Scalable Content-based Video Copy Detection

Michel Crucianu
(joint work with Sébastien Poullot and Olivier Buisson)

Conservatoire National des Arts et Métiers (Paris)
Vertigo research group
http://cedric.cnam.fr/vertigo/

Conservatoire National des Arts et Métiers

"Omnes docet ubique"

- Created on October 10, 1794, by the Convention
- Today
  - Lifelong education for about 88,000 students
  - Paris + 150 regional centers in France and other 40 countries
- 4 departments
  - Computer science, mathematics and electronics
  - Industrial sciences and techniques
  - Economics and management
  - Humanities and social sciences

Computer Science Lab (CEDRIC)

http://cedric.cnam.fr/?lang=en

- 120 people (65% being PhD students and post-docs)
- 5 teams
  - Certified design and programming
  - Interactive media and mobility
  - Information systems ⊃ Vertigo research group
  - Combinatorial optimization
  - Statistical methods for data-mining and learning
- Funding: mainly based on national projects

Vertigo research group

- How to deal with data having little or no structure (multimedia, data on the Web)?

→ Scalable methods for
  - structuring (making the relevant structure explicit)
  - querying
- Research directions
  - Large image and video databases
  - Data and services on the Web
- Today: 4 permanent staff, 5 PhDs, 1 post-doc
- Ongoing: 3 national projects, 4 national and 3 international collaborations

Content-Based Video Copy Detection: Outline

- What is a video “copy”
- Requirements: robustness and scalability
- Robust copy detection and local description
- Video stream monitoring
- Video mining by content-based copy detection
- Conclusion and perspectives

What is a Video “Copy”

- Copy = transformed version of an original video content
- What transformations are most frequently encountered
  - Photometric: contrast, gamma, color to B&W, noise, blur
  - Geometric: crop, change in scale or format
  - Temporal: tempo change, addition or suppression of images
  - Post-production: excerpt, compilation, logo or subtitle addition, video inlay, borders, re-encoding...
- Copies:

Not copies:

CBVCD: Motivation

- Video copy detection: what for
  - Protect the rights of content owners: detect potentially illicit transmission over video streams (Hertzian, satellite, cable TV, Internet) or in video databases (Web2.0, peer-to-peer)
  - Control broadcast agreements and/or (semi-)automatic billing
  - Video database mining using copy detection

1. Video copy detection by robust watermarking
   - Only if the original video was watermarked before any dissemination
   - Variable robustness to different types of "attacks"
   - Many alternative proposals ⇒ practical difficulties
2. Content-Based Video Copy Detection (CBVCD): to have some interest for a viewer, a "copy" should preserve the main visual information that is present in the original

CBVCD and retrieval by similarity

- "Keep the essential visual information" ⇒ similarity (and not just any type of similarity!) between the original and the "copy"
- Principle of Content-Based Video Copy Detection

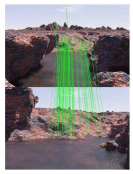
  The candidate video is a copy of the original if it is similar enough (with an adequate content representation and similarity measure) to the original

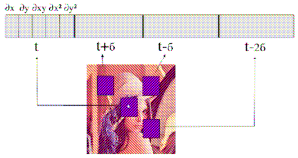
  → The representation of the candidate video is used as a query by similarity in a database containing the representations of all the original videos

CBVCD: what video description

- Global image description?
  - Not robust to the expected transformations...

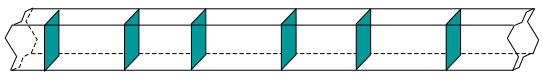
→ Local image description
  - SIFT, PCA-SIFT, GLOH, SURF?
    - Too invariant to changes in scale, viewpoint, etc.?
    - Computationally expensive (extraction, similarity-based retrieval)
  - → Improved Harris interest point detector, spatiotemporal differential description
    - Robustness (within tight bounds) to changes in scale and
    - Comparatively "light" (dimension 20)





CBVCD: General Processing Steps

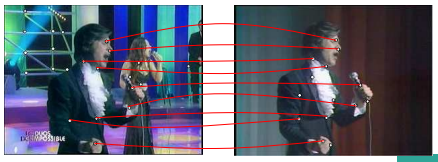
1. Extraction of keyframes from the video:



