Improving interactive video retrieval by exploiting automatically-extracted video structural semantics

Vasileios Mezaris, Panagiotis Sidiropoulos, Ioannis Kompatsiaris

Informatics and Telematics Institute / Centre for Research and Technology Hellas

IEEE ICSC 2011, Palo Alto, September 2011

Overview

- Introduction – problem formulation
- Related work
- Video structural semantics in interactive retrieval
- Automatic extraction of video structural semantics
- Experiments and results
- Conclusions

Introduction – problem formulation

- Semantic video retrieval is a key application
- Main challenge: bridge the semantic gap, between the possible video representations that are often:
  - machine-only-readable (e.g. low-level audio-visual features),
  - unreliable and incomplete (e.g. automatic visual concept detection results, user-assigned tags)
  - too specific to be meaningful when seen out of context (e.g. tag "Mary"),

  and the specific and diverse information needs of every possible user

Introduction – problem formulation

- Bridging the semantic gap in video retrieval is attempted by means of:
  - New low-level video features
  - More reliable video concept detectors
  - Event detection from audio-visual data
  - …
  - Better interaction strategies (putting the human in the retrieval loop)
- In this work
  - We focus on interaction strategies
  - Examine the possibility of using information about the structure of the video (video scenes) for guiding the user's interaction

