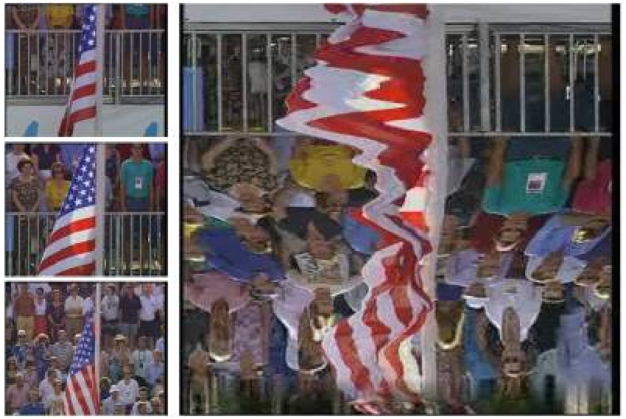
6-layer LR Bagging

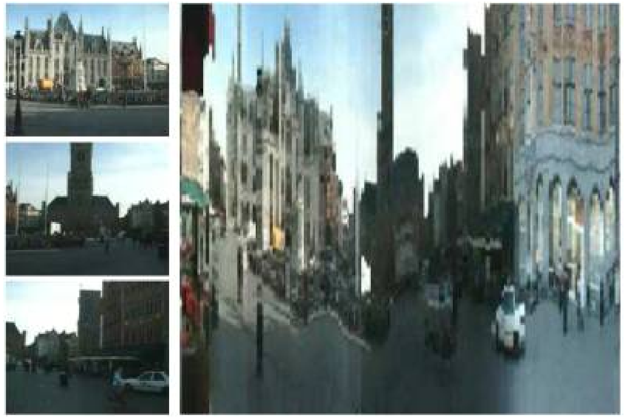


Indicative approach 1: going beyond keyframes for video indexing

- Video tomographs: 2-dimensional slices with one dimension in time and one dimension in space
- We extract two tomographs; use them together with keyframes
- The two tomographs are processed in the same way as keyframes [4]

  (in shown experiments, 3x SIFT/color-SIFT variants, with dense sampling and VLAD encoding)

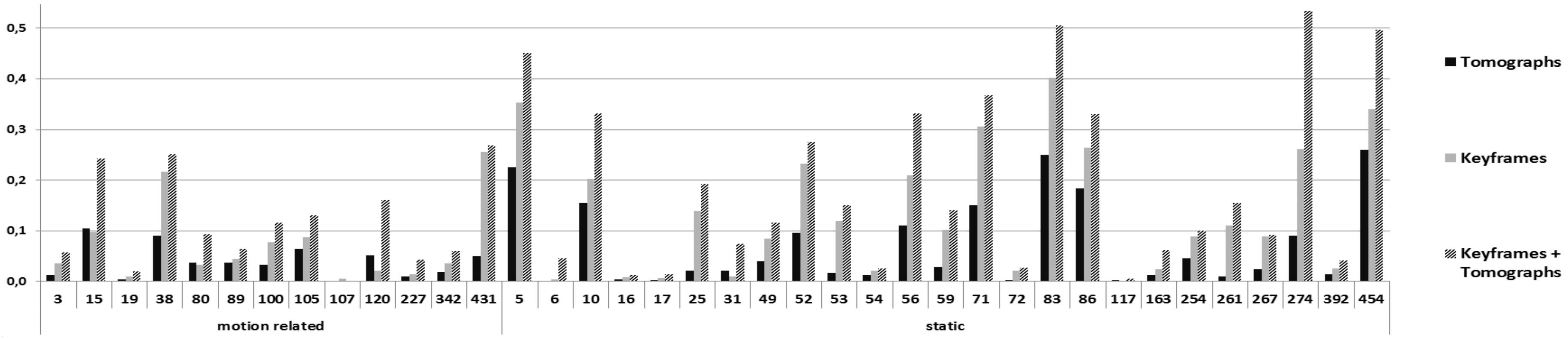




[4] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, "Video tomographs and a base detector selection strategy for improving large-scale video concept detection", IEEE Trans. on Circuits and Systems for Video Tech., 24(7), pp. 1251-1264, July 2014.

Video tomographs: indicative results

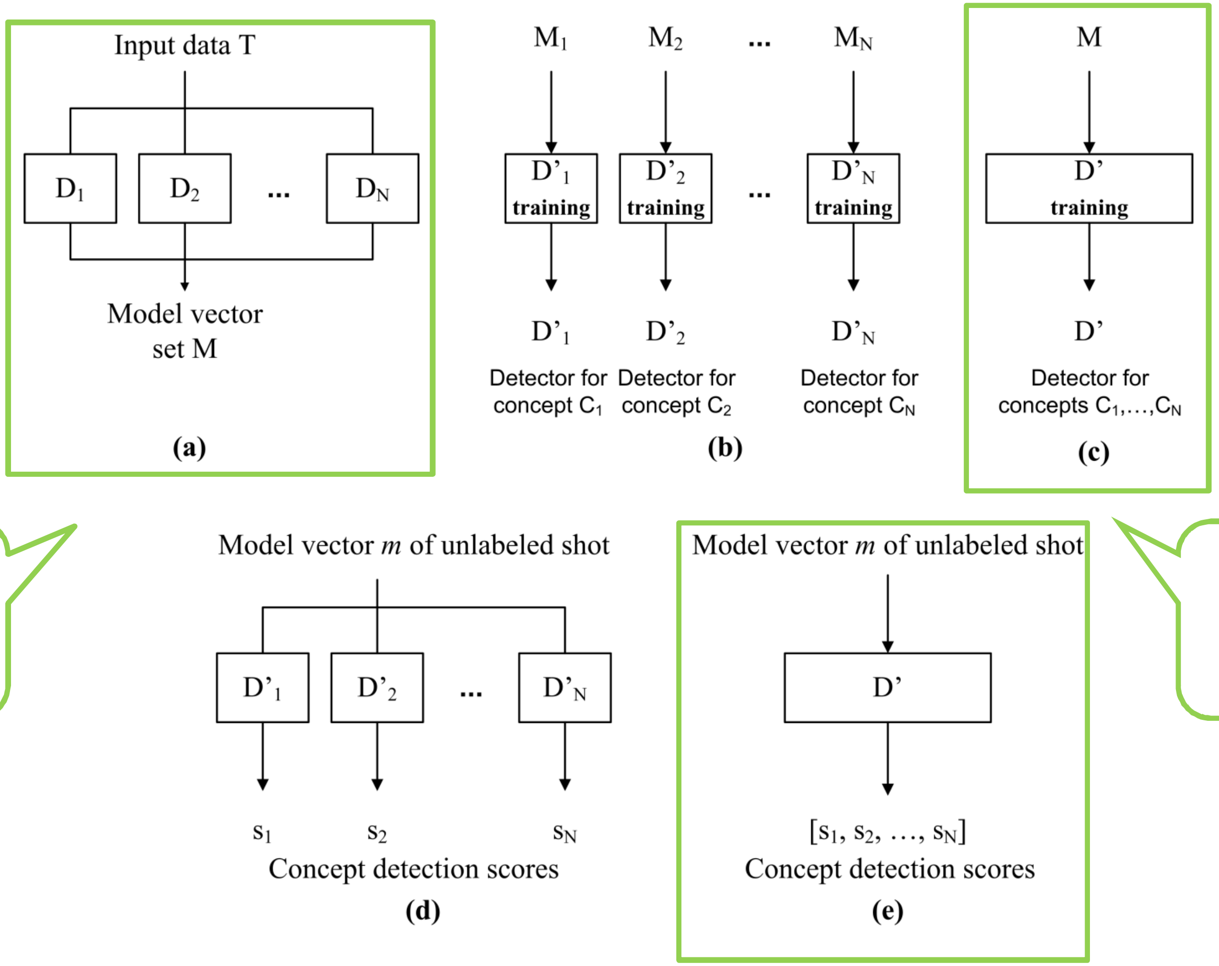
- Experimental setup
  - The TRECVID Semantic Indexing Task: Using the concept detectors retrieve for each concept a ranked list of 2000 test shots that are mostly related with it
  - Dataset: TRECVID 2013 (~800 and ~200 hours of internet archive videos for training and testing), 38 concepts (13 of them motion-related)
  - Evaluation: Extended Inferred Average Precision (xinfAP)