2. Extraction of points of interest in the keyframes:

3. Retrieval of candidates from the database:

4. Matching-based decision:



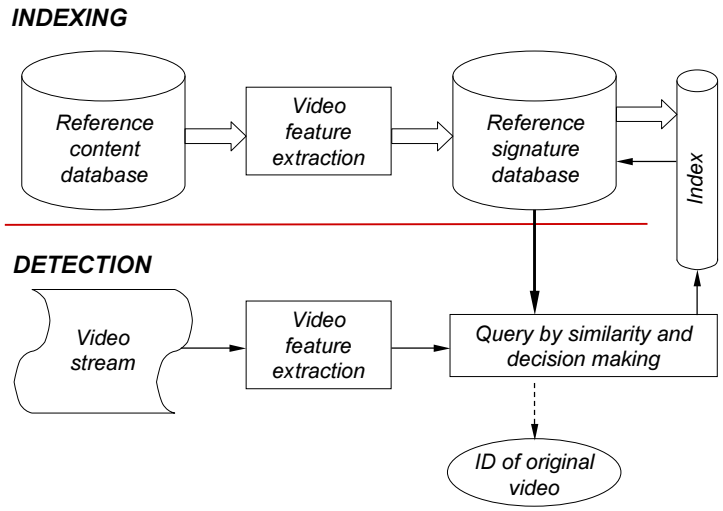
CBVCD: What About Scalability

- Every keyframe of the candidate video contains many points of interest, the signature of each point is a query in the database
- Size of the reference signature database:
  - 60 000 hours → 3 433 853 921 signatures
  - 280 000 hours → 16 354 748 143 signatures…
- CBVCD for video stream monitoring: our approach
  - Batch processing: accumulate a large number of queries, successively load into main memory parts of the database, process the accumulated queries with respect to the loaded part

  ⇒ Optimize throughput, not latency!

  ⇒ The index structure mainly serves to diminish the number of distance computations!

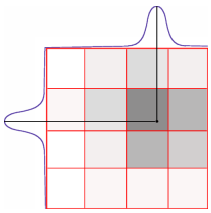
Stream Monitoring: Workflow



Stream Monitoring: Index

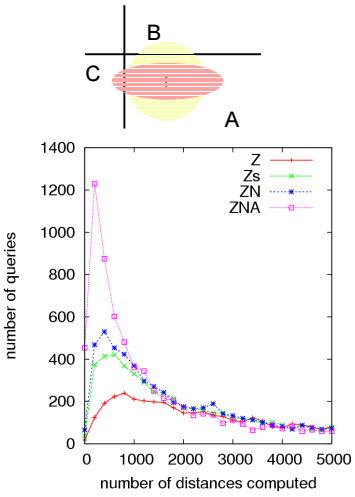
- k-d-B-tree (or LSDh-tree): for 60 000 hours of video, 8 Gb needed for storing the tree!
- Z-grid (Z): can be seen as a simplified k-d-B-tree, 2 bytes / level enough for storing the tree
- Partial balancing (ZN): adapted partitioning + most uniform dimensions partitioned first
- What type of retrieval
  - ε-range? Clear-cut separation, but when the amplitude of a transformation increases, its probability diminishes
  - *k*NN? In low-density regions *k*NNs can be too far

→ Probabilistic (followed by ε-range + *k*NN…): retain cells such that their cumulated probability (following an appropriate density function) is above an application-defined threshold



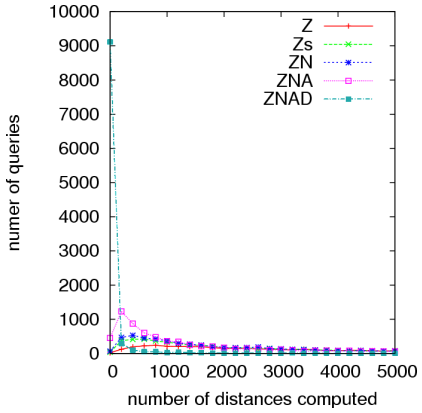
Stream Monitoring: Optimization

- Local models of signature distortions (ZNA)
  - Why: amplitude of signature transformations depends on location in feature space; better model → improved query selectivity (→ both higher speed and better quality)
  - How (hypothesis: independence between dimensions)
    - Estimate the models based on artificially generated copies
    - For each dimension: 1 model per abscissa value



