Similarity based event detection system for videos

Vojtěch Zavřel

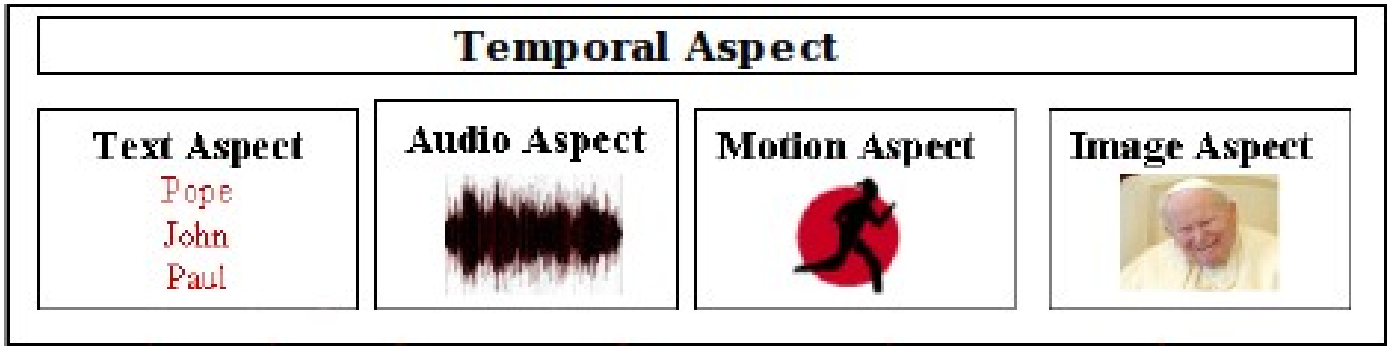
FI MUNI 2011

1. Motivation

- Growing number of public videos
- Copyright infringement
- Sport streams
- Surveillance
- Video archives
  - public
  - television
  - etc.

1. Video definition

- Different aspects
  - image - MPEG-7 descriptors
  - text -automatic Speach Recognition, captions
  - sound
  - motion
  - temporal



2. Event detection

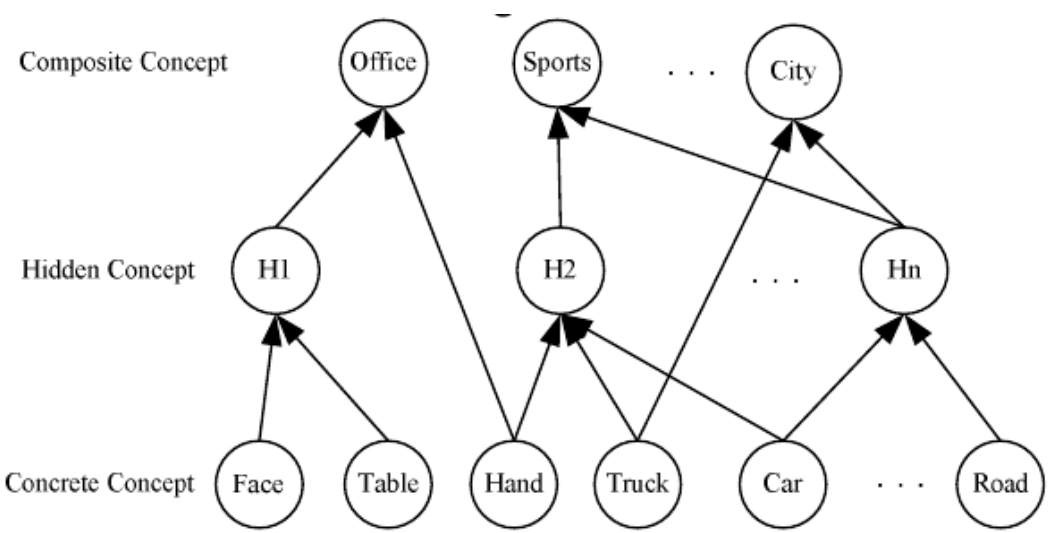
- Event
  - one or multiple defined aspects occurred in video in time interval
  - joined by operators AND, OR, SEQUENCE
- Example
  - TV news (by image) AND about IRAQ (by text) AND burning vehicles (by image) AND time interval < 1 minute (by temporal)