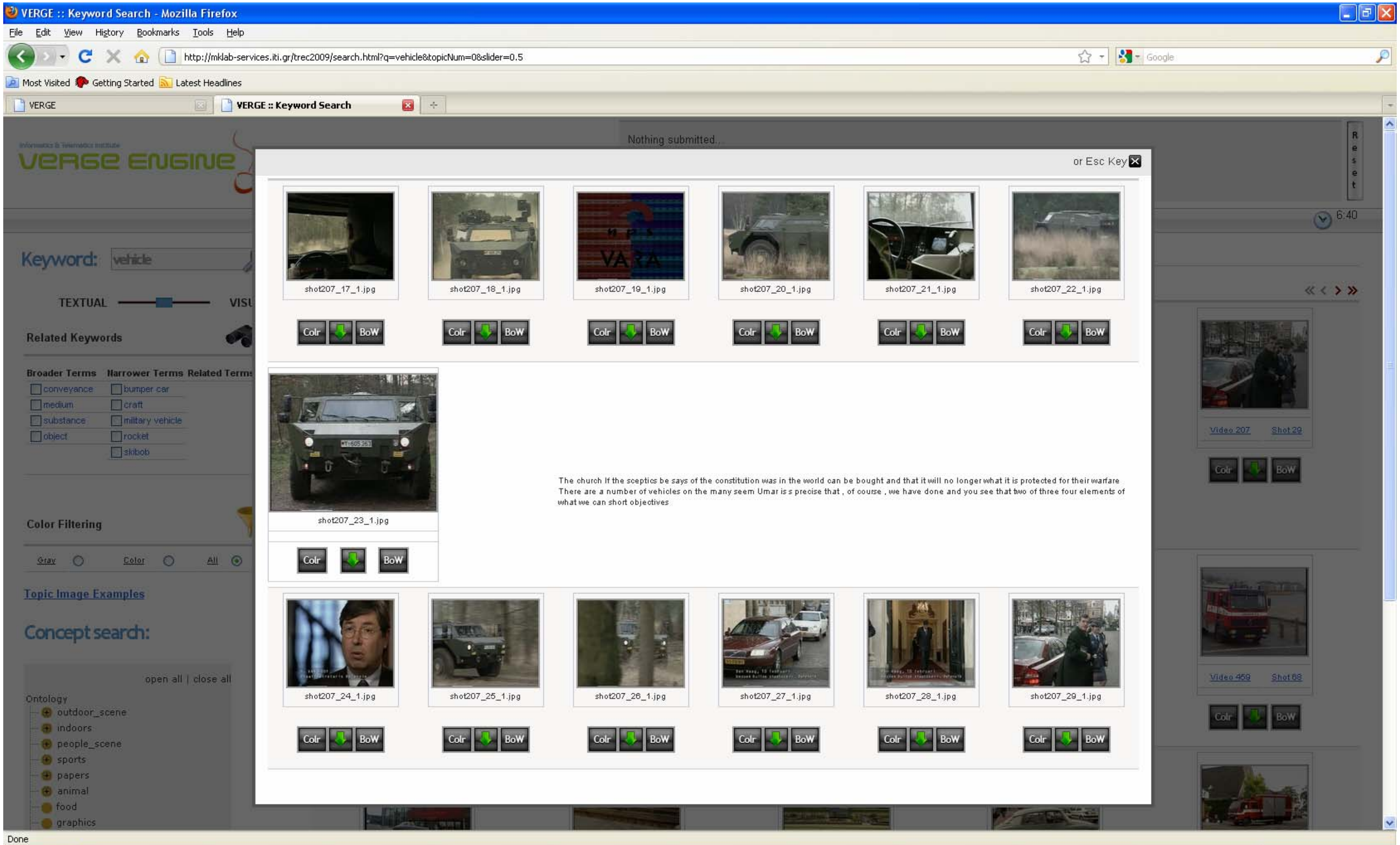
Related work

- Intelligent video retrieval is typically performed at the shot level, due to
  - Significant variability in the video content of an entire program
  - Need of users for retrieving only the bits of information that are of interest to them
- Thus, interactive video retrieval is all about assisting the user in searching and navigating within a large collection of video shots
- State-of-the-art
  - Different query formulations (e.g. query-by-text, query-by-example),
  - query expansion
  - relevance feedback
  - browsers for visualizing the collection or a subset of it according to different criteria (e.g. concept relevance, time), and others.

Related work

- Time information has been shown to be particularly important: issuing basic temporal queries, starting from a (found) positive sample
  - "Time treads" showing a sequential view of all the shots of a video (ForkBrowser and CrossBrowser)
  - Presentation of a fixed number of neighboring shots for each specified shot ("side shots")