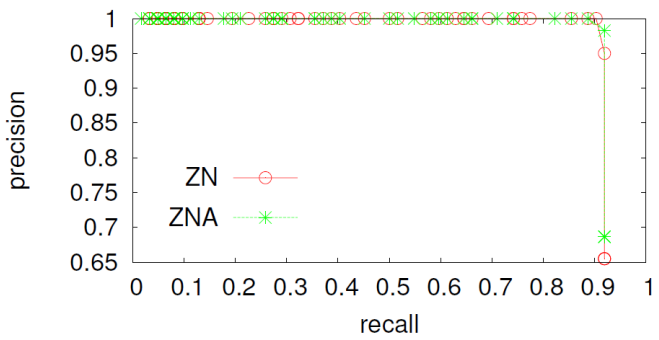
Stream Monitoring: Optimization

- Use of local signature density in description space (ZNAD)
  - Why: highly expensive but rather useless queries in locally very dense areas
  - How (hypothesis: independence between dimensions)
    - Estimate the local density (based on one-dimensional projections)
    - Modification of the cell selection thresholds



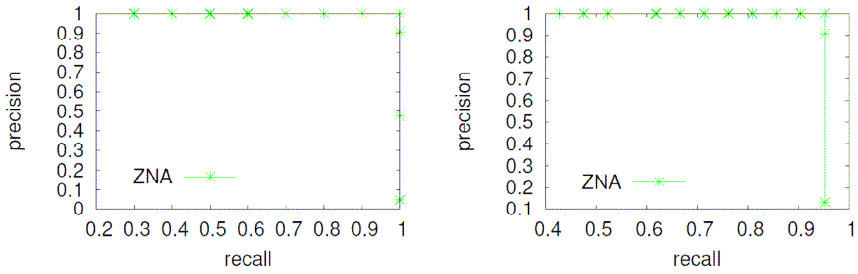
Stream Monitoring: Robustness

→ Precision and recall on INA ground truth (about 30 hours)



Stream Monitoring: Robustness

→ Precision and recall on CIVR 2007 ground truth (ST1, ST2)

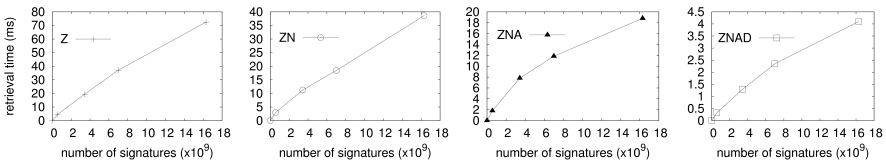


→ Precision diminishes by < 5% at same recall when the CIVR 2007 benchmark is inserted in the 280,000 hours database

Stream Monitoring: Speed

→ Mean cost (in milliseconds) per query

	HC	Z	ZN	ZNA	ZNAD
10,000 hours	17.1	4.5	2.7	1.8	0.3
60,000 hours	-	19.4	11.6	7.8	1.3
120,000 hours	-	37.1	18.4	11.8	2.35
280,000 hours	-	72.45	38.6	18.8	4.1

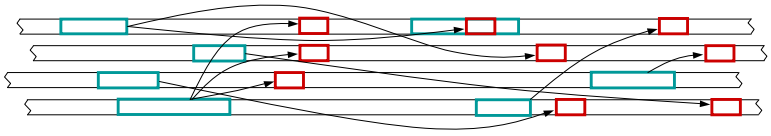


Video Mining by CBVCD

- Motivation: INA context, Web2.0 context
- Nature of the problem and challenges
- Compact description of local signatures: Glocal
- Indexing for similarity self-join
- Reconstruction of video sequences
- Evaluation and illustrated results
- Conclusion and perspectives

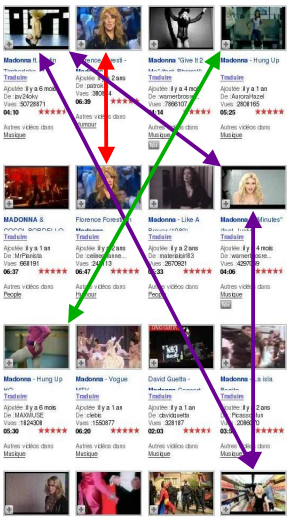
Video Mining: INA-like Context

- Goal: find, in a large video database, all the video sequences that occur several times (as more or less transformed versions)
- Applications
  - Segment and label content (detecting titles and credits)
  - Support librarians (extension of annotations)
  - Broadcast programming analysis
  - Media impact evaluation
  - Visual navigation in the database



