
Video Matching Using Spatiotemporal Volumes

Arslan Basharat, Yun Zhai, and Mubarak Shah

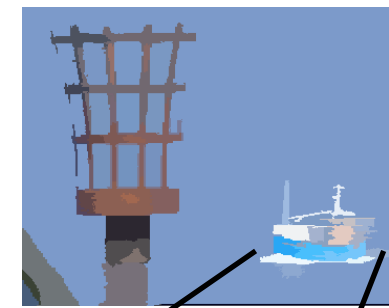
Computer Vision and Image Understanding (CVIU) 2008

Motivation

- Region based video matching and retrieval
- Typical technique for region detection
 - Color segmentation
 - Key-frame or video based
 - Illumination sensitive
 - Limited semantic
 - Motion segmentation
 - Optical flow commonly used
 - Limitations of small motion, aperture problem etc.



Matching video shots?



Color Segments



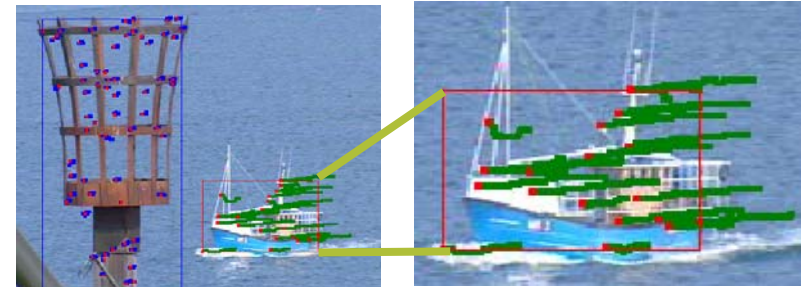
1 Object → 19 Color Segments



1 Object → 14 Color Segments

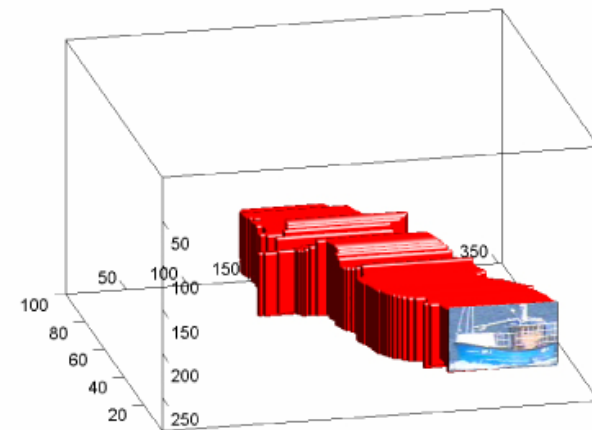
Our Approach

- Motion segmentation using
 - Reliable SIFT operator
 - SIFT based trajectories
- Extract motion volume
- Volume features
- Feature similarity
- Video similarity
- Application: Video Retrieval



SIFT Based Motion Segments

1 Object \rightarrow 1 Motion Segment



1 Object \rightarrow 1 Spatiotemporal Volume

Related Work

■ Region based image retrieval

□ Color segment based

- J. Z. Wang, J. Li, and G. Wiederhold. SIMPLcity: Semantics-sensitive integrated matching for picture libraries. IEEE TPAMI, vol. 23:947963, 2001
- F. Jing, M. Li, H. J. Zhang, and B. Zhang. Region-based relevance feedback in image retrieval. In Proc. IEEE ISCAS, 2002.

■ Video retrieval

□ Key-frame based

□ Color segment based regions

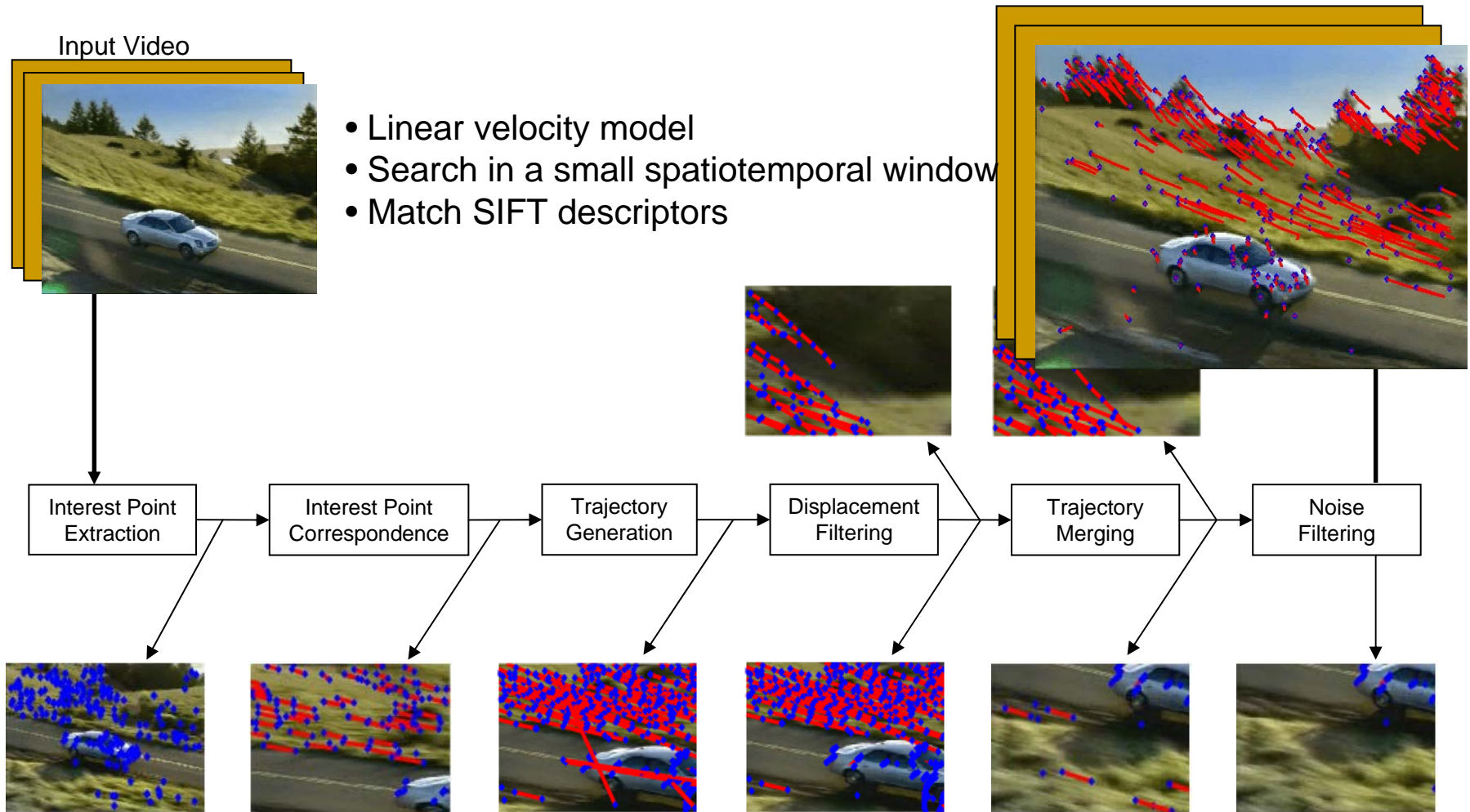
- J. Lee, J. Oh, and S. Hwang. Strg-index: Spatiotemporal region graph indexing for large video databases. SIGMOD, 2005.

