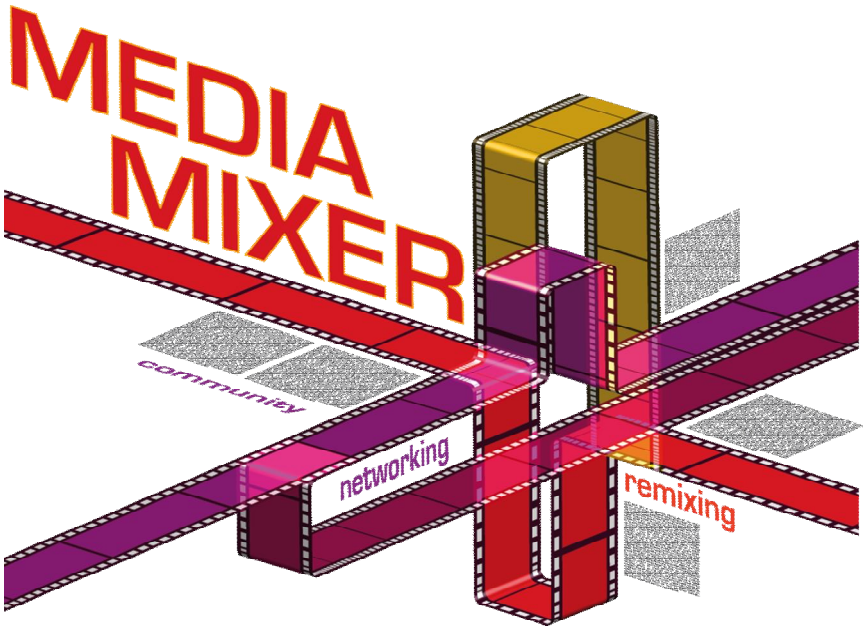
Fragmenting Media Assets: media analysis for fragment creation and description



Vasileios Mezaris, CERTH-ITI

November 2013

Information Technologies Institute
Centre for Research and Technology Hellas

Overview

- Introduction – overall motivation
- Technologies for media fragment creation and annotation
  - Video temporal segmentation to shots
  - Video temporal segmentation to scenes
  - Visual concept detection
  - Event detection
  - Object re-detection
- Demos

For each presented technology, we go through:

- More precise problem statement
- Brief overview of the literature
- A closer look at a MediaMixer-promoted approach
- Indicative experiments and results
- Conclusions
- Additional reading (references)

Introduction - motivation

- We have: media items
- We want to: enable fine-grained access to the media
- Thus, we need to:
  - Break down each media item to meaningful fragments
  - Annotate each fragment to make it searchable
- We could do this either
  - Manually: + accuracy, - speed, cost, effort → only feasible for low-volume, high-value media items
  - Automatically: - accuracy, + speed, cost, effort → the only option for handling high volume of content
- MediaMixer promotes technologies for the automatic fragmentation and annotation of video content
  - Temporal fragment creation: shot and scene detection
  - Class-level annotation: visual concepts, events
  - Instance-level annotation: object re-detection
  - Keep in mind: annotation is also useful for refining fragmentation

Fragment creation: shots

- What is a shot: a sequence of consecutive frames taken without interruption by a single camera
- Shot segmentation
  - temporal decomposition of videos by detecting the boundaries or the changes between the shots
  - foundation of most high-level video analysis approaches, such as video semantic analysis and fine-grained classification, indexing and retrieval.
- Shot change is manifested by a shift in visual content
  - Two basic types of transition

Fragment creation: shots

- What is a shot: a sequence of consecutive frames taken without interruption by a single camera
- Shot segmentation
  - temporal decomposition of videos by detecting the boundaries or the changes between the shots
  - foundation of most high-level video analysis approaches, such as video semantic analysis and fine-grained classification, indexing and retrieval.
- Shot change is manifested by a shift in visual content
  - Two basic types of transition **ABRUPT**

Fragment creation: shots

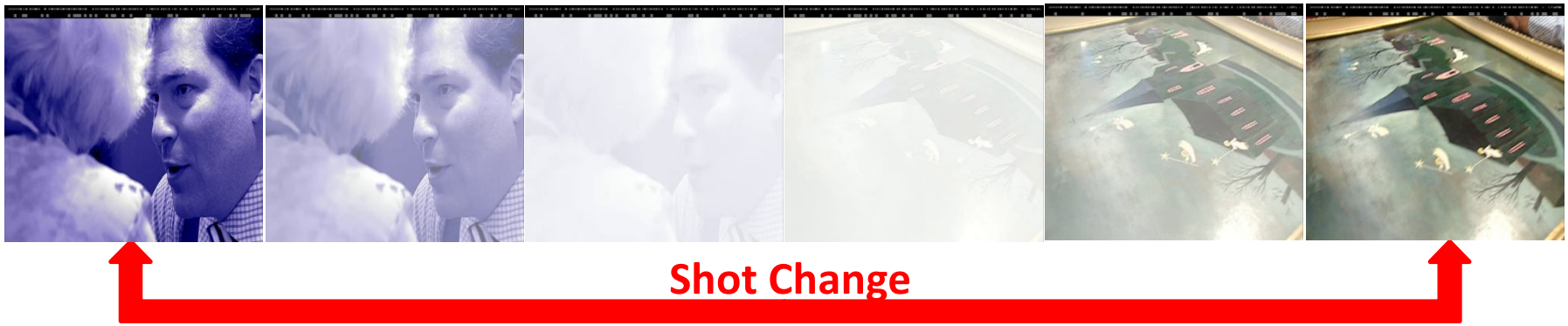
- What is a shot: a sequence of consecutive frames taken without interruption by a single camera
- Shot segmentation
  - temporal decomposition of videos by detecting the boundaries or the changes between the shots
  - foundation of most high-level video analysis approaches, such as video semantic analysis and fine-grained classification, indexing and retrieval.
- Shot change is manifested by a shift in visual content
  - Two basic types of transition **GRADUAL** (**dissolve**, wipe, fade in / fade out,…)

Fragment creation: shots

- What is a shot: a sequence of consecutive frames taken without interruption by a single camera
- Shot segmentation
  - temporal decomposition of videos by detecting the boundaries or the changes between the shots
  - foundation of most high-level video analysis approaches, such as video semantic analysis and fine-grained classification, indexing and retrieval.
- Shot change is manifested by a shift in visual content
  - Two basic types of transition **GRADUAL** (dissolve, **wipe**, fade in / fade out,…)

Fragment creation: shots

- What is a shot: a sequence of consecutive frames taken without interruption by a single camera
- Shot segmentation
  - temporal decomposition of videos by detecting the boundaries or the changes between the shots
  - foundation of most high-level video analysis approaches, such as video semantic analysis and fine-grained classification, indexing and retrieval.
- Shot change is manifested by a shift in visual content
  - Two basic types of transition **GRADUAL** (dissolve, wipe, **fade in / fade out**,…)



Challenges

- Challenge: avoid being mislead by
  - Illumination changes (e.g. due to camera flash-lights)
  - Fast camera movement
  - Rapid local (visual object) motion
  - …

Challenges

- Challenge: avoid being mislead by
  - **Illumination changes (e.g. due to camera flash-lights)**
  - Fast camera movement
  - Rapid local (visual object) motion
  - …

Example of camera flashlights

False alarm of abrupt shot change

Challenges

- Challenge: avoid being mislead by
  - Illumination changes (e.g. due to camera flash-lights)
  - **Fast camera movement**
  - Rapid local (visual object) motion
  - …

Example of fast camera movement

False alarm of gradual shot change

Challenges

- Challenge: avoid being mislead by
  - Illumination changes (e.g. due to camera flash-lights)
  - Fast camera movement
  - **Rapid local (visual object) motion**
  - …

