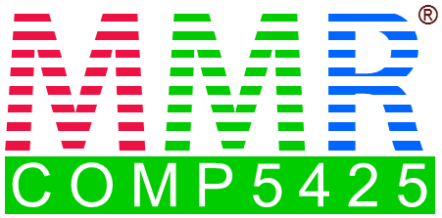


MULTIMEDIA RETRIEVAL

THE UNIVERSITY OF SYDNEY

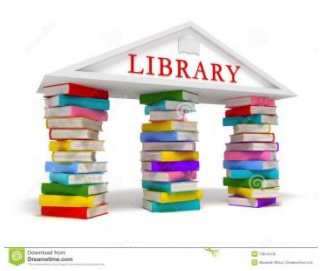
Week10

Semester 1, 2025

Information Summarization

- Text summarization
- Video Summarization
- Applications
  - LifeLogging
  - Scene summarization
  - StoryImaging

Information Deluge



Approximately 3.5 trillion photos have been taken since Daguerre captured Boulevard du Temple 174 years ago

http://blog.1000memories.com/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox

Information Deluge

6 billion (Aug 2011)

- 192 years to view all of them (1s per image)
- **3000+** uploads/minute
- 2% Internet users visit (2009)
- Daily time on site: 4.7 minutes (2009)

690 million (Mar 2012)

- 3,450 years to see all of them
- **48** hours uploaded/minute (2012)
- 20% Internet users visit (2009)
- Daily time on site: 23 minutes (2009)
- 2007 bandwidth = entire Internet in 2000
- 3B+ views per day (2012)

100 billion (Middle of 2011)

- 3,200 years to view all of them (1s per image)
- ~200M uploads/day; ~ 6B/month (2012)
- 800+M users (Dec 2011)
- Daily time on site: 30 minutes (2009)

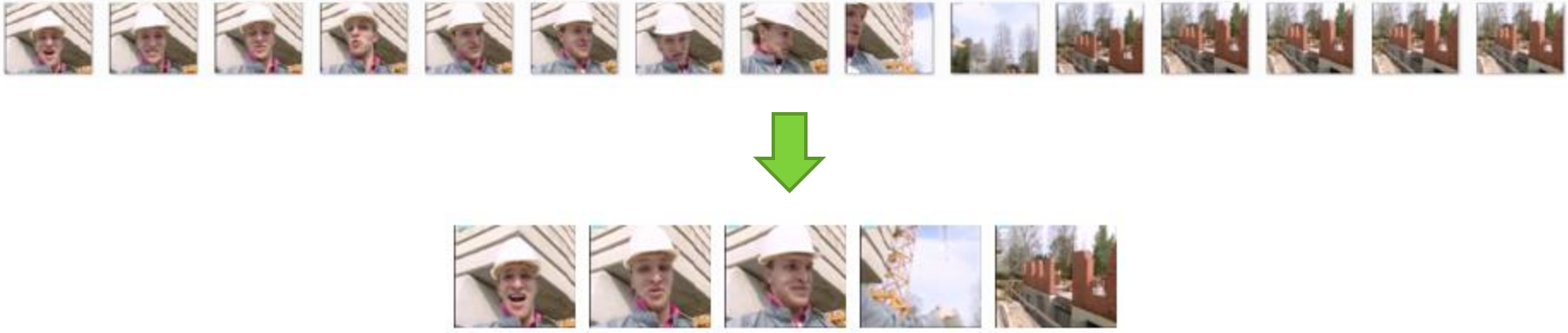
The box

Quantum TV DVR that records up to 12 channels at once

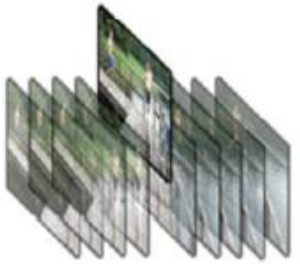
http://www.engadget.com/2014/04/01/verizon-fios-media-server-quantum-tv/

Summarization

- Distill the essence
- Provide a compact yet informative representation of a video
- Crucial for effective and efficient access of video content



Fun



A one-frame preview is often NOT enough to provoke interest.

A storyboard preview is more informative.