□ Motion features based

- S.-F. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong. "VideoQ: An automated content-based video search system using visual cues". ACM MM, 1997.
 - Sivic, J. , Schaffalitzky, F. and Zisserman, A. "Object Level Grouping for Video Shots". International Journal of Computer Vision (2006)
-

Trajectory Generation

Output: Interest points based Trajectories



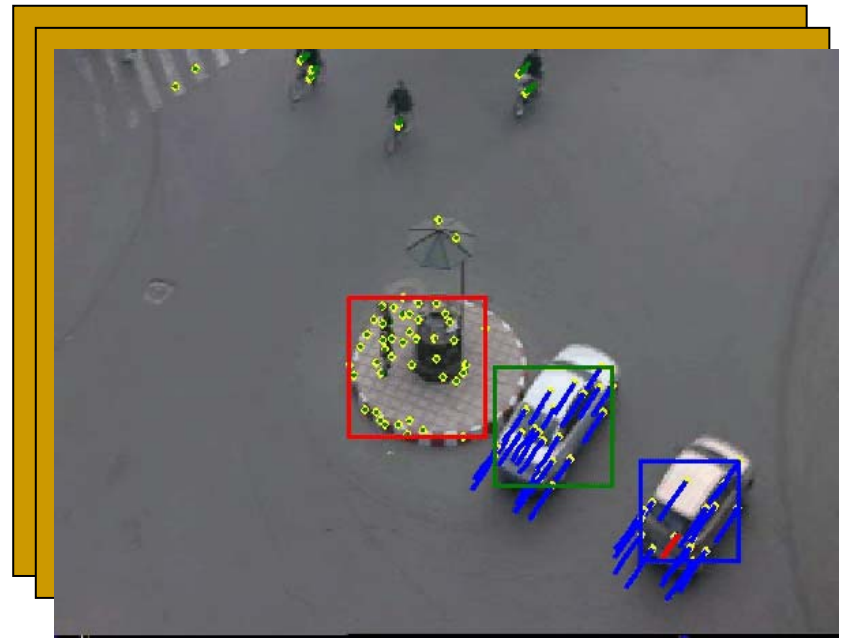
Temporal Trajectory Clustering

- Motion segmentation
- Use reliable trajectories
- For every frame
- Use RANSAC to estimate different homographies
 - Randomly choose 4 points
 - Compute homography H
 - Keep homography with largest set of inliers
 - Repeat until stopping condition met



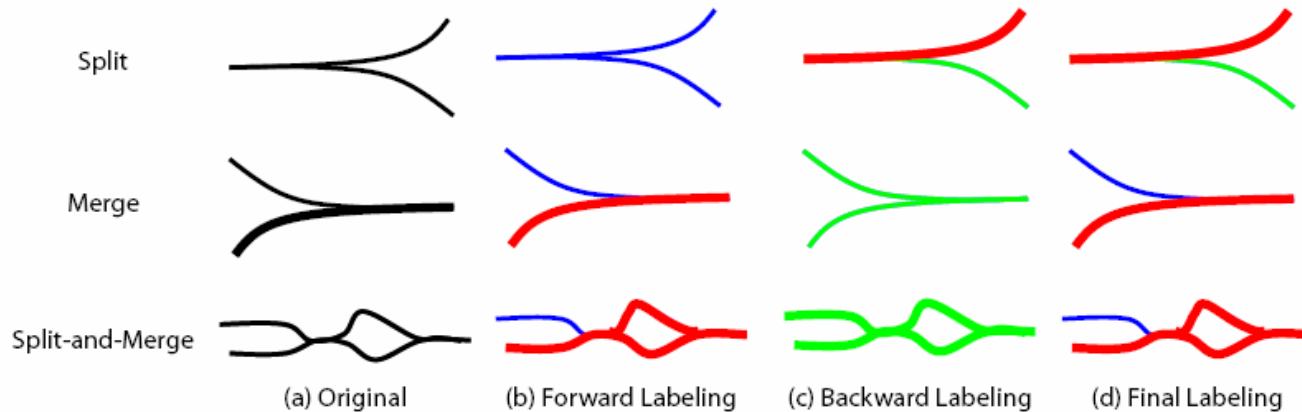
Spatial Trajectory Clustering

- Multiple objects with similar motion
- Spatial regions from motion segments
- Use interest points
- Spatial proximity constraint
- Output: Detected region

