GBGI9U07: multimedia document:
description and automatic retrieval

6. Video processing

Georges Quénot and Philippe Mulhem

Multimedia Information Indexing and Retrieval Group

Laboratory of Informatics of Grenoble

March 2018

Video specificities

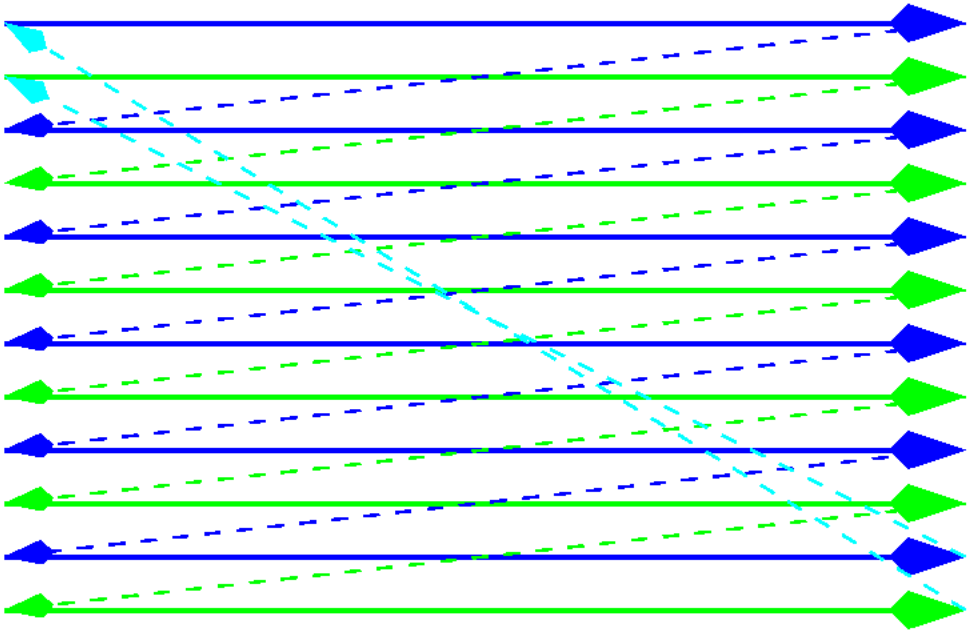
- Coherent and synchronized combination of simpler media
  - Image (animated)
  - Audio
  - Text (closed captions, subtitles)
  - Others (Virtual Reality, ...)
  - *Lots of* formats, resolutions and qualities
- Stream aspect
  - Temporal structure
- *Passage* retrieval from several media
- Various levels of semantic gap and uncertainties
- Contextual information

Animated image

- Interlaced and progressive schemes: images and fields
  - Image/field size
  - Image/field rate
- Compression
  - Image by image: spatial redundancy
  - Image sequence: spatial and temporal redundancies
  - From 28 kbit/s (176 x 144 x 12.5 Hz, sequence coding, realmedia) up to 28 Mbit/s (720 x 288 x 50 Hz, interlaced image coding, DV) or 270 Mbit/s (704 x 288 x 50 Hz, uncompressed, 10-bit D1).
- Very variable quality

Image interlacing

- Compromises resolution versus continuity
- *Complicates* image sequence processing
- Apply only to “full resolution” streams



Top field

Bottom field

Audio

- From 8 kHz up to 48 kHz sample rates
- Mono, stereo or more (5.1, ...)
- Multiple audio streams (multilingual DVDs)
- Compressed from ~5 kbit/s up to 256 kbit/s (1.41 Mbit/s for uncompressed, audio CD) for stereo streams
- Very variable quality

Text

- Subtitles or source speech transcripts
  - Good quality text inserted as separate streams
  - Accurate and relevant
  - When available...
- + Closed captions
  - Text actually contained in the image stream
  - Hard to recover, high miss and error rates
- + Field text (text in the scene)
  - Even worse...
- + Automatic Speech Recognition
  - Text actually contained in the audio stream
  - Significant word error rate but usable

Other stream types

- Animated Artifacts streams (MPEG-4)
  - Audio, text and image substreams
  - May contain accurate and relevant semantic information
  - Could be used of indexing and retrieval
  - Currently exotic and not widely used
- Musical Instrument Digital Interface streams
- Speech synthesis data streams
- ...