Related work

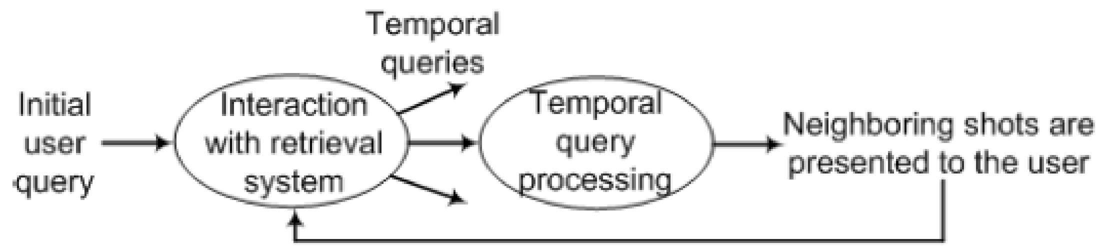


Informatics and Telematics Institute
Centre for Research and Technology Hellas

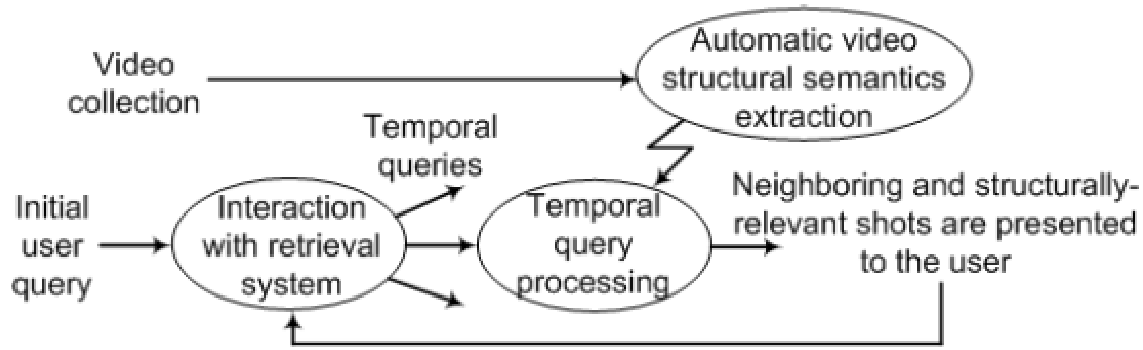
Structural semantics for retrieval

- Although temporal information is significant…
  - i.e, shots that are temporally close to a correctly retrieved shot are intuitively considered very likely to also be relevant to the query
- …its use is not governed by solid rules: basic temporal queries rely on
  - ad hoc rules (e.g. "show N side shots", where N is fixed)
  - no rules at all (e.g. "time threads", which are a sequential view of the entire video, shot-by-shot)
- Our hypothesis
  - Automatically-extracted video structural semantics, i.e. the outcome of algorithms for **video segmentation to scenes**, can intelligently guide the user in visually inspecting a variable number of temporally neighboring shots that are most likely to also satisfy the query criteria

Structural semantics for retrieval



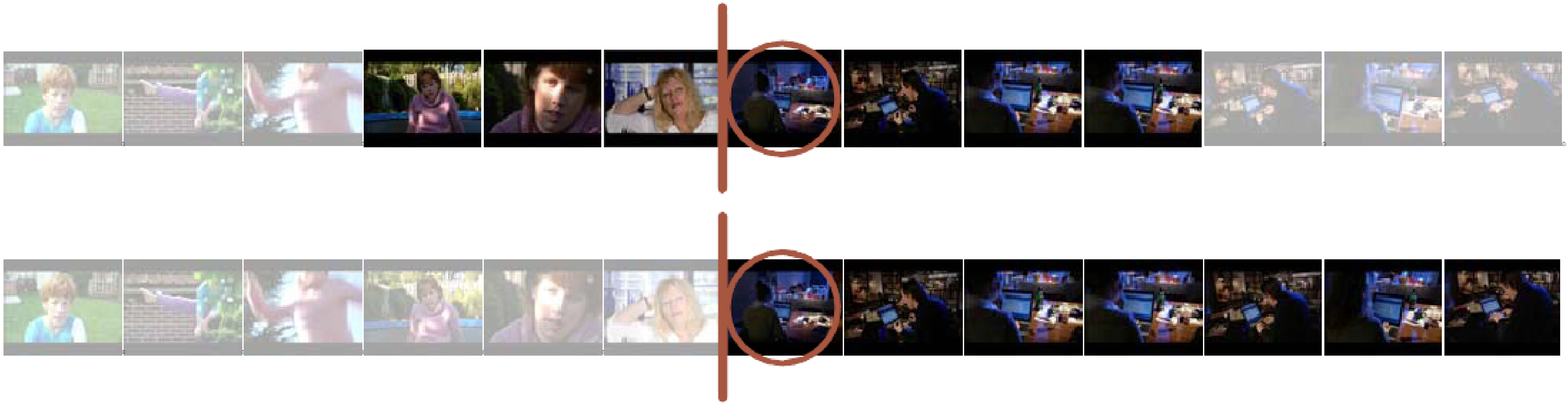
(a)



(b)

Structural semantics for retrieval

- Accepting this hypothesis could be straightforward if perfect scene segmentation results could be used…but this is not the case (assuming it does not exist, e.g. in documentaries, news,…)
  - Manually processing large collections of video is practically infeasible
  - SoA results of automatic techniques still deviate considerably from perfection
- So, the question is: can the results of existing SoA techniques for automatic video segmentation to scenes be useful in interactive retrieval?