2. Event detection

- Text
- Image
  - global – some of general features
  - local - spatio-temporal detection
- Sound
  - music sounds like ...
- Motion detection
  - based on background
  - picture deformation

3. Current approach

- Common principles
  - annotation based systems (manual vs. auto)
    - VARS, HASTAC, iVAS
  - learning-based systems
    - object-based
    - concept-based
- Domain
  - specific
  - general



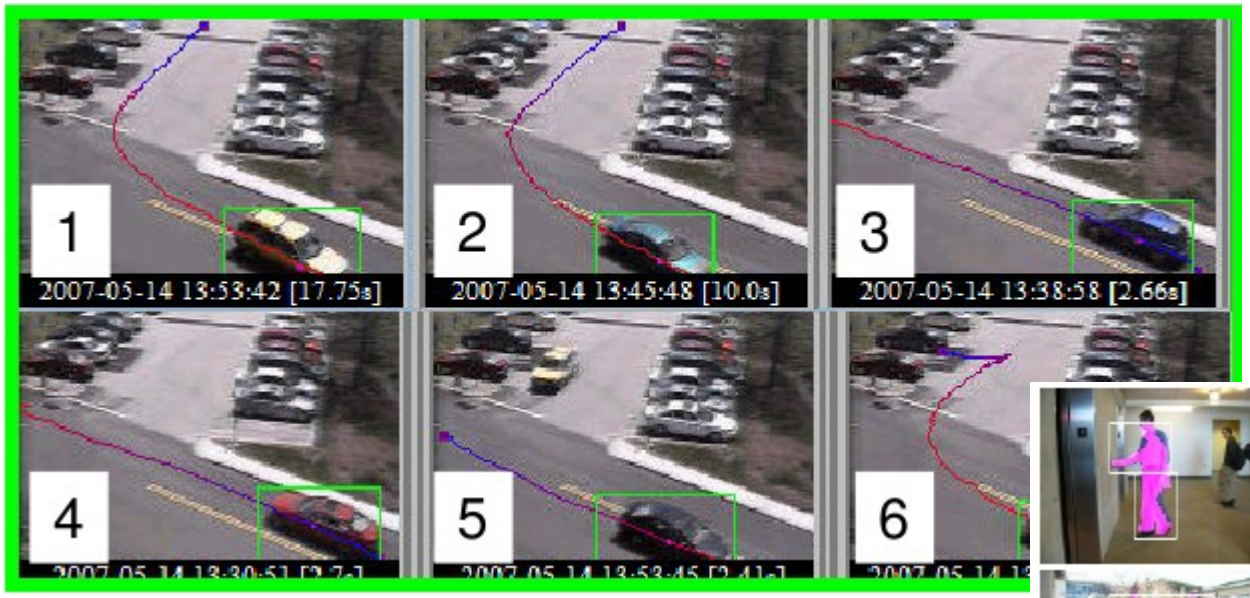
3. Current approach – domain specific

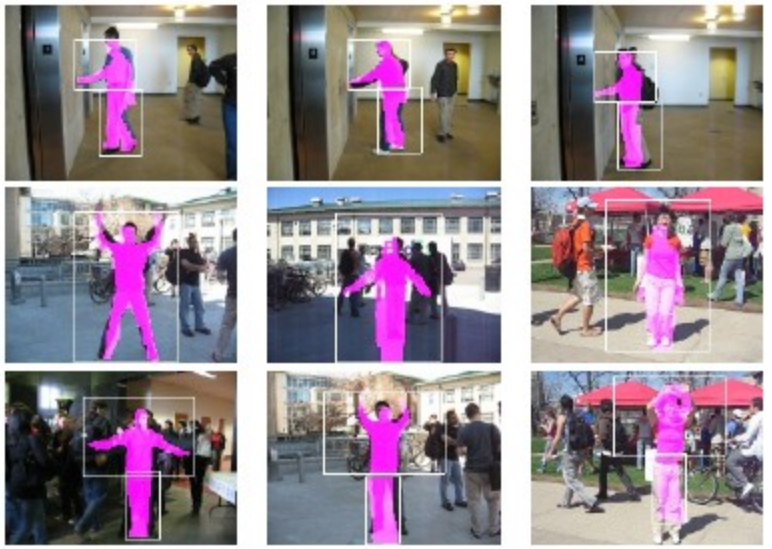
- Domain specific
  - well studied for limited domain like tenis, surveillance
  - well known objects
  - huge training sets
  - specialized structures
  - some are realtime