Video Mining: Web2.0 context

- Current situation: large number of copies, scattered annotations of uneven quality
- Applications
  - Video database cleanup (gains in storage capacity)
  - Annotation cleanup and sharing
  - Detection of specific types of videos (e.g. compilations)
  - Cleanup the answers to the queries
  - Visual navigation in the answers



Mining by Copy Detection

- Direct application of stream monitoring? The signature of every point of every keyframe serves as a query by similarity in the database, the returned signatures serve to identify the keyframes that occur several times

→ Very large volume of intermediate results (*k*NNs of the query signatures) that are
  - distant in the description space (so also in the indexed database)
  - useless most of the time because issued from very different keyframes

→ Inefficient parallel implementations

→ Decision (for every keyframe) based on expensive matching

→ Examples: 10,000 h → mining taking 22 days, 2 Tb of storage
300,000 h → mining taking 4 years, 60 Tb of storage

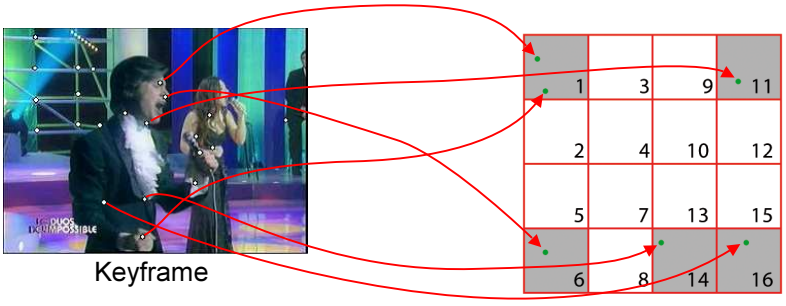
Principle of the Proposed Solution

1. Higher-level description: 1 signature per image
   - ⇒ Reduction in the size of the database
   - ⇒ More direct identification of similar images, without intermediate results and expensive matching

⇒ search for copies relies on similarity self-join

2. Segmentation of the database, with redundant indexing (allowed by the reduction in size)
   - ⇒ Computation cost diminishes
   - ⇒ Simple and efficient parallel implementation made possible

Compact Image Description



Keyframe

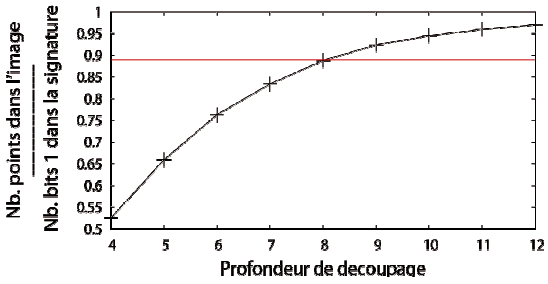
Description space

1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1

→ Glocal description + VideoID + TimeCode

Compact Image Description (2)

Preserves well local information:



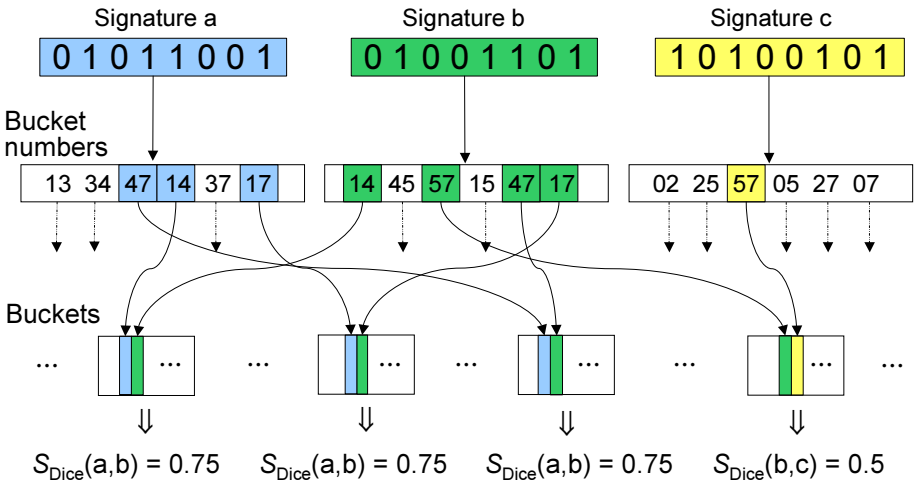
Similarity between Glocal: (Dice coefficient)

$$S_{\text{Dice}}(\mathbf{g}_1, \mathbf{g}_2) = \frac{2\,|\mathcal{G}_1 \cap \mathcal{G}_2|}{|\mathcal{G}_1| + |\mathcal{G}_2|}$$

