A Novel Video Thumbnail Extraction Method Using Spatiotemporal Vector Quantization

Junfeng Jiang and Xiao-Ping Zhang

Communication and Signal Processing Applications Laboratory (CASPAL)
Department of Electrical and Computer Engineering
Ryerson University, Toronto, Canada
{j7jiang, xzhang}@ee.ryerson.ca

Outline

- Introduction
- Video Thumbnail Extraction Using Spatiotemporal Vector Quantization
  - Temporal Quantization Using VTDF
    - VTDF (Video Time Density Function)
    - Temporal Quantization
  - Gaussian Mixture Vector Quantization
- Experimental Results
- Conclusion

Introduction

Objective

To find an optimal solution to generate a compact video summary by exploring both temporal and spatial characteristics.

Solution

- VTDF (Video Time Density Function) is applied to find the optimal number of temporal segments of video data.
- For each segment, Gaussian mixture is employed to explore the spatial characteristics and find the optimal number of sample frames.

Temporal Quantization Using VTDF

- Objective: VTDF is used for video segmentation
- Solution
  - Mutual information between two consecutive frames to model the video temporal density.
    - one video sequence is modeled as one dimensional time series signal to reduce computational cost.
  - VTDF is to measure the information passed from one frame to another frame. Therefore, it makes the boundary of temporal segments detection feasible.
  - VTDF-based temporal quantization is proposed to find the optimal number of segments in time domain.

Pixel Domain VTDF

- VTDF is calculated based on the inter-frame mutual information to measure the information passed from one frame to another.
- The inter-frame dependency is modeled by the difference of grey values of two consecutive frames in pixel level
- Two ways to model the grey values of each frame,

  128*D* HSV and 256*D* RGB color histogram

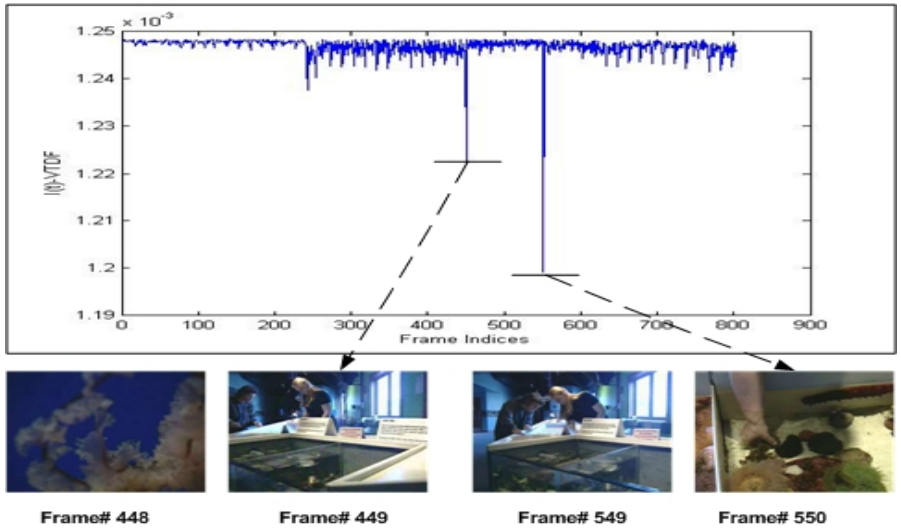
Pixel Value based VTDF

- Joint probability matrix *D*, each element *D* (*i*, *j*) is to model the inter-frame dependency in pixel level,
  - a pixel *k* has grey value *i* at frame *t* and *j* at frame *t*-1.
- Given a frame *t* (total frames-*N*), its VTDF, *I*(*t*) using interframe mutual information can be calculated as,

$$I(t) = -\frac{1}{U}\sum_{i=0}^{255}\sum_{j=0}^{255} D(i,j)\log\frac{D(i,j)}{D(i,)D(,j)}$$