3. Current approach – domain specific

- Surveillance – common use

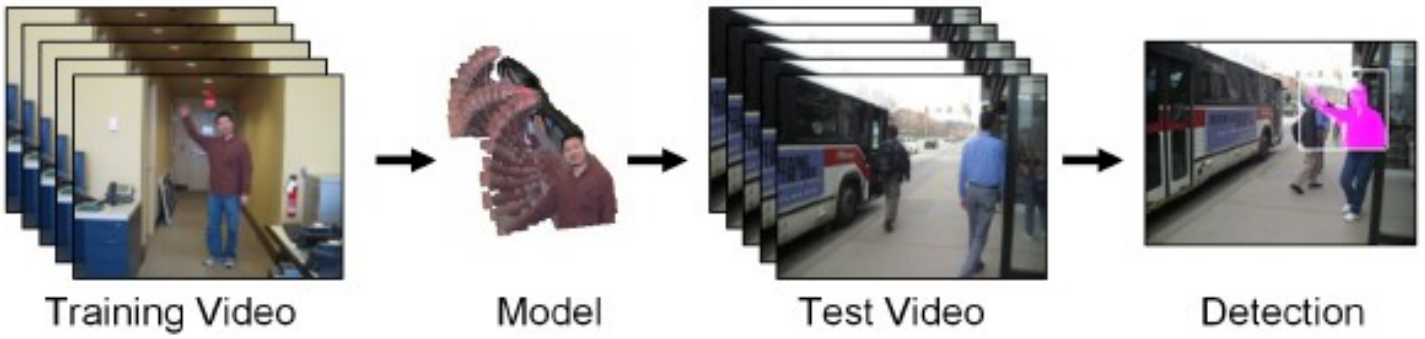




- Object tracking...
- Face detection...

3. Current approach – general domain

- Domain nonspecific (general)
  - based on learning algorithms (training necessary)
  - multi-aspect oriented
  - good results
  - concept-based
  - small datasets



3. Strengths & weaknesses

- Good results -> critical applications
- Usable for domain specific
- Combine multiple aspects

- Necesarry to have enough training set
  - usually described by people
  - usually 40% of whole database
- Often usable only on small datasets

4. Our aproach - ViMUF

- Similarity event detection video framework (ViMUF)
  - based on similarity principles not learning mechanisms
  - domain nonspecific
  - multi-aspect combination (image, sound, text, motion, temporal)
  - user-supplied aggregation function
  - usable on large video datasets

4. System goals

- Similarity based event detection
- Create general interfaces for different extractor
- UI for defigning events based on patterns and different aspects combination with operators AND, OR and SEQUENCE
- Usable on huge datasets