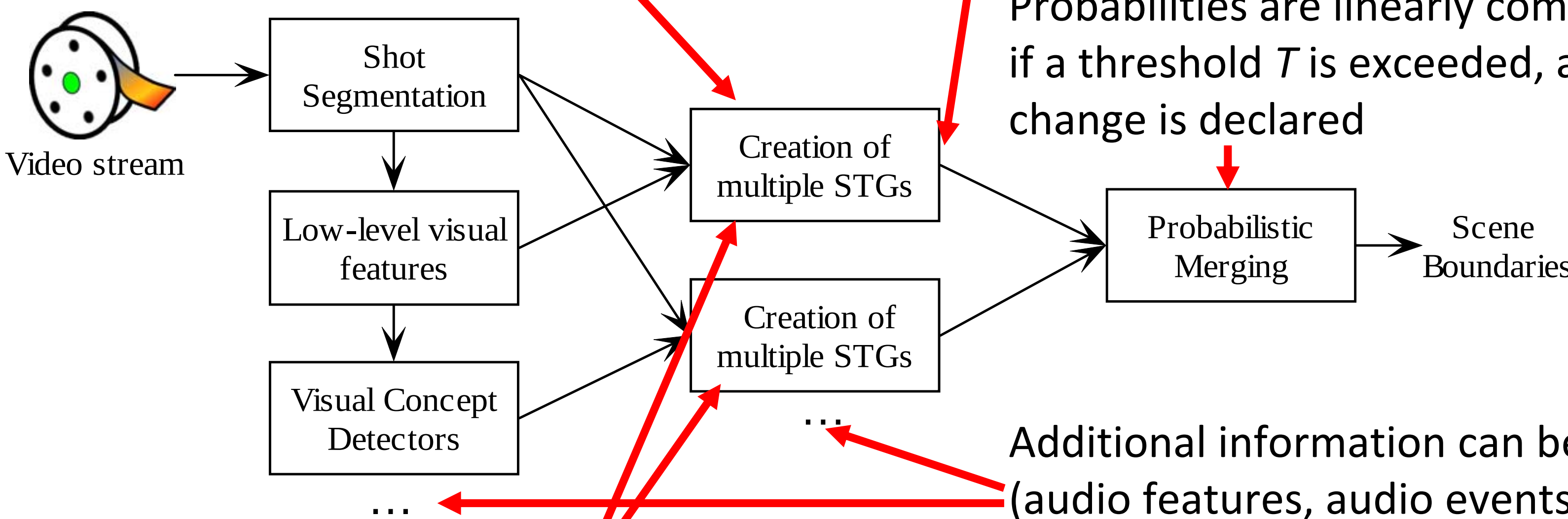
Extraction of structural semantics

- Overview

Each STG is created using a randomly selected parameter set

The probability of a shot boundary also being a scene boundary is calculated by counting the STGs

Probabilities are linearly combined; if a threshold *T* is exceeded, a scene change is declared



Additional information can be used (audio features, audio events), and more types of STGs can be created

Low- and high-level visual information is treated separately

- Advantages
  - Alleviates the need for STG construction parameter fine-tuning
  - Effectively combines heterogeneous information
  - The introduced parameters (*V*, *T*) can be easily optimized

Extraction of structural semantics

- Six variants of this algorithm are used in our experiments
  - Each was evaluated separately

M1 - Using low-level visual features only, optimal parameters
M2 - Using low-level visual features only, parameters favoring over-segmentation
M3 - Using low-level visual features only, parameters favoring under-segmentation
M4 - Combining low-level visual features and concept detector responses, all 101 detectors used
M5 - Combining low-level visual features and concept detector responses, 60 detectors selected according to AP
M6 - Combining low-level visual features and concept detector responses, 50 detectors selected according to ΔAP

Experiments and results

- Dataset

1 - One or more people walking up stairs
2 - A door being opened
3 - A person walking or riding a bicycle
4 - Hands at a keyboard typing or using a mouse
5 - A canal, river, or stream with some of both banks visible
6 - A person talking on a telephone
7 - A street market scene
8 - A street protest or parade
9 - A train in motion
10 - Shots with hills or mountains visible

3 - Bus	13 - People dancing
4 - Chair	14 - Person eating
5 - Cityscape	15 - Person playing musical instrum.
6 - Classroom	16 - Person playing soccer
7 - Demonstration or Protest	17 - Person riding a bicycle
8 - Doorway	18 - Singing
9 - Female human face closeup	19 - Telephone
10 - Hand	20 - Traffic intersection

Experiments and results

- Three types of basic temporal queries (TQ) were evaluated
- TQ were issued for all positive samples of the 20+24 queries
  - 3322 TQ in response to single concept queries
  - 4704 TQ in response to complex queries