Applications

- Search
  - General (web, ...)
  - Domain specific (medical, military, ...)
- Filtering
  - Technological (or scientific or political, ...) survey
  - Offending content
- Metrics
  - Commercials impact estimation (spots, logos)
- Copyright check
- Domains
  - Archives, web, ...
  - Conference or meeting records, ...

Video Segmentation

- Image Segmentation
  - Shot segmentation,
  - Object segmentation and tracking,
  - Camera and object motion,
  - Key frame extraction
- Audio Segmentation
  - Silence / music / noise / speech / ...
  - Male / female
  - High quality / telephone / ...
  - Speaker / known speaker
- Sub-shot (micro-) segmentation
- Story / topic (macro-) segmentation

Shot segmentation

- Direct image comparison or matching
  - Sum of square differences, ...
  - With or without motion compensation
- Descriptor extraction and comparison
  - Color moments or histograms
  - Texture, shapes, points of interest, ...
- Other methods
  - Rough contour tracking, ...
- Compressed domain methods
  - DC image comparison
  - Motion vectors, ratio of forward versus backward vectors
  - Bit rate variation, ratio of predicted versus intra-coded blocks
  - Specific to media encoding, fast but moderate quality

Shot segmentation

- “Cuts” versus gradual transitions
- Separated methods:
  - Cuts: search for discontinuities in images or descriptors
  - Other: ad’hoc methods, specific searches for wipes, dissolves, block changes, ...
  - Plus: photographic flashes filtering, ...
  - Need for detector outputs’ fusion
- Integrated multi-resolution methods:
  - Filtering of descriptors time derivatives at various time scales
  - Search for peaks with sophisticated filtering:
    - » Peak location -> transition (center) location
    - » Scale of the highest peak -> transition duration

Object segmentation and camera motion

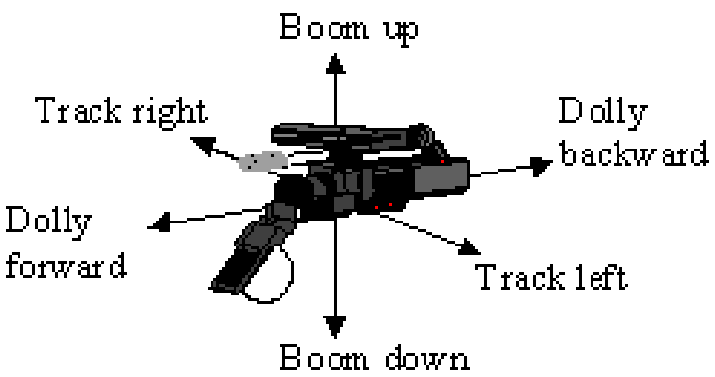
- Classical still image object segmentation (using color, texture, regions, contours, ...) plus:
- Extraction of objects moving relatively to a back-ground and of relative camera/background motion:
  - Objects are identified as not following the background motion
  - Background motion is identified by applying a (parametric) motion model to the part of the image excluding mobile objects
  - Reciprocal dependency can be broken using iterative methods if the background occupy a large enough part of the images
  - Several alternative motion models have to be considered
  - Speed versus accuracy compromises (MPEG vectors versus full optical flow computation)
- Combination of still image and motion based object segmentation
- Bonus: mosaic images or 3D views of the background

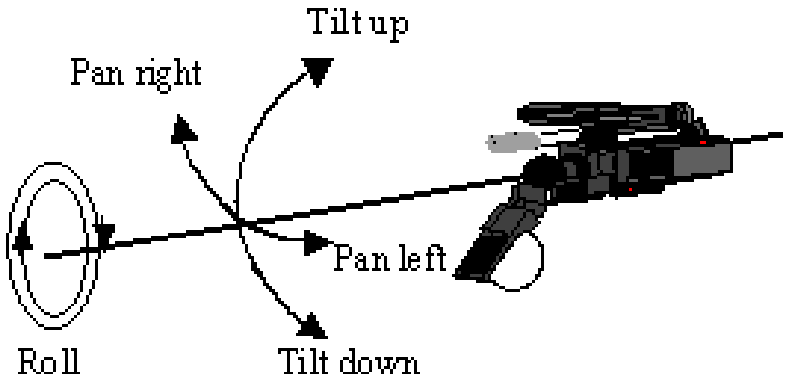
Iterative simultaneous extraction of mobile objects and of relative camera motion