Example of rapid local motion

False alarm of gradual shot change

Related work

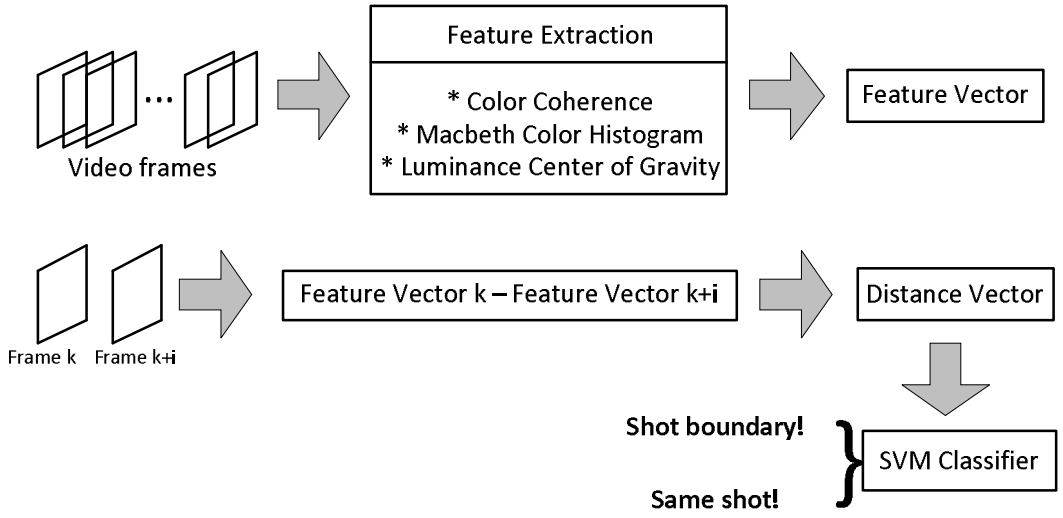
- Can generally be organized according to
  - Data to work with: uncompressed vs. compressed video
  - Features to use (also depends on the data)
  - Threshold-based vs. learning-based methods
- Compressed video methods
  - Reduce computational complexity by avoiding decoding, exploiting encoder results
    - 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 comparison
  - Global visual feature comparison (e.g. color histogram, color coherence) 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 → Machine learning (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

A MediaMixer-promoted approach

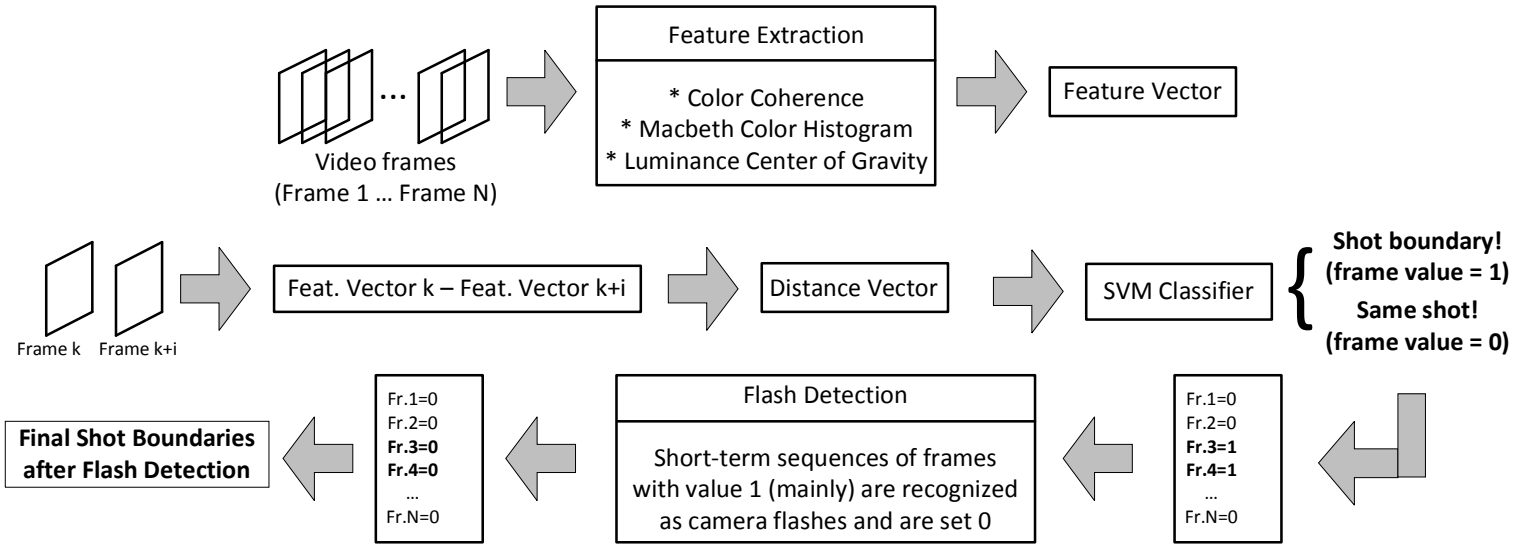
- Based on the uncompressed-domain approach introduced in [8]
- Detects both abrupt and gradual shot transitions, based on:
  - extracting visual features, (color coherence, Macbeth color histogram and luminance center of gravity) and forming a feature vector per frame
  - computing the distances between vectors of neighboring frames, composing distance vectors
  - evaluating distance vectors using one or more SVM classifiers



A MediaMixer-promoted approach

Further extension:

- Lightweight flash detection algorithm
- Changes within a short-term sequence of frames are recognized as camera flashlights



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
  - 18 camera flashlights
- System specifications
  - Intel Core i7 processor at 3.4GHz
  - 8GB RAM memory

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)
- Flash detectors performance
  - Precision: 100%
  - Recall: 78%
  - F-Score: 0.876
- Time performance
  - Runs in 1,25x real time (i.e. the video's actual duration)

Experimental Results	
Precision	**85.7 %**
Recall	**91.2 %**
F-Score	**88.4 %**

Online demo available at: http://www.youtube.com/watch?v=0IeVkXRTYu8

Shot detection conclusions

- Overall accuracy of shot detection methods is high (>90%), sufficient for any application
- Detection of gradual transitions & handling of intense motion still a bit more challenging
- Real-time or near-real-time processing is feasible (but faster processing may be needed in some applications)

Shot detection: additional reading

- E. Tsamoura, V. Mezaris, and I. Kompatsiaris, “Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework,”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,” 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,” in 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.
- S.-C. Pei and Y.-Z. Chou, “Effective wipe detection in mpeg compressed video using macro block type information,” Transactions on Multimedia, vol. 4, no. 3, pp. 309–319, Sept. 2002.
- 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,” Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing, 2006. ICASSP 2006, vol. 2, pp. II–II.
- A. Amiri and M. Fathy, “Video shot boundary detection using qr-decomposition and gaussian transition detection,” EURASIP Journal of Advanced Signal Processing 2009.