(a) Without considering scene boundaries:

query shot $s_i \quad \rightarrow \quad$ show $s_j,\ j \in [i-N,\ i+N],\ N = const$

(b) Based on scene boundary detection (considering a single scene):

query shot $s_i \quad \rightarrow \quad$ show all $s_j \in S_k | s_i \in S_k$

(c) Based on scene boundary detection (considering multiple scenes):

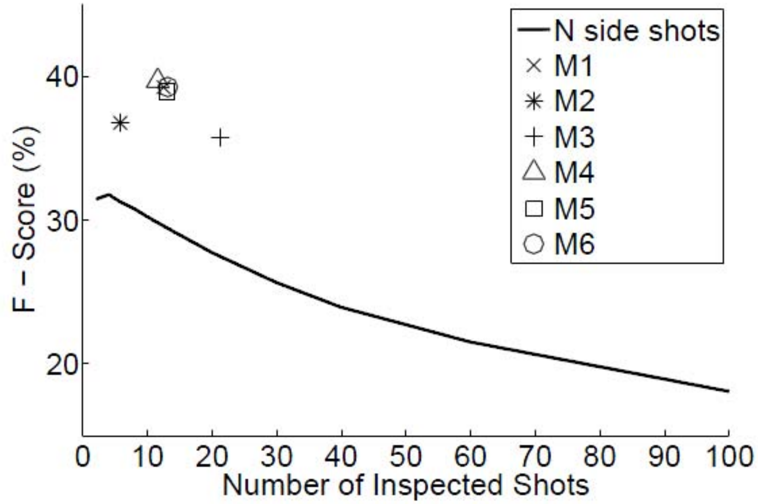
query shot $s_i \quad \rightarrow \quad$ show all $s_j \in \{S_{k-X}, ..., S_{k+X}\} | s_i \in S_k$
and X is a positive integer

Experiments and results

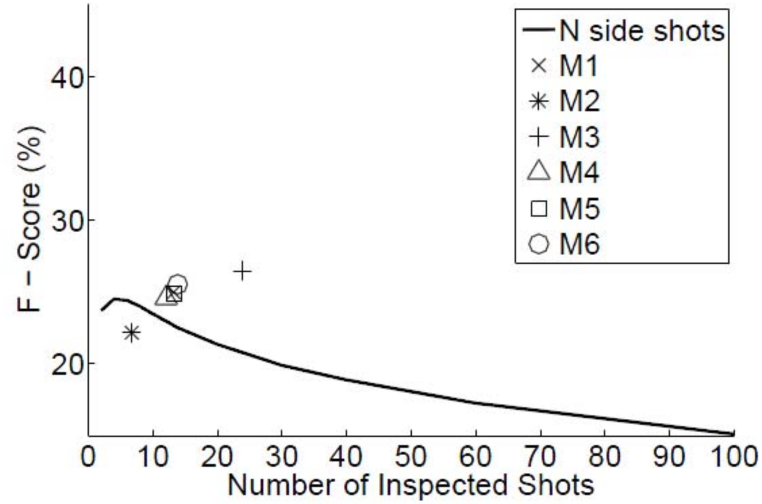
- Evaluation of the results of the basic temporal queries
  - Harmonic mean (F-score) of precision (P) and recall (R), $F=2PR/(P+R)$
  - Measures how successful each basic temporal querying strategy is in retrieving additional positive sample for the 20+24 considered queries, given that one such positive sample has already been found by the searcher and is used for launching a basic temporal query

Experiments and results

- F-score as a function of the number of shots returned by the temporal query (and thus inspected by the user)
  - Basic temporal query types (a) and (b), for (A) single-concept, and (B) complex queries



