Motion-based Approach for BBC Rushes Structuring and Characterization

Chong-Wah Ngo, Zailiang Pan

Department of Computer Science
City University of Hong Kong

BBC Rushes

- *Rushes*
  - Unedited videos
  - Similar to home videos, but with better capturing skills and visual quality
- *Always….*
  - *Pan* to have another view of scene
  - *Zoom-and-hold* to freeze the impression
  - *Search* for something
  - Long take without camera motion
  - Pan to have panoramic view

BBC Rushes

- Intentional
  - Another view of scene
  - Impression
  - Something
  - Long take, panoramic view
- Intermediate
  - *Pan* to have….
  - *Zoom-and-hold* to freeze ….
  - *Search* for …..
  - A series of search, pan, zoom
- Shaking

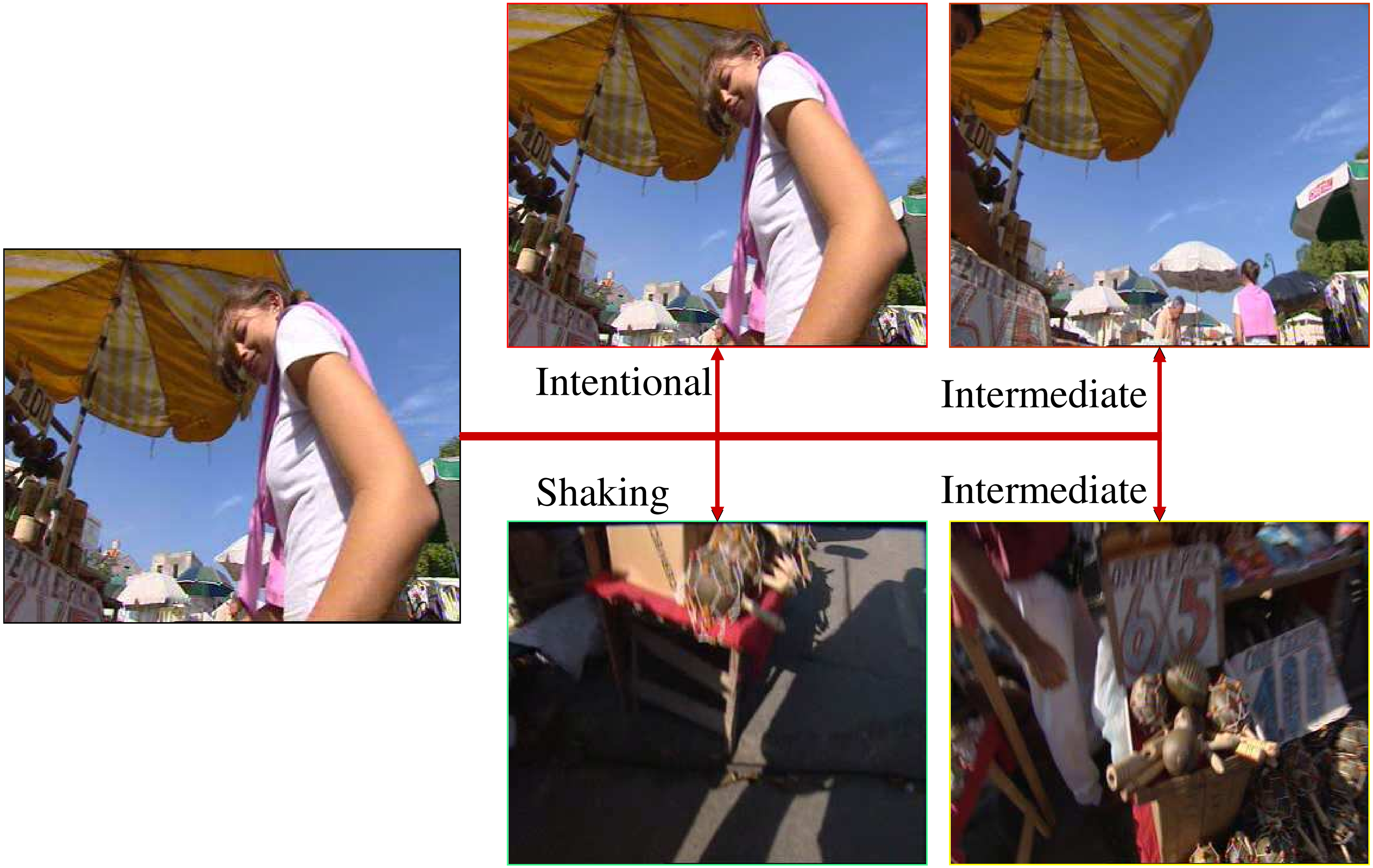
Our Intuition…

- Detecting intentions are useful for search, browsing and summarization
- Intermediate motions are not really meaningful for most tasks
- Shaking clips can be either useful or not useful