Fragment creation: 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 basic story-telling units
  - important prerequisite for summarization, indexing, video browsing,…
- Scene change is not manifested by just a change in visual content

Fragment creation: scenes

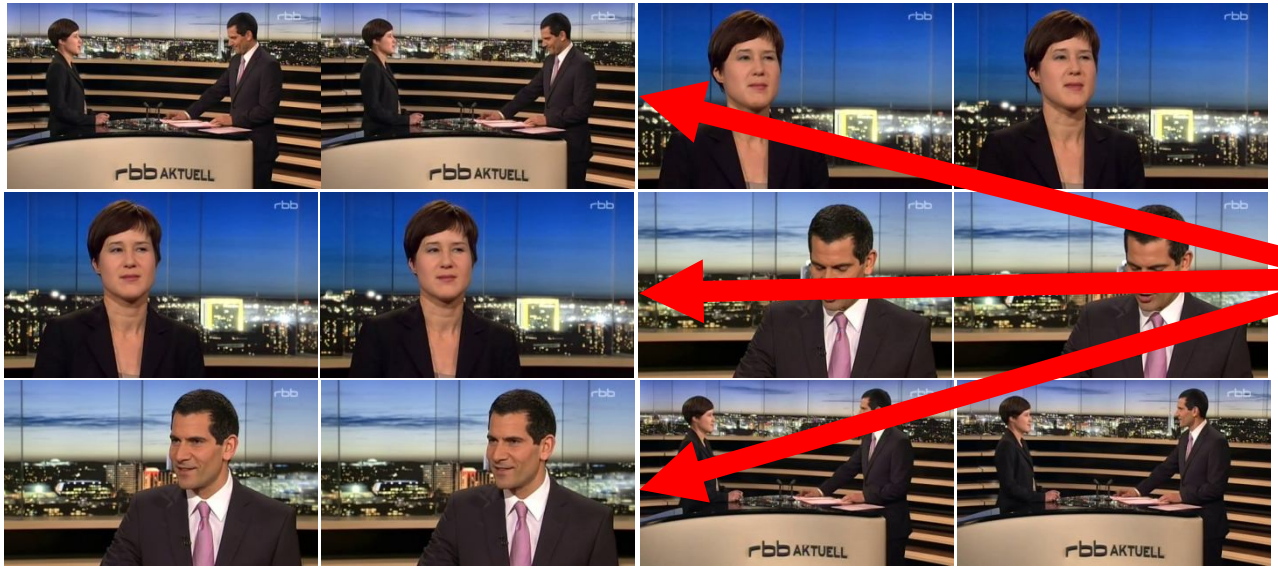
- What is a scene: a higher-level temporal video segment that is elementary in terms of semantic content, covering either a single event or several related events taking place in parallel
- Scene segmentation
  - temporal decomposition of videos into basic story-telling units
  - important prerequisite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

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
  - Scene segmentation is typically performed by

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
  - Scene segmentation is typically performed by
    - Shot segmentation, and



Shot Change

Problem statement

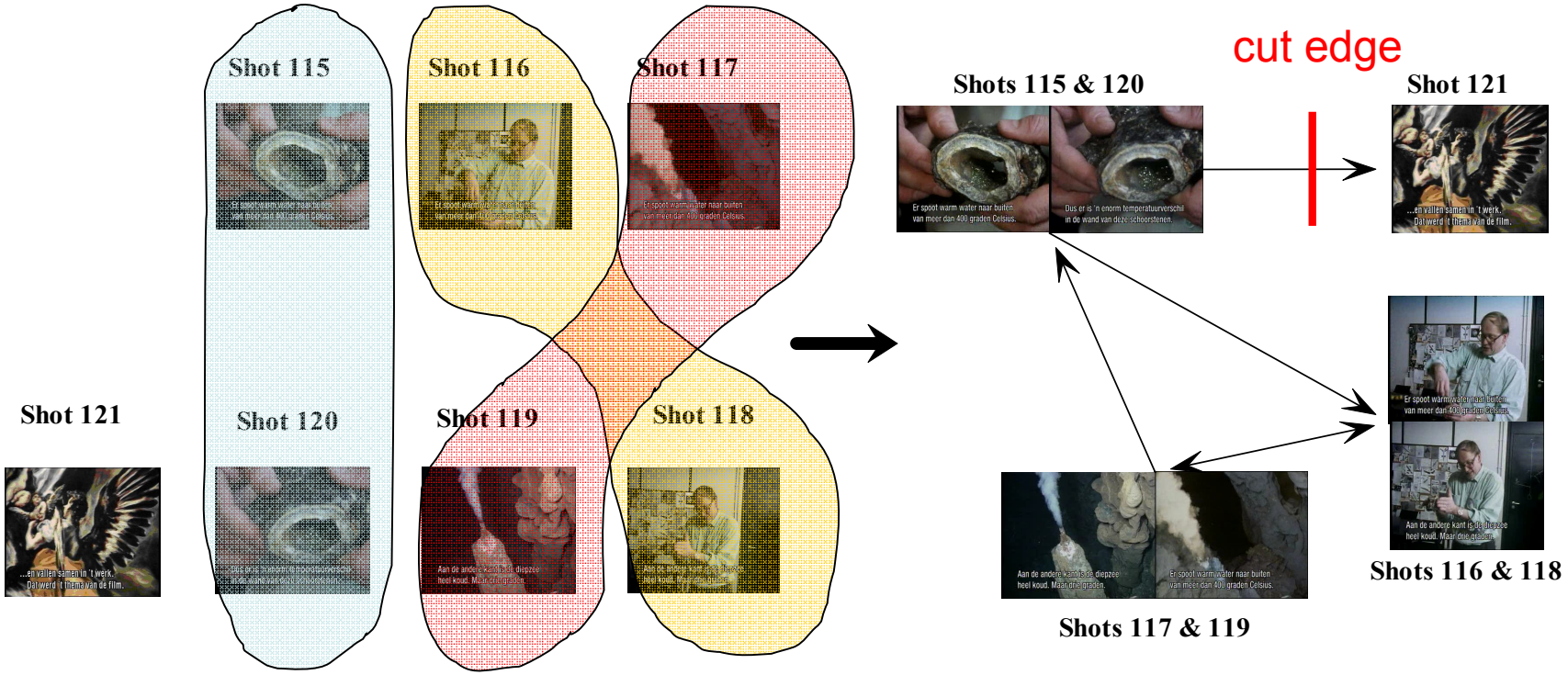
- Basic assumptions
  - A shot cannot belong to more than one scenes
  - Scene boundaries are a subset of the visual shot boundaries of the video
  - Scene segmentation is typically performed by
    - Shot segmentation, and
    - Shot grouping

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)

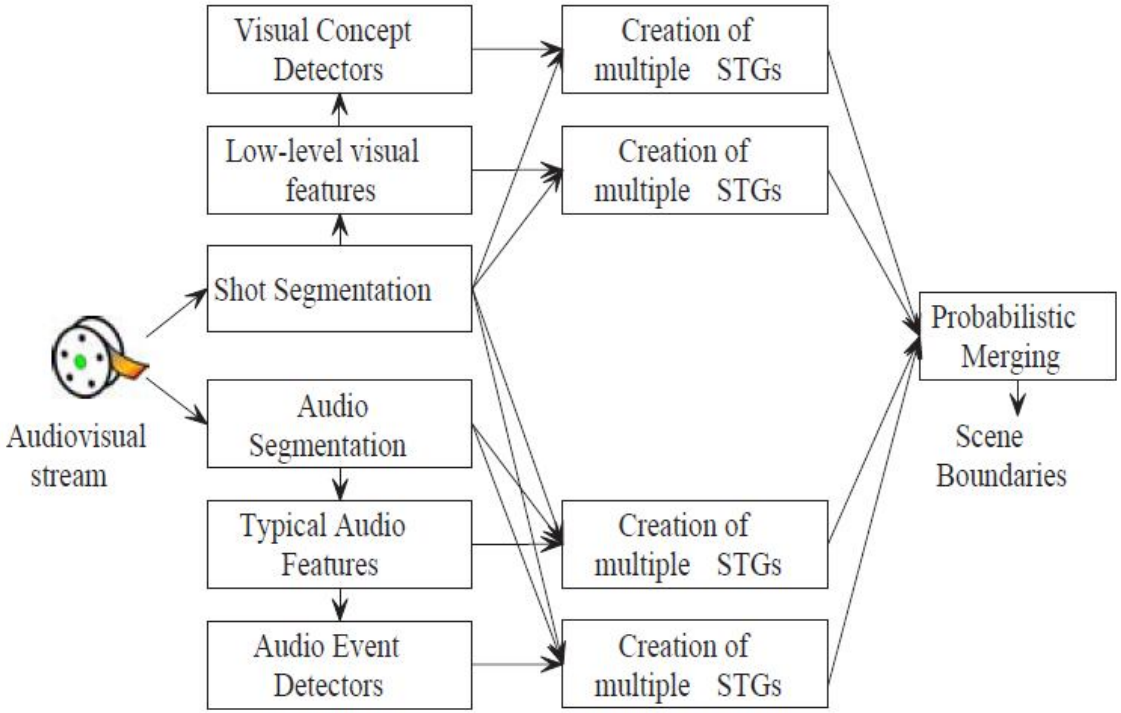
A MediaMixer-promoted approach

- Based on the Scene Transition Graph (STG) algorithm



A MediaMixer-promoted approach

- Introduces two extensions of the STG
  - 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)