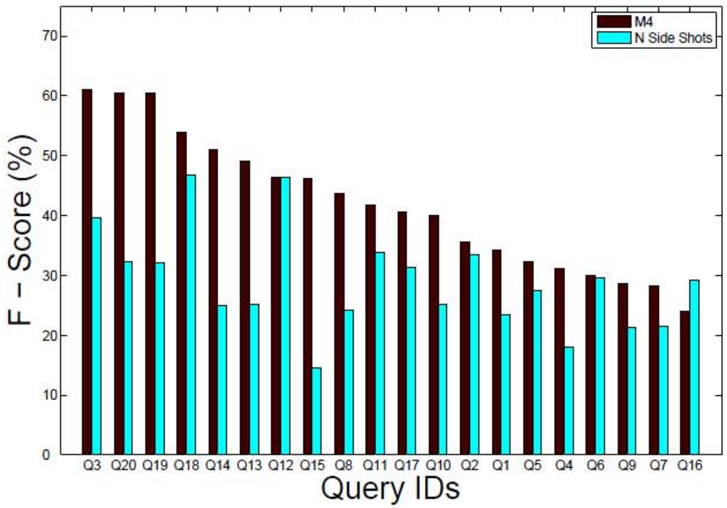
(A)



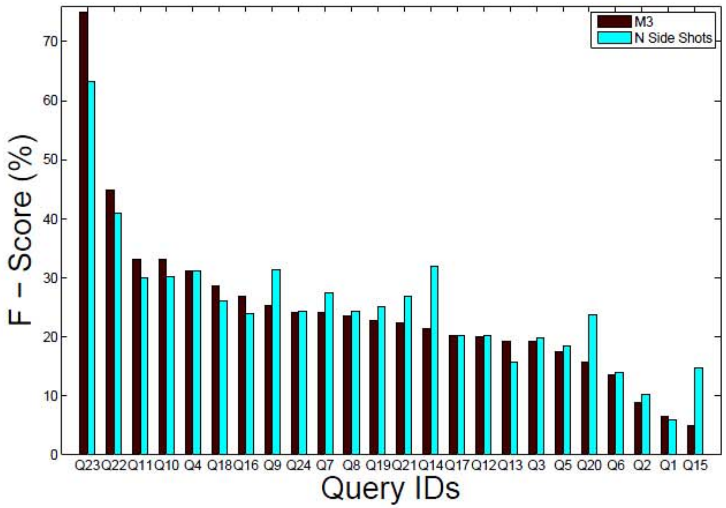
(B)

Experiments and results

- Results per query
  - Basic temporal query types (a) and (b), for (A) single-concept, and (B) complex queries



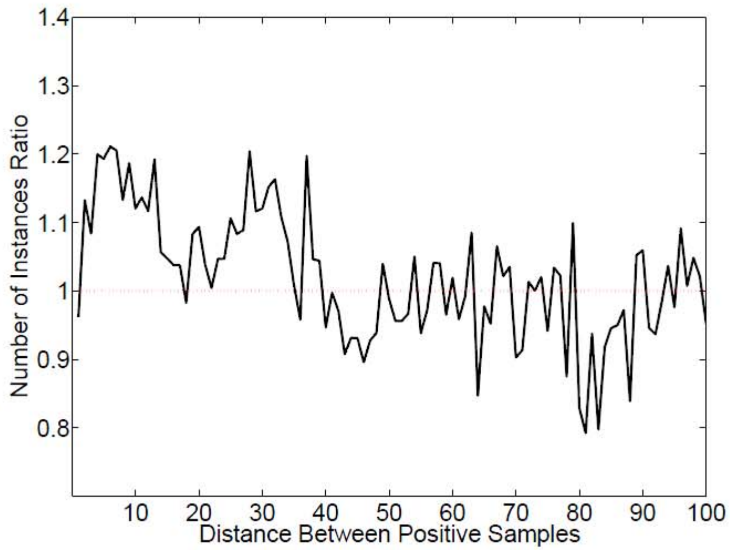
(A)