Objective

- To structure-and-characterize (or characterized-and-structure) video content, we propose
  - Finite State Machine (FSM)
  - Support Vector Machine (SVM)
  - Hidden Markov Model (HMM)

		FSM	*SVM*	*HMM*
I	Intentional	√	√	√
II	Intermediate	√	√	√
III	Shaking	√	√	√



Intentional

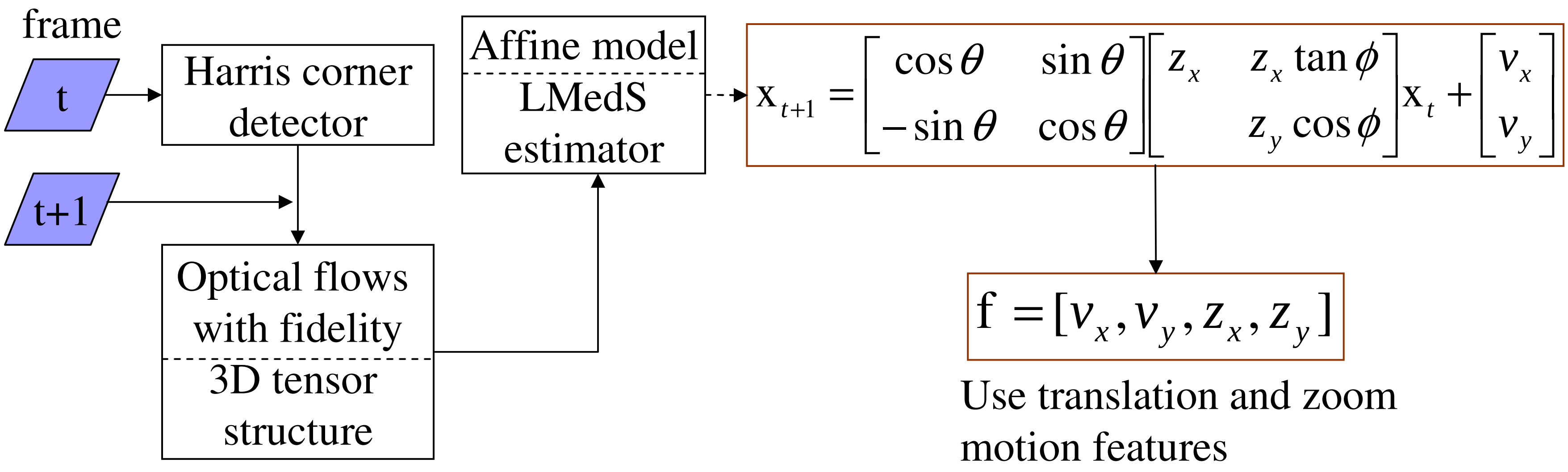
Intermediate

Shaking

Intermediate

Global Motion Estimation

- The motion-driven FSM, SVM and HMM are all based on the inter-frame global motion estimation. Considering the generalization and complexity, we choose to use the ***affine motion model***.



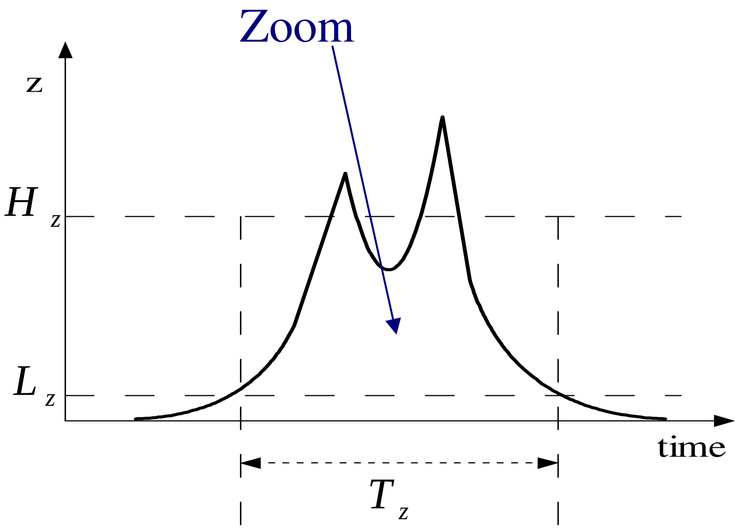
$$\mathrm{x}_{t+1} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} z_x & z_x \tan\phi \\ & z_y \cos\phi \end{bmatrix} \mathrm{x}_t + \begin{bmatrix} v_x \\ v_y \end{bmatrix}$$