4. Lifecycle - extraction

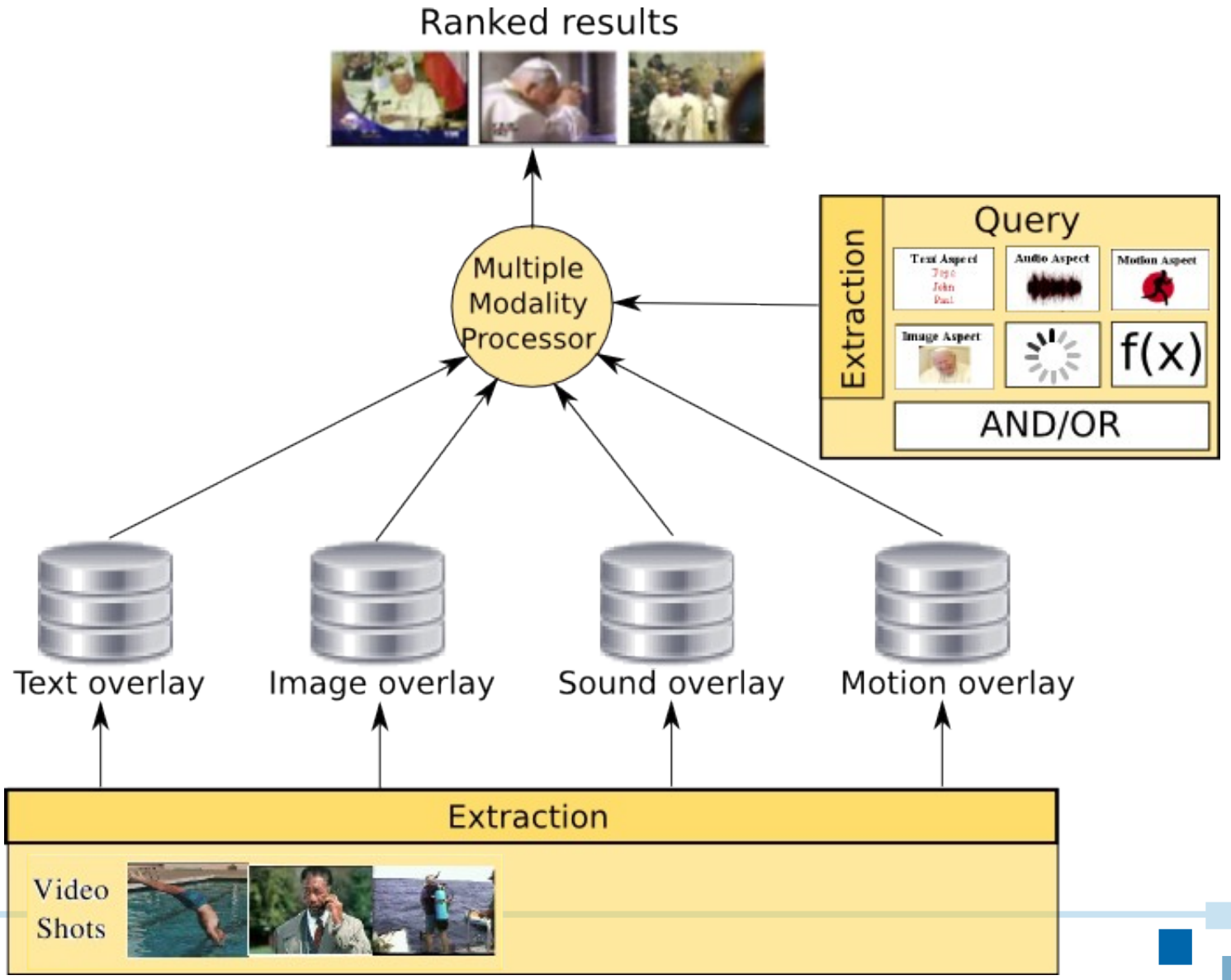
- Split video to scenes
  - extract keyframes and add temporal information (image aspect)
  - extract text (OCR, ASR) and add temporal information (text aspect)
  - extract sound (music) descriptors and add temporal information (sound aspect)
  - extract camera and motion vectors (motion aspect
  - put together temporal aspect

4. Query processing

- User defined function
-
  - Can be used without training set
  - Possibility to use multi-aspects query
  - Query function can be defined by user

- Multi Modality Processor

4. Query processing



6. Conclusion

- Video
  - Aspects: image, text, sound, motion, temporal
  - Event detection
    - domain specific, nonspecific
    - learning mechanisms
- ViMUF
  - Different approach based on similarity
  - Usable on large datasets