



Content-based indexing and retrieval in large collections of images and videos

Valérie Gouet-Brunet

valerie.gouet@cnam.fr

Conservatoire National des Arts et Métiers (CNAM), Paris France

CEDRIC Labs.

Vertigo Research Group

August 22, 2008 - UNICAMP



Context

o Democratization of digital images

- Consumers
 - Multimedia PC at home
 - Digital cameras, mobile phones, digital recorders, ...
- Professionals (audiovisual)
 - More perennial storage
 - Access more easy
 - CNN: 24 hours video storage
 - INA : 240 000h of digitalized videos since 70's, 800 000h in 2015

o Development of networks

- Internet
 - New applications: User-Generated-Content websites (YouTube, etc)
- HDTV



A huge
volume of
images and
videos





Solution: indexing by text

- How to index with text: keywords, conceptual graphs, ...



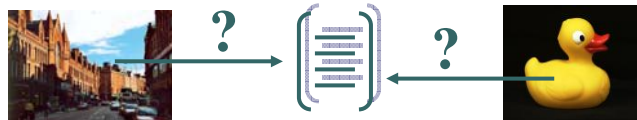
Keywords: sunflower, sun, south of France, ...

- Most classical approach
- Drawbacks
 - Language dependent
 - Possible ambiguity
 - Subjectiveness
 - Context/application dependent
 - Manual annotation expensive



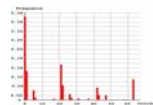
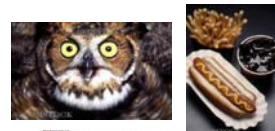
Alternative: indexing by visual content

- Automatic extraction of **descriptors** that will be used to search in database or to database structuring



- Representation of the visual content of images/objects (for an application)

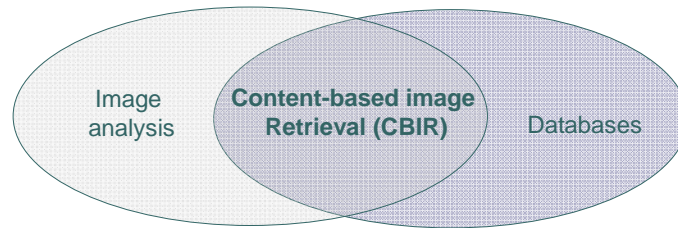
... but mind the **semantic gap**





Solution: indexing by visual content

- CBIR: domain at intersection of two domains of Computer Science



- Variants: CBVR (Content-Based Video Retrieval), CBCD (Content-Based Copy Detection), etc.



Some applications of CBIR

- Scientific applications
 - Medical images analysis
 - Ex: Finding images of pathological nature, for educational or diagnosis goal
 - Satellite images databases
 - Ex: Finding particular fields near rivers / watching over the evolution of fields
- Audiovisual
 - Copy detection for rights management
 - Automatic annotation of videos
- Authentication / Surveillance
 - Biometry: face, fingerprint, iris detection/recognition
 - Police investigations
 - Surveillance of areas, of traffic, ...
- Web
 - E-business (www.like.com)
 - Structuring of UGC websites



Outline

- Introduction to *local* visual descriptors
 - A typology of visual descriptors
 - Focus on *local descriptors*
 - Advantages / Drawbacks
 - Some examples
 - Recent improvements with local features
- Improving local description with global features
 - A synergy between heterogeneous visual descriptors
 - Application to video surveillance of truck traffic
- Dynamical behavior of local descriptors in video sequences
 - Motivation: video copy detection
 - Presentation of the ViCopT system



A typology of visual descriptors

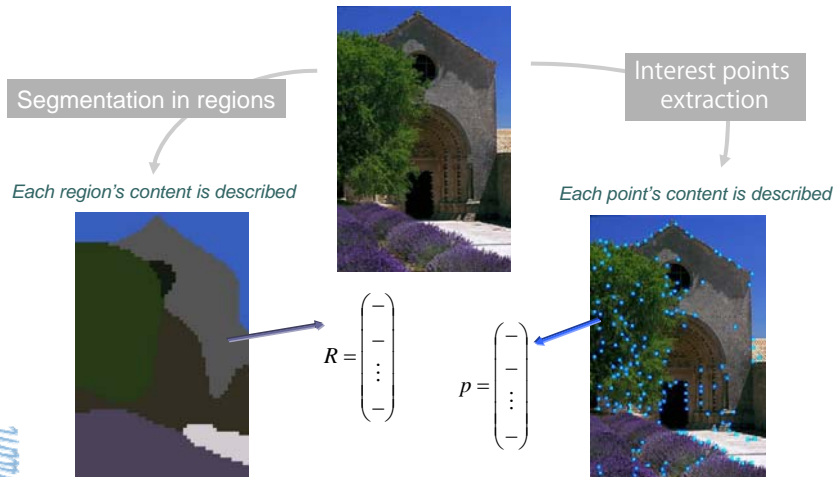
- **Global** description of the image
 - *Approximate* representation of the content
 - Solutions
 - Color, Texture and Shape



Focus on local descriptors

Principles

- Local description of interesting parts of the image



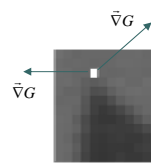
9

Focus on local descriptors

Extraction

- What is an interest point?

- Definition (Moravec): it is a site in the image where intensity varies a lot locally in several directions



- Background

- Interest points are very popular in Robotics and Computer Vision since 60's
 - Robot localization in the scene, Camera calibration, 3D reconstruction, etc
- Popular in CBIR since middle of 90's
 - Queries on image parts, Object recognition, etc
- Huge literature on point detectors
 - Moravec (1977)
 - Beaudet (1978)
 - Kitchen et Rosenfeld (1982)
 - Harris et Stephens (1988), Precise Harris (1996)
 - Deniche et Faugeras (1990)
 - Heltger (1992)
 - Förstner (1994)
 - Susan (Smith et Brady, 1997)
 - SIFT (1999, 2004)
 - ...

10



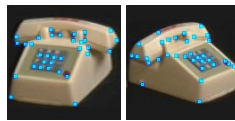
Focus on local descriptors

Extraction

- o Advantages of interest points
 - Visual attention is more caught by such sites (= sites where intensity varies)



- Interest points are repeatable through sequences of images



- No image segmentation required (can be hard)
- Useful for estimating transformations between images
 - Translation, rotation, scale, homography, etc



Focus on local descriptors

Description

- o Local description around the point
 - Principles

$$p = \begin{pmatrix} - \\ - \\ \dots \\ - \end{pmatrix}$$

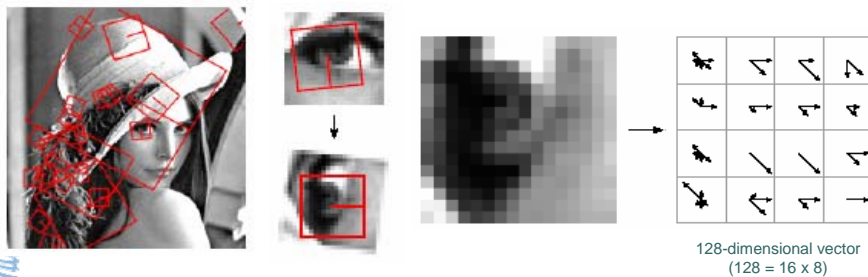
Feature vector
around the point



Focus on local descriptors

Description

- Local description around the point
 - Huge literature on point description
 - Distribution of gray level / color locally around the point (correlation, local jet, etc)
 - Texture description locally around the point
 - ...
 - Ex: SIFT (1999, 2004) → histogram of gradient orientations around the point