Indicative approach 2: going beyond independent concept detectors (while using handcrafted features)

- Concept detection using a two-layer stacking architecture
  - 1st layer: Build multiple fast and independent concept detectors
    - Use keyframes, and visual tomographs to capture the motion information
    - Extract high-dimensional feature vectors
    - Train Linear Support Vector Machine / Logistic Regression classifiers
    - Easily scalable but does not capture concept correlations
  - 2nd layer: Exploit concept correlations and refine the scores
    - Construct low-dimensional model vectors
    - Use multi-label learning to capture the concept correlations (e.g. ML-*k*NN or Label Powerset algorithms)
    - Use temporal re-ranking to exploit temporal information

2^{nd}-layer of stacking architecture



The first layer (baseline)

The proposed 2^{nd} layer

2^{nd}-layer of stacking architecture

- Exploit concept correlations
  - Construct model vectors by concatenating the responses of the concept detectors on a validation set
  - Use them to train 2^{nd}-layer classifier
- ML-*k*NN as the 2^{nd}-layer classifier [5]
  - A lazy style multi-label learning algorithm; uses label correlations in the neighbourhood of the tested instance to infer posterior probabilities.
- Label Powersets (LP) as the 2nd-layer classifier [5]
  - Searches for subsets of labels that appear together in the training set; considers each set as a separate class in order to solve a multi-class problem

[5] F. Markatopoulou, V. Mezaris, I. Kompatsiaris. A Comparative Study on the Use of Multi-Label Classification Techniques for Concept-Based Video Indexing and Annotation. Proc. 20th Int. Conf. on Multimedia Modeling (MMM'14), Jan. 2014.

Two-layer stacking: indicative results

- Experimental setup
  - Indexing: for a concept, assess how well the top retrieved shots relate to it
  - Annotation: for a shot, assess how well the top retrieved concepts describe it
  - Dataset: TRECVID 2013 (internet videos; ~800 hours for training; ~200 hours for testing)
  - Input: model vectors for 346 concepts
  - Output: refined scores for 38 concepts
  - Evaluation: Mean Extended Inferred Average Precision (MXinfAP), Mean Precision at depth 3 (MP@3)

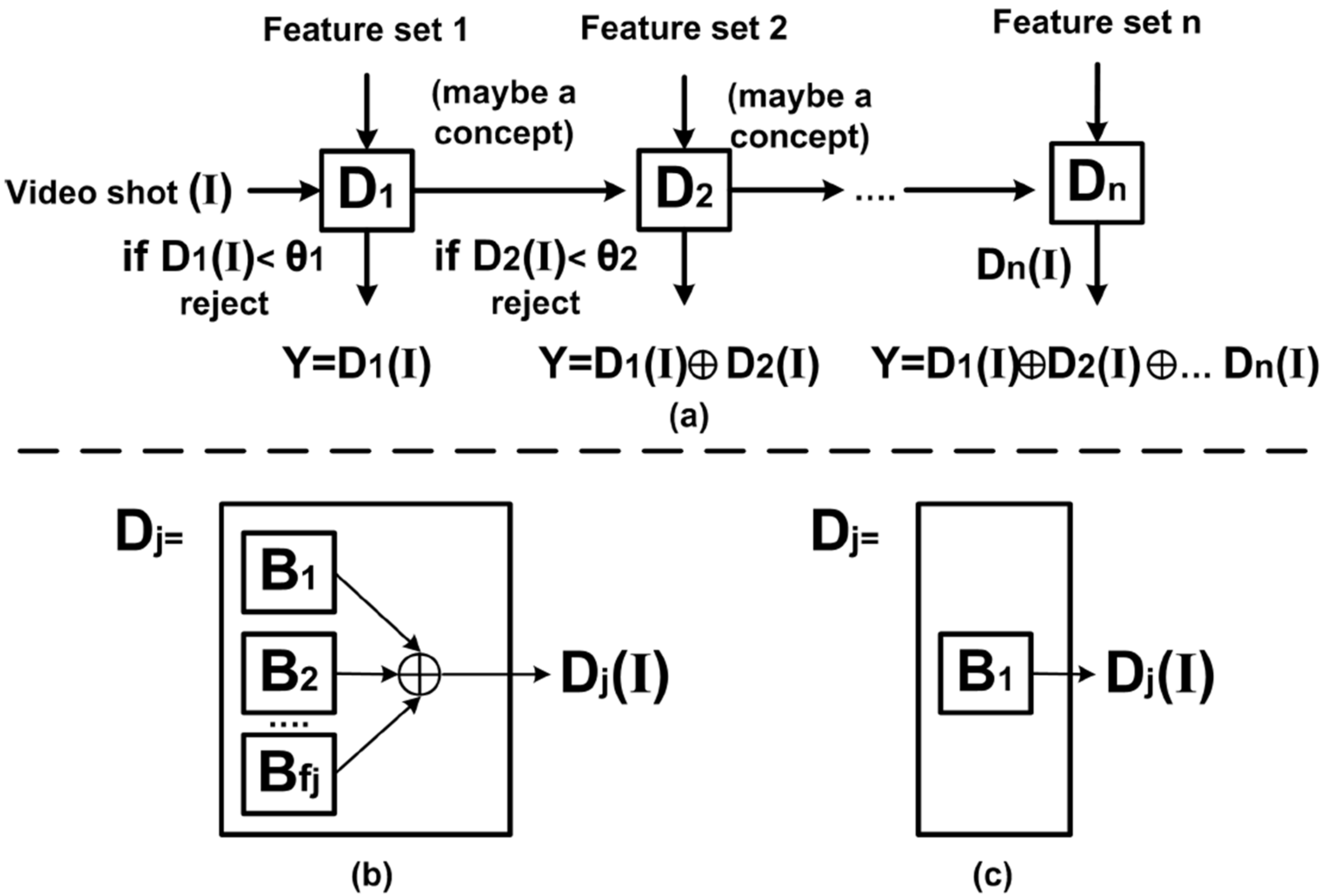
	MXinfAP (%) (indexing)		**MAP@3 (%) (annotation)**	
Method	2nd layer	1st and 2nd layer	2nd layer	1st and 2nd layer
Baseline	23.53	23.53	77.66	77.66
LP	24.2 (+2.8%)	24.72 (+5.1%)	80.98 (+4.3%)	79.1 (+1.9%)
ML-kNN	18.1 (-23.1%)	23.61 (+0.3%)	77.01 (-0.8%)	79.34 (+2.2%)

Indicative approach 3: efficiently combining DCNN-based and handcrafted features (for building independent concept detectors)