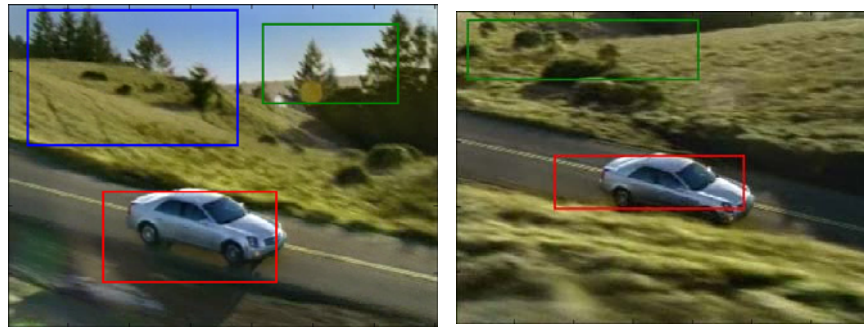


Motion Region Correspondence

- Temporal window per frame
- Common trajectories provide constraint
- Maximum common trajectories resolves correspondence
- Advantages:
 - Handles split and merge
 - Robust to noise

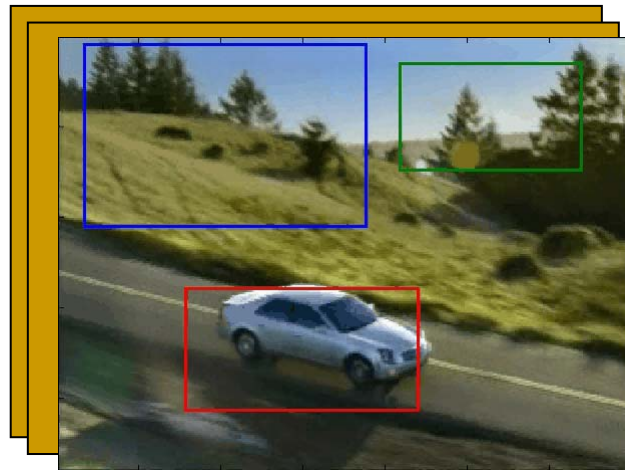


Tracked Regions



Detected Regions,
first frame

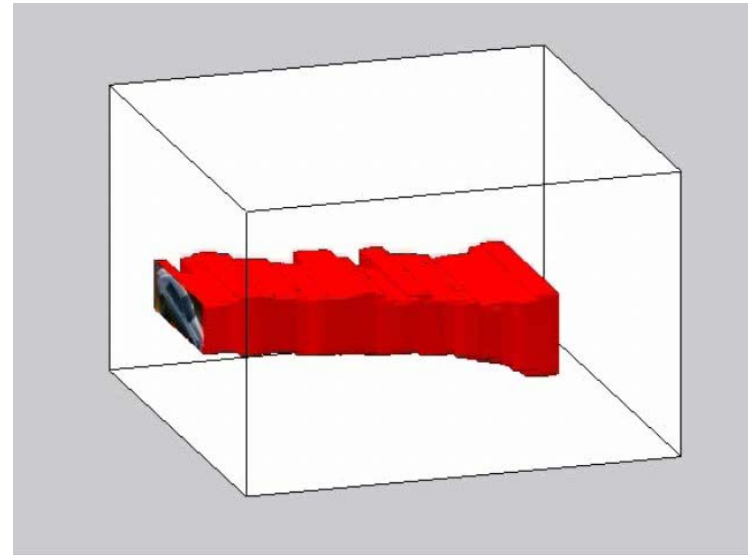
Similar Regions Merged,
last frame



Output

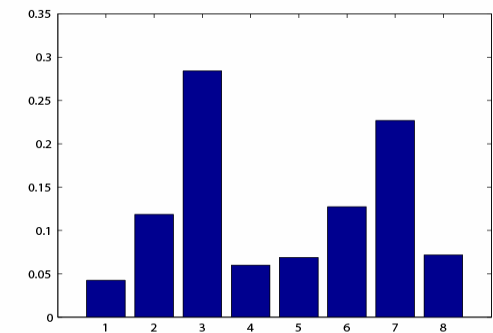
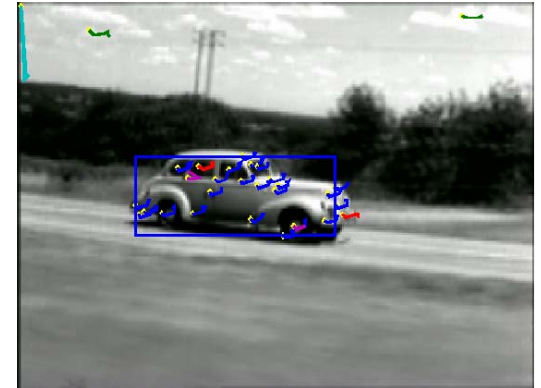
Volume Features

- Local volume features
- Invariant to volume size
- Features
 - Color
 - Texture
 - Motion
 - Interest point descriptors
- Earth Mover's Distance (EMD) for feature similarity



Volume Features

- Color
 - 3D HSV histogram
 - 64 normalized bins
- Texture
 - Gradient direction of Canny edges
 - Histogram of edge orientations
 - 8 bins and normalized
 - Similarity by EMD

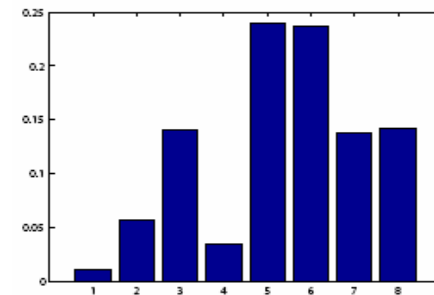


Volume Features

- Motion
 - Interest point based trajectories
 - Use velocity at every time instance
 - Histogram of velocity
 - 8 bins of direction and normalized
 - Similarity by EMD