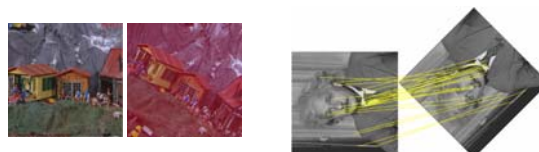
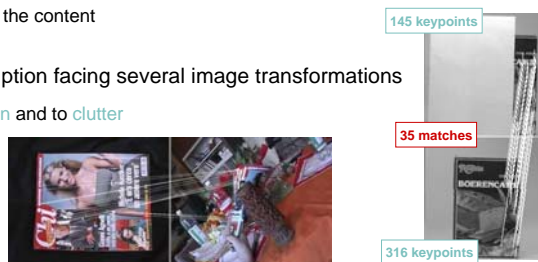


13

Focus on local descriptors

Description

- Advantages of local description
 - Distinctiveness of the photometric variability around the point
 - Relevant description of the content
 - Robustness of the description facing several image transformations
 - Robustness to occlusion and to clutter
 - Invariance or robustness to translation, rotation, scale, illuminations transformations



14

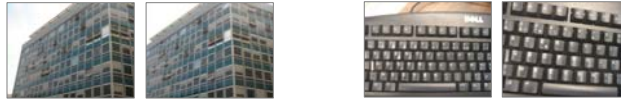


Focus on local descriptors

Drawbacks

o Drawbacks of local features

- Local description: possible ambiguities



- Low-level description: few semantics

- Does not take more global information into account



- Description expensive in storage and in CPU time during matching

- Ex: 100 000 images described by 500 points whose descriptors are 128-dimensional
→ 50 000 000 of points in a 128-dimensional feature space
- Curse of Dimensionality: requires the use of multidimensional index structures to perform NN search efficiently

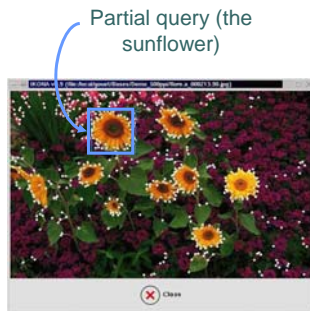


15

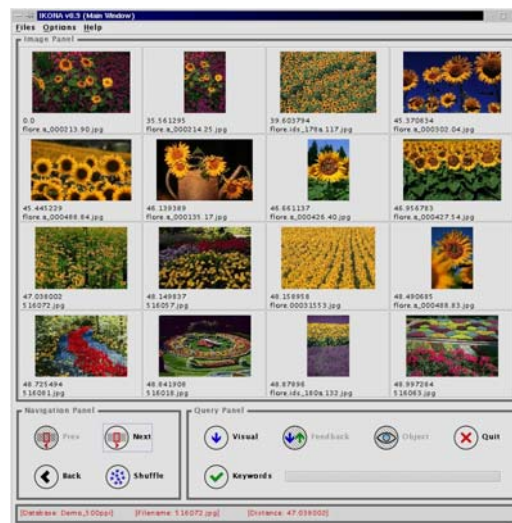


Focus on local descriptors

Some examples



Query by example



<http://www-rocq.inria.fr/imedia>

16



Focus on local descriptors

Some examples

Partial query
(the logo)



Query by example



<http://www-rocq.inria.fr/imedia>

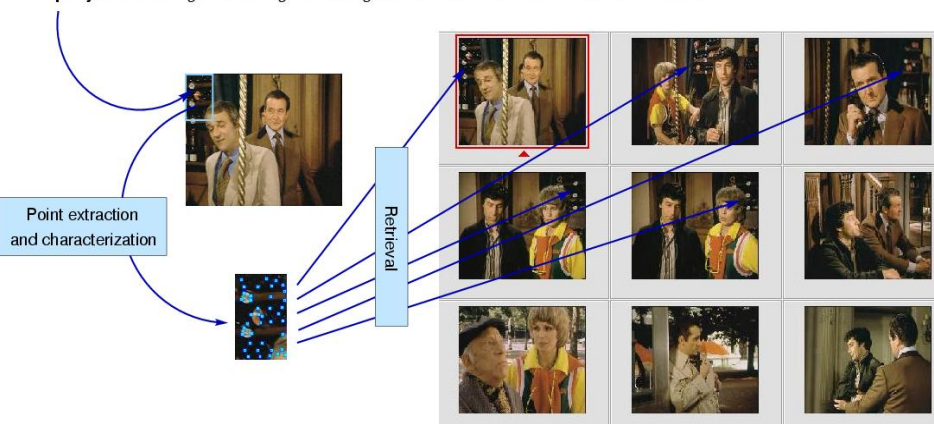
17

Focus on local descriptors

Some examples

- Aids to Police investigation
 - Contract with French Judicial Police (2000-2001) – European program « STOP »

The query : "I am looking for the images involving the room where there is this wine storeroom".



<http://www-rocq.inria.fr/imedia>

18



Focus on local descriptors

Some examples

- o Automatic logo detection
 - RIAM project MediaWorks with French TV channel TF1 (2001-2004)



-

- Catalogue



<http://www-rocq.inria.fr/imedia>

19



Focus on local descriptors

Some examples

- o Other examples of points extraction
 - Images from ACI project BIOTIM (2003-2006)





Recent improvements with local features

1. Content-based **Video** Indexing very popular since 3-4 years
 - Video databases are more easy to get
 - Democratization of video (on mobile, on internet, etc)
 - A lot of applications
 - Video surveillance, video copy detection, etc
 - A lot of images in a video → a lot of visual information [Sivic 03, Grabner 05, Law-To 06]
 - Videos show moving objects with visual appearance varying from one frame to another → a lot of training data!
 - Visual features (interest points) are easy to track in a video sequence
 - Motion is very informative

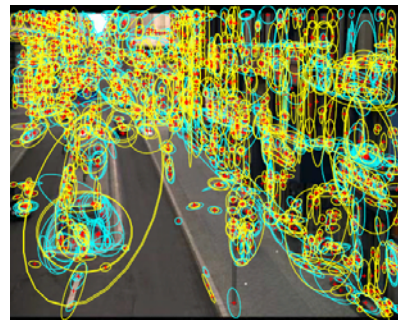


Recent improvements with local features

2. **Combination** of points of different natures [Sivic 03, Jurie 04, Opelt 05, Law-To 06]
 - Ex: Texture patches, Homogeneous regions, Local symmetries, etc.
 - Complementarity of points to better describe the image content



Symmetry points (x) and Harris points (+) (© Law-To thesis, 2007)



Maximally Stable regions in yellow / Shape Adapted regions in cyan (© Sivic and Zisserman, 2003)



Presentation of two works

- Improving local description with global features
 - Work done at CNAM (Bruno Lameyre's PhD thesis, 2005-2008) in collaboration with a French company specialized in video surveillance
 - Model-free object recognition in videos
 - Application to video surveillance of truck traffic
- Dynamical behavior of interest points in video sequences
 - Work done with INRIA (Julien Law-To's PhD thesis, 2004-2007) in collaboration with INA (Institut National de l'Audiovisuel)
 - Application to content-based video copy detection: ViCopT



23



Improving local description with global features

Motivation

- Global features
 - Global description of the object content
 - Solutions
 - *Active contours + shape descriptors*
 - *Region segmentation + region descriptors (color, shap, texture, etc)*
 - *Specific descriptors from object model (top-down)*
 - ...
- Advantages
 - High-level description of the object appearance
 - Easy to exploit (1 vector / object)
- Main Inconvenient
 - Requires a pre-processing of the image to extract the object from background
 - Image segmentation? Hard in general (specific applications)
 - Object detection? Requires a prior knowledge on the object model



24



Improving local description with global features

Principles of our approach

o Our objective

- Combination of **heterogeneous** visual features
 - Local descriptors for their **robustness** to image transformations, to clutter and to occlusions
 - Global descriptors for their **richness**, but without doing a pre-processing of the image (segmentation) to isolate the object

o How?