$$\mathrm{f} = [v_x, v_y, z_x, z_y]$$

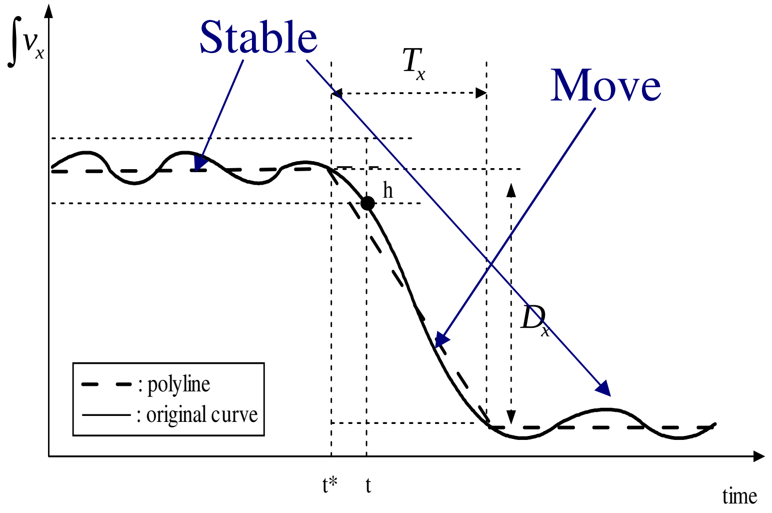
Use translation and zoom motion features

FSM—Partition

Zoom partition: The techniques of hysteresis thresholding are used for the zoom motion feature. Two thresholds are used: higher one for locating the position; lower one for the zoom partition ***Z***.

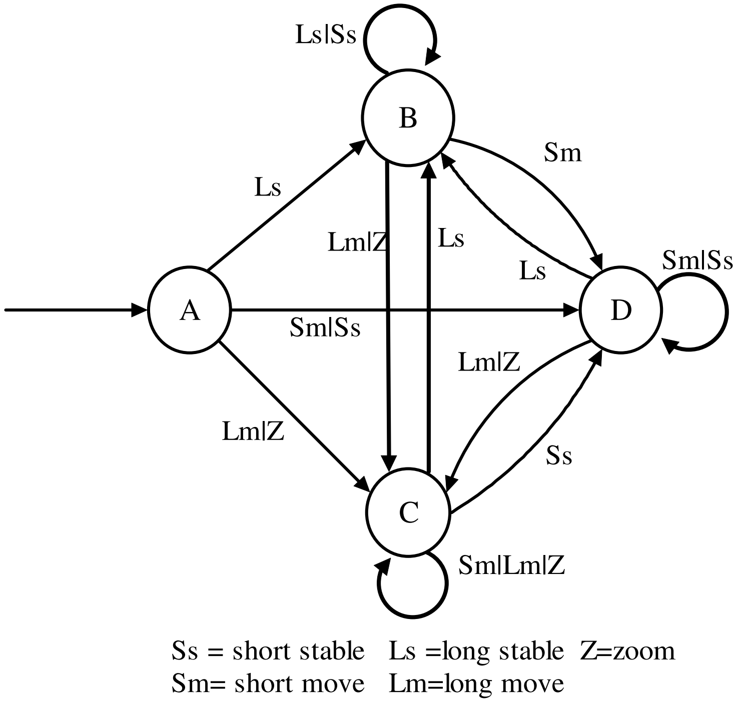


Static and move partition: A polyline is fitted to the camera trajectory using Kalman filter. Based on the properties of the lines, camera trajectory are partitioned into long stable ***Ls***, short stable ***Ss***, long move ***Lm*** and short move ***Sm***.