Input Sequence



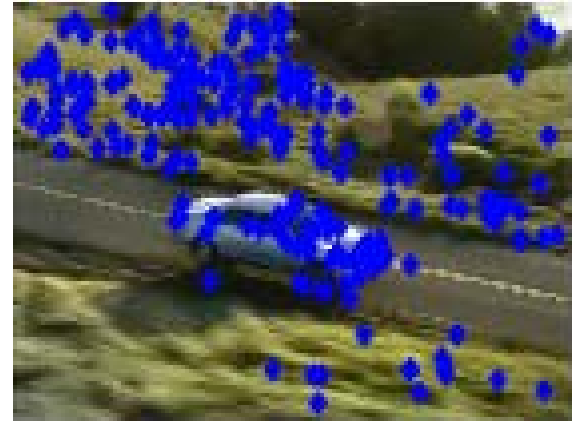
Velocity Histogram

Volume Features

- Interest point descriptors
 - 128 dimensional SIFT descriptors
 - One representative descriptor per trajectory
 - Generate signature using iso-data clusters

$$sig = (m_i, w_i)$$

where, m_i is cluster mean, w_i is cluster population



Earth Mover's Distance (EMD)

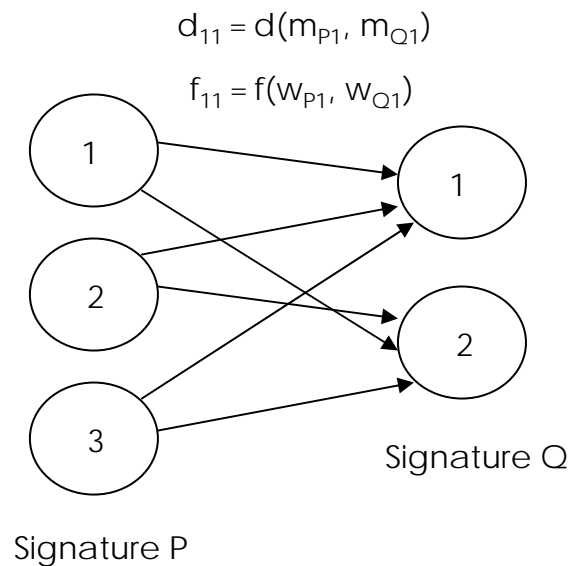
- Based on traffic flow problem
- A signature is group of clusters represented by mean and population
- Earth Mover's Distance (EMD) between signatures

$$EMD(P, Q) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}}$$

where, d_{ij} is cost and f_{ij} is flow between clusters i and j

- Signature similarity from EMD

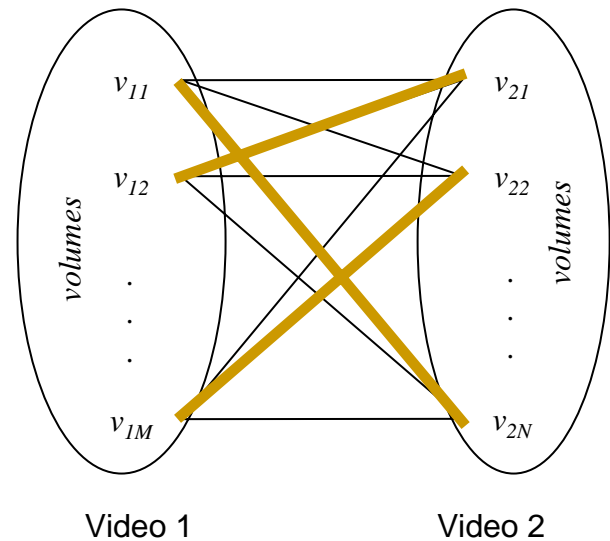
$$sim(P, Q) = \exp\left(-\frac{EMD(P, Q)^2}{2\sigma^2}\right)$$



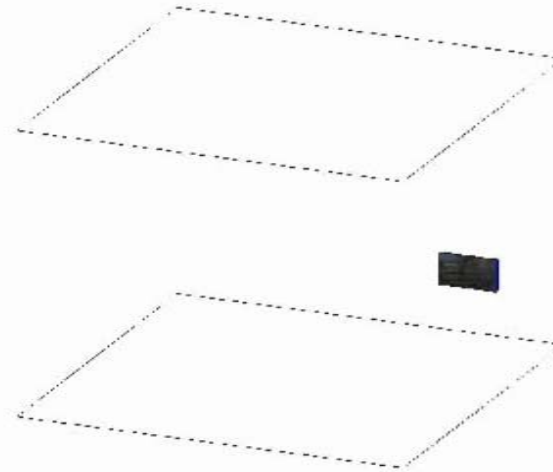
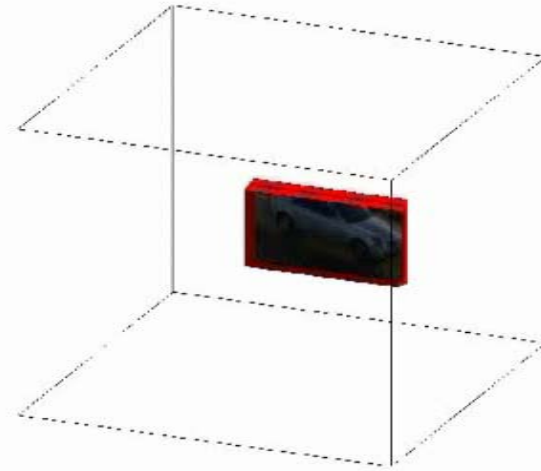
$$sig = (m_i, w_i)$$