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

Experiments and Results

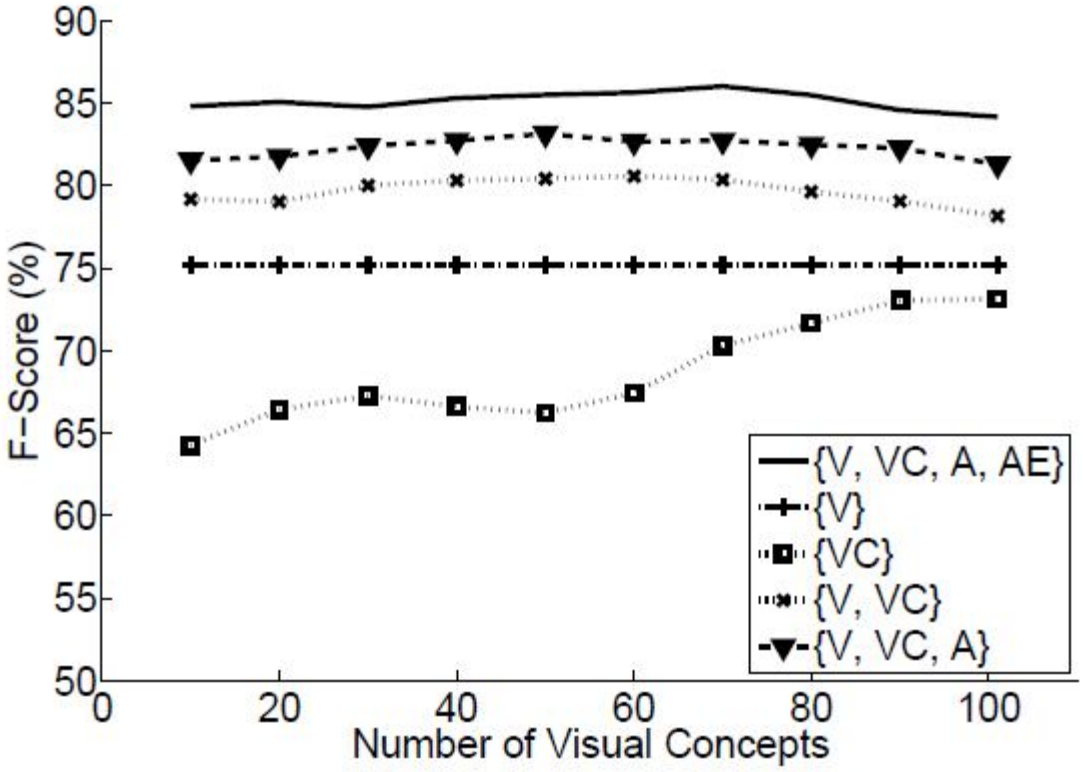
- Detection accuracy expressed in terms of:
  - Coverage (C): to what extent frames belonging to the same scene are correctly grouped together (optimal value 100%)
  - Overflow (O): the quantity of frames that, although not belonging to the same scene, are erroneously grouped together (optimal value 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. video's actual duration), as long as the features have been extracted

Experiments and Results

- Contribution of different modalities (on a different dataset)



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, 71(1):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, 21(8):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, 7(6):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, 15(2):296–305, February 2005.
- Zhao, Y., Wang, T., Wang, P., Hu, W., Du, Y., Zhang, Y., & Xu, G. (2007). Scene segmentation and categorization using N-cuts. IEEE Conference 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, 7(4):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, 13(1):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, 14(5):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. In Proceedings of 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, 8(4):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. In Proceedings of the TREC Video Retrieval Evaluation (TRECVID). Washington D.C.: NIST, pages 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., 25(7):767– 775, May 2004.
- H. Lu, Z. Li, and Y.-P. Tan. Model-based video scene clustering with noise analysis. In Proceedings of the 2004 International Symposium on Circuits and Systems, ISCAS '04, volume 2, pages 105–108, May 2004.
- Y. Ariki, M. Kumano, and K. Tsukada. Highlight scene extraction in real time from baseball live video. In Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, MIR '03, pages 209–214, New York, NY, USA, 2003. ACM.
- U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of tv news for automatic topic retrieval. In Proceedings of the 2001 IEEE International Conference on the Acoustics, Speech, and Signal Processing, volume 3 of ICASSP '01, pages 1397–1400, Washington, DC, USA, 2001. IEEE Computer Society.
- Y. Cao, W. Tavanapong, K. Kim, and J.H. Oh. Audio-assisted scene segmentation for story browsing. In Proceedings of the 2nd international conference on Image and video retrieval, CIVR'03, pages 446–455, Berlin, Heidelberg, 2003. Springer-Verlag.
- A. Velivelli, C.-W. Ngo, and T. S. Huang. Detection of documentary scene changes by audio-visual fusion. In Proceedings of the 2nd international conference on Image and video retrieval, CIVR'03, pages 227–238, Berlin, Heidelberg, 2003. Springer-Verlag.

Fragment annotation: visual concept detection

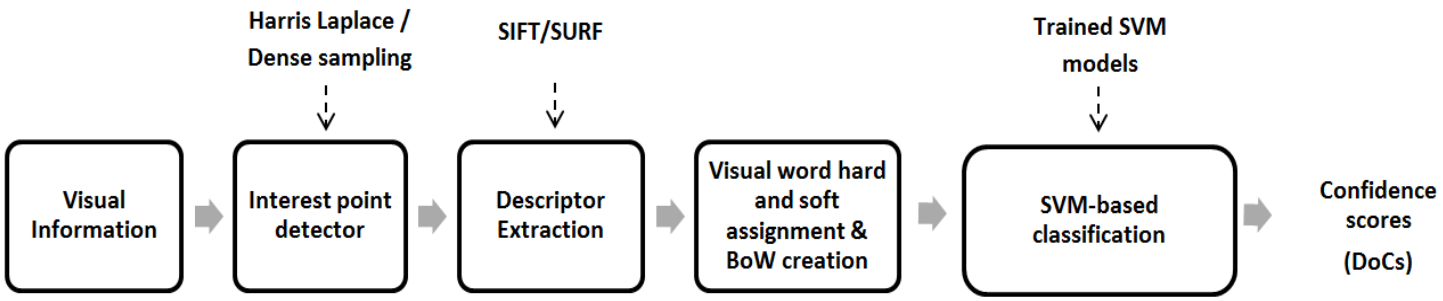
- Goal: assign one or more semantic concepts to temporal video fragments (typically, shots), from a predefined concept list
  - Input: visual fragment or representative visual information (e.g. keyframes)
  - Output: concept labels and associated confidence scores (DoC)
- Applications: concept-based annotation, image/video search and retrieval, clustering, summarization, further analysis (e.g. event detection)
- 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