1. Start with assumption of no mobile objects
2. Estimate motion parameters from whole image
3. Estimate the probability of belonging to the background for each pixel
4. Re-estimate motion parameters from pixels with a high probability of belonging to the background
5. Re-estimate the probability of belonging to the background for each pixel
6. Iterate 4. - 5. until convergence or given count

- No need to make any binary decision
- Provides both camera motion and background segmentation

Background / camera motion models
7 degrees of freedom (MPEG-7 descriptors)





Parametric camera model, 7 parameters:

- Translation: track, boom, dolly,
- Rotation: tilt, pan, roll,
- Zoom or focal length

One- or two-level camera motion search

- Direct “3D” camera motion search or:
- “3D” camera motion search from a previously extracted “2D” image motion model
- Parametric image motion models:

Geometric Model	Params	X’	Y’
Translational	2	$x+a$	$y+b$
Similitude	4	$ax-by+c$	$bx+ay+d$
Affine	6	$ax+by+c$	$dx+ey+f$
Homographic	8	$(ax+by+c)/(gx+hy+1)$	$(dx+ey+f)/(gx+hy+1)$
Quadratic	12	$ax^2+by^2+cxy+dx+ey+f$	$gx^2+hy^2+ixy+jx+ky+l$

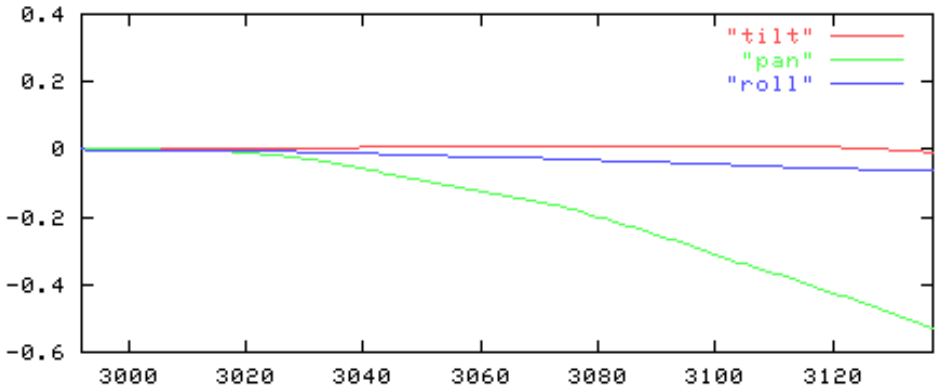
Techniques for Motion Field Extraction

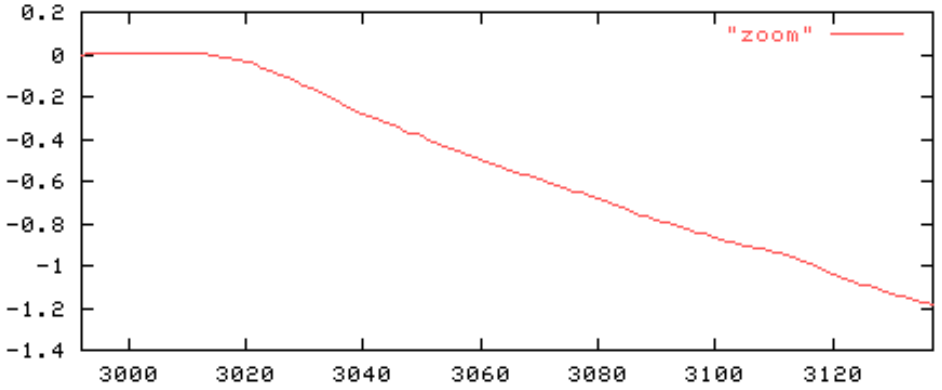
Method	Type	Description
Dense Optical Flow Calculation	Pixel or Sub Pixel Level (very slow)	***Exhaustive pixel correspondance search between frames giving a dense motion field for each pixel***
Phase Correlation Motion Estimation	Block Based (relatively fast)	***Using the shift property of the Fourier Transform, fairly precise motion of image blocks can be calculated with acceptable speed***
Block Matching for Intensity Difference	Block Based (fast, depending on search strategy)	***Motion fields extracted on a block level by finding the best maching area that minimizes the total intensity difference (squared or absolute)***
Readymade Motion Vectors (eg. MPEG)	Block Based (available)	***Assuming the motion vectors represent the approximate intensity flow, they can be read directly from the MPEG stream and used***