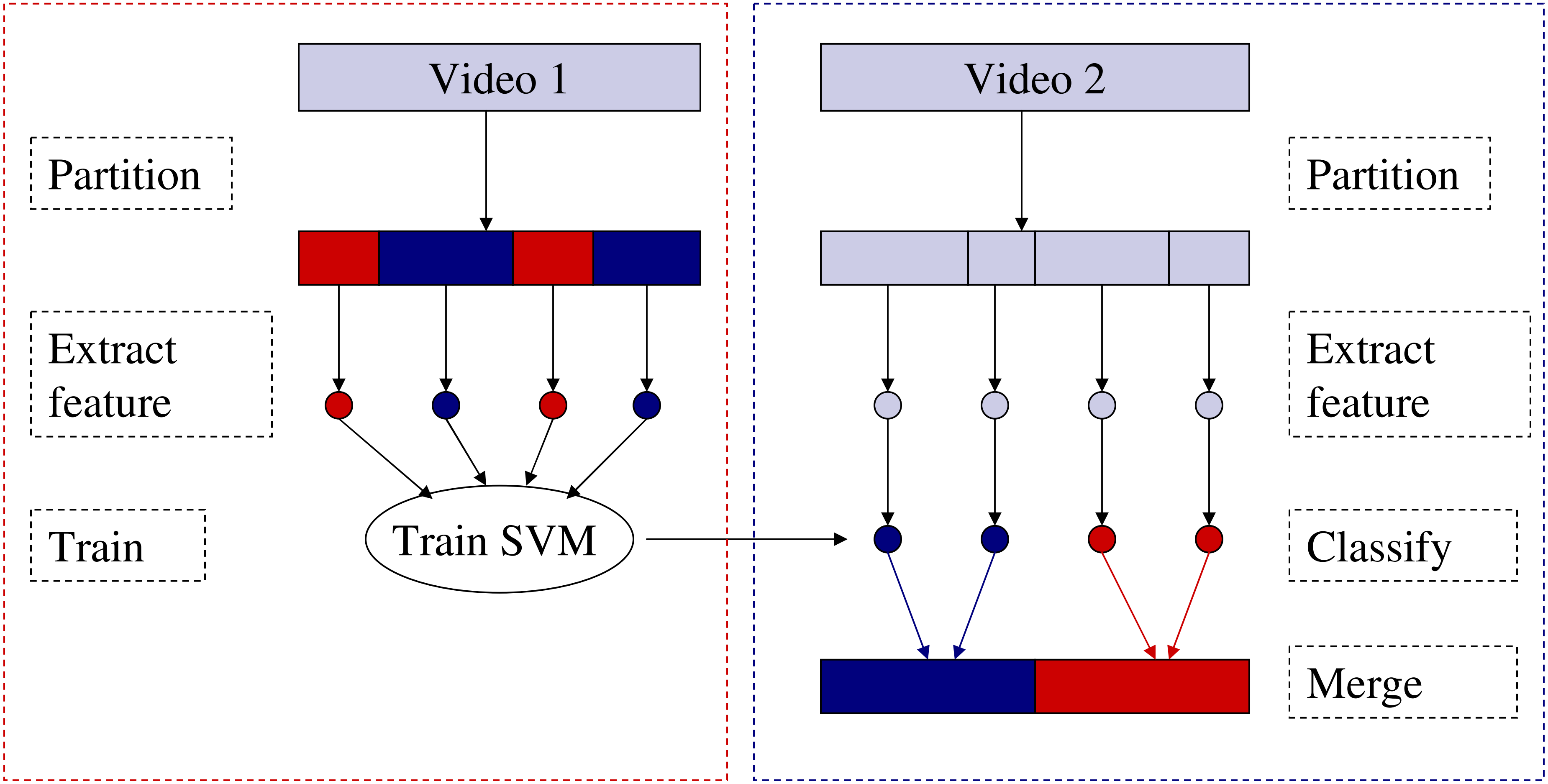
FSM—Classification

- A 4-state FSM is employed to refine the partition and characterize video.
  - A: initial state.
  - B: intentional motion.
  - C: intermediate and shaky motions. They are further separated by the rate of camera direction changes.
  - D: temporarily undetermined short segments.



Z. Pan and C.-W. Ngo, “Structuring home video by snippet detection and pattern parsing,” in *ACM SIGMM Int’l Workshop on MIR*, 2004.

Flowchart of SVM


Video 1

Partition

Extract feature

Train

Train SVM

Video 2

Partition

Extract feature

Classify

Merge

Train

Classify

SVM Implementation

- Partition: video is divided into segments of equal fixed duration.
- Feature extraction: 9 features from motion are extracted for each video segment. They are:

Speed: $M_x = \operatorname*{mean}_{i=1}^{N}(|v_i^x|), \quad M_y = \operatorname*{mean}_{i=1}^{N}(|v_i^y|)$

Zoom: $Z_x = \operatorname*{mean}_{i=1}^{N}(|z_i^x|), \quad Z_y = \operatorname*{mean}_{i=1}^{N}(|z_i^y|)$

Acceleration: $D_x = \operatorname*{mean}_{i=1}^{N-1}(|v_{i+1}^x - v_i^x|), \quad D_y = \operatorname*{mean}_{i=1}^{N-1}(|v_{i+1}^y - v_i^y|$

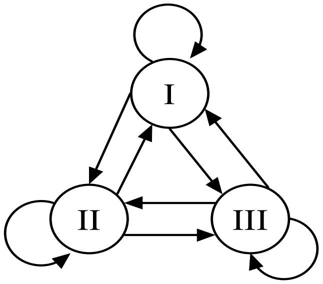
Acceleration variance: $V_x = \operatorname*{var}_{i=1}^{N-1}(|v_{i+1}^x - v_i^x|), \quad V_y = \operatorname*{var}_{i=1}^{N-1}(|v_{i+1}^y - v_i^y|)$

Motion change: $S = \operatorname*{mean}_{i=1}^{N-1}(|\mathbf{v}_{i+1}||\mathbf{v}_i| - \mathbf{v}_{i+1} \cdot \mathbf{v}_i)$

Motion change feature actually is $|\mathbf{v}_{i+1}||\mathbf{v}_i|(1 - \cos\theta)$, which considers both the angle change and motion magnitude.

