Spatiotemporal segmentation of compressed video sequences

Informatics and Telematics Institute / Centre for Research and Technology Hellas

Spatiotemporal segmentation of compressed video sequences

Vasileios Mezaris

Postdoctoral Research Fellow
Multimedia Knowledge Group
Informatics and Telematics Institute / Centre for Research and Technology Hellas

Presentation Overview

- Introduction: problem, motivation, objectives
- MPEG-2 compression basics
- Compressed-domain versus Pixel-domain segmentation
- Spatiotemporal segmentation of MPEG-2 video – literature
- Spatiotemporal segmentation of MPEG-2 video – proposed approach
  - Compressed-domain information extraction
  - Segmentation algorithm
    - Moving object segmentation and tracking
    - Background segmentation
- Pixel-domain boundary refinement
- Experimental results and comparisons
- Conclusions
Introduction

- Segmentation
  - partitioning a piece of information into meaningful elementary parts termed segments

- Video segmentation
  - range of different processes for partitioning the video to meaningful parts at different granularities
    - *temporal*, partitioning the video to *scenes* or *shots*
    - *spatial*, independently partitioning each video frame to arbitrarily shaped regions
    - *spatio-temporal*, generating temporal sequences of arbitrarily shaped spatial regions
    - *Foreground/background separation*, can be seen as special case of spatio-temporal segmentation
Introduction

- Effective video manipulation
  - Capture
  - Storage
  - Content-based indexing and retrieval
  - Coding and transmission
  - Understanding

- Content-based Indexing and retrieval
  - Require the extraction of characteristic features
  - Segmentation enables the extraction of features at the region/object level
  - Extraction at the region level has been documented to be advantageous for still image retrieval
  - Querying in any kind of video collections (internet, stock video collections for commercial purposes, personal collections)

- Coding and transmission
  - Can benefit from an object-based approach

- Video understanding
  - Segmentation and feature extraction a first step towards region identification and understanding of semantics
  - Major step towards more efficient indexing, retrieval, coding and transmission, personalized delivery etc.
Introduction

- Segmentation
  - Ill-posed problem – no unique solution exists
  - The best solution depends on
    - The considered application
    - Subjective judgment of human observer
  - In video, this subjectivity is (partially) alleviated by considering differently moving objects

- Desired features of segmentation process
  - Resulting regions/objects should represent well enough (in terms of their number and their localization) to the most important of the depicted objects
  - No need for human intervention (e.g. for parameter tuning)
  - Good time-efficiency: in video, real-time processing without use of specialized hardware
**MPEG-2 compression basics**

- **MPEG-2 exploits information redundancy**
  - Intra-frame (spatial): adjacent pixels often have similar intensity values
  - Inter-frame (temporal): adjacent frames are usually very similar

- **Intra-frame redundancy**
  - Discrete Cosine Transform (DCT)
    - 8x8 intensity value blocks are transformed into 8x8 DCT coefficient blocks
    - DC DCT coefficient corresponds to mean intensity over all block pixels
    - AC DCT coefficients correspond to higher-level frequency information – depending on their significance, they may be heavily quantized
    - Several AC coefficients are ~0 and turn to 0 after quantization
  - Blocks coded as a mean value + additional levels of higher frequency info

- **Inter-frame redundancy**
  - Residual information
    - MB motion vectors (MVs) estimated, associating each MB of the frame with a MB-sized area of one (two) reference frame(s), so that the MB Displaced Frame Difference (DFD) is minimized
    - DCT-coding intensity differences and quantization results in few non-zero coefficients
    - Instead of DCT-coding original intensity values, only MB motion vectors and DCT-coded intensity differences (residual information) are used
MPEG-2 compression basics

- **MPEG-2 structure**
  - Succession of Group of Pictures (GOPs) - each GOP made of:
    - **I-frames**
      - Coded exploiting intra-frame redundancy only: DCT
    - **P-frames**
      - Coded exploiting inter- and intra-frame redundancy: previous I or P-frame used as reference frame for estimating MVs
      - Intra-coded MBs can be found in P-frames
    - **B-frames**
      - Coded exploiting inter- and intra-frame redundancy: previous and following I or P-frame used both as reference frames for estimating MVs
      - Intra-coded and forward-predicted MBs can be found in B-frames
MPEG-2 compression basics

- MPEG-2 stream coding / decoding
  - Motion vectors rather easy to decode
  - Intensity information: DCT coded at best – computationally intensive IDCT required to decode, but…
  - Coarse intensity information (DC DCT coefficients) readily available – rather easy to decode

- MPEG-2 information typically used for segmentation
  - Motion vectors
  - Coarse intensity information (DC DCT coefficients)
  - Texture information (AC DCT coefficients)
Compressed-domain versus Pixel-domain segmentation