Background / camera relative motion search

- No background motion (simple check)
- Motion without parallax (two-level search):
  - Rotations (3) and focal length,
  - Search for an homographic transform,
  - Mosaicing (panoramic view) and mobile objects,
- Motion with parallax (one-level search):
  - Rotations (3), translations (3) and focal length,
  - “Motion and structure from motion”, “paraperspective decomposition” method from Poelman et Kanade (1993).
  - Three-dimensional view of the background
- Irregular motion (crowd, waves, ...)

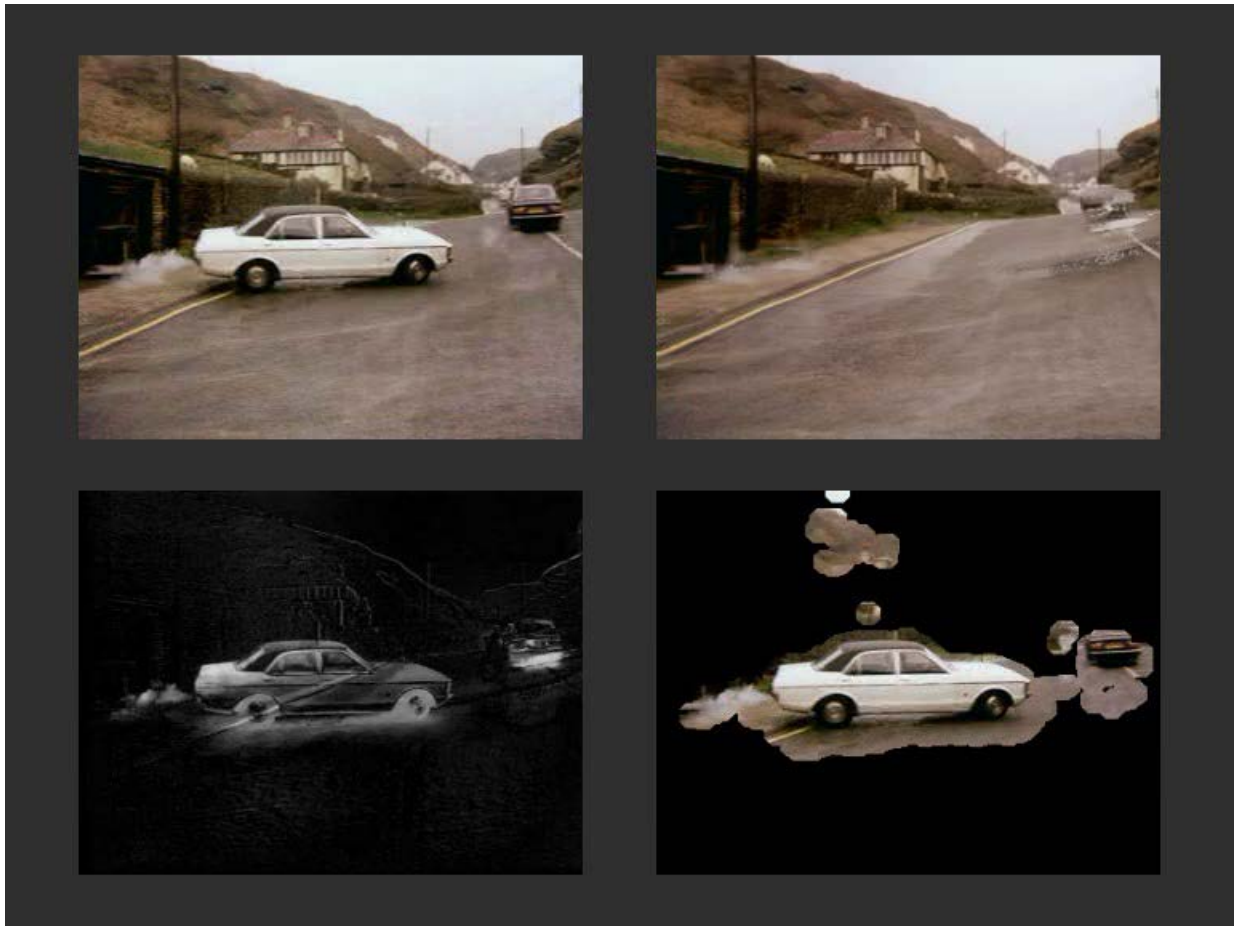
Camera motion, aim1mb08.mpg document, 2992-3137 sequence