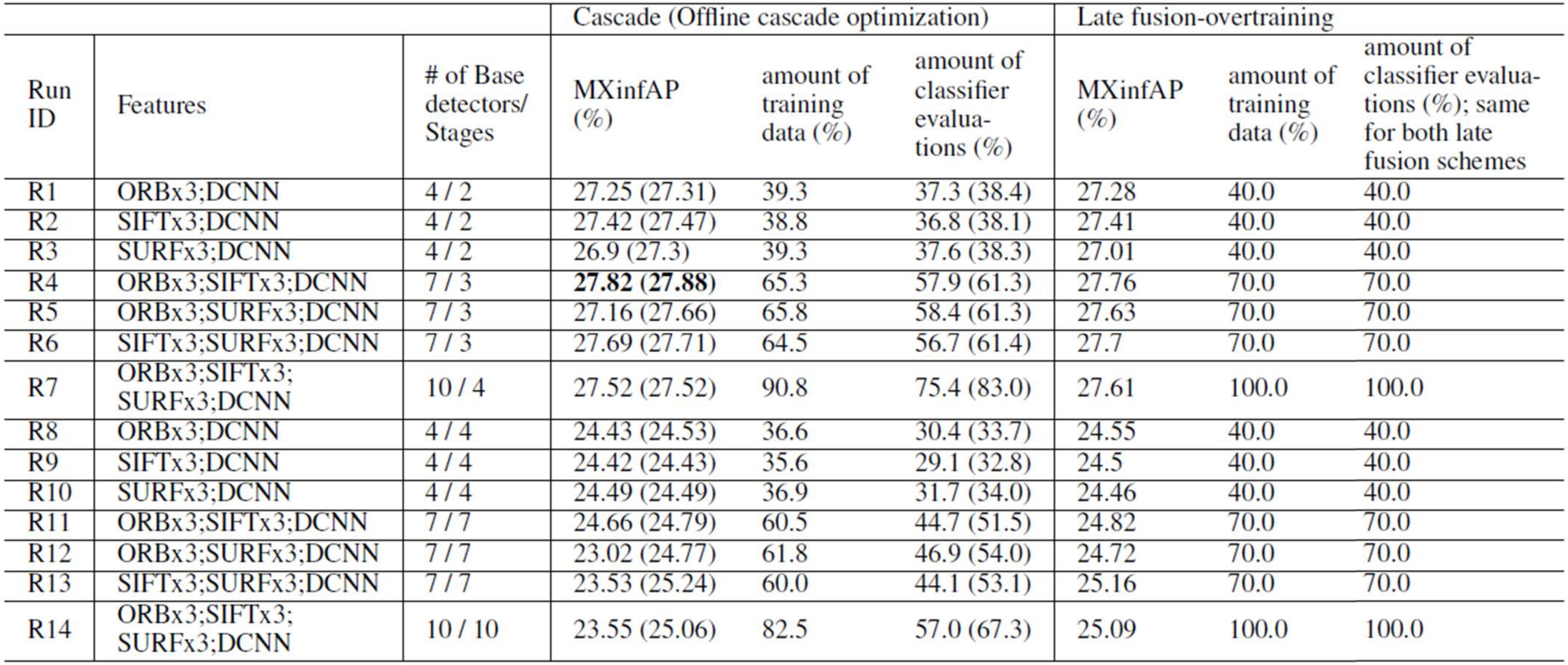
- Typical approach for combining multiple features: late fusion
  - Fusion of classifier scores, e.g. by computing their arithmetic, harmonic or geometric mean
  - Requires that all classifiers are evaluated for all media items
- Cascades of classifiers
  - Reject “easy” samples by evaluating just a subset of the available classifiers; still, evaluate and combine the scores of all classifiers for the potentially relevant samples [6]

[6] F. Markatopoulou, V. Mezaris, I. Patras, "Cascade of classifiers based on Binary, Non-binary and Deep Convolutional Network descriptors for video concept detection", Proc. IEEE Int. Conf. on Image Processing (ICIP 2015), Quebec City, Canada, Sept. 2015.

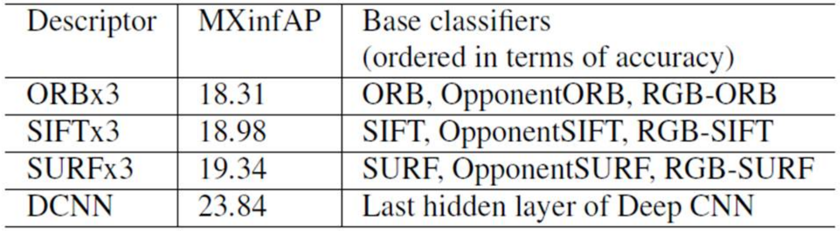
Cascade architecture



Cascade experiments and results



			Cascade (Offline cascade optimization)			Late fusion-overtraining		
Run ID	Features	# of Base detectors/ Stages	MXinfAP (%)	amount of training data (%)	amount of classifier evaluations (%)	MXinfAP (%)	amount of training data (%)	amount of classifier evaluations (%); same for both late fusion schemes
R1	ORBx3;DCNN	4 / 2	27.25 (27.31)	39.3	37.3 (38.4)	27.28	40.0	40.0
R2	SIFTx3;DCNN	4 / 2	27.42 (27.47)	38.8	36.8 (38.1)	27.41	40.0	40.0
R3	SURFx3;DCNN	4 / 2	26.9 (27.3)	39.3	37.6 (38.3)	27.01	40.0	40.0
R4	ORBx3;SIFTx3;DCNN	7 / 3	**27.82 (27.88)**	65.3	57.9 (61.3)	27.76	70.0	70.0
R5	ORBx3;SURFx3;DCNN	7 / 3	27.16 (27.66)	65.8	58.4 (61.3)	27.63	70.0	70.0
R6	SIFTx3;SURFx3;DCNN	7 / 3	27.69 (27.71)	64.5	56.7 (61.4)	27.7	70.0	70.0
R7	ORBx3;SIFTx3; SURFx3;DCNN	10 / 4	27.52 (27.52)	90.8	75.4 (83.0)	27.61	100.0	100.0
R8	ORBx3;DCNN	4 / 4	24.43 (24.53)	36.6	30.4 (33.7)	24.55	40.0	40.0
R9	SIFTx3;DCNN	4 / 4	24.42 (24.43)	35.6	29.1 (32.8)	24.5	40.0	40.0
R10	SURFx3;DCNN	4 / 4	24.49 (24.49)	36.9	31.7 (34.0)	24.46	40.0	40.0
R11	ORBx3;SIFTx3;DCNN	7 / 7	24.66 (24.79)	60.5	44.7 (51.5)	24.82	70.0	70.0
R12	ORBx3;SURFx3;DCNN	7 / 7	23.02 (24.77)	61.8	46.9 (54.0)	24.72	70.0	70.0
R13	SIFTx3;SURFx3;DCNN	7 / 7	23.53 (25.24)	60.0	44.1 (53.1)	25.16	70.0	70.0
R14	ORBx3;SIFTx3; SURFx3;DCNN	10 / 10	23.55 (25.06)	82.5	57.0 (67.3)	25.09	100.0	100.0



Descriptor	MXinfAP	Base classifiers (ordered in terms of accuracy)
ORBx3	18.31	ORB, OpponentORB, RGB-ORB
SIFTx3	18.98	SIFT, OpponentSIFT, RGB-SIFT
SURFx3	19.34	SURF, OpponentSURF, RGB-SURF
DCNN	23.84	Last hidden layer of Deep CNN

Concept-based image/video indexing conclusions

- Image/video concept detection has progressed a lot
- Results far from perfect; yet, already useful in a variety of applications (retrieval, further analysis of fragments, video hyperlinking,…)
- DCNN-based features are almost a must (but this doesn’t mean that others do not contribute to further improvement)
- Motion information can help (but, traditional motion descriptors are more computationally expensive than keyframes / tomographs)
- Linear SVMs very popular for building independent detectors (due to the size of the problem) – but other learning methods are applicable (e.g. 2-layer stacking; Discriminant Analysis [discussed in the sequel])
- Exploiting concept correlations can be important
- Understanding which features + learning method are most suitable for the problem at hand is very important; and, one does not need to stick to a single set of concept detection results!

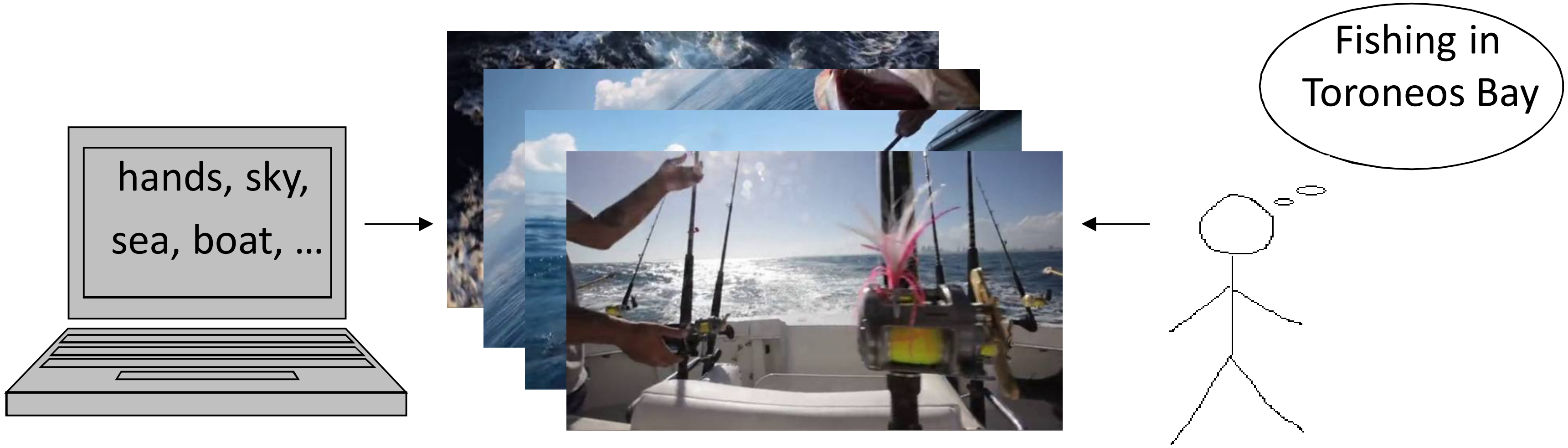
Demos: http://multimedia.iti.gr/mediamixer/demonstrator.html
http://multimedia2.iti.gr/onlinevideoanalysis/service/start.html