- **Pixel-domain video segmentation:**
  - **Produces pixel-accuracy segmentation masks**
    - Color information available at pixel-granularity
    - Texture, motion, etc information can be computed at pixel-granularity
    - The amount of information to be processed is a significant issue
  - Requires that the sequence is fully decoded
    - No video is stored in raw format
    - Decoding a computationally-intensive task
    - Storage of decoded video poses significant storage requirements (one hour of raw DVD-resolution video would take up over 100GBs)
  - Often requires block matching for motion estimation
    - Block matching is computationally-intensive
    - Other motion estimation methods (e.g. optical flow) also computationally-intensive
  - Suffers of high computational complexity
    - Can be alleviated only by making restrictive assumptions about camera motion, background uniformity etc.
Compressed-domain versus Pixel-domain segmentation

- MPEG-2 compressed-domain video segmentation:
  - Produces coarse-grained segmentation masks
    - Motion information available at MB-level granularity (16x16 pixels)
    - Chromaticity information directly available (DC DCT coefficients) only at MB-level granularity
    - Luminance information directly available (DC DCT coefficients) only at block-level granularity (8x8 pixels)
    - AC DCT coefficients difficult to interpret at pixel-level without IDCT
  - Does not require full sequence decoding before processing
    - Typically only motion information and DC (AC) DCT coefficients are used
  - Does not require block-based motion estimation
    - Motion information, although noisy and coarse, is readily available in the stream
  - Features low computational complexity (real-time operation is possible)
    - Small volume of information to be processed
    - No complete MPEG-2 decoding
Compressed-domain versus Pixel-domain segmentation

- **Pixel-domain video segmentation:**
  - Necessary when pixel-level accuracy is required
  - Feasible in restricted cases

- **MPEG-2 compressed-domain video segmentation:**
  - Suitable for application in large video archives
  - Could serve as a preprocessing step even if pixel-domain accuracy is required
Spatiotemporal segmentation of compressed video sequences

Informatics and Telematics Institute / Centre for Research and Technology Hellas

Spatiotemporal segmentation of MPEG-2 sequences – literature review

- [Sukmarg00]
  - Uses DC and AC DCT coefficients

- **Algorithm**
  - Initial clustering by k-means application to DC DCT coefficients
  - Spatiotemporal segmentation
    - Estimation of spatial cluster similarity based on energy estimated from the AC-coefficients
    - Estimation of temporal similarity from the results of a 3D Sobel operator
    - Spatio-temporal merging following two criteria:
      - Form regions with strong spatiotemporal similarity
      - Form regions with lower spatiotemporal similarity but high average temporal change within the region
    - Foreground/background separation based on the evaluation of the average temporal change of each region
      - High average temporal change signifies moving objects

- **Does not make use of motion information**
  - Average temporal change may be affected by object homogeneity and global camera motion
Spatiotemporal segmentation of MPEG-2 sequences – literature review

- [Jamrozik02]
  - Uses MB motion vectors of P-frames and DCT coefficients of I-frames

- Algorithm
  - Watershed segmentation on a DC+2AC image
  - Motion vector accumulated over a few frames
  - Motion segmentation
    - Sum of absolute displacement in x- and y-direction estimated for each MB
    - Uniform quantization in 8 levels
  - Adjacent regions created by the watershed segmentation are merged if they undergo similar motion

- Simple use of motion information
  - Method can account for restricted types of global motion, e.g. NOT zoom
    - Under zooming, the motion map will appear as concentric rings – motion based merging will be erroneous
Spatiotemporal segmentation of compressed video sequences

Informatics and Telematics Institute / Centre for Research and Technology Hellas

Spatiotemporal segmentation of MPEG-2 sequences – literature review

- [Babu04]
  - Uses MB motion vectors of I-, P- and B-frames

- Algorithm
  - Motion vector accumulated over a few frames
  - Accumulated motion vectors spatially interpolated and smoothened
  - Pixels with zero motion assigned to background layer
  - Number of motion models estimated
    - Affine parameter estimation for non-overlapping square regions whose variance is less than a threshold
    - Affine vectors clustered using K-means, tried for different values of K
    - K selected by evaluating the resulting Mean Square Error
  - Pixels assigned to layers using Expectation-Maximization algorithm
  - Edge refinement by decoding of edge MBs
    - sub-block matching by evaluating MAD
    - classification to objects based on the direction of motion

- Non-real-time segmentation
  - Background layer formation may fail in the presence of global motion

Introduction
MPEG-2 compression basics
Compressed-domain vs. Pixel-domain
Spatiotemporal segmentation of MPEG-2 sequences – literature review
Proposed compressed-domain approach
Pixel-domain boundary refinement
Experimental results and comparisons
Conclusions
References