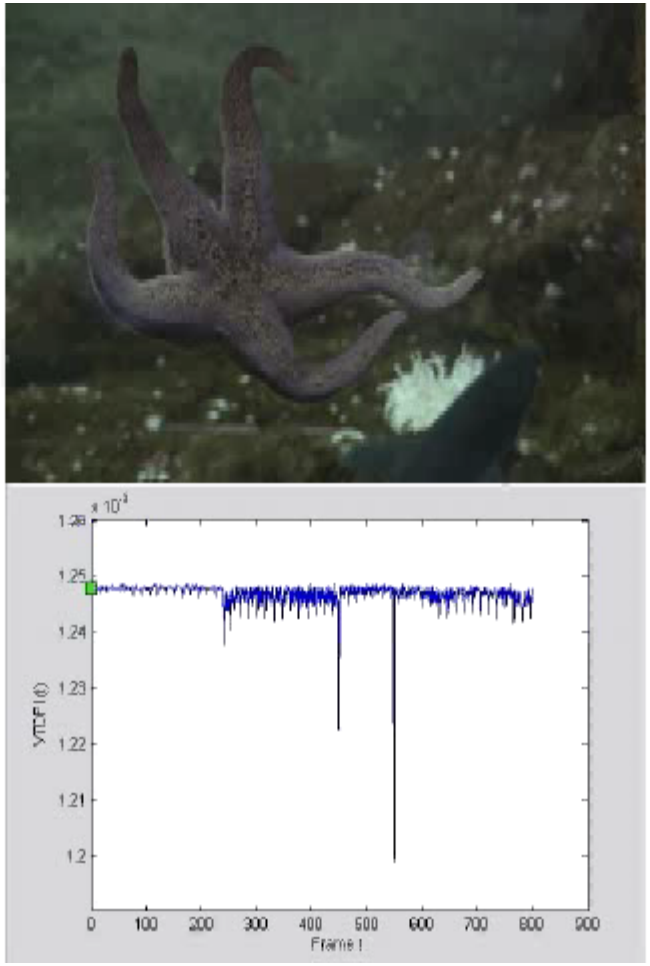
  - *D*(*i*,) and *D*(,*j*) are the marginal probabilities
  - *U* is normalization factor to make $\sum_{t=1}^{N} I(t) = 1$
  - *I*(1)=0
- A large difference in content between two frames shows a weak inter-frame dependency and leads to a small value of VTDF