Similarity self-join:

$$\mathcal{K}_\theta = \{(\mathbf{g}_i, \mathbf{g}_j) \mid \mathbf{g}_i, \mathbf{g}_j \in \mathcal{D}, i < j, S_{\text{Dice}}(\mathbf{g}_i, \mathbf{g}_j) > \theta\}$$

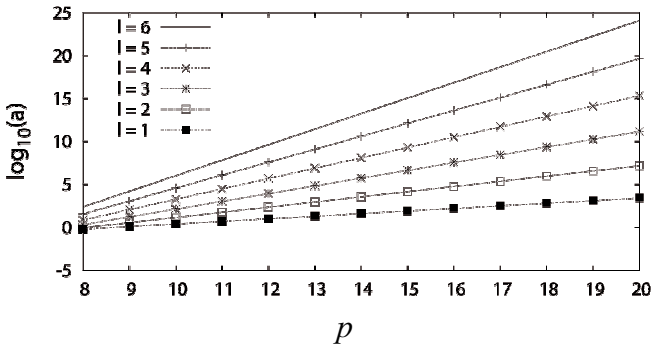
Glocal Indexing



Glocal Indexing (2)

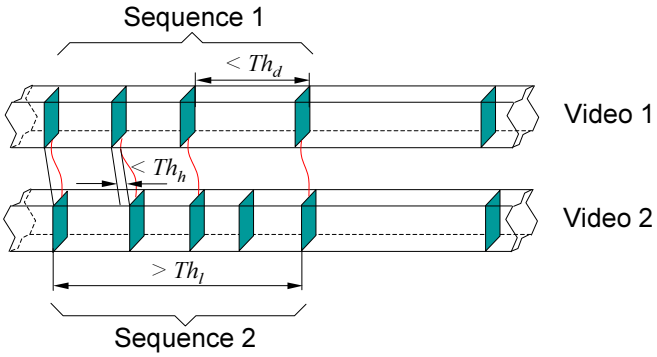
→ Controllable reduction of computation cost: $a = \frac{C_{2^p}^l}{(C_M^l)^2}$

- l = "sentence" length, M = nb. of bits set to 1
- p = partitioning depth



Linking Video Sequences

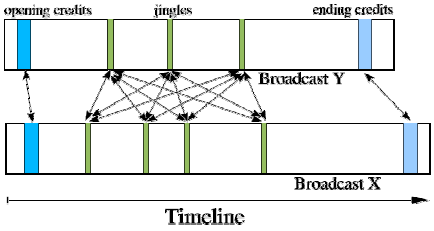
- Links between keyframes → correspondences between video sequences



Video Mining: Post-processing

→ Distinguish different types of content links using link structure and characteristics of the sequences

- Identify broadcast design sequences (titles and other show-specific sequences)



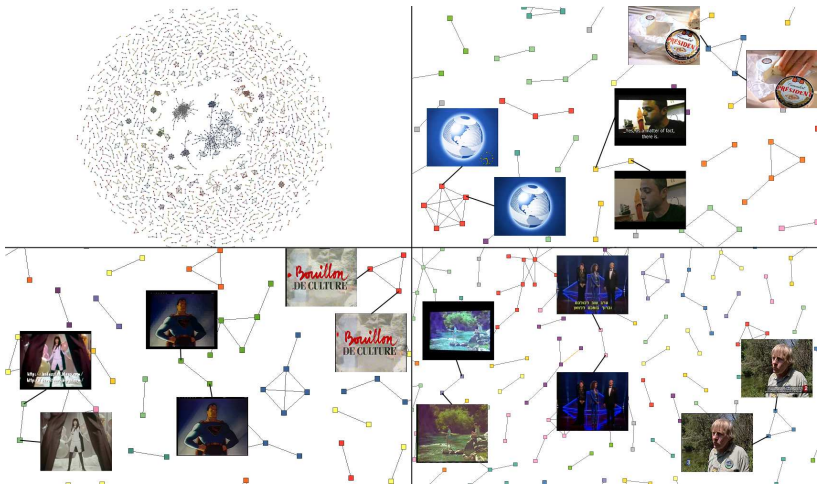
- Identify advertisements
- Find show or movie retransmissions
- Identify compilations of other video sequences