Spatiotemporal segmentation of MPEG-2 sequences – literature review

- \([Zhu06]\)
  - Uses MB motion vectors of P-frames and DCT coefficients of I-frames

- Algorithm
  - Motion detection (MB-level accuracy)
    - Motion of moving objects assumed to be non-Gaussian
    - Noise assumed to be zero-mean Gaussian
    - Moving object detection by fourth moment calculation and thresholding
  - Watershed applied to DCT coefficient vectors (Block-level accuracy)
  - Simple motion & color mask fusion
  - Edge correction and morphological post-processing after decoding of edge Blocks
  - Object tracking in subsequent frames based on matching object contour blocks
    - Edge correction and morphological post-processing required for each frame

- Non-real-time segmentation
  - Susceptible to initialization errors – if the first frame segmentation fails, tracking in subsequent frames will also fail
  - Does not handle the appearance / disappearance of moving objects
Spatiotemporal segmentation of compressed video sequences

Introduction

MPEG-2 compression basics

Compressed-domain vs. Pixel-domain

Spatiotemporal segmentation of MPEG-2 sequences – literature review

Proposed compressed-domain approach

Pixel-domain boundary refinement

Experimental results and comparisons

Conclusions

References

Spatiotemporal segmentation of MPEG-2 sequences – literature review

- [Liu07]
  - Examines H264 instead of MPEG-2 video
  - Uses MB motion vectors

- Algorithm
  - MVs can correspond to blocks of variable size – when a block is larger than 4x4, the same MV is assigned to its constituent 4x4 blocks
  - Smoothing – 3x3 median filtering of motion field
  - Sobel operator applied to the motion magnitude matrix for each motion direction
  - Gradient map simplified using morphological operators
  - Watershed algorithm applied to simplified gradient map
  - The motion of each region is modeled using a six-parameter affine model
  - Mergings based on motion similarity are performed – Binary Partition Tree used for representing the merging process

- Information available different to some extent that in MPEG-2

- Non-real-time segmentation
Spatiotemporal segmentation of MPEG-2 sequences – literature review

- Tracking of objects in the compressed domain
  - [Favalli00]
    - Based on motion vectors of P-frames
    - Performs Block-matching to estimate motion vectors in I-frames
    - MB-level accuracy
  - [Aggarwal06]
    - DCT domain background subtraction in Y plane of I-frames to find candidate objects
    - DCT domain histogram matching using Cb and Cr planes for final object selection
    - Interpolation for object localization in intermediate P- and B-frames
    - MB-level accuracy
  - [DeSutter07]
    - Based on motion vectors; encoder slightly modified to perform motion estimation even for intra-coded frames
    - MB-granularity motion information translated to finer-granularity motion information by examining the MB-Block overlapping
    - Finer-granularity tracking using blocks of any size, demonstrated for 8x8 blocks
Algorithm overview

- [Mezaris04]
- Non-supervised approach
- Novelty
  - Use of iterative rejection procedure for initial foreground/background separation
  - Enforcement of the temporal consistency of the output of iterative rejection
**Feature extraction**

- **MPEG-2 compressed sequences**
  - **P-frames**
    - Macroblock motion vectors
  - **I-frames**
    - DC coefficients of the DCT for each macroblock, corresponding to the Y, Cb and Cr coordinates of the MPEG color space
    - Macroblock motion vectors, extracted by linear interpolation of the corresponding motion vectors in the two adjacent P-frames
  - **B-frames**
    - Not considered
  - Feature extraction using properly modified MPEG reference decoder
Iterative rejection

- Proposed for global motion estimation [Rath99, Yu01]
- Here used for foreground / background separation
  - Bilinear motion model (less vulnerable to noise than others, e.g. the affine motion model)
    \[
    \hat{u}_t(b_i) = a_0 + a_1 b_x^i + a_2 b_y^i + a_3 b_x^i b_y^i, \quad \hat{v}_t(b_i) = a_4 + a_5 b_x^i + a_6 b_y^i + a_7 b_x^i b_y^i
    \]
  - Model parameters estimated from macroblock motion vectors by means of Least Squares Estimation
  - Macroblocks with motion estimation error higher than average are rejected
  - Motion model parameters and fg/bg mask iteratively refined until no new macroblocks are rejected during an iteration

Introduction
MPEG-2 compression basics
Compressed-domain vs. Pixel-domain
Spatiotemporal segmentation of MPEG-2 sequences – literature review
Proposed compressed-domain approach
Pixel-domain boundary refinement
Experimental results and comparisons
Conclusions
References

Spatiotemporal segmentation of compressed video sequences
Informatics and Telematics Institute / Centre for Research and Technology Hellas