Video Time Density Function - Example

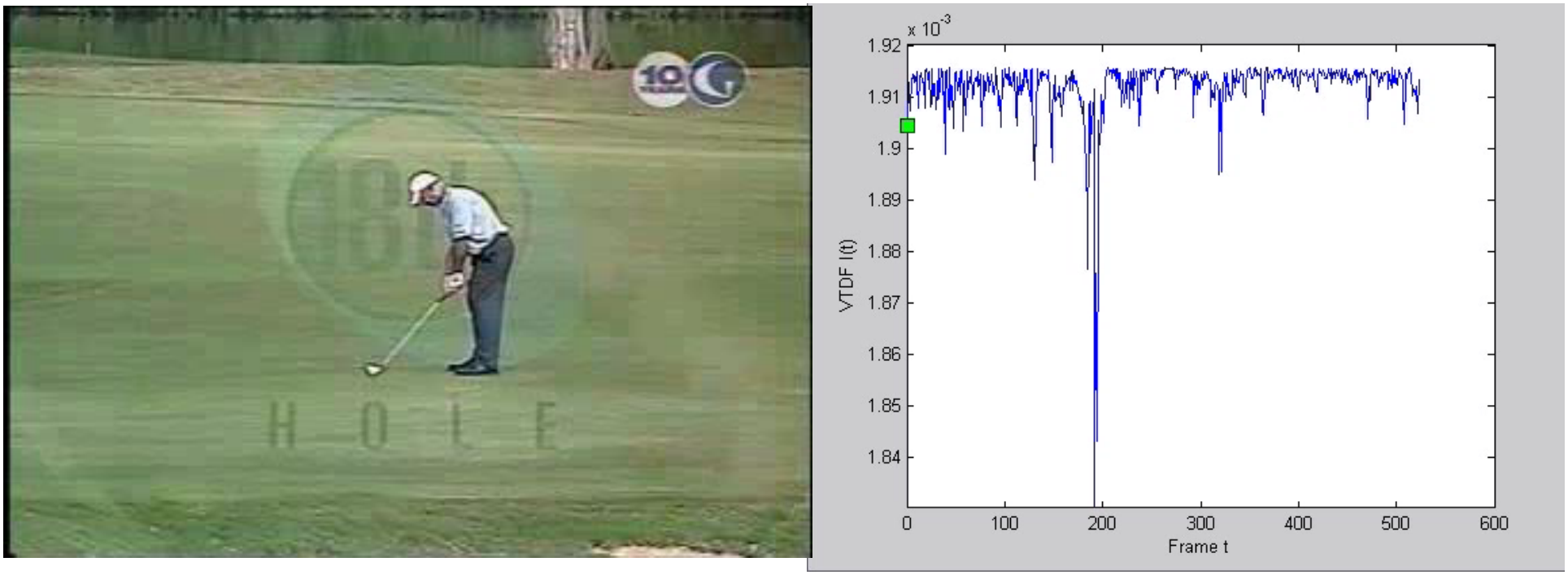


VTDF not only determines the quantity of information passed from one frame to another but also provides us a simple quantitative representation of video temporal dynamics

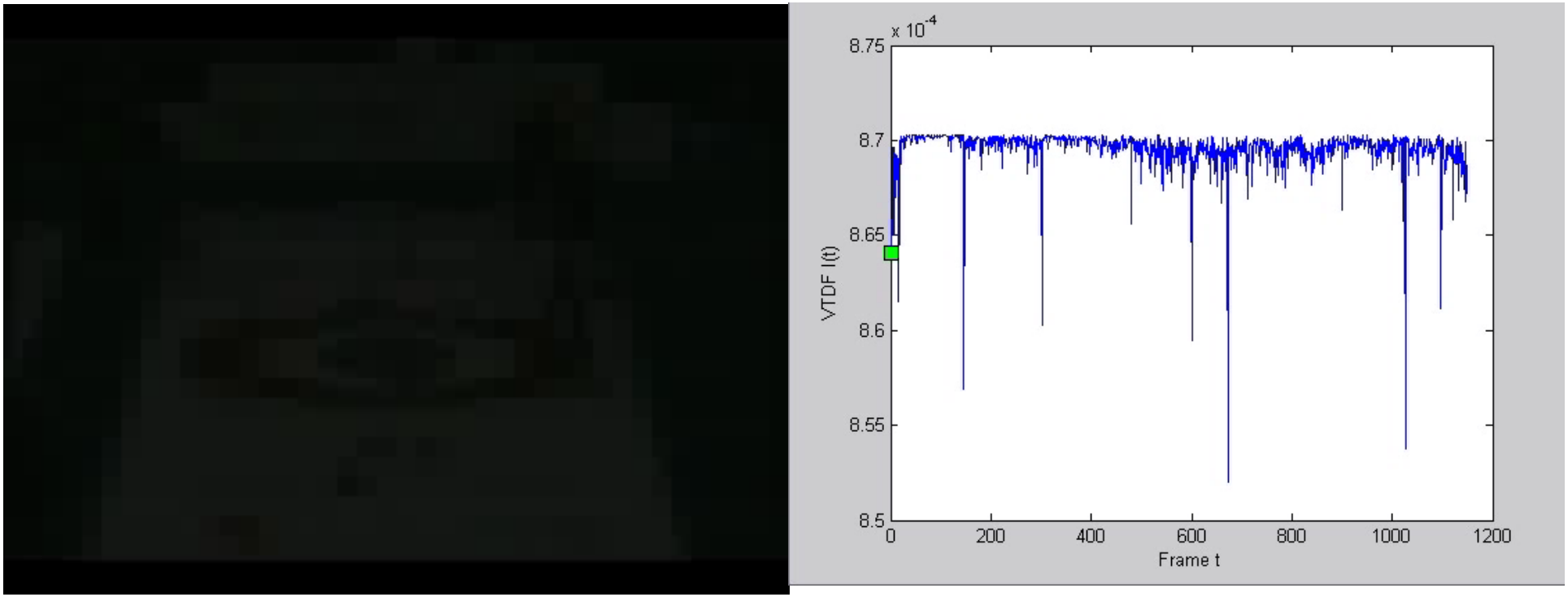
VTDF Demo


x 10^-3
1.26
1.25
1.24
1.23
1.22
1.21
1.2
VTDF(t)
0
100
200
300
400
500
600
700
800
900
Frame t

VTDF Demo


x 10^-3
1.92
1.91
1.9
1.89
1.88
1.87
1.86
1.85
1.84
VTDF I(t)
0
100
200
300
400
500
600
Frame t