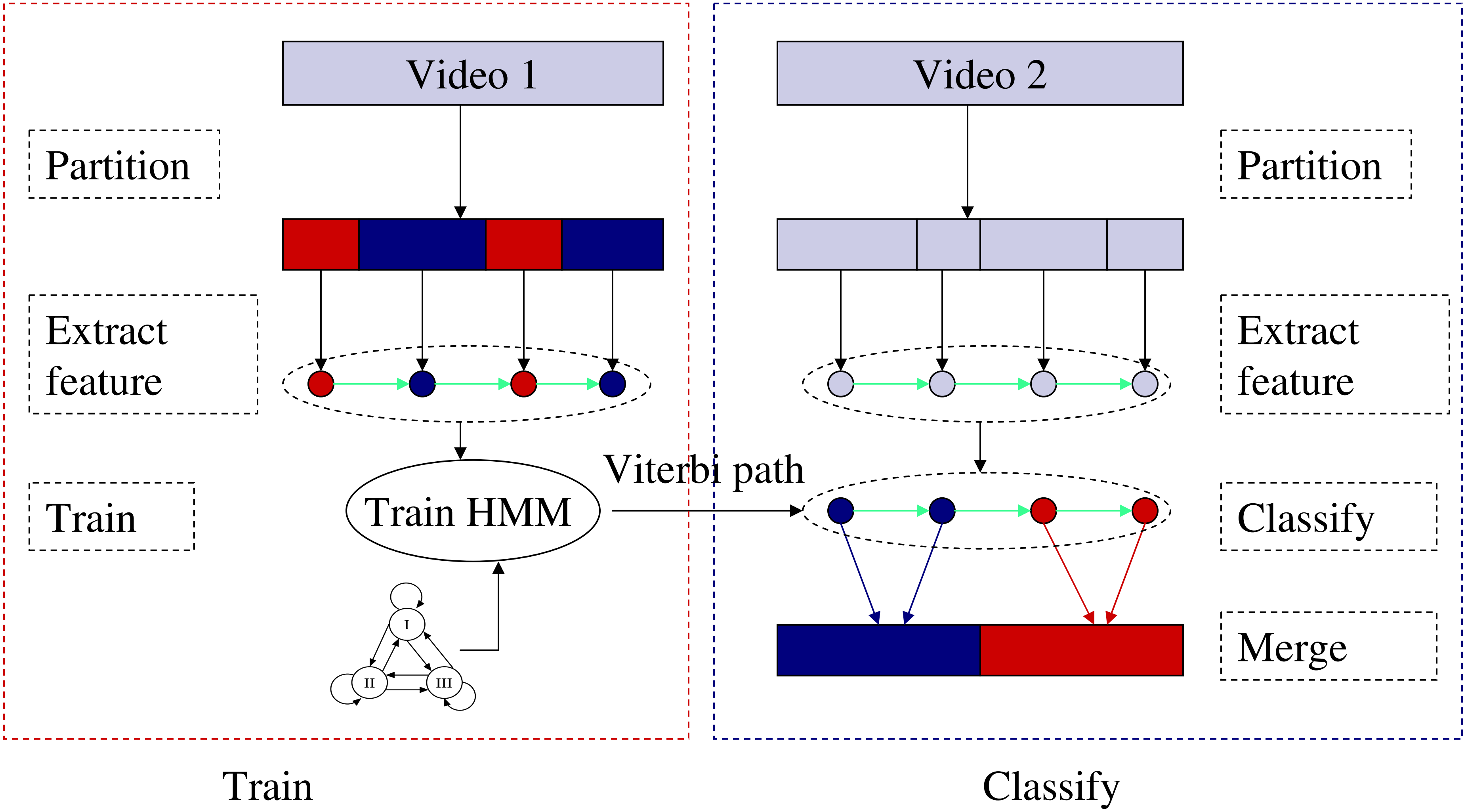
HMM-based Approach

- Motivation: ***First order decision*** (look at one sample and make decision at a time) may not be sufficient, ***Second order decision*** (look at multiple samples to make decision) should be better in principle.
- **Hidden Markov Model (HMM)** is then used as second order decision for video structuring and characterization.
  - HMM State transition → video structuring
  - HMM State prediction → video characterizing
- **3-state** hidden Markov model is used to represent respectively the intentional, intermediate and shaky motions.



HMM

Flowchart of HMM



MHMM & SHMM

- We investigate two kinds of HMM, called ***MHMM*** and ***SHMM***. The difference is,
  - MHMM (*m*otion-based):

    Partition: Video is divided into segments of equal fixed duration.