Concept-based image/video indexing: additional reading

- P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris, “Enhancing video concept detection with the use of tomographs”, Proc. IEEE Int. Conference. on Image Processing (ICIP 2013), Melbourne, Australia, 2013.
- F. Markatopoulou, V. Mezaris, I. Kompatsiaris, “A Comparative Study on the Use of Multi-Label Classification Techniques for Concept-Based Video Indexing and Annotation”, Proc. 20th Int. Conf. on Multimedia Modeling (MMM'14), Jan. 2014.
- C. Snoek, M. Worringm, “Concept-Based Video Retrieval”, in Foundations and Trends in Information Retrieval, vol. 2, no. 4, pp. 215–322, 2009.
- A. W. M. Smeulders, M.Worring, S. Santini, A.Gupta, and R. Jain, “Content- based image retrieval at the end of the early years”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1349–1380, 2000.
- G. Nasierding, A. Kouzani, “Empirical Study of Multi-label Classification Methods for Image Annotation and Retrieval”, Proc. 2010 Int. Conf. on Digital Image Computing: Techniques and Applications, China, pp. 617–622.
- G.J. Qi, X.S. Hua, Y. Rui, J. Tang, T. Mei, H. Zhang, “Correlative multi-label video annotation”, Proc. 15th Int. Conf. on Multimedia, pp. 17–26.
- M.-L. Zhang and Z.-H. Zhou, “ML-KNN: A lazy learning approach to multi-label learning”, Pattern Recognition, vol. 40, no. 7, 2007.
- A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid”, Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval (MIR '06), pp. 321-330, 2006.
- B. Safadi and G. Quenot, “Re-ranking by Local Re-Scoring for Video Indexing and Retrieval”, in C. Macdonald, I. Ounis, and I. Ruthven, editors, CIKM, pp. 2081-2084, 2011.
- J. C. Van Gemert, C. J. Veenman, A. Smeulders, J.-M. Geusebroek, “Visual word ambiguity”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 7, pp. 1271–1283, 2010.
- G. Csurka and F. Perronnin, “Fisher vectors: Beyond bag-of visual-words image representations”, Computer Vision, Imaging and Computer Graphics. Theory and Applications. Springer CCIS vol. 229, 2011.
- H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1704–1716, 2012.

Event-based video indexing

- Extending visual concept detection results with more elaborate annotations: event labels



Problem statement

- Objective:
  - Automatically detect high-level events in large video collections
  - Events are defined as “purposeful activities, involving people, acting on objects and interacting with each other to achieve some result”

“Changing a vehicle tire”

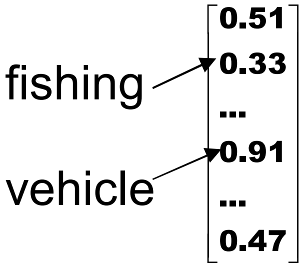
“Skiing”

“Working on a sewing project”

- Exploit an annotated video dataset
- Represent videos with suitable feature vectors {(**x** ,y) ϵ X x {-1,1}}
- Learn an appropriate event detector f: X → [0,1]

Related work

- Low-level feature-based approaches
  - Extract one or more low-level features (SIFT, MoSIFT, LFCC, ASR-based, etc.)
  - Combine features (late fusion, early fusion, etc.)
  - Motion visual features usually offer the most significant information
- Model vector-based approaches
  - Exploit a semantic model vector (i.e., automatic visual concept detection results) as a feature
  - Intuition behind this approach: high-level events can be better recognized by looking at their constituting semantic entities
  - This video representation can also be useful for explaining why a specific event was detected
  - DCNN-feature-based approaches: they are by definition model vector-based approaches, when the output of the last DCNN layer is used (but others can be used, too, similarly to the concept detection problem)
- Hybrid approaches: combination of low-level features and model vectors



An indicative approach (for start, the rather generic part)

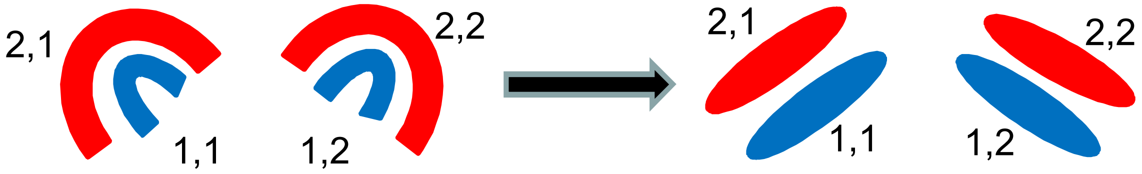
Hybrid approach

- Temporal sampling or video segmentation
  - keyframes at fixed time intervals (or perform shot/scene segmentation)
- Sample representation
  - Low-level visual feature extraction (dense trajectories; static SIFT, SURF)
    - VLAD encoding (separately for each kind of features)
  - Application of trained visual concept detectors (SVM- or DCNN-based)
    - Many of the concepts are seemingly irrelevant to the sought events
    - For DCNNs, responses of the last, or the last and second-last, layer(s) are used
- A concatenated feature vector is formed for each video keyframe
- Video representation: Averaging of the feature vectors for all keyframes of a video (or could work with e.g. a sequence of model vectors)
- Problem: high-dimensional video feature vectors (>100k elements), if all the above features are used

An indicative approach (the more specific part)

Event detection using Discriminant Analysis (DA)

- Use a DA algorithm to extract the most significant concept information for the detection of the target events [6, 7]
  - Mixture Subclass DA (MSDA): Linear or kernel-based method
  - Generalized subclass DA (GSDA): fast kernel-based method
- Kernel subclass methods
  - Identify a nonlinear feature transformation and a suitable subclass partition in order to map the data into a new space
  - In this space subclasses belonging to different classes are expected to be linearly separable

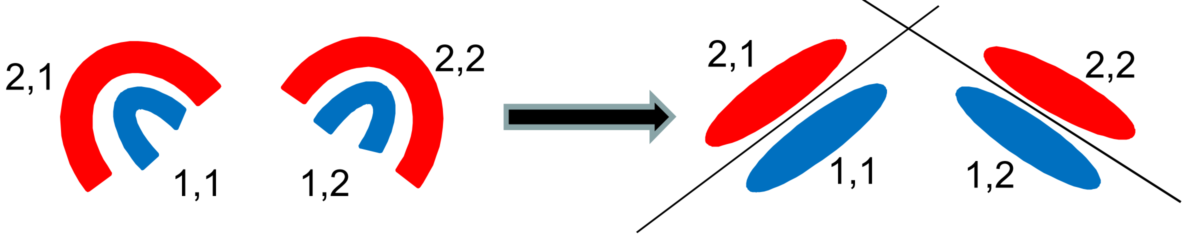


[6] N. Gkalelis, V. Mezaris, I. Kompatsiaris, T. Stathaki, "Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations", IEEE Trans. on Neural Networks and Learning Systems, vol. 24, no. 1, pp. 8-21, Jan. 2013.

[7] N. Gkalelis, V. Mezaris, "Video event detection using generalized subclass discriminant analysis and linear support vector machines", Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR), Glasgow, UK, April 2014.