- Typical concept detection system: consisting of independent concept detectors
  - Feature extraction (typically local features; choices of IP detectors/ descriptors
  - Global representation (Bag of Words, Fisher vectors, …)
  - Training/classification (supervised learning; need for annotated training examples)
  - Confidence score extraction



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

Related work

- Feature extraction
  - Visual features (global vs. local; most popular local descriptors SIFT, Color SIFT , SURF; interest point detection: Harris-Laplace, Hessian, dense sampling)
  - Motion features (STIP, MoSIFT, feature trajectories,…)
  - Others (text, audio): of limited use
- Feature encoding
  - Pyramidal decomposition
  - Bag-of-words (BoW): codebook construction (K-means); hard/soft assignment to codewords
  - Fisher vectors: extension of BoW; characterize each keyframe by a gradient vector
- Early / late fusion
- Machine learning
  - Binary classification: Support Vector Machines (SVMs),…
  - Multi-label learning approaches – Limited use due to time requirements
- Concept correlation
  - Inner-learning approaches
  - Stacking-based approaches

A MediaMixer-promoted approach

- Concept detection using a two-layer stacking architecture
  - 1st layer: Build multiple fast and independent concept detectors
    - Use keyframes, and tomographs to capture the motion information
    - Extract high-dimensional feature vectors
    - Train Linear Support Vector Machines
    - 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 (ML-*k*NN algorithm)
    - Use temporal re-ranking to exploit temporal information

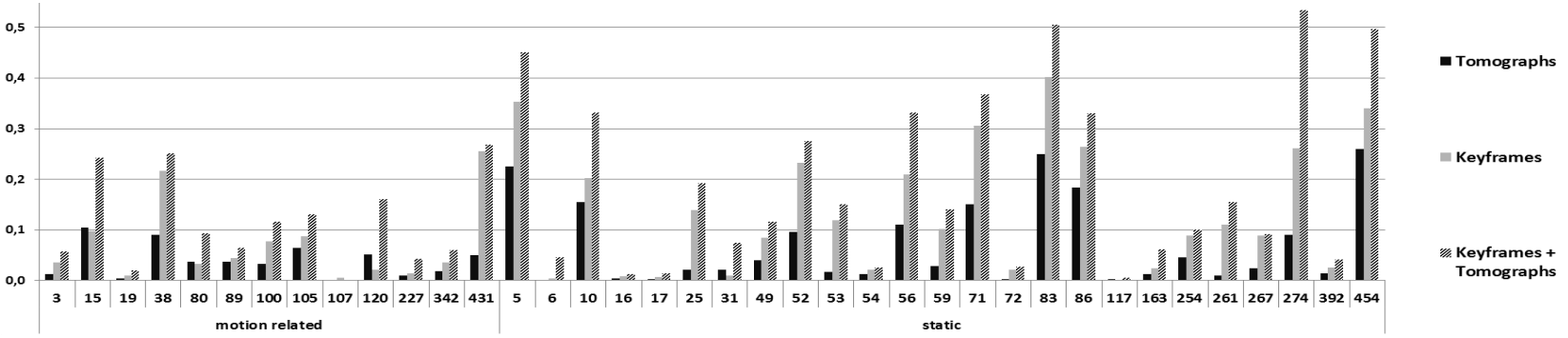
A MediaMixer-promoted approach

- 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

Representation	Feature extraction procedure
Keyframe	**12 keypoint-, keyframe-based** feature extraction procedures (3 descriptors (SIFT, Opponent-SIFT, RGB-SIFT) x 2 sampling strategies (Dense, Harris-Laplace) x 2 BoW strategies (soft-, hard-assignment)) **1 global-image** feature extraction procedure (color histograms)
Tomograph	**12 keypoint-, tomograph-based** feature extraction procedures (2 types of video tomographs (horizontal, vertical) x 3 descriptors (SIFT, Opponent-SIFT, RGB-SIFT) x 2 BoW strategies (soft-, hard-assignment))

Experiments and Results

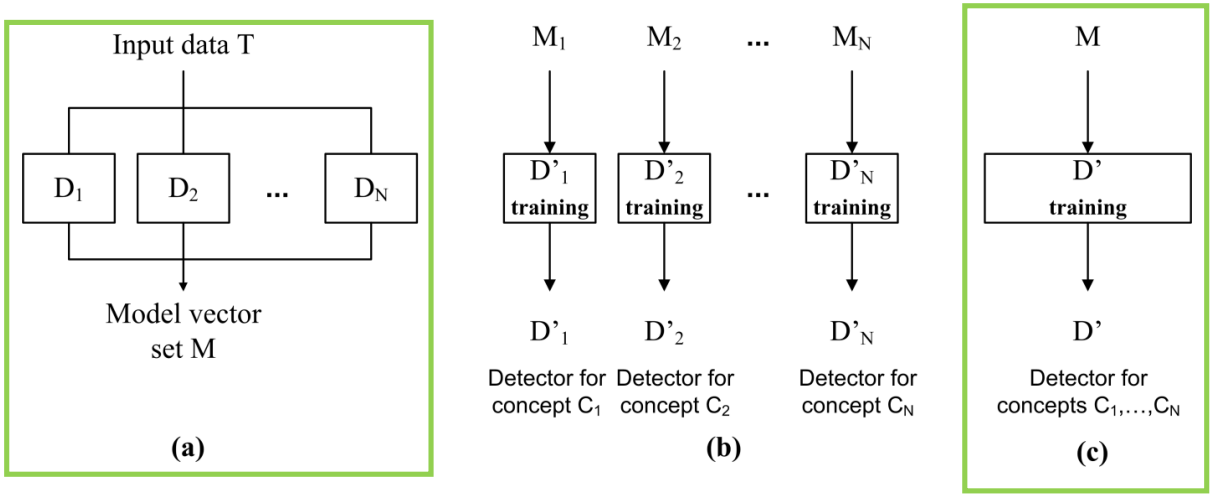
- Experimental setup
  - The TRECVID Semantic Indexing Task: Using the concept detectors retrieve for each concept a ranked list of 2000 test shots that are mostly related with it
  - Dataset: TRECVID 2013 (~800 and ~200 hours of internet archive videos for training and testing), 38 concepts (13 of them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



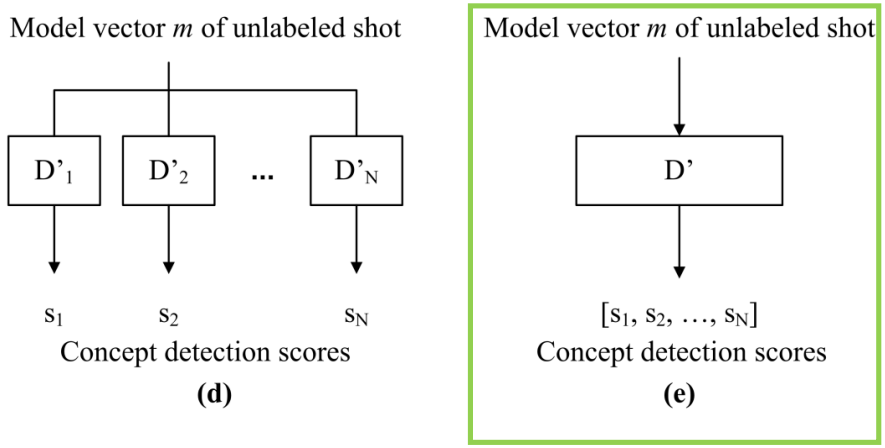
A MediaMixer-promoted approach

- Stacking-based ML-*k*NN to exploit concept correlations
  - Construct model vectors by concatenating the responses of the concept detectors on a separate validation set
  - Use the model vectors to train a ML-kNN model
  - ML-*k*NN : A lazy style multi-label learning algorithm
  - ML-*k*NN uses label correlations in the neighbourhood of the tested instance, to infer posterior probabilities.