Video Mining: Evaluation

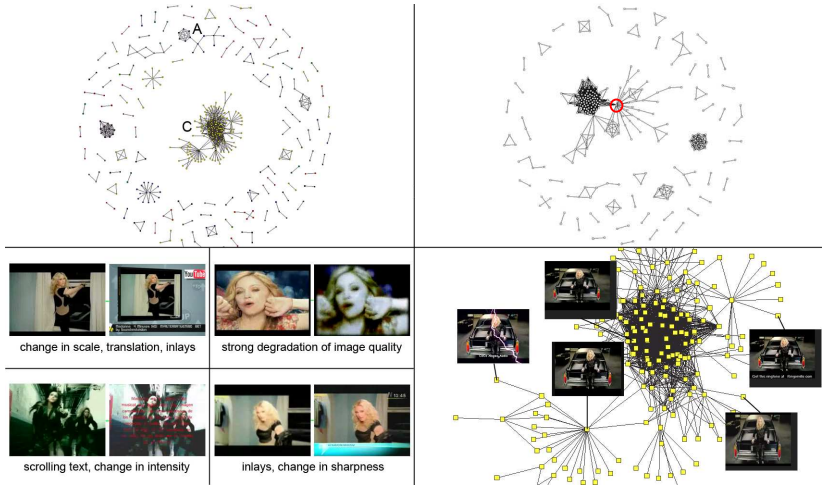
- Quality of detection: measured on ground truths
  - INA (30 h): recall 0.84 for precision 0.95 (θ = 0,55)
  - CIVR2007 (80 h): recall 0.8 for precision 0.96 (θ = 0,55)
- Mining Web2.0 data (pre-computed video signatures)
  - Example: 63 h (925 first videos in answer to a text query): 42 s
- Scalability

Size of the database	2,000 hours	5,000 hours	10,000 hours
Number of keyframes	5.8×10^6	14.5×10^6	28.7×10^6
Building the database	2 h 35 min	3 h 38 min	7 h 00 min
Similarity self-join	5 h 40 min	14 h 59 min	55 h
Linking video sequences	1 h 15 min	7 h 15 min	17 h 35 min

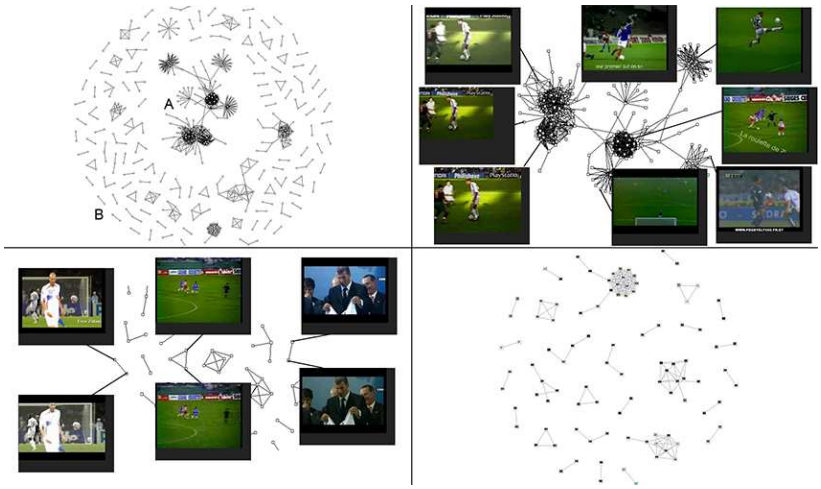
Video Mining: 1000 hours INA



Video Mining: Madonna query



Video Mining: Zidane query



General Conclusion

- Simple but optimized indexing for point of interest (PoI) signatures
  - Impact of local models of signature distortions and signature density

→ Monitor in deferred real time, with 1 PC, 1 video stream against a database of 280,000 hours of video

- Compact keyframe signature (Glocal) from set of PoI signatures
  - Direct evaluation of keyframe similarity
  - Reduction in the volume of data to store and process
  - Maintains good discriminating abilities (good precision)
- Similarity-based segmentation for similarity self-join
  - Controllable reduction of computation cost
  - Simple and efficient parallel processing made possible