- Construction and structuring of a catalogue of these features
 1. Computation of the two (local and global) features spaces **separately**
 2. Definition of **connexions** between these spaces
 - *Many-to-many relationships*
 - *Semantic connexions: a given local pattern is not associated to every kinds of shapes (and vice versa)*
 - Human eyes and mouth does not appear inside a duck's shape.

CNAM

25



Improving local description with global features

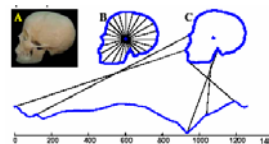
Choice of a visual global descriptor

o Global descriptor employed: **shape of the object**

- Detection implementation: Discrete active contours (*snakes*) [Kass 88]



- Shape descriptor implementation: Distance centroid method + DFT [Zhang 01]



- Why snakes?

- High-level description of the object shape
- Complementarity with local features that describe the object's inside
- Snakes enable to localize the object precisely after recognition
- With videos, snakes can help during points tracking and vice versa [Lameyre 04]



CNAM



Improving local description with global features

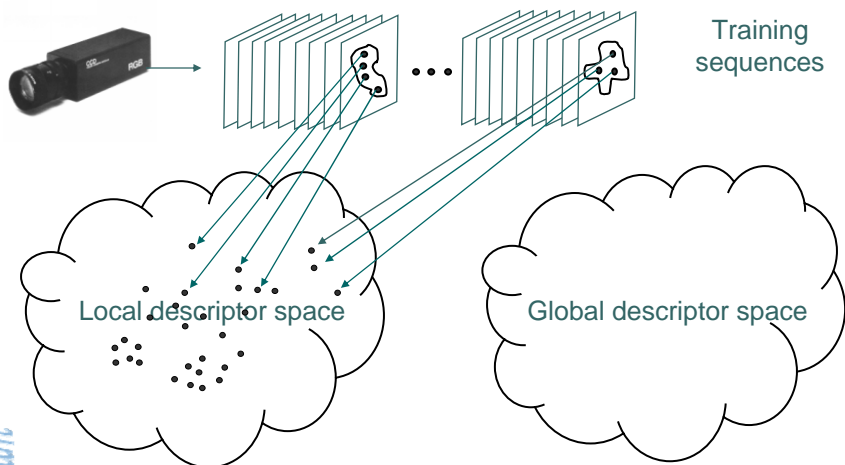
Principles of our approach

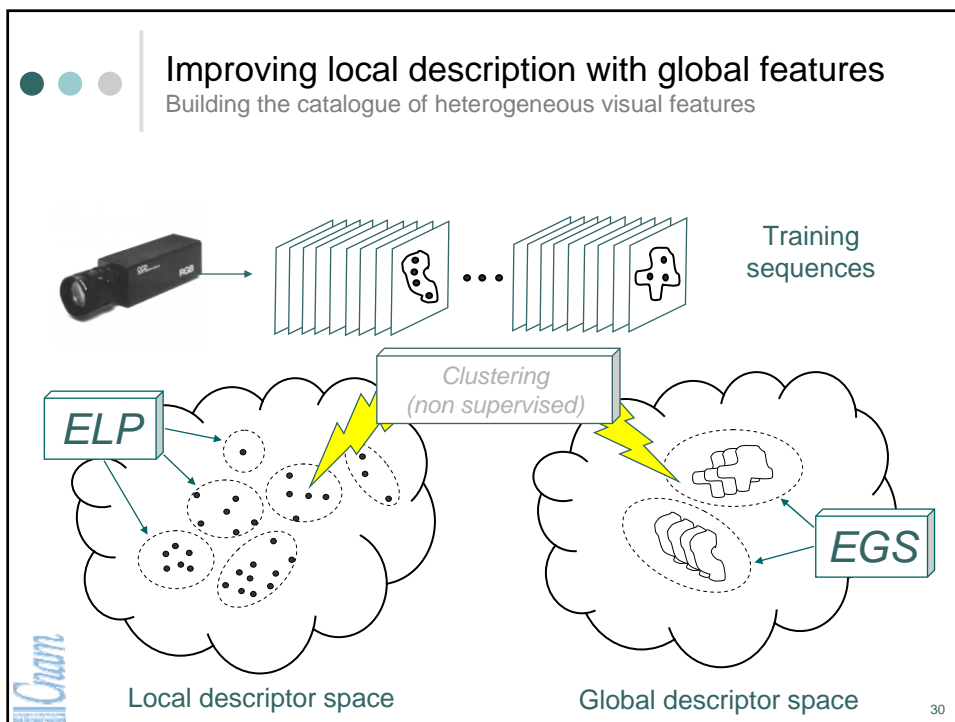
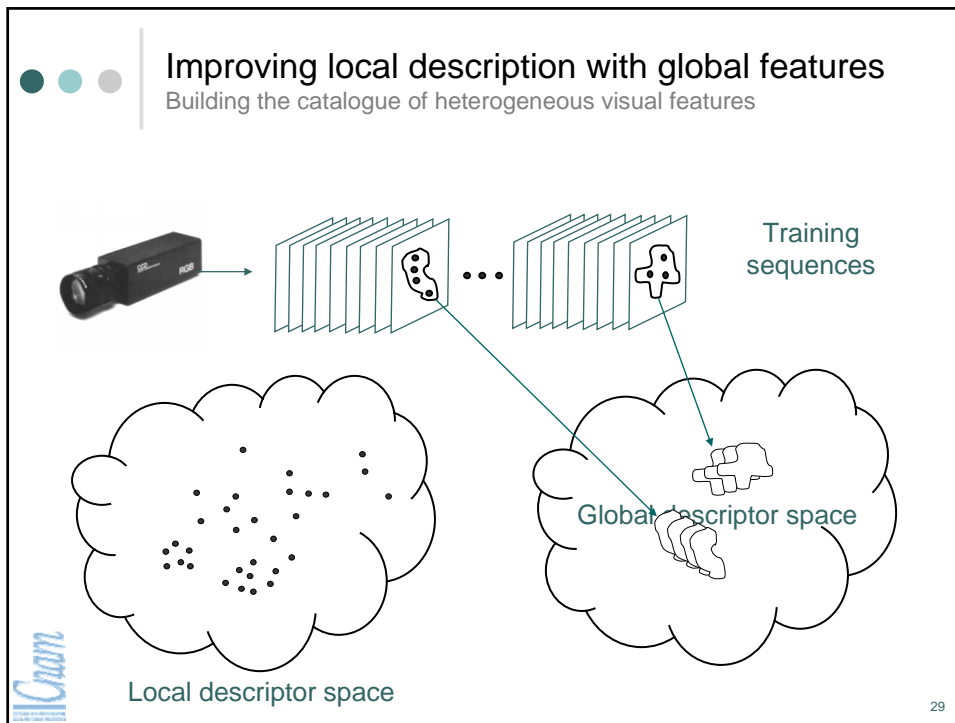
- General algorithm of recognition in two steps
 - Step #1: Local descriptors as **primary source of pruning** (anchors)
 - If present, anchors give a first approximation of its location
 - Step #2: **Refinement** of recognition and localization with global features
 - Anchors are connected to global features: they can be seen as **indexes** for the global descriptors