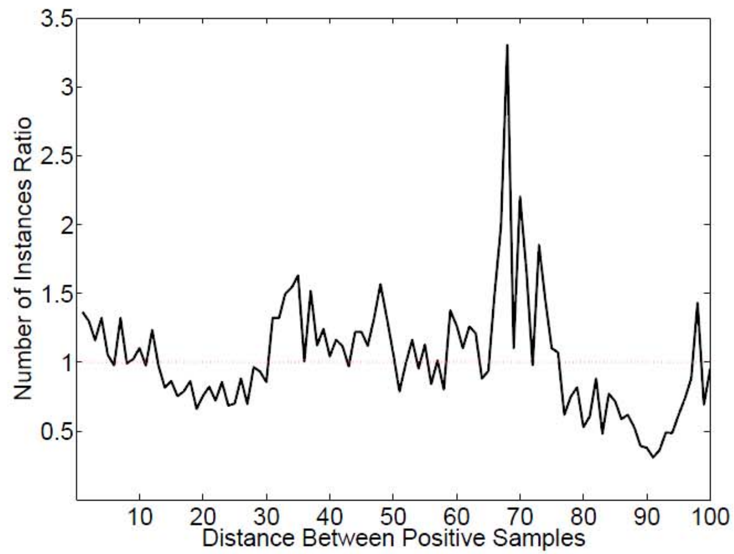
(B)

Experiments and results

- Qualitative differences between single-concept and complex queries
  - (A) # positive samples for single-concept queries / # number of positive samples for complex queries, for every given distance between the samples
  - (B) similar ratio for the positive samples of two individual complex queries



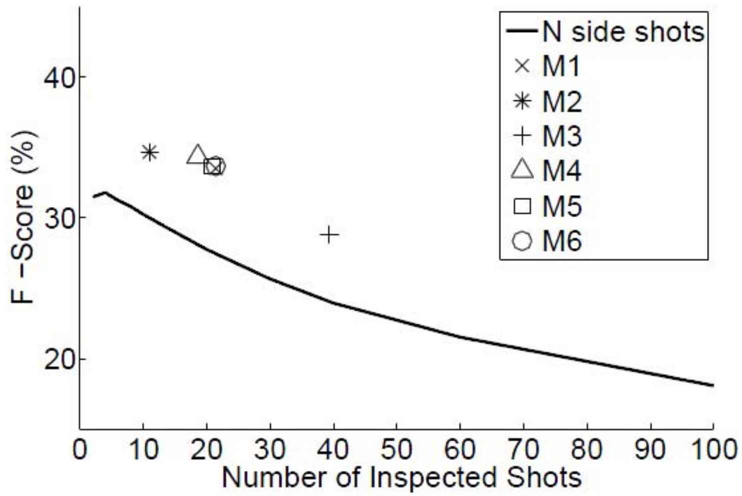
(A)



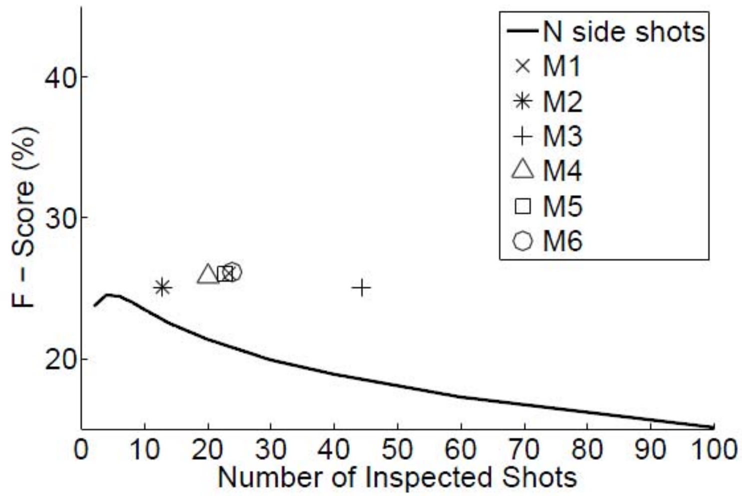
(B)

Experiments and results

- F-score as a function of the number of shots returned by the temporal query (and thus inspected by the user)
  - Basic temporal query types (a) and (c), for (A) single-concept, and (B) complex queries