Feature Matching

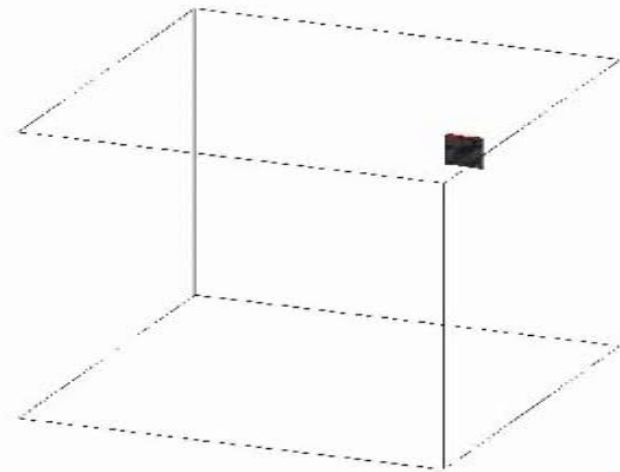
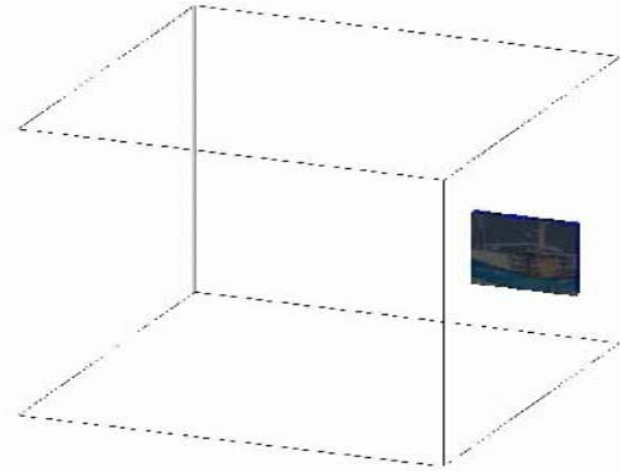
- Volume similarity matrix S_{12}^k for feature k
- Weighted combination of K features
$$S_{12} = \sum_{k=1}^K S_{12}^k w_k$$
- Bi-partite graph of volumes as vertices, and similarity as edge weights
- Maximum matching \rightarrow Volume correspondence
- Mean of maximum matches \rightarrow Video matching Score



Results - Volume extraction

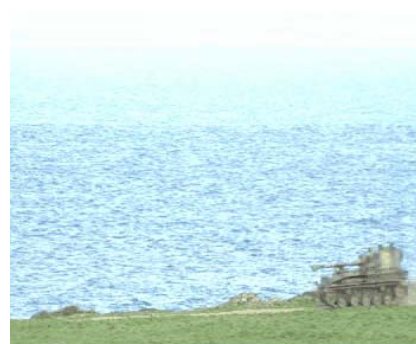
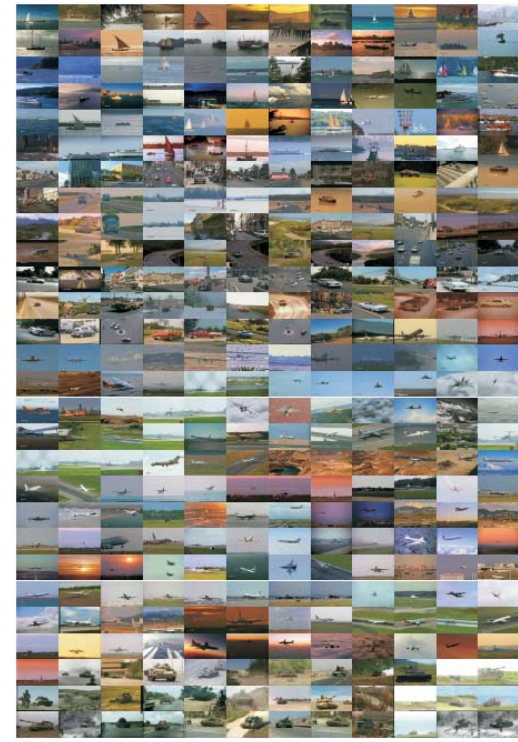


Results - Volume extraction



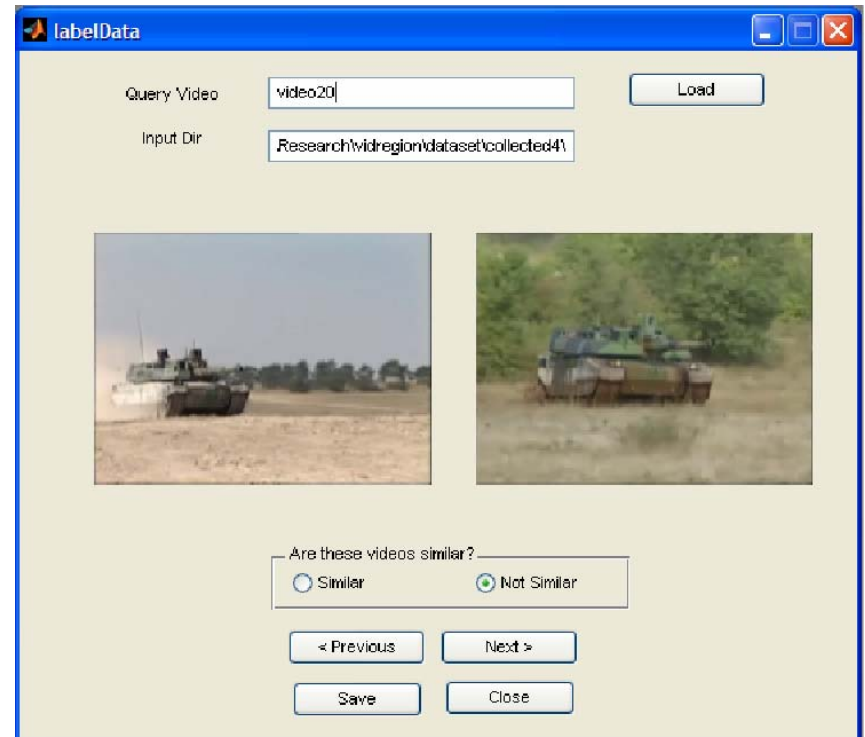
Application – Video Shot Retrieval

- Experiments performed on footage available online
- 337 video shots in the dataset
- 150 frames per shot on average
- Rigid moving objects like vehicles, planes etc.