Improving local description with global features

Building the catalogue of heterogeneous visual features





Improving local description with global features

Building the catalogue of heterogeneous visual features

Training sequences

Feature spaces structuring by clustering

- Reduction of the spatial and temporal redundancy of the visual features
- Makes available dynamic improvement of the catalogue as recognition proceeds in new sequences

Local descriptor space

Global descriptor space

31

Improving local description with global features

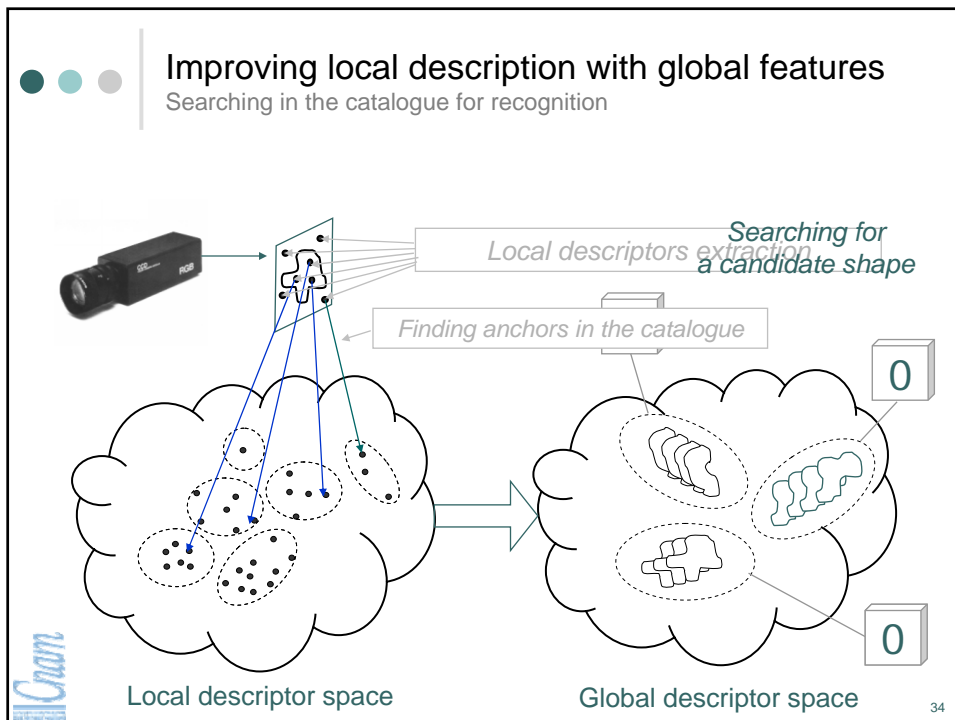
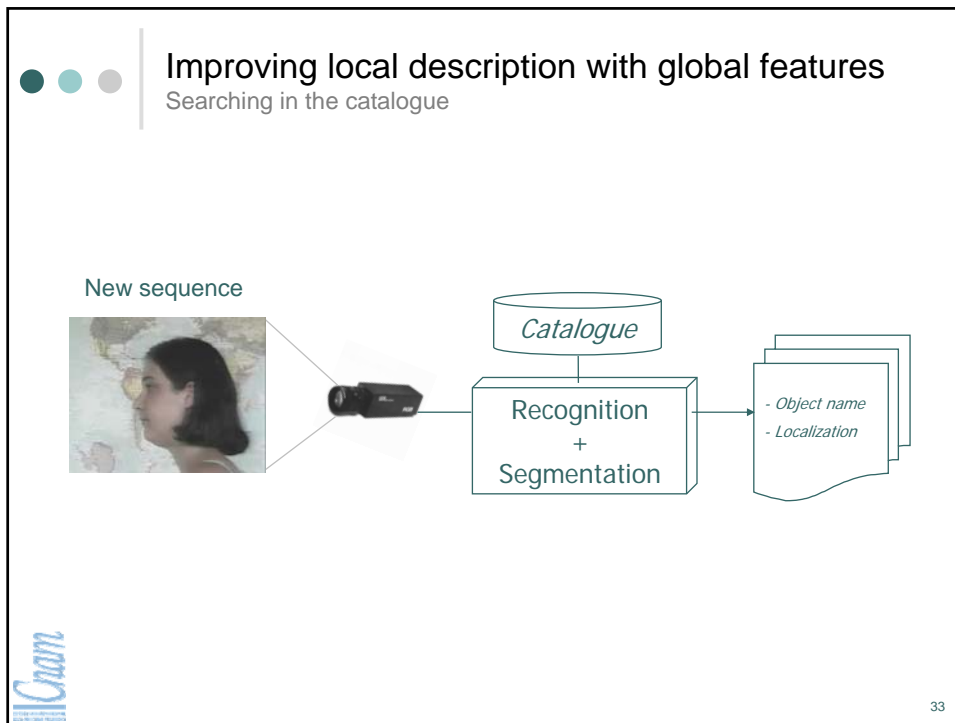
Building the catalogue of heterogeneous visual features