    Feature: Extract 9 features from motion.
  - SHMM (*s*hot-based):

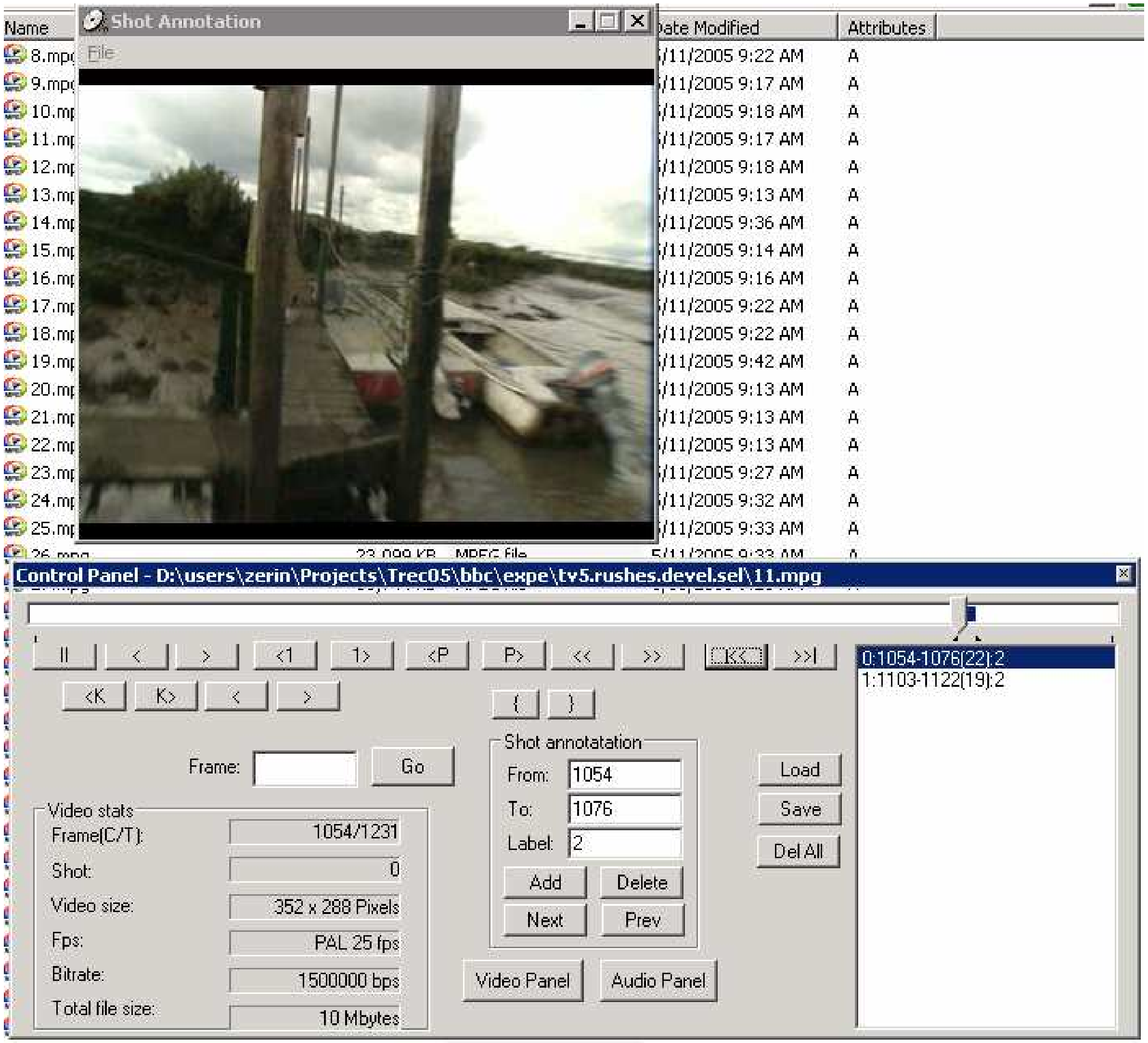
    Partition: Video is divided into shots by cut detector.

    Feature: Extract shot duration
  - *Note*: We use SHMM as baseline
    - Intuition: Short shots correspond to shaking/intermediate motion

Experiments – Data Set and Training

- 60 videos (337K frames) from the development set
- Manually annotate sub-shots and their characteristics
- 768 shots and 1135 sub-shots
- 30 videos for training and 30 videos for testing.