Application – Video Shot Retrieval

- User specific scene interpretation
- Annotation Tool
- Used for quantitative analysis of retrieval
- DEMO



Application – Video Shot Retrieval

- Query Video



- Ranked results



Application – Video Shot Retrieval

■ Mean Average Precision

$$Precision = \frac{\{SimilarVideos\} \cap \{RetrievedVideos\}}{\{RetrievedVideos\}},$$

$$Recall = \frac{\{SimilarVideos\} \cap \{RetrievedVideos\}}{\{SimilarVideos\}},$$

$$AveragePrecision = \frac{\sum_{i=1}^m Precision(r)\delta(r)}{\{SimilarVideos\}},$$

Annotator ID	Average precision					Mean average precision
	Video20	Video18	Video51	Video117	Video5	
User1	0.69	0.59	0.75	0.71	0.74	0.70
User2	0.70	0.56	0.75	0.71	0.75	0.69
User3	0.68	0.61	0.72	0.69	0.76	0.69
User4	0.71	0.49	0.71	0.70	0.77	0.68
User5	0.71	0.55	0.77	0.72	0.76	0.70

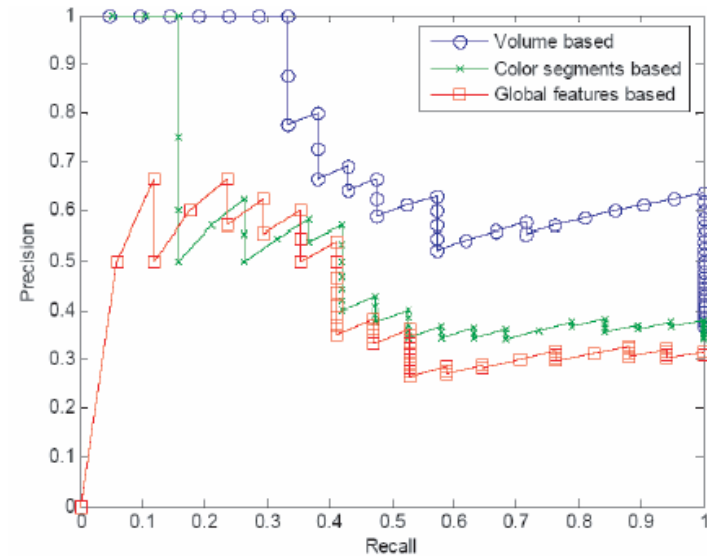
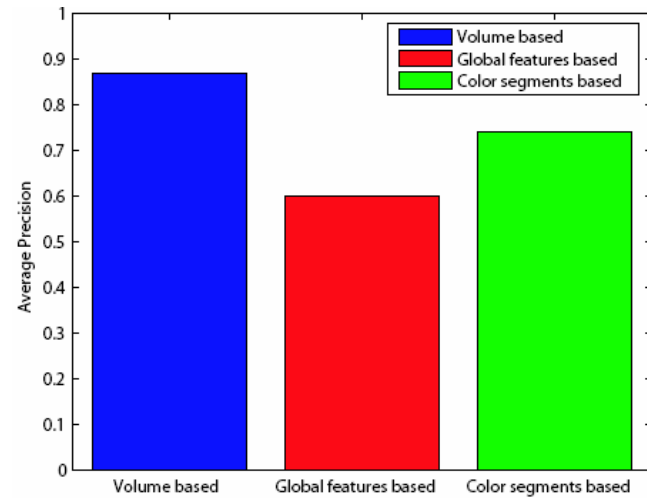
Comparison

- Comparison with two other approaches
 - Key-frame based approaches
 - Global features based
 - Color and texture features of complete image
 - Color segments based
 - Color and texture features of color regions
 - Region correspondence computed like volume correspondence
-