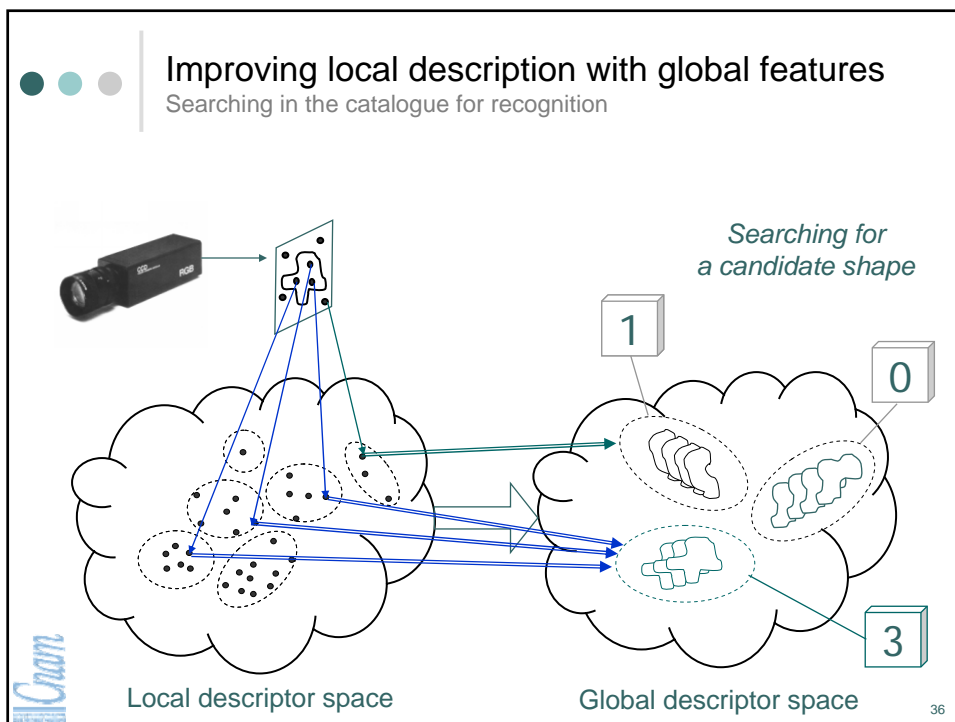
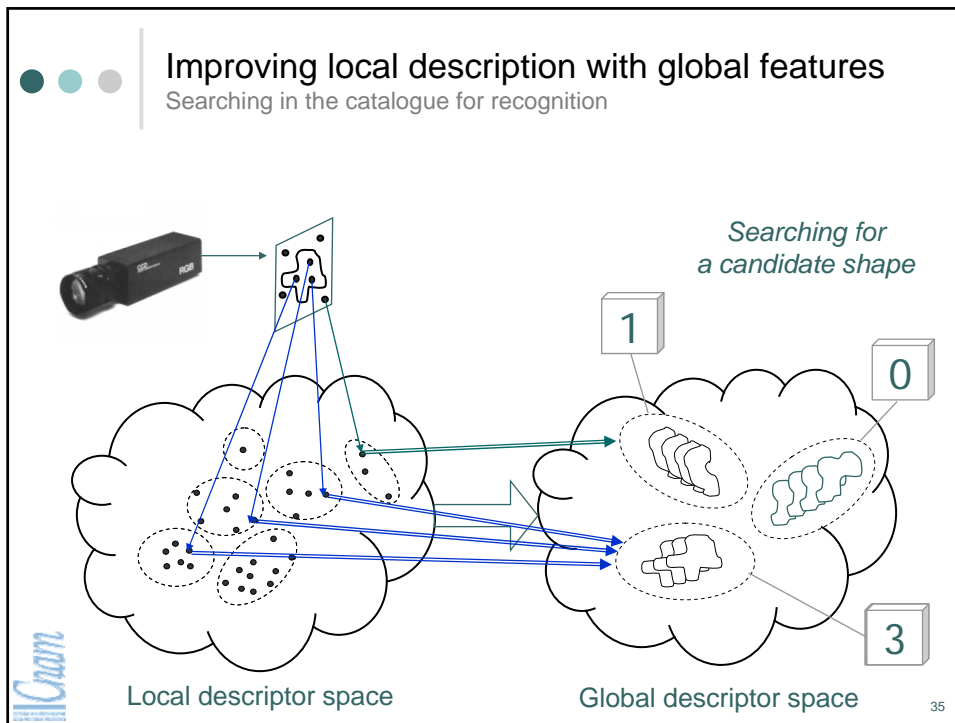
Training sequences

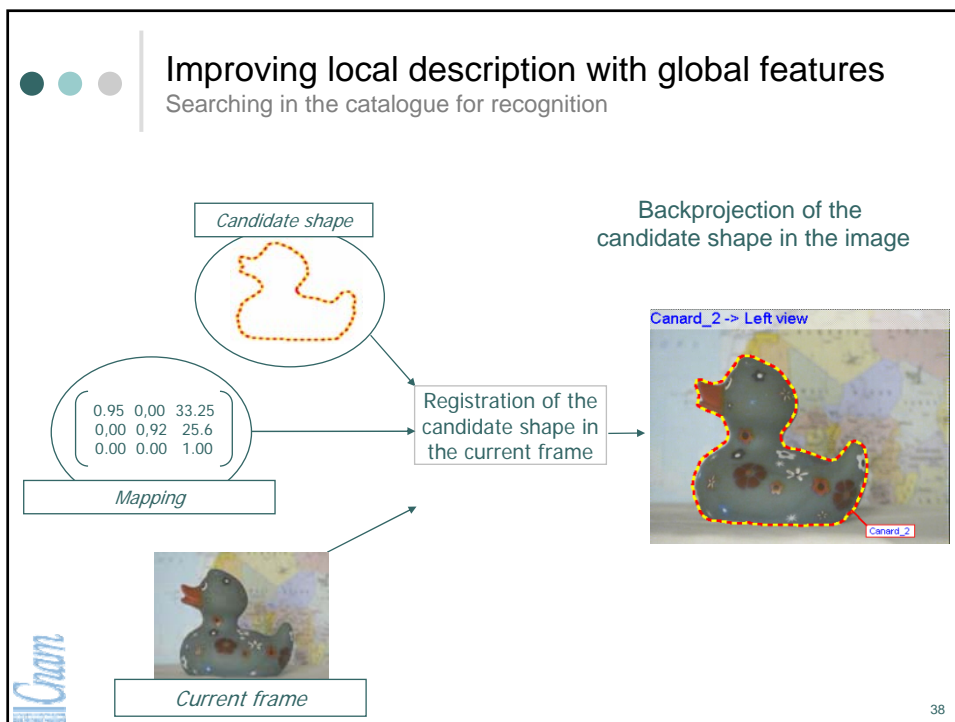
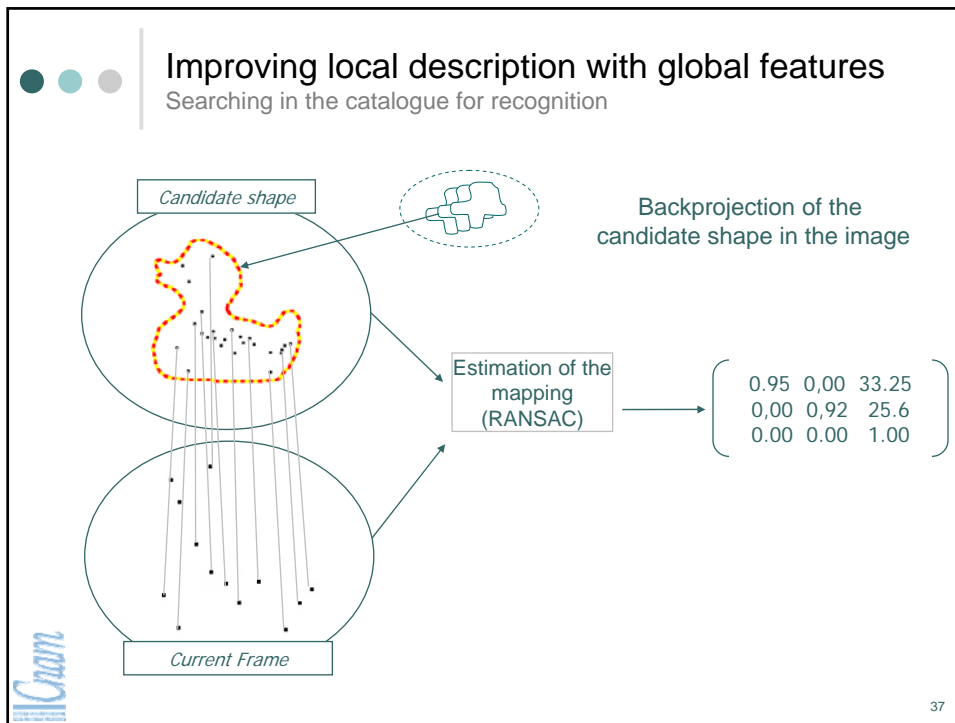
Local descriptor space