Annotation Tool



Shot Annotation

Control Panel - D:\users\zerin\Projects\Trec05\bbc\expe\tv5.rushes.devel.sel\11.mpg

Name	Date Modified	Attributes
8.mpc	/11/2005 9:22 AM	A
9.mpc	/11/2005 9:17 AM	A
10.mp	/11/2005 9:18 AM	A
11.mp	/11/2005 9:17 AM	A
12.mp	/11/2005 9:18 AM	A
13.mp	/11/2005 9:13 AM	A
14.mp	/11/2005 9:36 AM	A
15.mp	/11/2005 9:14 AM	A
16.mp	/11/2005 9:16 AM	A
17.mp	/11/2005 9:22 AM	A
18.mp	/11/2005 9:22 AM	A
19.mp	/11/2005 9:42 AM	A
20.mp	/11/2005 9:13 AM	A
21.mp	/11/2005 9:13 AM	A
22.mp	/11/2005 9:13 AM	A
23.mp	/11/2005 9:27 AM	A
24.mp	/11/2005 9:32 AM	A
25.mp	/11/2005 9:33 AM	A
26.mpg 23,099 KB MPEG file	5/11/2005 9:33 AM	A

Video stats

Frame(C/T):	1054/1231
Shot:	0
Video size:	352 x 288 Pixels
Fps:	PAL 25 fps
Bitrate:	1500000 bps
Total file size:	10 Mbytes

Shot annotatation

From: 1054

To: 1076

Label: 2

0:1054-1076(22):2
1:1103-1122(19):2

Approaches

	Segment Unit	Feature Number	Feature Types	Training	Decision
FSM	Sub-shot	4	Motion	No	1st
SVM	Equal duration	9	Motion	Yes	1st
MHMM	Equal duration	9	Motion	Yes	2nd
SHMM	Cut	1	Time	Yes	2nd

1st : look at one sample and make decision at a time

2nd: look at multiple samples to make decision

Experiment – Structuring

- Sub-shot boundary detection
- A sub-shot boundary is counted as correct as long as we can find a matched ground-truth boundary within 1 second.

	Training		Testing	
	Recall	Prec.	Recall	Prec.
FSM	0.614	0.282	0.593	0.279
SVM	**0.769**	0.281	**0.763**	0.289
MHMM	0.461	**0.419**	0.395	**0.379**
SHMM	0.060	0.355	0.056	0.322

Results of structuring BBC rushes

Experiment – Characterization

- Sub-shot classification
- Use frame as basic unit for evaluation

	Intentional		Intermediate		Shaky	
	Recall	Prec.	Recall	Prec.	Recall	Prec.
FSM	0.815	0.981	**0.802**	0.118	0.011	0.050
SVM	0.827	**0.990**	0.701	**0.162**	**0.715**	0.239
MHMM	**0.927**	0.970	0.329	0.137	0.311	**0.339**

Results of characterizing BBC rushes (training videos)

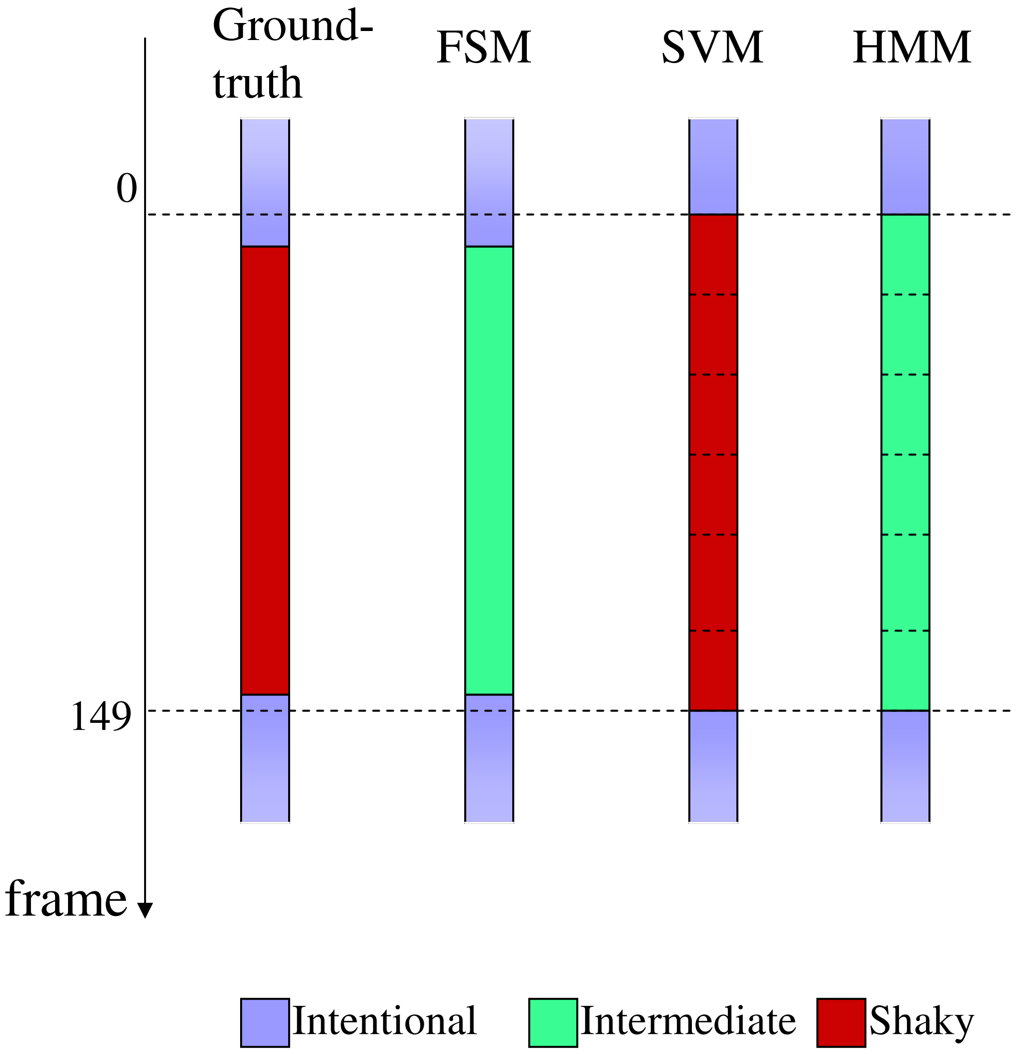
Experiment – Characterization Cont’