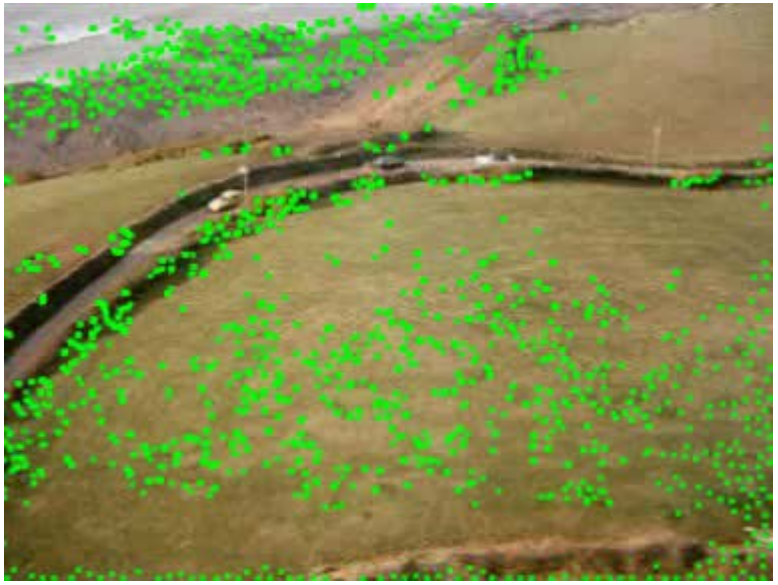


Panoramic view, aim1mb08.mpg document, 2992-3137 sequence

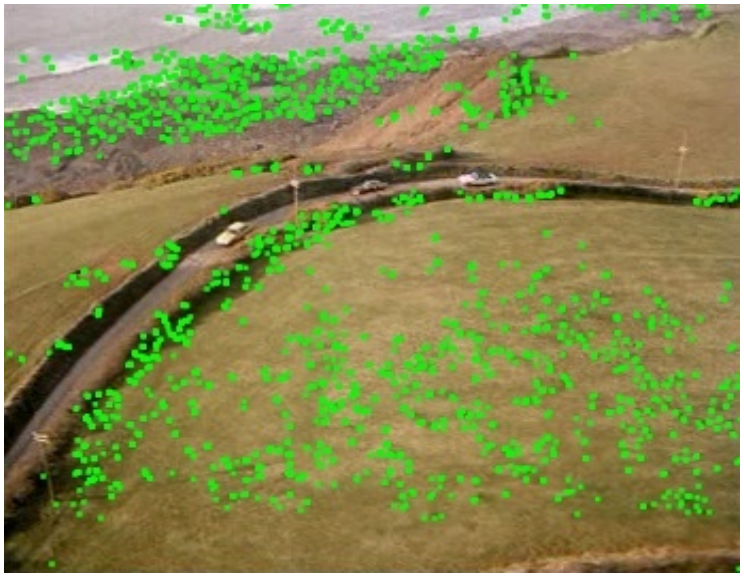
Mobile objects segmentation and tracking aim1mb08.mpg, 1671-1715 sequence