Global descriptor space

32

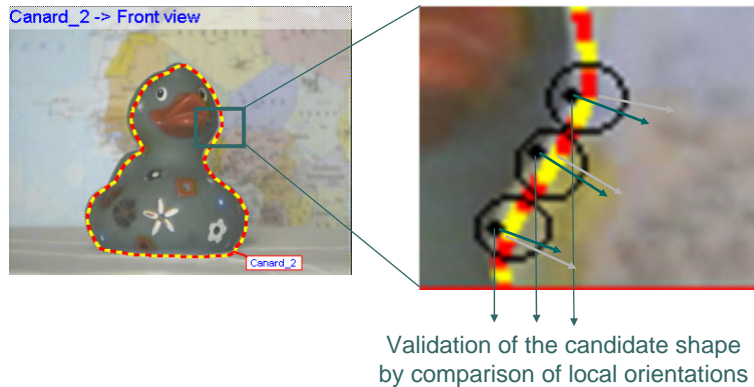






Improving local description with global features

Searching in the catalogue for recognition



Improving local description with global features

Principles of our approach

- o Structuring of the visual features to perform **real-time recognition**
 - Use of multidimensional index structure to accelerate retrieval of interest points (PhD thesis Nouha Bouteldja, CNAM 2004-2008)
 - Multiple queries
 - *Processing sets of query points jointly and not in sequence [Bouteldja et al. 2006]*
 - HiPeR: a hierarchical model for accelerating retrieval in high-dimensional metric spaces
 - *Exact and approximate retrieval of nearest neighbors [Bouteldja et al. 2008]*



Improving local description with global features

Evaluation of the approach

o Framework of the evaluation

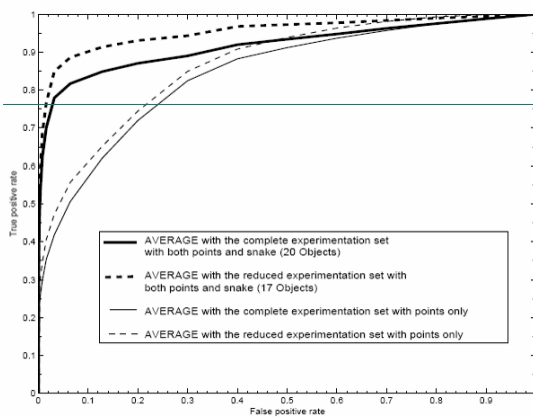
- 20 objects with different visual appearance (toys, faces, boxes, etc)
- Reference technique to compare: local descriptors alone
- Criteria: ROC curves (Receiver Operating Characteristic curves)
 - Ratio of False Positives / Ratio of True Positives, according to a parameter
- Scenarios
 - *Still-to-video: 1 image / video sequences*
 - *Video-to-video: 1 video / video sequences*



Improving local description with global features

Evaluation of the approach

o Recognition with both local and global features



✓ Results globally better

✓ Curves shifted to the left:
for a given TP rate, the FP
rate is greatly reduced

→ Good for surveillance
purposes

Improving local description with global features

Examples of recognition

- Examples of recognition and precise localization
 - Multiple-object recognition
 - Occlusions
 - Moving camera
 - Different levels of recognition possible

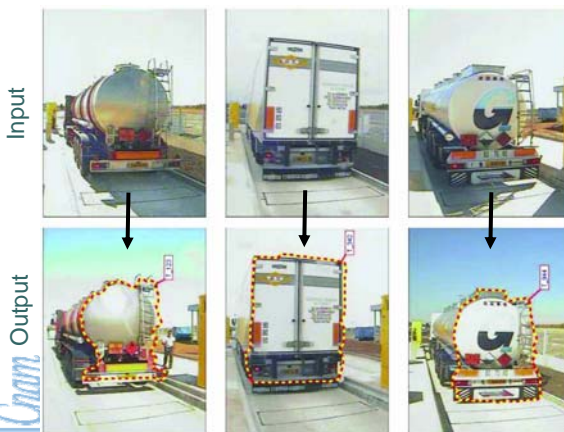


43

Improving local description with global features

Application: surveillance of truck traffic

- Industrial contract with a French company
 - Surveillance of truck traffic on secured parking areas of motorways, to increase the security of trucks and their trailers in the parking area



First objective

Detection of trailers swapping in the parking area, observable when a truck leaves the area:

- with a trailer different from the one it had when entering,
- without trailer.

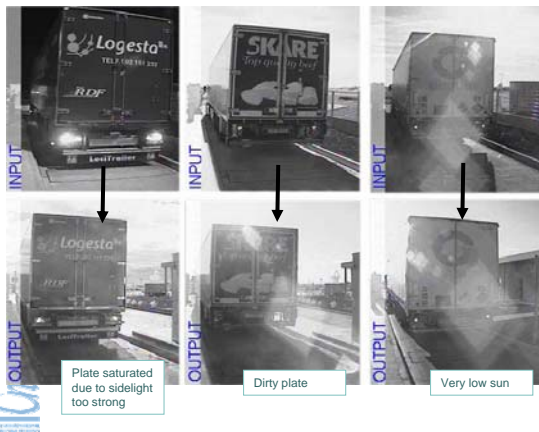
Such events may indicate that a trailer is going to be stolen, or that a trailer has been warehoused in the parking area during a period that does not correspond (longer) to the arrival and leaving of the truck.

44

Improving local description with global features

Application: surveillance of truck traffic

- o Industrial contract with a French company
 - Surveillance of truck traffic on secured parking areas of motorways, to increase the security of trucks and their trailers in the parking area



Second objective

Help when license plate recognition fails.

In 30% of the cases, license plate detection on front and/or back views failed. In such a case, the system switches to the visual recognition mechanism. With it, 70% of the missed trucks are recognized (21% of all the filmed vehicles). In other words, **the complete approach allows reaching a percentage of detections of 91%, while it was 70% with license plates detection only.**

45

Improving local description with global features

Publications

- o For details, see publication to journal CVIU 2008

Object recognition and segmentation in videos by connecting heterogeneous visual features, Valerie Gouet-Brunet and Bruno Lameyre, Computer Vision and Image Understanding, February 2008.

46

INRIA
ina

Dynamical behavior of interest points

Motivation

- The challenge of content-based video copy detection (CBCD)
- Motivated by
 - Large diffusion of audiovisual content
 - TV channels
 - 24 hours video stream
 - Internet
 - UGC websites: videoGoogle, YouTube, MySpace, etc
 - Web TV, Video Blogs
 - Banks of multimedia contents
 - Example of INA: 240 000h of digital videos, 800 000h in 2015
 - Tracability of large video databases
 - Piracy
 - Statistical informations

47

Dynamical behavior of interest points

Definition of copy

- What is a copy?
 - The two videos are made from the same video source
- Difference between CBCD and watermarking

Content Based Copy Detection

Watermarking

© A. Joly et al. (2007)

48

Dynamical behavior of interest points

Definition of *copy*

Difficulty #1: transformations can occur

- Post production (TV)
 - Insert, crop, shift, gamma, brightness, etc.
- Involuntary (web)
 - Noise, gamma, color, encoding, etc.