VTDF Demo


x 10^-4
8.75
8.7
8.65
8.6
8.55
8.5
VTDF I(t)
0
200
400
600
800
1000
1200
Frame t

VTDF-based Temporal Quantization

- Quanta and boundary (2*M*-1 members):

$$q_1 < t_1 < q_2 < t_2 < \ldots < q_{M-1} < t_{M-1} < q_M$$

- Partition set: $Q_i = \{t \mid t_{i-1} \le t < t_i\}$
- Relation between quanta and partition:

$$t_1 = \frac{1}{2}(q_1 + q_2),$$
$$t_2 = \frac{1}{2}(q_2 + q_3)$$
$$\ldots$$
$$t_{M-1} = \frac{1}{2}(q_{M-1} + q_M)$$

- Evaluation -- TMSE (Temporal Mean Square Error):

$$TMSE = \frac{1}{M}\sum_{i=1}^{M} e(q_i) \qquad e(q_i) = \sum_{t \in Q_i} (t - q_i)^2 I(t)$$

Step 1: Find the Number of Segments

- VTDF-based video segmentation
  - VTDF-based temporal quantization is applied to find the optimal number of temporal segments
  - TMSE-based criterion

$$\frac{TMSE(M-1)-TMSE(M)}{TMSE(M-1)} < \varepsilon$$

where *TMSE*(*M*) is the quantization error to sample *M* frames, ε is a predefined small value.

  - The value of M to terminate the above formula will be chosen as the optimal number of temporal segments.

Step 2: Find Sample Frames

- Problem:

  For each temporal segment, how many frames to sample?

- Solution: Spatial quantization using Gaussian mixture
  - Gaussian mixture is effective to explore the mixture characteristics of video data in spatial domain.
  - Mixture model (Gaussian mixture)-based vector quantization
  - BIC criteria is used to determine how many frames will be sampled for each temporal segment

GMVQ

- Mixture model $f(x_t \mid \theta) = \sum_{j=1}^{K} \pi_j p_j(x_t \mid \theta_j)$
- Spatial segmentation $Q_{ij} = \{x_t \mid j = \arg\max_j (p_j(x_t \mid \theta_j))\}$
- Gaussian mixture model and EM

$$p_j(x_t \mid \theta_j) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} e^{-\frac{1}{2}(x_t - \mu_j)^T \Sigma_j^{-1} (x_t - \mu_j)}$$

$$\pi_j = \frac{N(Q_{ij})}{N(Q_i)}$$

$$\mu_j = \frac{1}{N(Q_{ij})} \sum_{x_t \in Q_{ij}} x_t$$

$$\Sigma_j = \frac{1}{N(Q_{ij})} \sum_{x_t \in Q_{ij}} (x_t - \mu_j)(x_t - \mu_j)^T$$

- Quantization error and termination

$$Q_{error} = \frac{1}{K} \left(\sum_{j=1}^{K} \frac{\sum_{x_t \in Q_{ij}} \left\| q_{ij} - x_t \right\|^2}{N(Q_{ij})} \right)$$

$$q_{ij} = \frac{1}{N(Q_{ij})} \sum_{x_t \in Q_{ij}} x_t$$

$$\frac{Q_{error}^{(l-1)} - Q_{error}^{(l)}}{Q_{error}^{(l-1)}} < \varepsilon$$

BIC Criterion

- BIC Criterion to find the optimal number of sample frames

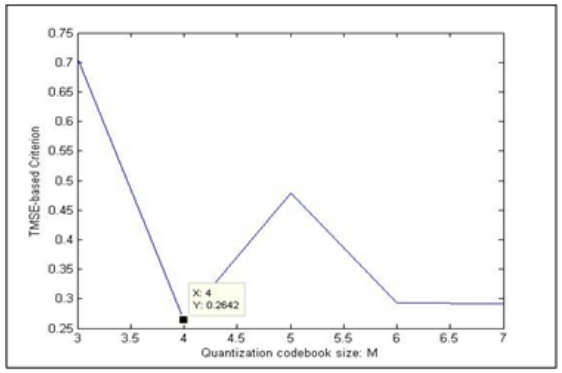
$$BIC = -2 \times \log f(x_t \mid \theta) + p \times \log K$$