Results – Video Shot Retrieval



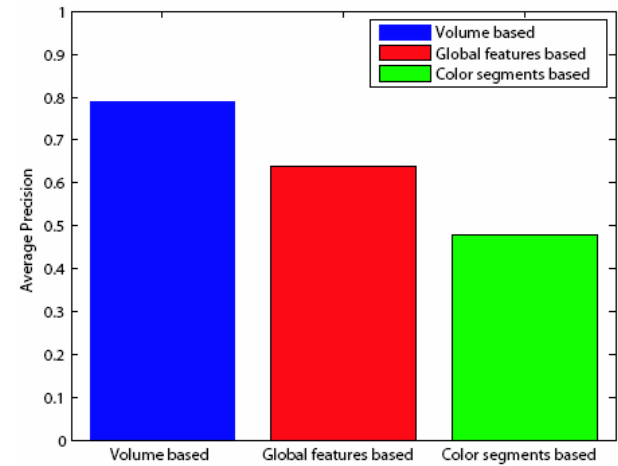
Query Video



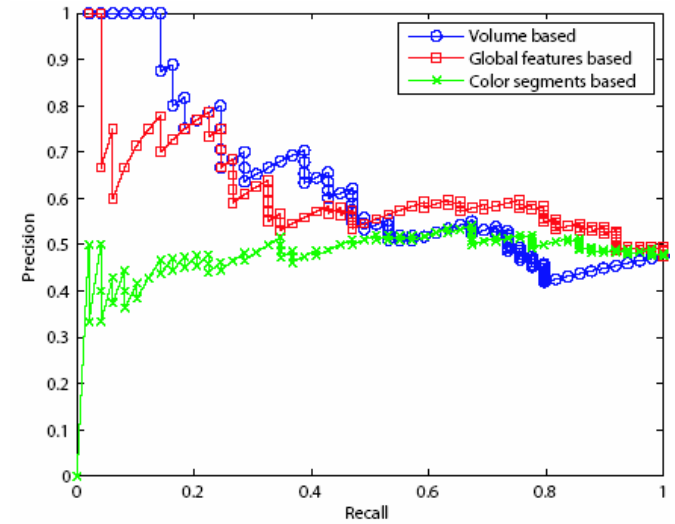
Results – Video Shot Retrieval



Query Video



(d) Average precision

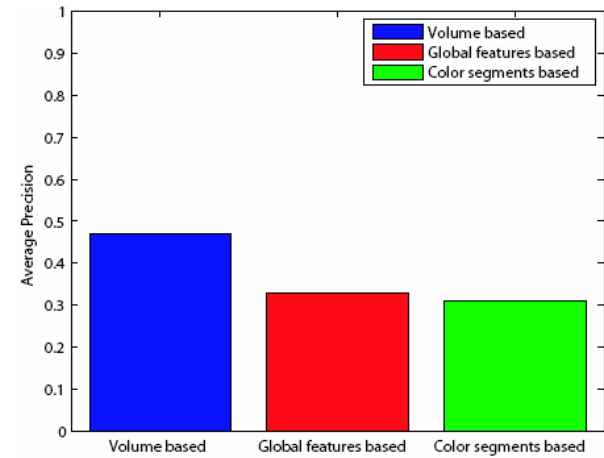


(c) Precision recall curves

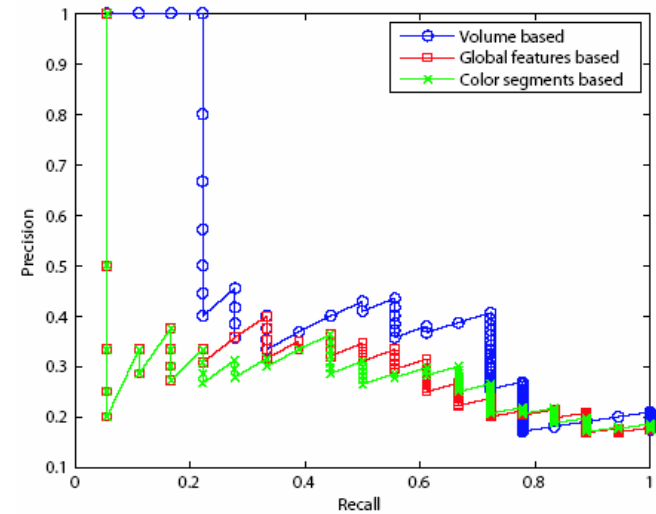
Results – Video Shot Retrieval



Query Video



(d) Average precision

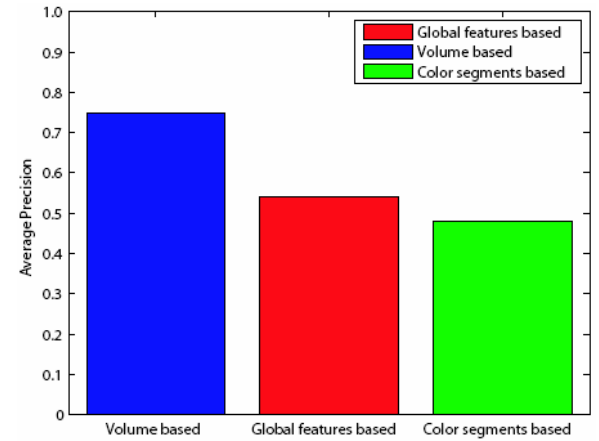


(c) Precision recall curves

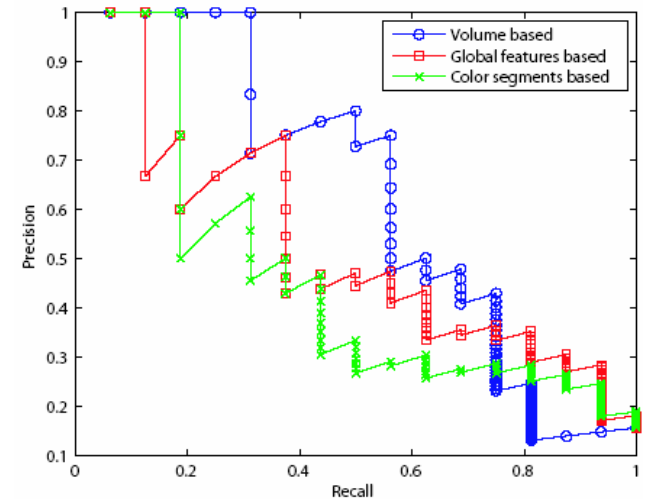
Results – Video Shot Retrieval



Query video