→ Fast mining of the answers to online queries, realistic processing time for large databases

Perspectives

- Improve the indexing solution developed for mining
- Inexpensive solution for taking into account the spatial configuration of points of interest for mining, in order to further improve precision ← *ongoing*
- Application of the improved mining solution to video stream monitoring → potential for immediate replies
- Extension to more invariant (and higher-dimensional) descriptions in order to go beyond copy detection, to more general similarity measures ← *ongoing*
- Evaluation: how to measure detection quality (e.g. precision and recall) on very large databases?

References

A. Arasu, V. Ganti, and R. Kaushik. Efficient exact set-similarity joins. In Proc. 32nd intl. conf. on Very Large Data Bases (VLDB'06), pp. 918-929. VLDB Endowment, 2006.

R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In Proc. 16th intl. conf. on World Wide Web (WWW'07), pp. 131-140, New York, NY, USA, 2007. ACM.

O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In Proc. 6th ACM intl. Conf. on Image and Video Retrieval (CIVR'07), pp. 549-556, Amsterdam, The Netherlands, 2007. ACM Press.

J. M. Gauch and A. Shivadas. Finding and identifying unknown commercials using repeated video sequence detection. Computer Vision and Image Understanding, 103(1):80-88, 2006.

T. Quack, V. Ferrari, and L. J. V. Gool. Video mining with frequent itemset configurations. In H. Sundaram, M. R. Naphade, J. R. Smith, and Y. Rui, editors, CIVR, Lecture Notes in Computer Science vol. 4071, pp. 360-369. Springer, 2006.

S. Sarawagi and A. Kirpal. Efficient set joins on similarity predicates. In Proc. 2004 ACM SIGMOD intl. conf. on Management of data (SIGMOD'04), pp. 743-754, New York, NY, USA, 2004. ACM.

S. Satoh. News video analysis based on identical shot detection. In Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME'02), pp. 69-72, 2002.

S. Satoh, M. Takimoto, and J. Adachi. Scene duplicate detection from videos based on trajectories of feature points. In Proc. intl. workshop on Multimedia Information Retrieval (MIR'07), pp. 237-244, New York, NY, USA, 2007. ACM.

References

J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In Proc. 9th IEEE Intl. Conf. on Computer Vision (ICCV'03), pp. 1470-1477, Washington, DC, USA, 2003. IEEE Computer Society.

M. Takimoto, S. Satoh, and M. Sakauchi. Identification and detection of the same scene based on flash light patterns. In Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME'06), pp. 9-12, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

X. Wu, A. G. Hauptmann, and C.-W. Ngo. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In Proc. 15th ACM intl. conf. on Multimedia, pp. 168-177, New York, NY, USA, 2007. ACM.

X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In Proc. 15th ACM intl. conf. on Multimedia, pp. 218-227, New York, NY, USA, 2007.

F. Yamagishi, S. Satoh, and M. Sakauchi. A news video browser using identical video segment detection. In K. Aizawa, Y. Nakamura, and S. Satoh, editors, PCM (2), Lecture Notes in Computer Science vol. 3332, pp. 205-212. Springer, 2004.

Y. Zhai and M. Shah. Tracking news stories across different sources. In Proc. 13th ACM intl. conf. on Multimedia, pp. 2-10, New York, NY, USA, 2005.

References to Our Work

[Poullot, S., Buisson, O., Crucianu, M. Scaling Content-Based Video Copy Detection to Very Large Databases. Submitted.]

Poullot, S., Crucianu, M., Buisson, O. (2008) Fast Content-Based Mining of Web2.0 Videos, In *Pacific-Rim Conference on Multimedia.* Taiwan, December 2008.

Poullot, S., Crucianu, M., Buisson, O. (2008) Scalable Mining of Large Video Databases Using Copy Detection. In *Proceedings of ACM Multimedia 2008*, pp. 61-70. Vancouver, Canada, October 27-30, 2008.

Poullot, S., Buisson, O., Crucianu, M. (2007) Z-grid-based Probabilistic Retrieval for Scaling Up Content-Based Copy Detection. In *ACM International Conference on Image and Video Retrieval.* Amsterdam, July 9-11, 2007, pp. 348-355.