- Sampling Frames

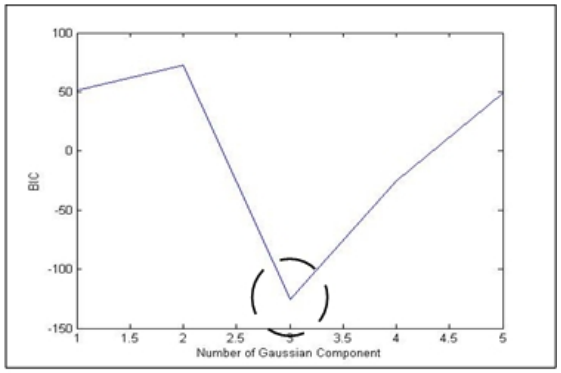
$$s_i = \{s_{i1}, \quad s_{i2}, \quad \ldots, \quad s_{iK}\} \qquad s_{ij} = \arg\min_{x_t \in Q_{ij}} \| x_t - q_{ij} \|^2$$

- The above Gaussian mixture vector quantization enables a compact and optimal representation of each temporal segment

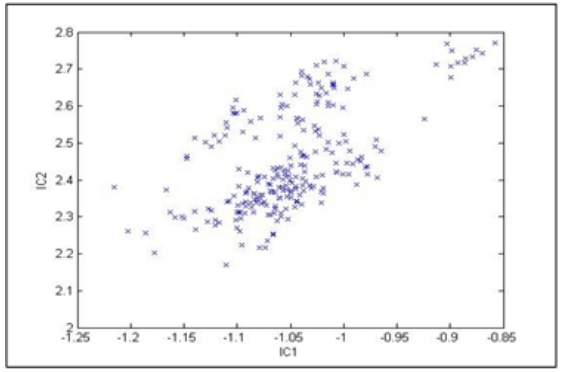
Experiment - Video Thumbnail Extraction



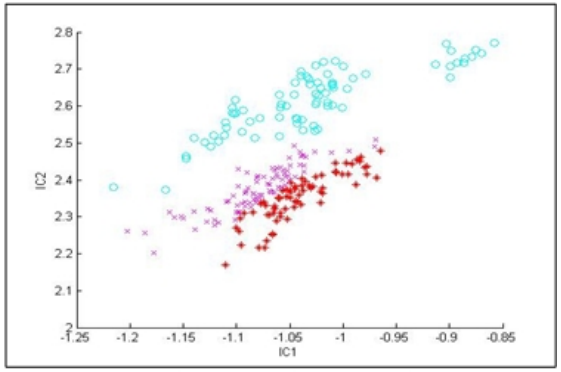
(a) TMSE-criterion for optimal number of temporal segments



(b) BIC-criterion for optimal number of Gaussian mixture components



(c) 2D ICA feature space for temporal segment 1



(d) GMVQ for temporal segment 1

(e) Video thumbnail for temporal segment1

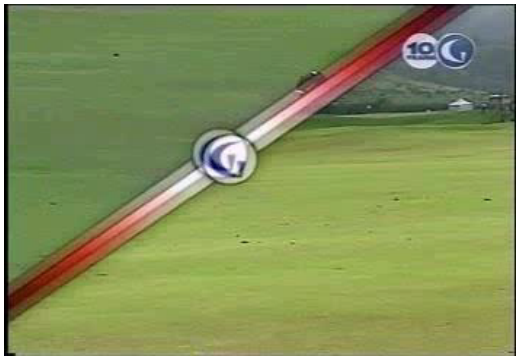
Summary:

The whole video has optimal number of temporal semgnents-4, optimal number of sample frames-3 for the first segment.

Experimental Results - 2

- Video thumbnails using key frame layout:

  First two rows--our method, last row--existing shot-based method-SVS, one frame per shot



Experimental Results-3

- Another Example-fireworks video

  Our method-4 frames, SVS method-1 frames

Experimental Results-4

- Video thumbnails example (file format: gif)
  - Example1-aquarium video (10 frames)

Experimental Results-5

- Example2-golf video (6 frames, gif format)

Original Golf Video

Video Thumbnails

Summary

- A new video thumbnail extraction method using spatiotemporal vector quantization is presented
- VTDF is used to model video data in time domain
- VTDF-based temporal quantization is employed to explore the temporal characteristics of video data and find the optimal number of temporal segments
- Gaussian mixture vector quantization is proposed to explore the spatial characteristics of video data and find the optimal number of sample frames for each temporal segment

Thank You