A MediaMixer-promoted approach



The baseline



The proposed approach

Experiments and Results

- Experimental setup
  - Indexing task: given a concept, measure how well the top retrieved video shots truly relate to it
  - Annotation task: given a shot, measure how well the top retrieved concepts describe it
  - Dataset: TRECVID 2011 and 2012 (~800 hours of internet videos each)
  - ~700 hours for training; ~100 hours for testing
  - Input: model vectors from 346 concepts
  - Output: refined scores for 50/46 concepts (for TRECVID 2011, 2012)
  - Evaluation: Mean Average Precision (MAP), Mean Precision at depth k (MP@k)
- Comparison
  - System_1: baseline system consisting of independent concept detectors
  - System_2: two-layer stacking architecture with ML-*kNN*

	TRECVID 2011				TRECVID 2012			
Method	MAP	MP@100	MAP	MP@3	MAP	MP@100	MAP	MP@3
System_1	0.340	0.660	0.615	0.370	0.205	0.371	0.601	0.325
System_2	0.496	0.808	0.681	0.415	0.318	0.528	0.770	0.411

Concept detection conclusions

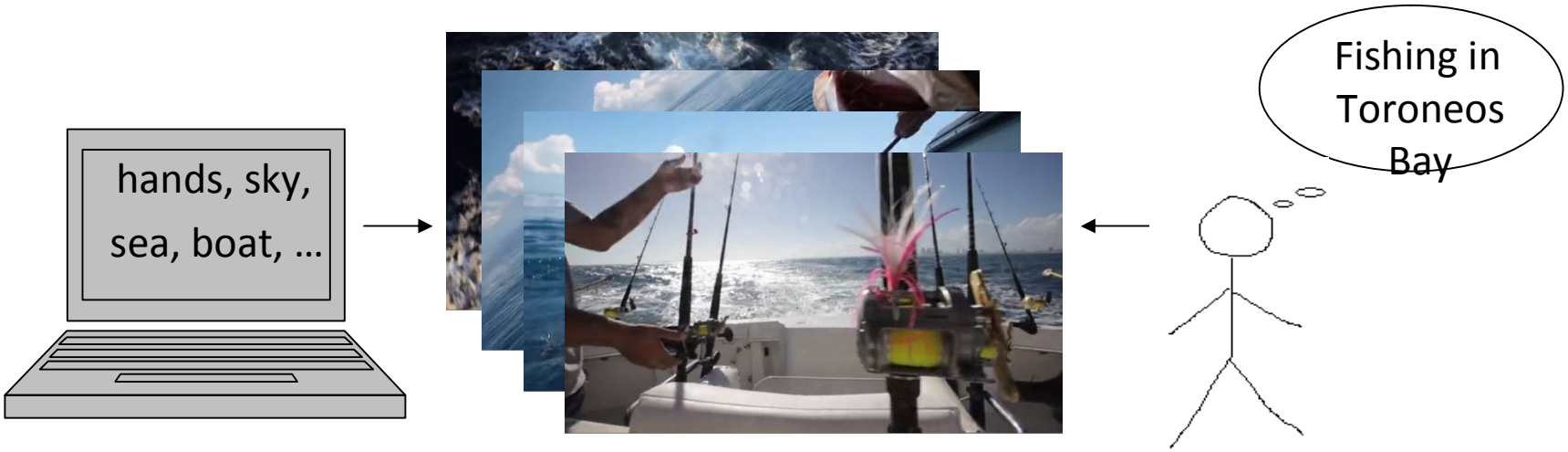
- Concept detection has progressed a lot
- Results far from perfect; yet, already useful in a variety of applications (retrieval, further analysis of fragments)
- Motion information is important (but, extracting traditional motion descriptors more computationally expensive than working with keyframes / tomographs)
- Linear SVMs very popular (due to the size of the problem)
- Exploiting concept correlations is very important
- Computationally-efficient concept detection, considering hundreds or thousands of concepts, is another major challenge

Concept detection: additional reading

- P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Enhancing video concept detection with the use of tomographs. In IEEE International 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, to appear.
- C. Snoek, M. Worring. Concept-Based Video Retrieval. Foundations and Trends in Information Retrieval 2(4) (2009) 215–322.
- 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. In: 2010 Int. Conf. on Digital Image Computing: Techniques and Applications, China, IEEE (2010) 617–622.
- G.J. Qi, X.S. Hua, Y. Rui, J. Tang, T. Mei, H. Zhang. Correlative multi-label video annotation. In: 15th international conference on Multimedia. MULTIMEDIA ’07, New York, ACM (2007) 17–26.
- J. Smith, M. Naphade, A. Natsev. Multimedia semantic indexing using model vectors. In: 2003 Int. Conf. on Multimedia and Expo. ICME ’03., New York, IEEE press (2003) 445–448.
- M.-L. Zhang and Z.-H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 40(7), 2007.
- A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In MIR '06: Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval, pages 321-330, New York, NY, USA, 2006. ACM Press.
- 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, pages 2081-2084. ACM, 2011.
- J. C. Van Gemert, C. J. Veenman, A. Smeulders, J.-M. Geusebroek. Visual word ambiguity. IEEE transactions on pattern analysis and machine intelligence, 32(7), 1271–83, 2010.
- G. Csurka and F. Perronnin, Fisher vectors: Beyond bag-of visual-words image representations, Computer Vision, Imaging and Computer Graphics. Theory and Applications, 2011

Fragment annotation: event detection

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



Problem statement

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

“Getting a vehicle unstuck”

“Grooming an animal”

“Making a sandwich”

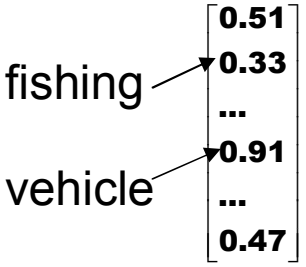
- Building an event detection system
  - Exploit an annotated video dataset
  - Represent videos with suitable feature vectors $\{(\boldsymbol{x}, y) \in X \times \{-1,1\}\}$
  - Learn an appropriate mapping (event detector) $f: X \to [-1,1]$

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
  - The inspiration behind this approach is that high-level events can be better recognized by looking at their constituting semantic entities
  - Experimental results sho improved event detection performance when model vectors are used
- Hybrid approaches: combination of low-level features and model vectors

A MediaMixer-promoted approach

Model vector representation



- Temporal video segmentation
  - Shot segmentation, keyframes at fixed time intervals, etc.
- Low-level feature extraction
  - Spatial pyramid decomposition scheme, Keypoint detection, keypoint descriptors, BoW model, soft/hard assignment
- Application of a set of trained visual concept detectors
  - TRECVID SIN task, SVM-based
  - Detectors may be seemingly irrelevant to the sought events
- A visual model vector is formed for each video keyframe
  - Concatenate the responses (confidence scores) of all the detectors
- Video representation
  - Sequence of model vectors or an overall model vector, e.g., averaging the model vectors for all shots of a video