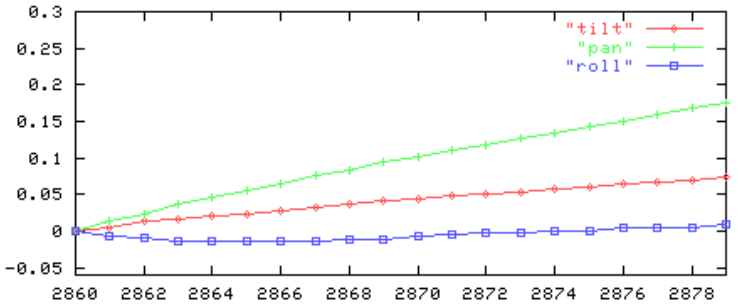


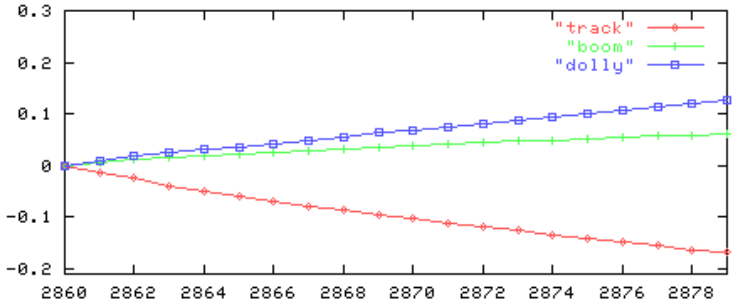
Motion with parallax: Tracking of feature points aim1mb08.mpg, 2860-2879 sequence



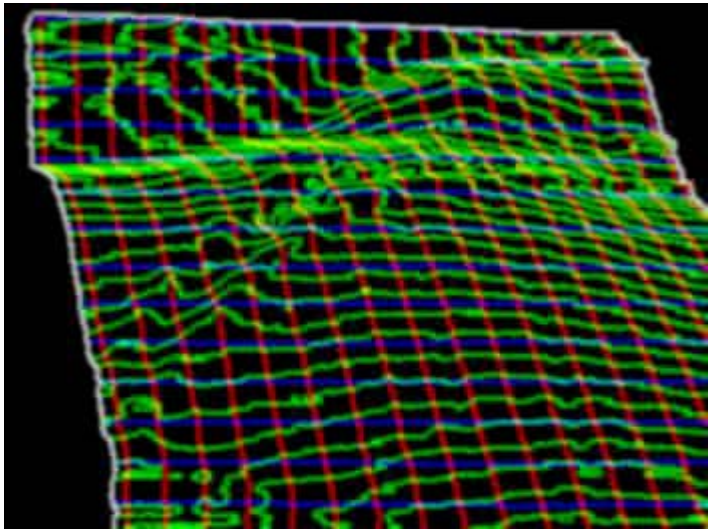
Motion with parallax







Three-dimensional scene structure



Key frame extraction

- Choice of one or more representative key frame per continuous shot
- First or center frame
- Frame with low motion, following a zoom, ...
- Frame with high contrast
- Weighted combination of criteria
- Panoramic view when available
- Frame showing a best view of an object (or frontal face)
- Multiple key frame selection according to content change detection and/or important feature detection
- Maximization of inter-shot key frame dissimilarity
- Used for content display (browsing) or indexing

Audio Segmentation

- Silence / music / noise / speech / language / male / female / high quality / telephone / speaker / known speaker / emotion / ...
- Feature extraction: spectral analysis (MFCC, LPC, plus ...) on 10-20 ms windows.
- Class modeling or clustering using Gaussian mixture densities.
- Bayesian Information Criterion:
  - build models for a whole segment and for its left and right parts independently with the same parameter count,
  - Compare predictions using both strategies.
- Known classes (music, speech, male, anger, ...) or classes to be built (one for each unknown speaker).

Micro-segmentation

- Segmenting into units shorter than a shot
- May be hierarchical
- Change in background / camera relative motion
  - Start / stop of a zoom or a pan
- Change in object motion
  - Start, stop or direction change of objects
  - Appearance or disappearance of objects/persons
- Partial / local transitions
  - Small image appearance, disappearance or change
- Speaker or topic change
- Useful for key frames selection and content analysis

Macro-segmentation

- Segmenting into units longer than a shot
- May be hierarchical
- Not necessarily aligned with image or audio transitions
- Generally according to semantic changes like switch of topic within a TV journal
- Use of various clues:
  - Visual or audio jingles, black or blue frames,
  - Topic detection and tracking from audio transcription,
  - Pattern detection from audio transcript,
  - Detection of text or small image appearance or change.
- Useful for determining appropriate boundaries of retrieved passages