An indicative approach (the more specific part)

- Very efficient GPU implementations of GSDA extensions (AGSDA) exist [8]
- Following DA, Linear SVM classifiers are used for learning each event [9]
- Advantages
  - The effect of noise or irrelevant features to the classification problem is minimized → classification performance may be improved
  - Classifier is applied in the lower dimensional subspace → classifier training and testing times are reduced
  - The application of kernel subclass DA yields linear separable classes (or subclasses) → linear classifiers may be used (better generalization, faster)

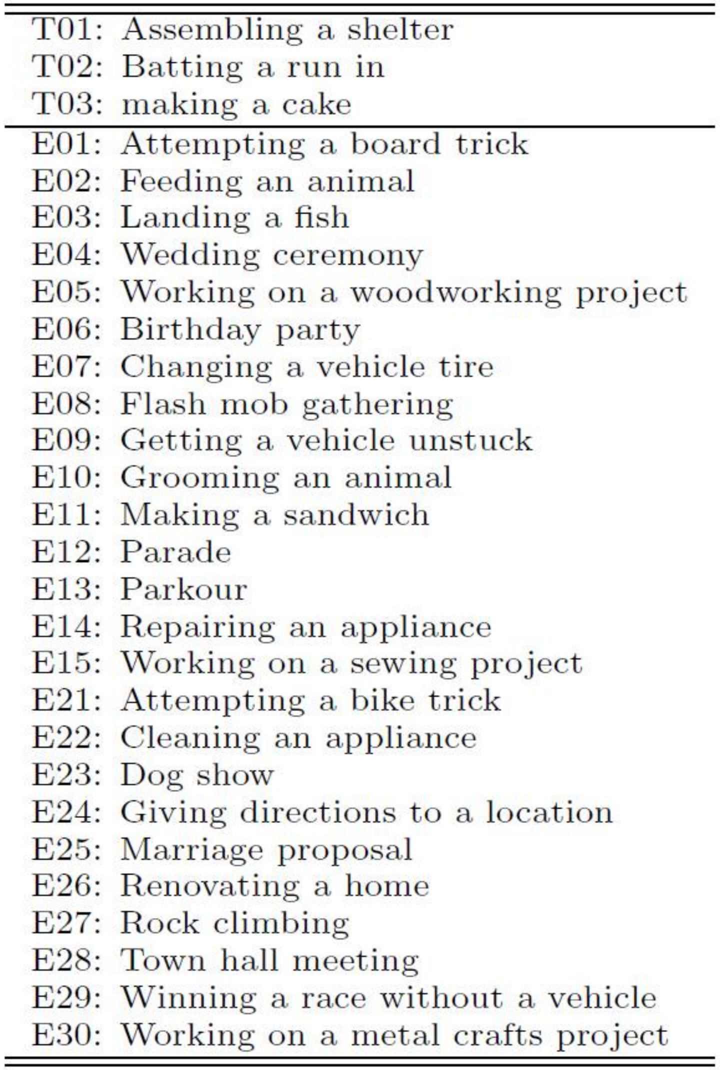


[8] N. S. Arestis-Chartampilas, N. Gkalelis, V. Mezaris, "GPU accelerated generalised subclass discriminant analysis for event and concept detection in video", Proc. ACM Multimedia 2015, Brisbane, Australia, October 2015.

[9] N. Gkalelis, V. Mezaris, "Video event detection using generalized subclass discriminant analysis and linear support vector machines", Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR), Glasgow, UK, April 2014.

1st experimental setup and results

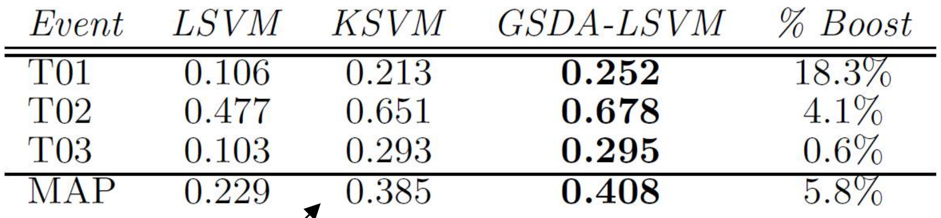
- TRECVID 2010
  - 3 target events
  - Development: 1745 videos
  - Evaluation: 1742 videos
- TRECVID 2012
  - A publicly available annotated subset is used [8]
  - 25 target events
  - Development: 8840 videos
  - Evaluation: 4434 videos
- Simplified, low-dimensional model vector representation (346 SVM-based concept detectors)



T01:	Assembling a shelter
T02:	Batting a run in
T03:	making a cake
E01:	Attempting a board trick
E02:	Feeding an animal
E03:	Landing a fish
E04:	Wedding ceremony
E05:	Working on a woodworking project
E06:	Birthday party
E07:	Changing a vehicle tire
E08:	Flash mob gathering
E09:	Getting a vehicle unstuck
E10:	Grooming an animal
E11:	Making a sandwich
E12:	Parade
E13:	Parkour
E14:	Repairing an appliance
E15:	Working on a sewing project
E21:	Attempting a bike trick
E22:	Cleaning an appliance
E23:	Dog show
E24:	Giving directions to a location
E25:	Marriage proposal
E26:	Renovating a home
E27:	Rock climbing
E28:	Town hall meeting
E29:	Winning a race without a vehicle
E30:	Working on a metal crafts project

[8] A. Habibian, K.E.A. van de Sande, C.G.M. Snoek , "Recommendations for video event recognition using concept vocabularies", Proc. ICMR 2013, pp. 89-96.

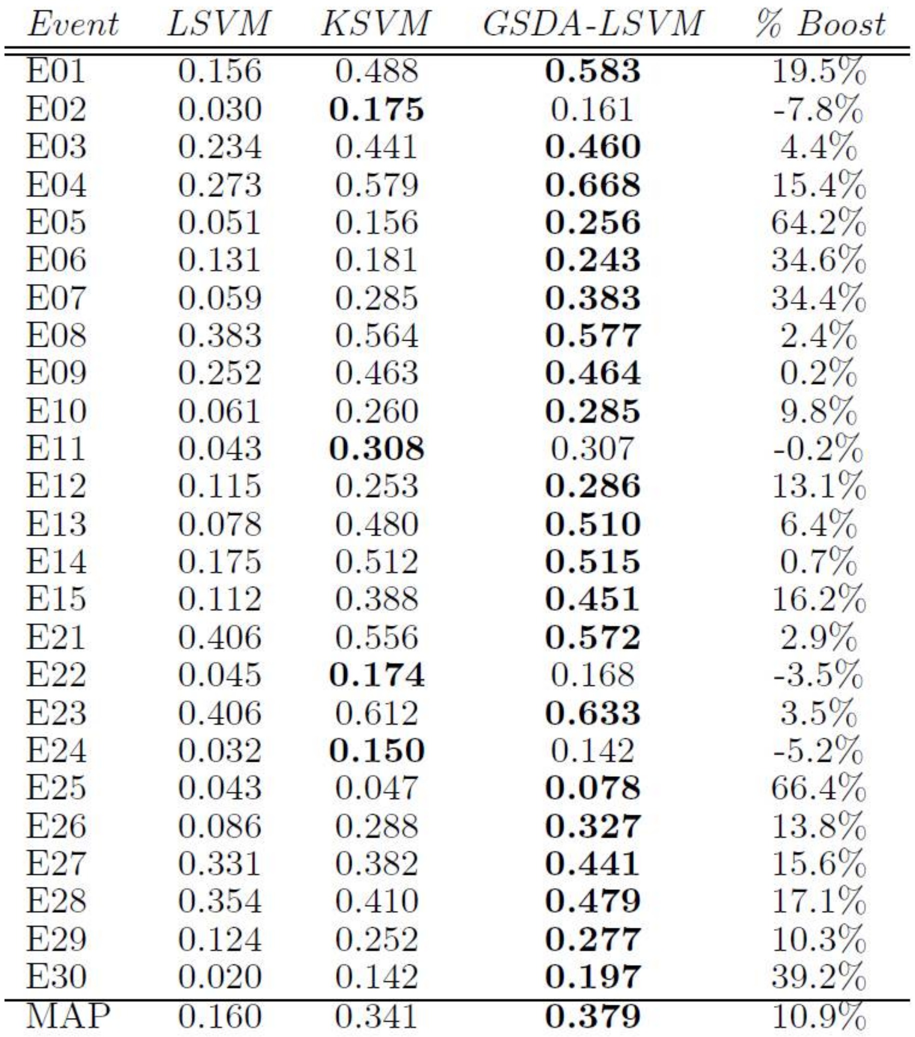
1st experimental setup and results



Event	*LSVM*	*KSVM*	*GSDA-LSVM*	*% Boost*
T01	0.106	0.213	**0.252**	18.3%
T02	0.477	0.651	**0.678**	4.1%
T03	0.103	0.293	**0.295**	0.6%
MAP	0.229	0.385	**0.408**	5.8%