- 30 testing videos

	Intentional		Intermediate		Shaky	
	Recall	Prec.	Recall	Prec.	Recall	Prec.
FSM	0.756	0.968	**0.844**	0.128	0.000	0.000
SVM	0.778	**0.975**	0.456	0.120	**0.362**	**0.182**
MHMM	**0.909**	0.929	0.375	**0.196**	0.043	0.067

Results of characterizing BBC rushes (testing videos)

Example


Ground-truth
FSM
SVM
HMM
0
149
frame
Intentional
Intermediate
Shaky

Summary

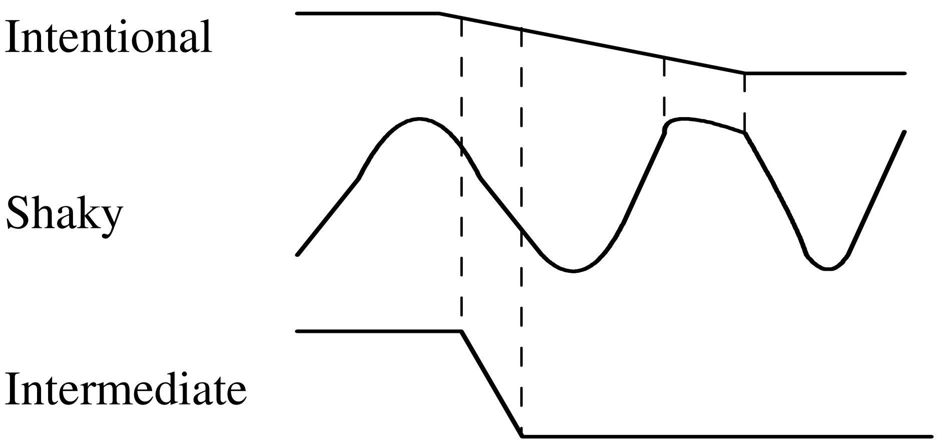
- For structuring, SVM gives the best recall (above 75%), followed by FSM (about 60%); the performances of MHMM and SHMM are poor.
- For characterization:
  - HMM performs best for extracting intentional motion
  - FSM performs best for intermediate motion detection
  - On average, SVM is best for three characteristics.
- Several problems remain difficult and challenging

FSM—Limitation

- For FSM, the following issues should be considered.
  - The threshold is difficult to set empirically to distinguish between intentional and intermediate. For example, “panorama view” or “pan to search”?
  - The use of rate of directional changes as features for separating shaky and intermediate motions is poor.

SVM—Limitation

- For SVM, the following sorts of segments are ambiguous by just looking at small time frame:
  - A panoramic or “pan to search”?
  - “Pan to search” or one part of a shaky?
  - A relative stable part of a shaky or intentional?



MHMM—Limitation

- More works can be done in HMM:
  - Only one state is not enough to represent the intentional, intermediate or shaky characteristic, e.g.
    - “Intermediate” may have two sub-state: “pan to search” and “zoom-and-hold”
    - “Shaky” may have sub-states such as “shake left”, “shake right”, “shake up” , “ shake down”.
  - State “intentional: is over trained since sequences has more intentional than intermediate/shaky segments. Over-trained “intentional” state compresses the detection of other two types, especially shaky.

More on Characteristic of BBC Rushes…

I. **Intentional**

II. **Intermediate Motion**

III. **Shaky Motion**

IV. **Blur**

- motion blur, defocusing blur

V. **Illumination Change**

Challenge in Motion Estimation

- Camera motion estimation is difficult for cases like blur, illumination and large foreground objects

Blur

Illumination

Foreground object

Future Work

- Detecting segments with blur and sharp/inconsistent illumination changes –
    - facilitate browse/search/summarization
    - Motion estimation can be an easier task
- Consider variants of SVM and HMM models for more accurate structuring and characterization.