Dynamical behavior of interest points

Definition of *copy*

Difficulty #2: copy detection is not finding near duplicate



Very similar videos but not copies



Very different contents but copies

Dynamical behavior of interest points

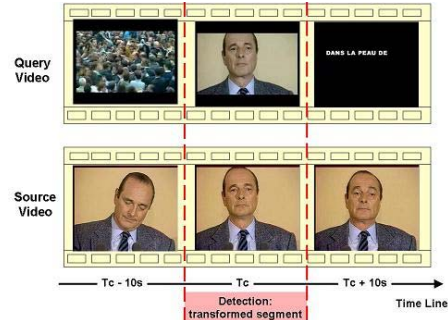
Definition of *copy*

- Difficulty #3: two kinds of scenarios

The query is a single video



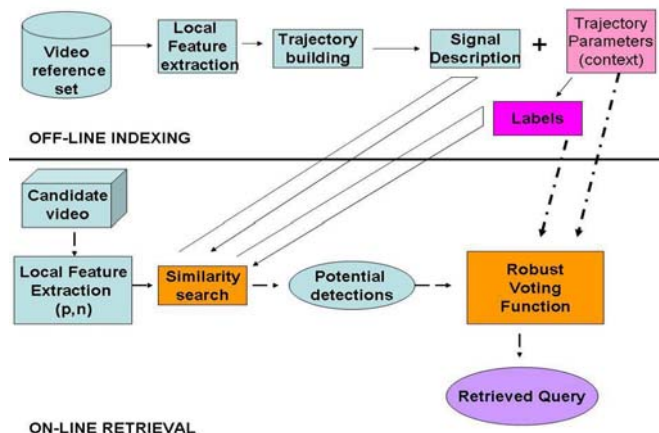
The query is a video stream



Dynamical behavior of interest points

Principles of the approach

- The proposed approach: ViCopT (Video Copy Tracking)





Dynamical behavior of interest points

Description of the content

- A **low-level** description of the content
 - Implemented technique
 1. Extraction of interest points in all frames, combination of several natures of points (Harris + Symmetry)
 2. **Tracking** of such points in the sequence (KLT algorithm)
 3. Point description (with Normalized Local Jet: invariance to translation and illumination) **averaged over the trajectory**
 4. For each point, **description of its trajectory**: 3D box + statistical properties
 - Justifications
 - Why a low-level (bottom-up) description?
 - *To be independent of the application: you can't re-index 800 000 hours of video when you change of application!*
 - Why tracking points?
 - *To reduce **redundancy** of the description along the sequence*
 - *To obtain more **robust** points, i.e. points that survive along frames*
 - *To characterize their **kinematic behavior***



Dynamical behavior of interest points

Description of the content

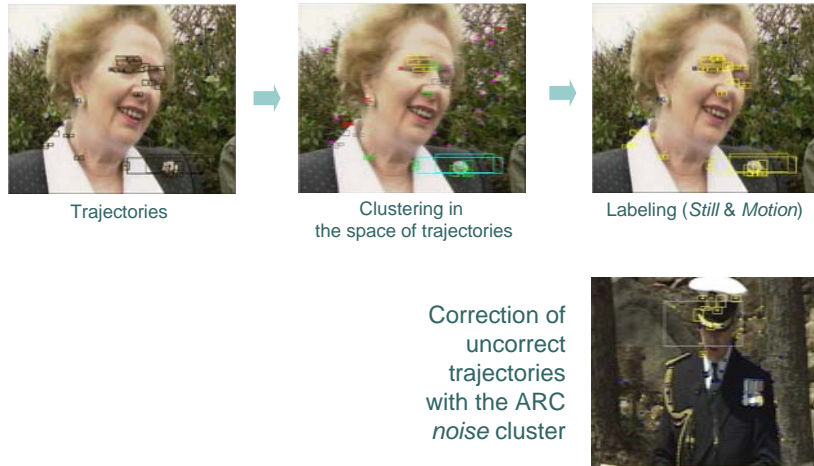
- A **high-level** description of the content: **labels of behavior**
 - Labels of behavior = a more high-level description of the temporal behavior of points along the sequence (**kinematic context**)
 - Set depending on the trajectory parameters
 - Determined by an unsupervised clustering process in the space of trajectory parameters (ARC = Adaptative Robust Competition, [Le Saux et al. 2001])
 - Analyse of the clusters to determine labels
 - Outcome
 - Reduction of the number of features
 - Selection of a priori interesting features for a given application
 - Labels chosen for copy detection purposes
 - *Still* for robustness along the frames (background if motionless camera)
 - *Motion* for perceptual saliency, distinctiveness



Dynamical behavior of interest points

Description of the content

- o A high-level description of the content: labels of behavior



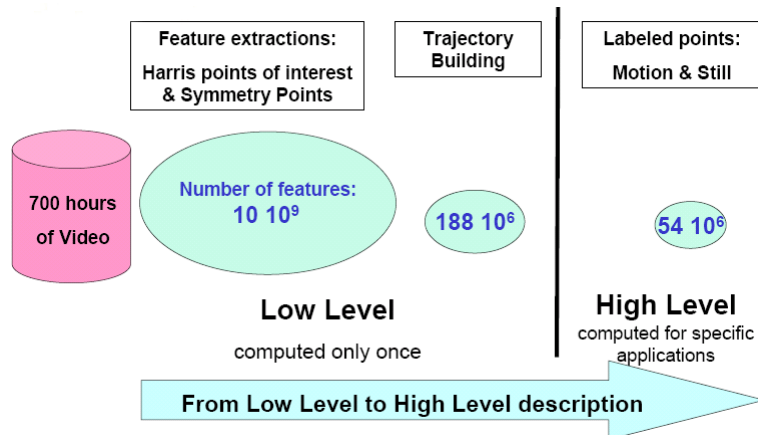
55



Dynamical behavior of interest points

Description of the content

- o A high-level description of the content: labels of behavior



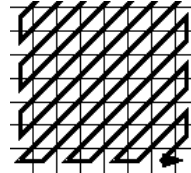
56



Dynamical behavior of interest points

Description of the content

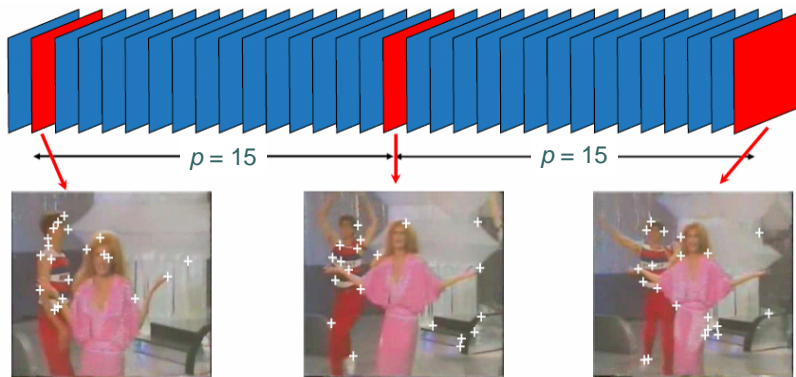
- Structuring of the visual features to perform **real-time detection**
 - 700 hours of indexed videos: 54,000,000 of features in a 20-dimensional space !!!
 - Literature of databases is very abundant on multidimensional index structures
 - Our solution:
 - Linearization of the feature space with spaces filling curves (Z order)
 - Approximate search: compromise between precision and speed