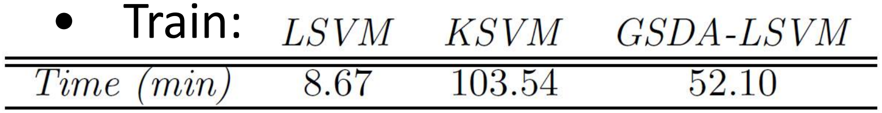
TRECVID 2010

TRECVID 2012



Event	*LSVM*	*KSVM*	*GSDA-LSVM*	*% Boost*
E01	0.156	0.488	**0.583**	19.5%
E02	0.030	**0.175**	0.161	-7.8%
E03	0.234	0.441	**0.460**	4.4%
E04	0.273	0.579	**0.668**	15.4%
E05	0.051	0.156	**0.256**	64.2%
E06	0.131	0.181	**0.243**	34.6%
E07	0.059	0.285	**0.383**	34.4%
E08	0.383	0.564	**0.577**	2.4%
E09	0.252	0.463	**0.464**	0.2%
E10	0.061	0.260	**0.285**	9.8%
E11	0.043	**0.308**	0.307	-0.2%
E12	0.115	0.253	**0.286**	13.1%
E13	0.078	0.480	**0.510**	6.4%
E14	0.175	0.512	**0.515**	0.7%
E15	0.112	0.388	**0.451**	16.2%
E21	0.406	0.556	**0.572**	2.9%
E22	0.045	**0.174**	0.168	-3.5%
E23	0.406	0.612	**0.633**	3.5%
E24	0.032	**0.150**	0.142	-5.2%
E25	0.043	0.047	**0.078**	66.4%
E26	0.086	0.288	**0.327**	13.8%
E27	0.331	0.382	**0.441**	15.6%
E28	0.354	0.410	**0.479**	17.1%
E29	0.124	0.252	**0.277**	10.3%
E30	0.020	0.142	**0.197**	39.2%
MAP	0.160	0.341	**0.379**	10.9%

- In learning one MED 2012 event, GSDA was ~25% faster than traditional Kernel SDA
- Train:



	LSVM	*KSVM*	*GSDA-LSVM*
Time (min)	8.67	103.54	52.10

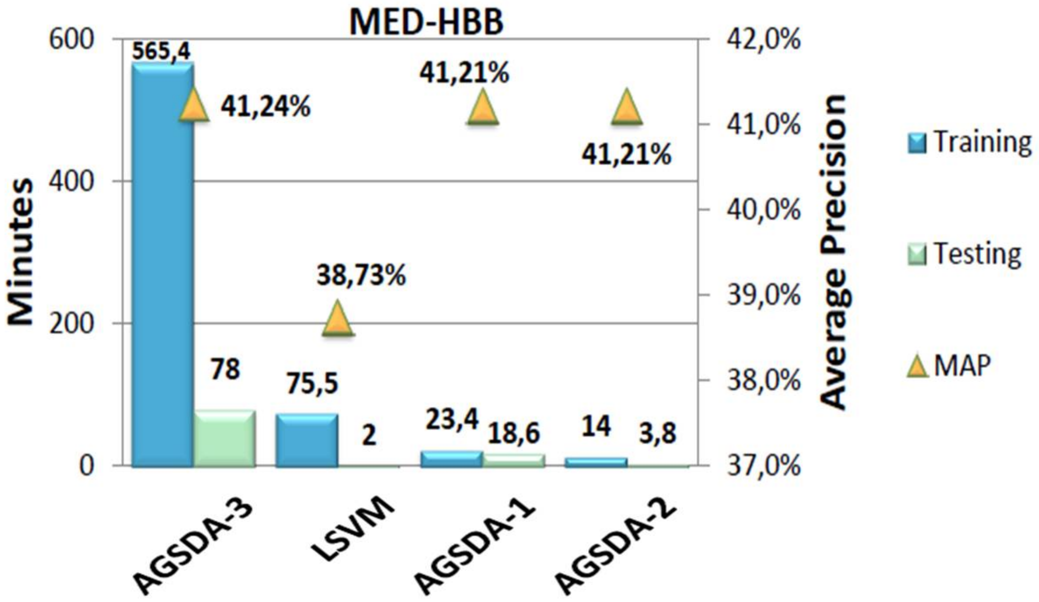
- Testing times: similar for GSDA-LSVM and KSVM
- We have developed a GSDA-LSVM extension 60x faster than LSVM in training, for large-scale problems*

*per CV cycle

2nd experimental setup and results

- Dataset: TRECVID 2012 (same as in previous experiment)
  - 25 target events; 8840 development videos; 4434 evaluation videos
- Improved dense trajectories are used to represent each video with a feature vector of size 101376 (instead of 346-element model vectors).
- We compare four approaches:
  a) AGSDA-1: GPU-AGSDA with LSVM (using LibSVM), on GTX 650 (low-end GPU)
  b) AGSDA-2: GPU-AGSDA with LSVM (using LibSVM), on Tesla K40 (high-end GPU)
  c) AGSDA-3: the Matlab version of AGSDA with LSVM (using LibSVM).
  d) LSVM: C++ liblinear implementation (state-of-the-art in training time, up to 100 times faster than LibSVM) of LSVM operating directly in the input feature space.
- All experiments were performed on a PC (Intel i7 3770K @3.5GHz processor, 32GB RAM, Win7 x64 OS)

2nd experimental setup and results

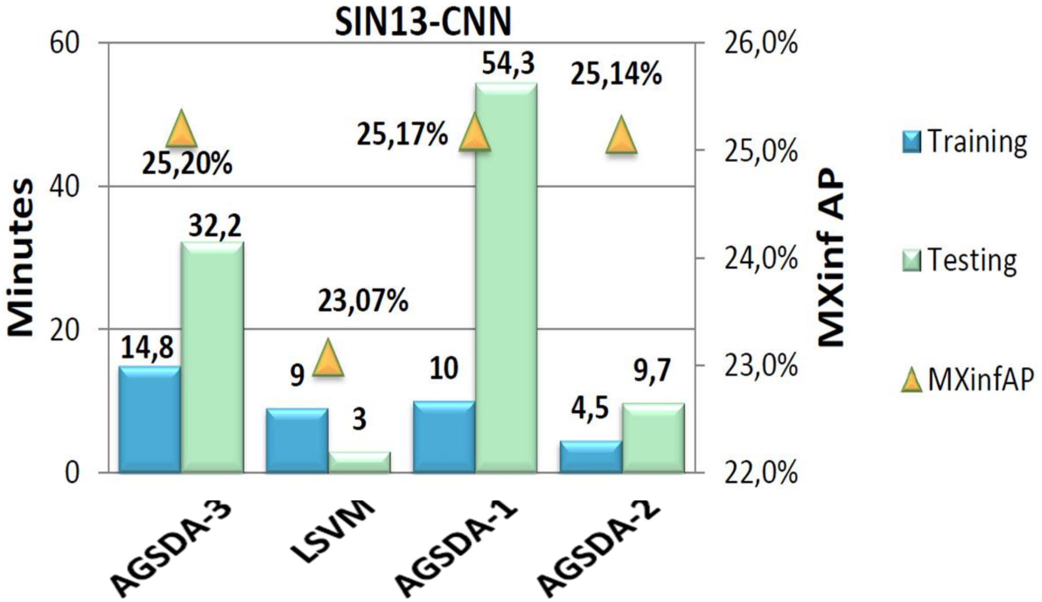


- Detection performance increase over LSVM by 2.5 percentage points, requiring only 1/3rd (AGSDA-1) down to 1/5th (AGSDA-2) of the training time.
- Significantly less memory requirements. AGSDA-3 utilized the HDD as swap memory, while AGSDA-1 & -2 required 1/8th of the RAM

3rd experimental setup and results (going back to concept detection)

- Dataset: TRECVID SIN 2013
  - 38 target concepts; 546529 development video shots; 112677 evaluation shots
  - Up to 36000 shots per concept used as training data
- DCNN–based features are employed for video representation
  - 16-layer ConvNet network, created using the 2009 ImageNet dataset
  - The last layer is utilized as output, providing a 1000-dimensional model vector for each keyframe
  - One keyframe per shot is used
- Same comparisons as in previous experimental setup