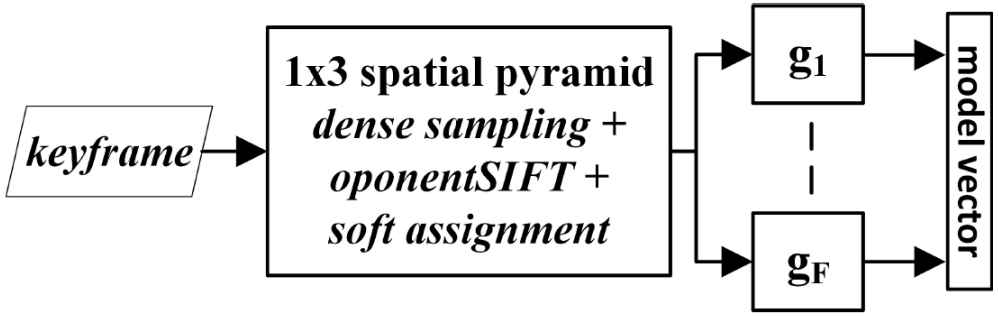
A MediaMixer-promoted approach

Event detection using discriminant concepts

- Use a DA algorithm to extract the most significant concept information for the detection of the target events
  - MSDA is used to derive a discriminant concept subspace
- Classification is done using LSVM in the discriminant subspace
- Advantages
  - In comparison to using the “raw” model vectors, MSDA improves performance (accuracy, speed, storage) by discarding noise or irrelevant concept detections
  - MSDA is much faster than Kernel DA variants typically used to solve nonlinearity problems

Experiments and Results

- Model vectors at video-level are used
- A video is decoded and 1 keyframe every 6 seconds is extracted
- A keyframe is represented with a 4000 BoW vector: 1x3 spatial pyramid decomposition scheme, dense sampling, opponentSIFT descriptor, 4000 BoW model (1000 visual words per pyramid cell), soft assignment
- Concept detectors: 346 TRECVID SIN 2012 concepts, LSVM-based



- A model vector at video-level is created by averaging model-vectors at keyframe-level

Experiments and Results

- Using the TRECVID 2010 dataset
  - 3 target events: assembling a shelter (E01), batting a run in (E02), making a cake (E03)
  - 3487 videos (development: 1745, evaluation: 1742)
  - Evaluation measure: MAP
- Event detection using discriminant concepts
  - Model vectors are projected in discriminant subspace using MSDA
  - Classification of test videos is done using an LSVM classifier
  - Comparison with LSVM classifier trained using “raw” model vectors
- Results:

	E01	***E02***	***E03***	***MAP***
LSVM	0.106	0.477	0.103	0.229
MSDA+LSVM	**0.180**	**0.648**	**0.106**	**0.311**
Boost	70%	36%	3%	36%

Event detection conclusions

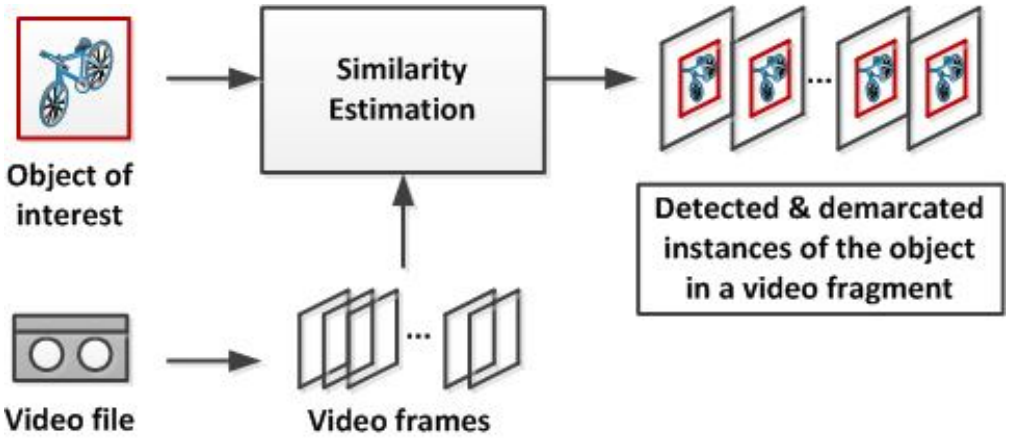
- Performance of the event detection system increases by
  - Discarding irrelevant or noisy concept detections
  - Effectively combining multiple classifiers
  - Exploiting the subclass structure of the event data
- General hints
  - Importance of low level features: Visual motion features are the most important followed by visual static features; for some events, audio features provide complementary information
  - However, the use of motion features in large-scale video databases has high computational cost (associated with their extraction)
  - Combining low-level features with model vectors provides small but noticeable performance gains

Event detection: additional reading

- P. Over, J. Fiscus, G. Sanders et. al., "TRECVID 2012 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics", Proc. TRECVID 2012 Workshop, November 2012, Gaithersburg, MD, 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, and I. Kompatsiaris, “High-level event detection in video exploiting discriminant concepts,” in Proc. 9th Int. Workshop CBMI, Madrid, Spain, June 2011, pp. 85–90.
- 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 Multimedia 2013 (MM’13), Barcelona, Spain, October 2013.

Spatiotemporal fragment creation and annotation: object re-detection

- 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 file
  - Processing: similarity estimation by means of image matching
  - Output: detected instances of the object of interest
  - If input includes a label for this object, this label can also be propagated



Related work

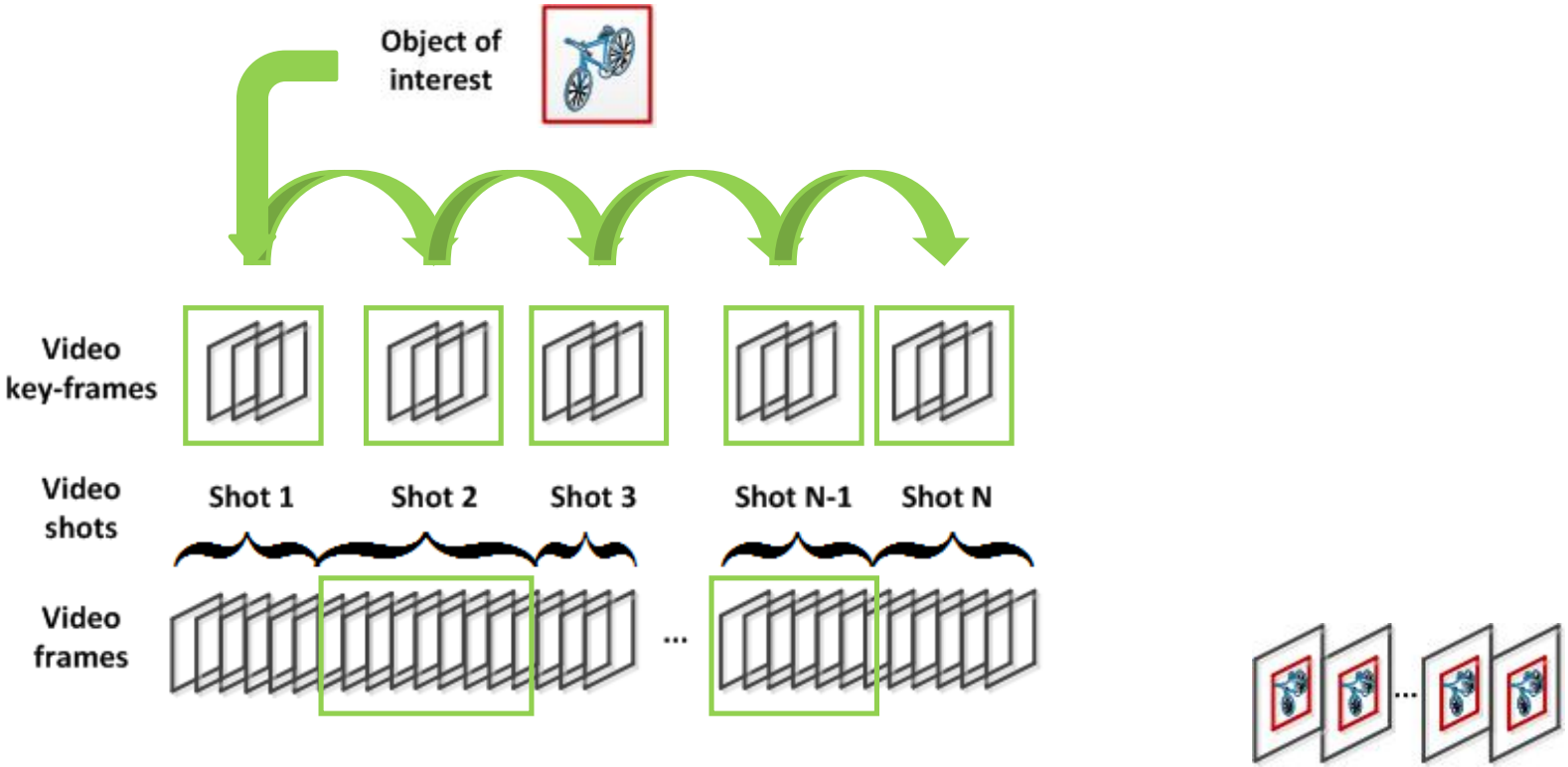
- Extraction and matching of scale- and rotation-invariant local descriptors is one of the most popular SoA approaches for similarity estimation between pairs of images
  - Interest point detection (e.g. Harris-Laplace)
  - Local feature extraction (e.g. SIFT, SURF)
  - Matching of local descriptors (e.g. k-Nearest Neighbor search between descriptor pairs using brute-force, hashing)
  - Filtering of erroneous matches
    - Symmetry test between the pairs of matched descriptors
    - Ratio test regarding the distances of the calculated nearest neighbors
    - Geometric verification between the pair of images using RANSAC
- Various extensions, e.g.
  - Combined use of keypoints and motion information
  - Bag-of-Words (BoW) representation and matching for pruning
  - Graph matching approaches