Dynamical behavior of interest points

On-line copy retrieval

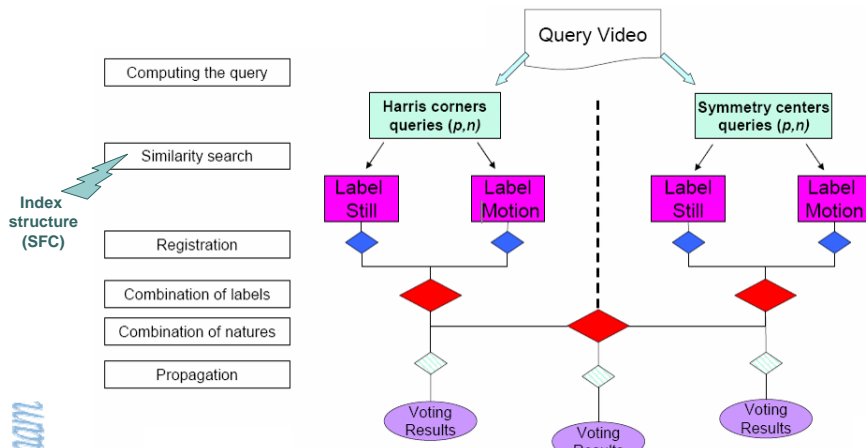
- On-line retrieval: **computation of the query**
 - Asymmetrical technique
 - Precision p chosen on-line



Dynamical behavior of interest points

On-line copy retrieval

- On-line retrieval: a robust voting function

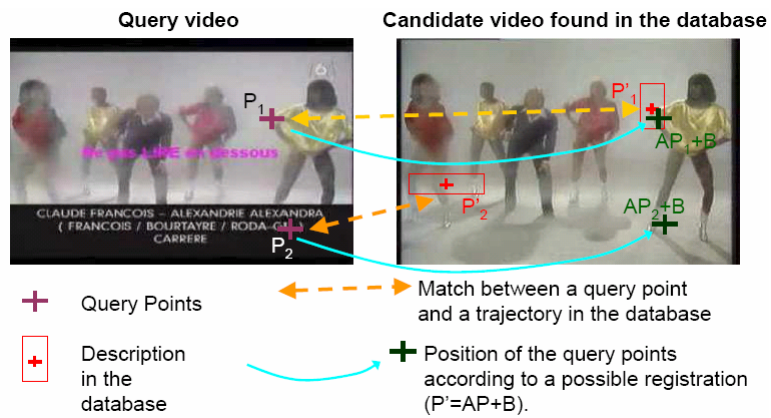


59

Dynamical behavior of interest points

On-line copy retrieval

- On-line retrieval: the spatio-temporal registration step



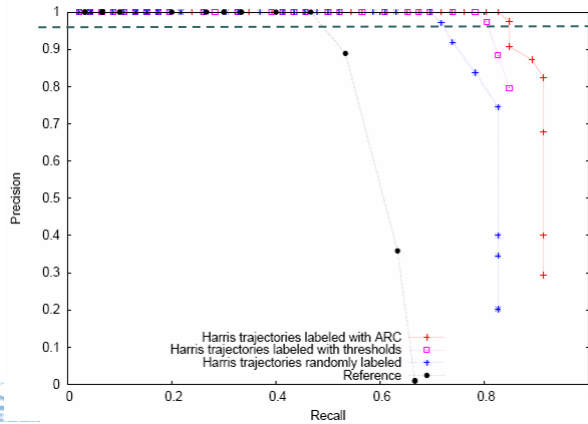
60

Dynamical behavior of interest points

Evaluation

○ Evaluation #1: Precision and recall **by segment**

- Comparison with technique [Joly et a. 2006] that exploits interest points without trajectories and labels of behavior
- 1000 hours of video



At precision 97%, recall is 85% with ARC clustering, 80% with thresholds, 71% with random labeling and 50% for the reference technique.

ViCopT is real time with (constant delay) detection on large collections (1000 hours)

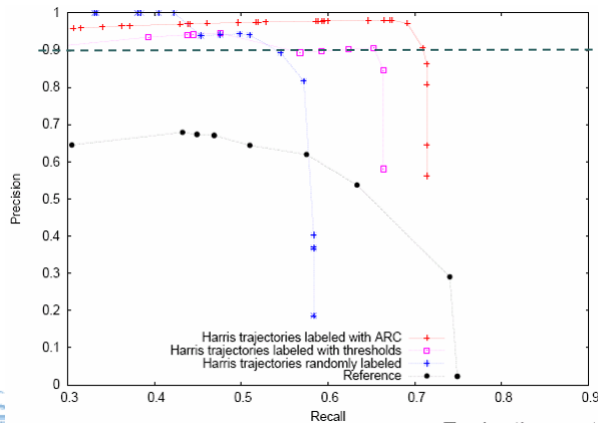
61

Dynamical behavior of interest points

Evaluation

○ Evaluation #2: Precision and recall **by frame**

- Comparison with technique [Joly et a. 2006] that exploits interest points without trajectories and labels of behavior
- 1000 hours of video



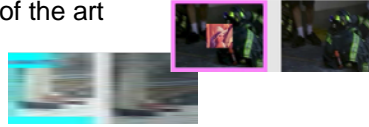
At precision 90%, recall is 71% with ARC, 67% with thresholds, 53% with random labeling.

62

Dynamical behavior of interest points

Evaluation

- Evaluation #3: comparison with state of the art
 - European NoE Muscle (2004-2007)
 - 3 hours of video / Mixed transformations

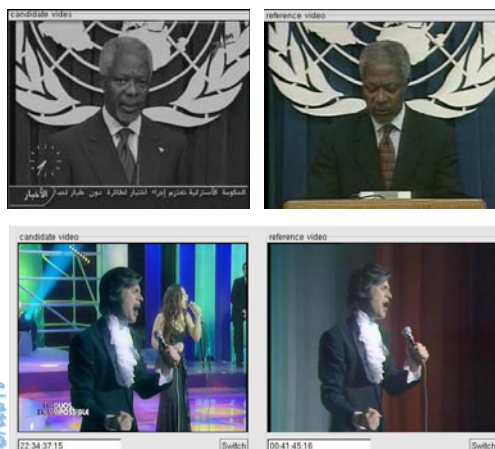


Technique	Average precision
ViCopT	0.86
AJ_SpatioTemp	0.79
AJ_Temp	0.68
Temporal Ordinal Meas.	0.65
STIP	0.55
Temporal	0.51
Ordinal Meas.	0.36
Color histograms	0.24

Dynamical behavior of interest points

Some results

- Examples of detection: False alarms removal



Not a match:
only points with label *Still*
are matched (motions are
different)

Copy detected:
points with label *Motion*
are matched

Dynamical behavior of interest points

Some results

- Other results and other applications



Classification / linkage of videos: here, the background is relevant (labels Still)


65

Dynamical behavior of interest points

Publications

- For details, see publications
 - Local Behaviours Labelling for Content Based Video Copy Detection*, J. Law-To, V. Gouet-Brunet, O. Buisson and N. Boujemaa, ICPR 2006
 - Labeling complementary local descriptors behavior for video copy detection*, J. Law-To, V. Gouet-Brunet, O. Buisson and N. Boujemaa, Int. Workshop MRCS 2006
 - Robust voting algorithm based on labels of behavior for video copy detection*, J. Law-To, O. Buisson, V. Gouet-Brunet and N. Boujemaa, ACM Multimedia 2006
 - Video Copy Detection on the Internet: the challenges of copyright and multiplicity*, J. Law-To, V. Gouet-Brunet, O. Buisson and N. Boujemaa, ICME 2006
 - Video copy detection: a comparative study*, J. Law-To, L. Chen, A. Joly, Y. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa and F. Stentiford, CIVR 2007


66




Thank you!

<http://cedric.cnam.fr/~gouet>

<http://cedric.cnam.fr/vertigo>



67



68