3rd experimental setup and results



- Detection performance increase over LSVM by 2.2 percentage points. The training time required was equivalent (AGSDA-1), or even half (AGSDA-2).
- AGSDA-1 achieved a training time equal to 2/3rds of the Matlab (AGSDA-3) approach

Event detection conclusions

- Importance of low level features: visual motion features are the most important followed by visual static features; for some events, audio features provide complementary information
- Model vectors / DCNN-based features are also very effective
- Combining multiple features (low-level and higher-level) is beneficial, similar to the concept detection problem
- The machine learning methods that one uses can make a difference in both detection speed and accuracy
- Learning an event detector from a few or even zero positive video examples is a difficult and interesting problem

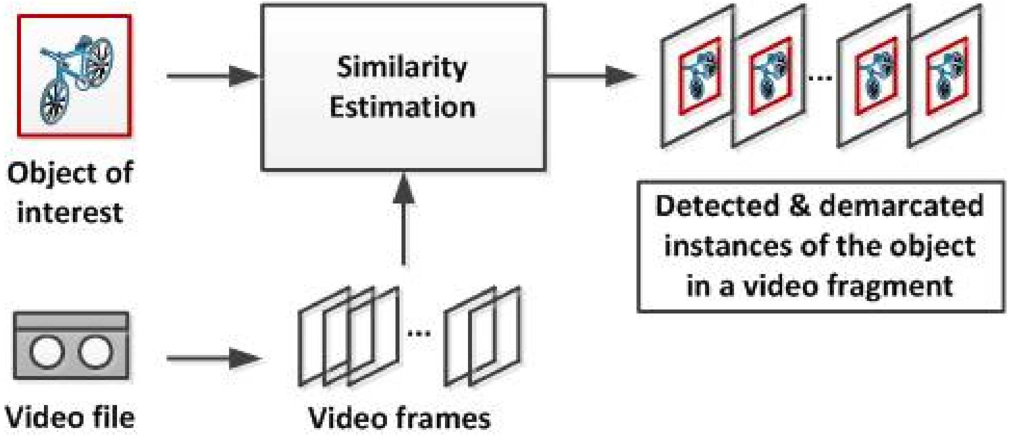
Event detection: additional reading

- P. Over, J. Fiscus, G. Sanders et. al., "TRECVID 2014 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics ", Proc. TRECVID 2014 Workshop, Nov. 2014, Orlando, FL, USA.
- N. Gkalelis, V. Mezaris, I. Kompatsiaris, T. Stathaki, "Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations", IEEE Transactions on Neural Networks and Learning Systems, vol. 24, no. 1, pp. 8-21, January 2013.
- N. Gkalelis, V. Mezaris, "Video event detection using generalized subclass discriminant analysis and linear support vector machines", Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR), Glasgow, UK, April 2014.
- M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, “Semantic model vectors for complex video event recognition,” IEEE Trans. Multimedia, vol. 14, no. 1, pp. 88–101, Feb. 2012.
- N. Gkalelis, V. Mezaris, M. Dimopoulos, I. Kompatsiaris, T. Stathaki, "Video event detection using a subclass recoding error-correcting output codes framework", Proc. IEEE Int. Conf. on Multimedia and Expo (ICME 2013), San Jose, CA, USA, July 2013.
- Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, “High-level event recognition in unconstrained videos,” Int. J. Multimed. Info. Retr., Nov. 2013.
- A. Habibian, K. van de Sande, C.G.M. Snoek, "Recommendations for Video Event Recognition Using Concept Vocabularies", Proc. ACM Int. Conf. on Multimedia Retrieval, Dallas, Texas, USA, April 2013.
- Z. Ma, Y. Yang, Z. Xu, N. Sebe, A. Hauptmann, "We Are Not Equally Negative: Fine-grained Labeling for Multimedia Event Detection", Proc. ACM Multimedia 2013 (MM’13), Barcelona, Spain, October 2013.
- C. Tzelepis, N. Gkalelis, V. Mezaris, I. Kompatsiaris, "Improving event detection using related videos and Relevance Degree Support Vector Machines", Proc. ACM MM 2013 (MM’13), Barcelona, Spain, Oct.2013.
- S. Arestis-Chartampilas, N. Gkalelis, V. Mezaris, "GPU accelerated generalised subclass discriminant analysis for event and concept detection in video", Proc. ACM MM 2015, Brisbane, Australia, Oct. 2015.

Other relevant analysis technologies

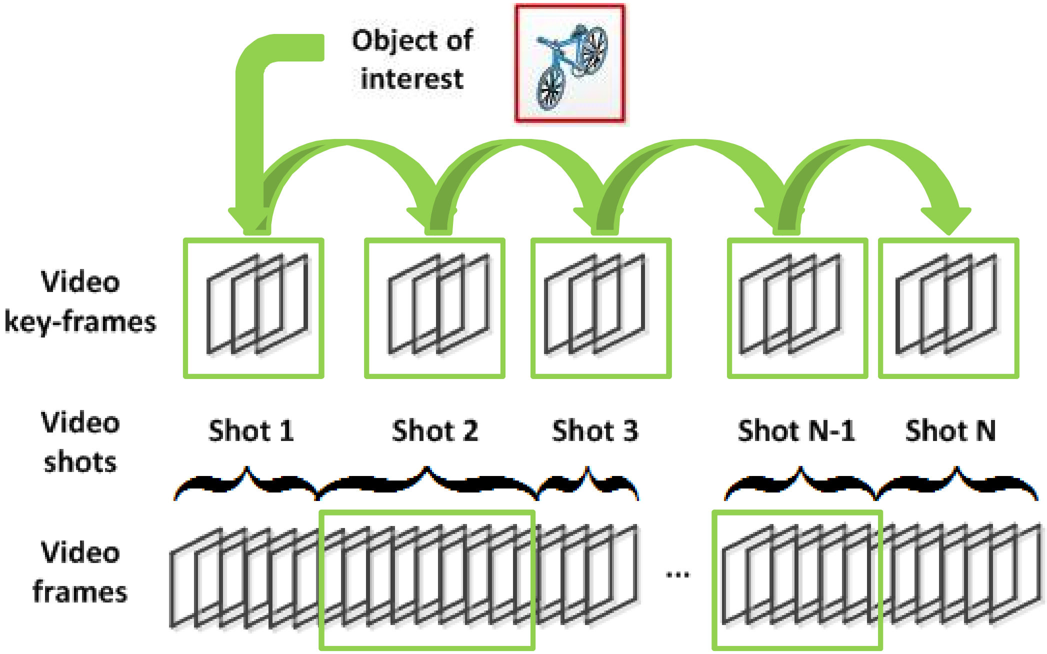
- Object re-detection: a particular case of image matching
- Main goal: find instances of a specific object within a single video or a collection of videos
  - Input: object of interest + video files
  - Processing: similarity estimation by means of image matching (using local features)
  - Output: detected instances of the object of interest

Semi-automatic process!


Object of interest
Similarity Estimation
Detected & demarcated instances of the object in a video fragment
Video file
Video frames

Other relevant analysis technologies

- An indicative object re-detection approach: the sequential processing of video frames is replaced by a structure-based one, using the results of shot segmentation [9]



Example

Check shot 1

Check shot 2

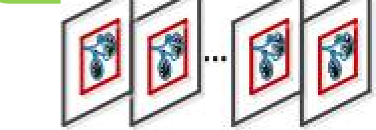
No detection!

Detection!