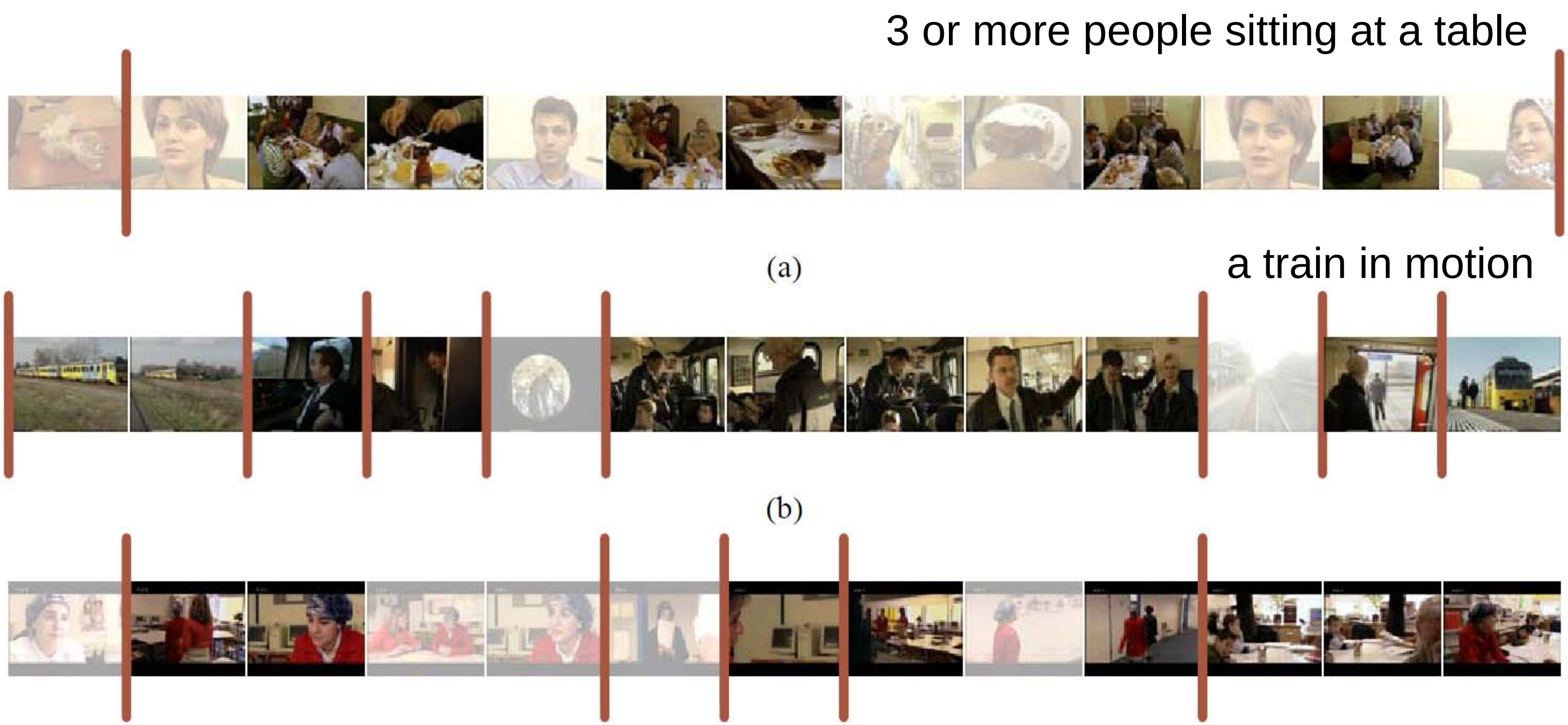
(A)



(B)

Experiments and results

- Examples (success, failure)



Conclusions

- Using existing state-of-the-art scene segmentation algorithms for responding to basic temporal queries can indeed improve the efficiency and effectiveness of interactive retrieval
  - Demonstrated here on a large dataset
  - Considering heterogeneous single-concept and complex queries
  - Using 6 variations of a scene segmentation technique
- The gains are affected by
  - The nature of the queries and the dataset (which result in qualitative differences in the distribution of the distances between positive samples of a query)
  - The quality of scene segmentation (over-segmentation is a problem)

Questions?

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