A MediaMixer-promoted approach

- Starting from a baseline,
  - Improve detection accuracy
  - Reduce the processing time
- Work directions:
  - GPU-based processing
  - Video-structure-based sampling of frames
  - Enhancing robustness to scale variations

A MediaMixer-promoted approach

- Sequential processing of video frames is replaced by a structure-based one, using the analysis results of a shot segmentation method



A MediaMixer-promoted approach

Problem: major changes in scale may lead to detection failure due to the significant limitation of the area that is used for matching (see figures (b) and (c) as zoomed-in and -out instances of figure (a))


a


b


c

Solution: we automatically generate a zoomed-out and a centralized zoomed-in instance of the object of interest and we utilize them in the matching procedure

Original image

Zoomed-in instance

Zoomed-out instance

Experiments and Results

- Dataset
  - 6 videos* of 273 minutes total duration
  - 30 manually selected objects

- Ground-truth (generated via manual annotation)
  - 75.632 frames contain at least one of these objects
  - 333.455 frames do not include any of these objects

Examples of sought objects

- Robustness to scale variations was quantified using two specific sets of frames where the object was observed from a very close (2.940 frames) and a very distant (4.648 frames) viewing position

** The videos are episodes from the "Antiques Roadshow" of the Dutch public broadcaster AVRO (http://avro.nl/)*

Experiments and Results

Detection accuracy

- The algorithm is robust against a wide range of different scales and orientations and for partial visibility or partial occlusion
  - Overall data-set: Precision 99.9%, Recall 87.2%
  - Zoomed-in instances: Precision 100%, Recall = 99.2%
  - Zoomed-out instances: Precision 100%, Recall = 91.4%

Processing time

- 10 times faster than real-time (i.e. about 10% of the video's duration)

Example of a 2D object of interest

Single instance

→

Online demo available at: http://www.youtube.com/watch?v=0IeVkXRTYu8

Experiments and Results

Detection accuracy

- The algorithm is robust against a wide range of different scales and orientations and for partial visibility or partial occlusion
  - Overall data-set: Precision 99.9%, Recall 87.2%
  - Zoomed-in instances: Precision 100%, Recall = 99.2%
  - Zoomed-out instances: Precision 100%, Recall = 91.4%

Processing time

- 10 times faster than real-time (i.e. about 10% of the video's duration)

Example of a 3D object of interest

Multiple instances

→

Online demo available at: http://www.youtube.com/watch?v=0IeVkXRTYu8

Object re-detection conclusions

- Accurate re-detection of pre-defined objects in video is possible
- Choice of objects plays important role
  - Complex objects can be detected more reliably than simpler ones
  - True 3D objects more challenging than “2D” ones (e.g. paintings)
- Faster-than-real-time processing of video is possible
  - Re-detection can be used in interactive applications
- Several possible uses
  - Instance-level annotation
  - Finding and linking related videos or fragments of them
  - Supporting other analysis tasks, e.g. scene detection

Object re-detection: additional reading

- L. Apostolidis, V. Mezaris, I. Kompatsiaris, "Fast object re-detection and localization in video for spatio-temporal fragment creation", Proc. 1st Int. Workshop on Media Fragment Creation and reMIXing (MMIX'13) at the IEEE Int. Conf. on Multimedia and Expo (ICME 2013), San Jose, CA, USA, July 2013.
- M. Muja and D. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in In VISAPP International Conference on Computer Vision Theory and Applications, 2009, pp. 331–340.
- Z. He and Q. Wang, “A fast and effective dichotomy based hash algorithm for image matching,” in Proceedings of the 4th International Symposium on Advances in Visual Computing, Berlin, Heidelberg, 2008, ISVC ’08, pp. 328–337, Springer-Verlag.
- B. Kulis and K. Grauman, “Kernelized locality-sensitive hashing for scalable image search,” in ICCV, 2009, pp. 2130–2137.
- D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli, “Surftrac: Efficient tracking and continuous object recognition using local feature descriptors,” in IEEE Conf. on Computer Vision and Pattern Recognition, CVPR’09, June.
- J. Fauqueur, G. Brostow, and R. Cipolla, “Assisted video object labeling by joint tracking of regions and keypoints,” in 11th IEEE Int. Conf, on Computer Vision, ICCV 2007, Oct., pp. 1–7.
- J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in Proc. of the 9th IEEE Int. Conf. on Computer Vision, Washington, USA, 2003, ICCV ’03, pp. 1470–1477.
- S. Hinterstoisser, O. Kutter, N. Navab, P. Fua, and V. Lepetit, “Real-time learning of accurate patch rectification,” in IEEE Conf. on Computer Vision and Pattern Recognition, CVPR’09, June, pp. 2945–2952.
- J. Mooser, Q. Wang, S. You, and U. Neumann, “Fast simultaneous tracking and recognition using incremental keypoint matching,” in Proceedings of the 4th Int.l Symposium on 3D Data Processing, Visualization and Transmission, 3DPVT’08, Atlanta, USA, June 2008.
- H. Y. Kim, “Rotation-discriminating template matching based on fourier coefficients of radial projections with robustness to scaling and partial occlusion,” Pattern Recogn., vol. 43, no. 3, pp. 859–872, Mar. 2010.
- O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce, “A tensor-based algorithm for high-order graph matching,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 12, pp. 2383–2395, 2011.

Concluding remarks

- We discussed different classes of techniques for media fragment creation and annotation, but several others also exist, e.g.
  - Object recognition
  - Face detection, tracking, clustering, recognition
  - Quality assessment
  - Sentiment / emotion detection
- Not all of techniques for media fragment creation and annotation are suitable for every possible problem!
- Understanding the problem at hand and the volume, value and variability of the data is key to selecting appropriate methods
- Is some cases the automatic analysis results remain far from perfect (manual) annotations; yet, these results may still be very useful in the right domain or for solving the right problem

Thank you! Questions?

More information:
http://www.iti.gr/~bmezaris
bmezaris@iti.gr