Check the frames of this shot

Move to the next shot

Detect and highlight the object of interest



Demo video: http://www.youtube.com/watch?v=0IeVkXRTYu8

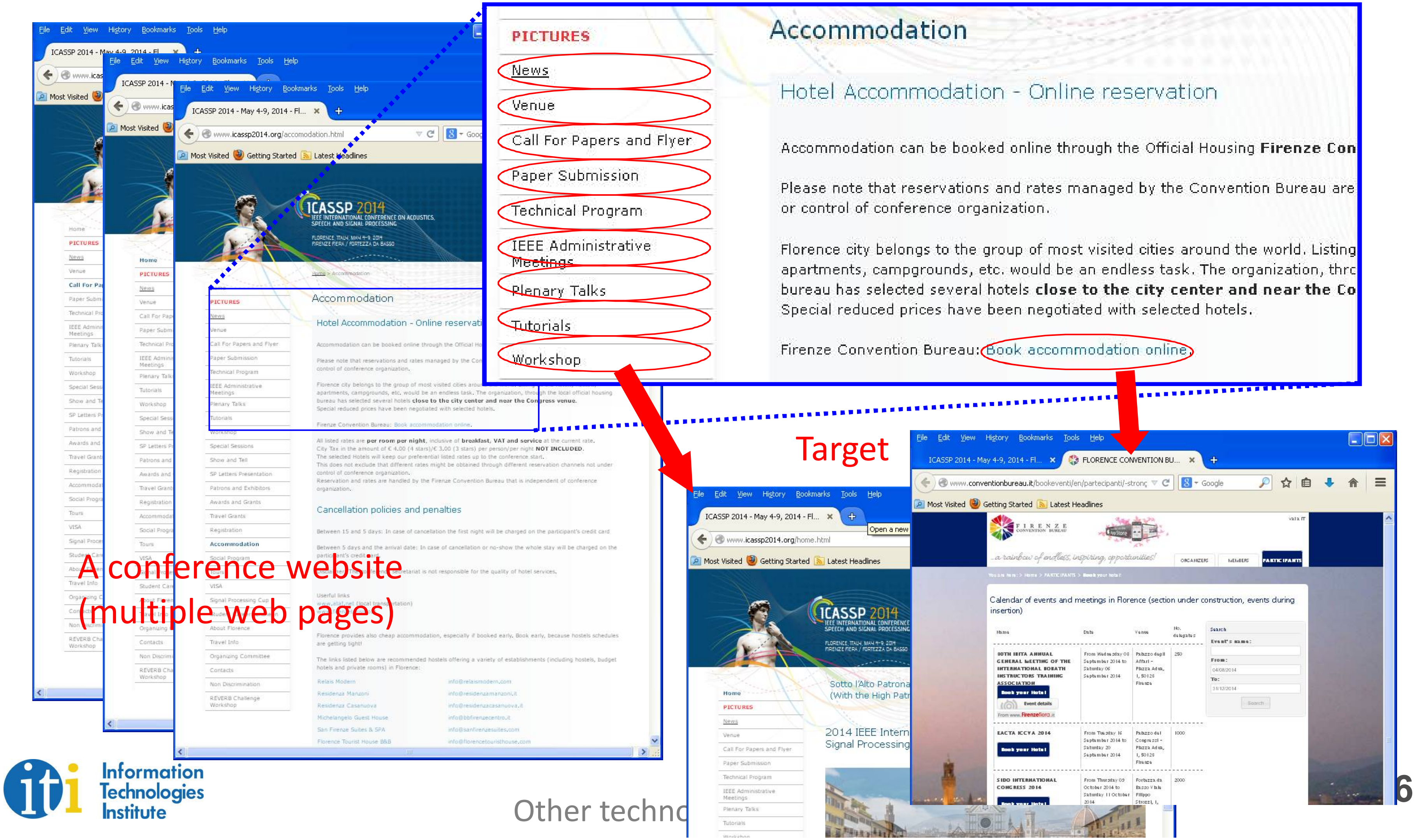
[9] L. Apostolidis, V. Mezaris, I. Kompatsiaris, "Fast object re-detection and localization in video for spatio-temporal fragment creation", Proc. MMIX'13 at IEEE ICME 2013, San Jose, CA, USA, July 2013.

Other relevant analysis technologies

- Near-duplicate frame / video segment detection
- Rationale: often the repetition of the same frames indicates related content
  - E.g. in News, video coverage of an evolving story will often repeat the same iconic footage of the event over a period of days
  - Can help us find pieces of related video content (as long as it's not e.g. just a frame showing the anchorperson…)
- Typical detection process
  - Local feature extraction (e.g. SIFT), encoding (e.g. VLAD), indexing (e.g. KD-trees)
  - Given a query frame, find the N closest matches to it
  - Then, perform geometric verification to decide if they are indeed near-duplicates or not

A relevant application: video hyperlinking

- The text hyperlinking analogy



Origin of hyperlinks

PICTURES

News

Venue

Call For Papers and Flyer

Paper Submission

Technical Program

IEEE Administrative Meetings

Plenary Talks

Tutorials

Workshop

Accommodation

Hotel Accommodation - Online reservation

Accommodation can be booked online through the Official Housing **Firenze Con**

Please note that reservations and rates managed by the Convention Bureau are or control of conference organization.

Florence city belongs to the group of most visited cities around the world. Listing apartments, campgrounds, etc. would be an endless task. The organization, thro bureau has selected several hotels **close to the city center and near the Co** Special reduced prices have been negotiated with selected hotels.

Firenze Convention Bureau: Book accommodation online

Target

ICASSP 2014 - May 4-9, 2014 - Fl... | FLORENCE CONVENTION BU...

www.conventionbureau.it/bookeventi/en/partecipanti/-stronc

FIRENZE CONVENTION BUREAU

...a rainbow of endless, inspiring, opportunities!

ORGANIZERS | MEMBERS | PARTICIPANTS

You are here: > Home > PARTICIPANTS > Book your hotel!

Calendar of events and meetings in Florence (section under construction, events during insertion)

Name	Date	Venue	No. delegates
30TH IBITA ANNUAL GENERAL MEETING OF THE INTERNATIONAL BOBATH INSTRUCTORS TRAINING ASSOCIATION Book your Hotel Event details From www.Firenzefiera.it	From Wednesday 03 September 2014 to Saturday 06 September 2014	Palazzo degli Affari - Piazza Adua, 1, 50123 Firenze	250
EACTA ICCVA 2014 Book your Hotel	From Tuesday 16 September 2014 to Saturday 20 September 2014	Palazzo dei Congressi - Piazza Adua, 1, 50123 Firenze	1000
SIDO INTERNATIONAL CONGRESS 2014 Book your Hotel	From Thursday 09 October 2014 to Saturday 11 October 2014	Fortezza da Basso Viale Filippo Strozzi, 1,	2000

Search

Event's name:

From: 04/08/2014

To: 31/12/2014

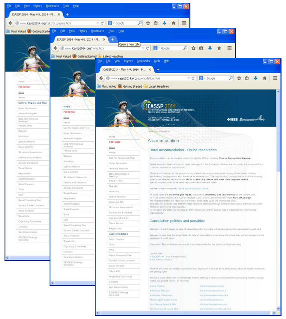
Search

A conference website (multiple web pages)

A relevant application: video hyperlinking

- The text hyperlinking analogy
  - Web sites accommodating hyperlinks, and the origins and targets of these hyperlinks represent information at different levels of granularity

Web site: a collection of interlinked web pages, typically by the same creator(s), altogether serving the same purpose



An entire video

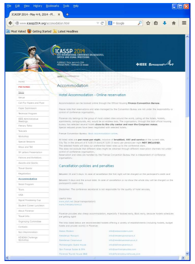
*Practical restriction: how do you select a 2D image that appears on your screen for e.g. 1/25 sec?

Origin of a hyperlink: a very elementary text (a single word or a short phrase); an iconic representation of a topic

Book accommodation online.

A very elementary part of the video (a 2D image? a shot?)*

Target of a hyperlink: a focused, yet explanatory and understandable even on its own piece of information



A meaningful story-telling part of the video (scene / chapter / topic)

A relevant application: video hyperlinking

- To realize video hyperlinking, given a set of videos, we need to:
  - Break down each video to fragments that can serve as the origin and the target of hyperlinks →
    - Video temporal segmentation to shots
    - Video temporal segmentation to scenes
    - Object re-detection
  - Identify meaningful pairs of fragments (i.e., identify two fragments as the origin and the target of the same link) →
    - By annotating them
      - Visual concept detection
      - Event detection
    - By finding visually similar fragments
      - Object re-detection
      - Near-duplicate frame/fragment detection

Concluding remarks

- We discussed different classes of techniques for video fragmentation and annotation, to support video organization and retrieval, but others also exist, e.g.
  - Face detection, tracking, clustering, recognition
  - ASR, OCR, and textual transcript analysis
- In some cases the automatic analysis results remain far from perfect (manual) annotations; yet, these results may still be useful in retrieval and other applications (e.g. video hyperlinking)

Thank you!

More information (including links to various demos):
http://www.iti.gr/~bmezaris
bmezaris@iti.gr