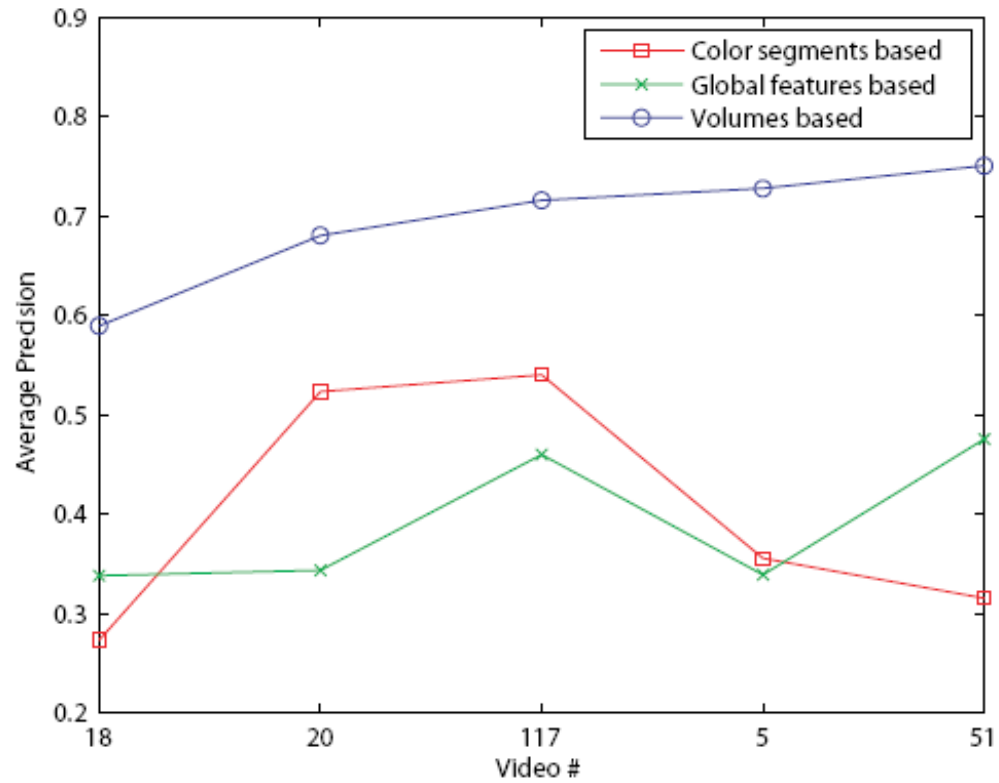


(d) Average precision

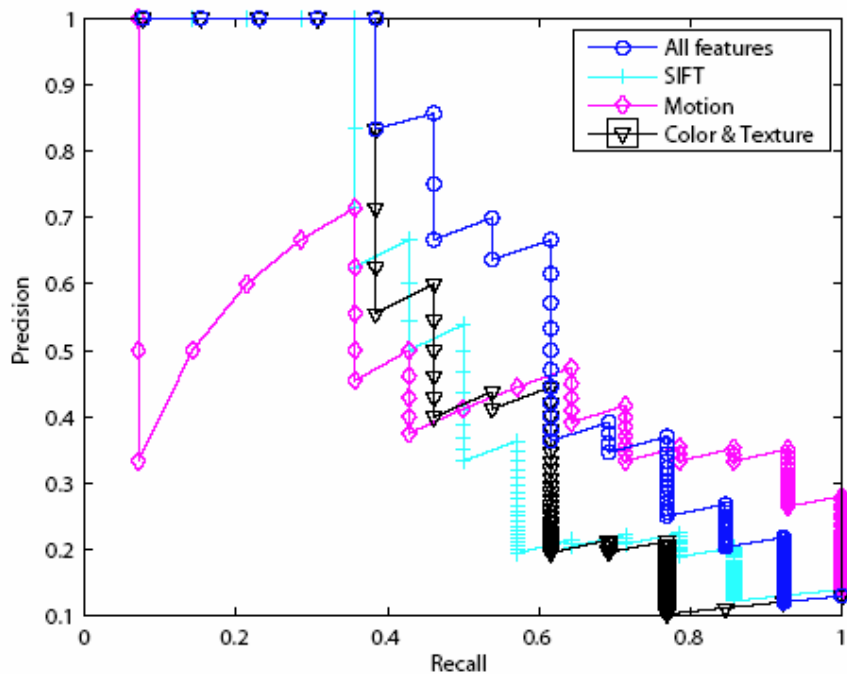


(c) Precision recall curves

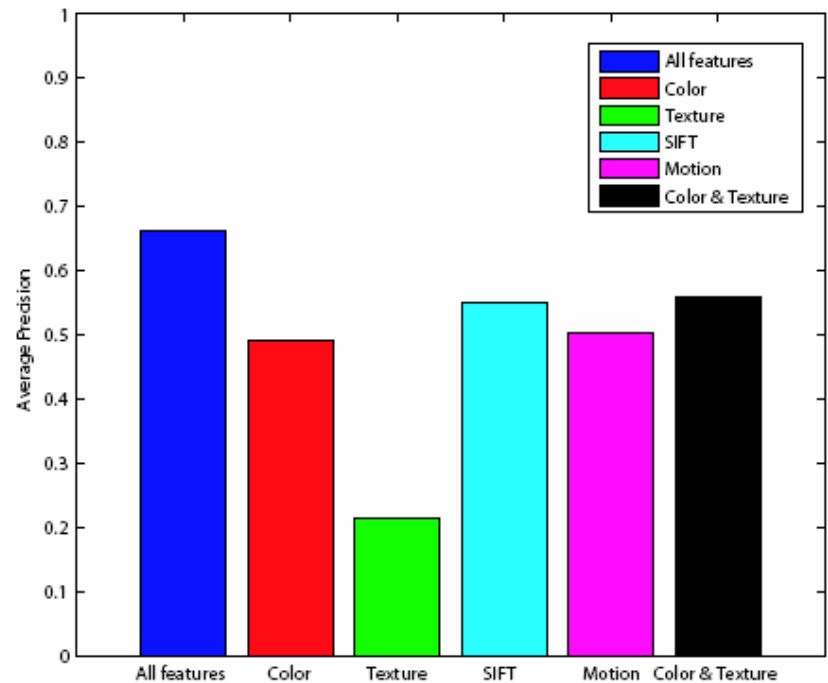
Comparison



Volume Features



(a) Precision recall curves



(b) Average precision

Conclusion

- SIFT based motion segmentation is robust
 - Volumes correspond to real world objects
 - Video matching framework useful for
 - Unsupervised video clustering
 - Content based video indexing and retrieval
 - Future directions
 - Intra cluster similarity analysis
 - Explore stronger visual features
-

Thank You!