Iterative rejection

- Results in a fg/bg mask
  - Shows which macroblocks have been rejected – these may belong to the foreground
- Fast procedure, but suffers from
  - Inaccurate macroblock motion vectors
  - Inability of the motion model to represent all possible motions
  - Application to each frame independently – may result in temporal inconsistencies
Macroblock-level tracking

- Tracking based on macroblock motion vectors
- Tracking operator defined as in [Favalli00]
  \[ \tau(t) \]
  Foreground macroblock tracking results in slight expansion of the foreground regions

**temporal tracking operator example**
Macroblock-level tracking

- Examines and enforces the temporal consistence of the output of iterative rejection
Object formation

**Rejected macroblocks**
- Are clustered to connected spatial regions ($\mu$σκα $R^I_t$)
  - Recursive connected component labeling procedure based on 4-connectivity
- Are assigned to differently moving spatiotemporal objects
  - In the first frame of the sequence, each foreground connected component corresponds to a different foreground object
  - In any subsequent frame, the macroblock temporal tracking operator is applied to the final segmentation mask of the previous frame and the spatial overlapping of regions is examined – the previously generated connected components fall into 3 distinct categories:
  - Category 1: $s^t_2$
  - Category 2: $s^t_1$
  - Category 3: $s^t_3$
Background segmentation

- **First frame of each shot**
  - Estimation of the number of different background spatiotemporal object by application of maximin algorithm to the intensity features of the frame

- **I-frames**
  - Application of k-means algorithm to the intensity features of the frame, k being pre-set based on the output of maximin
  - Enforcement of region connectivity

- **P-frames**
  - Temporal tracking of background macroblocks using motion information
Pixel-domain boundary refinement

- Further processing of the coarse-grained segmentation masks
  - Use of pixel intensity information
  - Partial pixel reclassification using Bayes binary classifier

- Simple method originally proposed for still image segmentation refinement
  - Does not make use of motion information
Experimental results

Comparison, Table-tennis sequence

- [COST211]
- [Sifakis01]
- Proposed approach
- Proposed approach + pixel domain boundary refinement
Experimental results

Comparison, Table-tennis sequence

- [COST211]
- [Sifakis01]
- Proposed approach
- Proposed approach + pixel domain boundary refinement
- [Mezaris04b]

Table-tennis sequence error vs. frame number
Experimental results

Comparison, Coast-guard sequence

- [COST211]
- [Sifakis01]
- Proposed approach
- Proposed approach + pixel domain boundary refinement

Introduction
MPEG-2 compression basics
Compressed-domain vs. Pixel-domain
Spatiotemporal segmentation of MPEG-2 sequences – literature review
Proposed compressed-domain approach
Pixel-domain boundary refinement
Experimental results and comparisons
Conclusions
References
Spatiotemporal segmentation of compressed video sequences

Informatics and Telematics Institute / Centre for Research and Technology Hellas

Introduction
MPEG-2 compression basics
Compressed-domain vs. Pixel-domain
Spatiotemporal segmentation of MPEG-2 sequences – literature review
Proposed compressed-domain approach
Pixel-domain boundary refinement
Experimental results and comparisons
Conclusions
References


Experimental results

Comparison, Coast-guard sequence

- [COST211]
- [Sifakis01]
- Proposed approach
- Proposed approach + pixel domain boundary refinement
- [Mezaris04b]

Coast-guard sequence graph showing error vs. frame number.
Spatiotemporal segmentation of compressed video sequences

Informatics and Telematics Institute / Centre for Research and Technology Hellas

Experimental results

Comparison, Penguin sequence

- [COST211]
- [Sifakis01]
- Proposed approach
- Proposed approach + pixel domain boundary refinement

Introduction
MPEG-2 compression basics
Compressed-domain vs. Pixel-domain
Spatiotemporal segmentation of MPEG-2 sequences – literature review
Proposed compressed-domain approach
Pixel-domain boundary refinement
Experimental results and comparisons
Conclusions
References

Experimental results

Comparison, Stair sequence

- [COST211]
- [Sifakis01]
- Proposed approach
- Proposed approach + pixel domain boundary refinement
Experimental results

- **Computational efficiency**
  - Compressed-domain method performs segmentation in real-time, without use of specialized hardware
    - 5.02 msec for each CIF frame (352x288) on a 800MHz Pentium III
  - Simple fast process for pixel domain boundary refinement
    - 0.48 sec for each CIF frame (352x288) on a 800MHz Pentium III
Conclusions

- MPEG-2 compressed segmentation
  - Produces results of quality comparable with that of raw domain segmentations
  - Performs in real-time
  - Even segmentation of MB-accuracy has been shown to be beneficial for retrieval [Mezaris04]

- Further research needed
  - More elaborate processing of motion information
  - Use of coarse color information in combination with motion information for coarse segmentation
  - More elaborate methods for pixel-domain boundary refinement, combining color and motion information
Questions
References

References