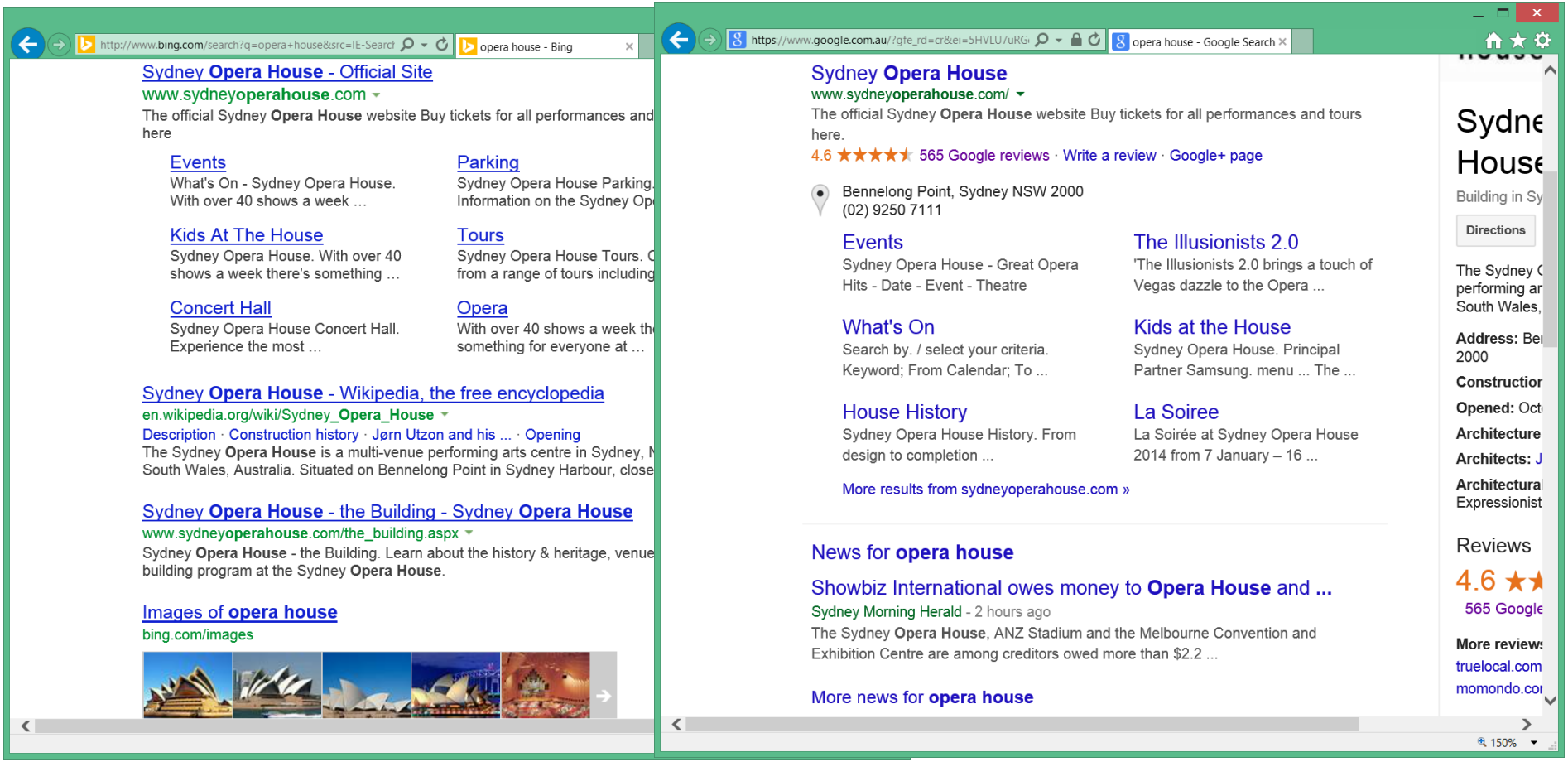
Incentive

- Summly
  - http://summly.com/index.html
  - Founded by 17-year-old Nick D'Aloisio
  - Acquired by Yahoo in 26/03/2013

- 30 Million!!!

http://www.smh.com.au/digital-life/digital-life-news/teens-multimilliondollar-yahoo-payday-before-18th-birthday-20130326-2gqvg.html

Text Summarization



Human summarization and abstracting

- What professional abstractors do
  - “To take an original article, understand it and pack it neatly into a nutshell without loss of substance or clarity presents a challenge which many have felt worth taking up for the joys of achievement alone. These are the characteristics of an art form” - Ashworth.

Text Summarization

- Purpose
  - Indicative, informative, and critical summaries
- Form
  - Extracts (representative paragraphs/sentences/phrases)
  - Abstracts: “a concise summary of the central subject matter of a document”
- Dimensions
  - Single-document vs. multi-document
- Context
  - Query-specific vs. query-independent

Dragomir R. Radev, Text summarization, SIGIR 2004 Tutorial.

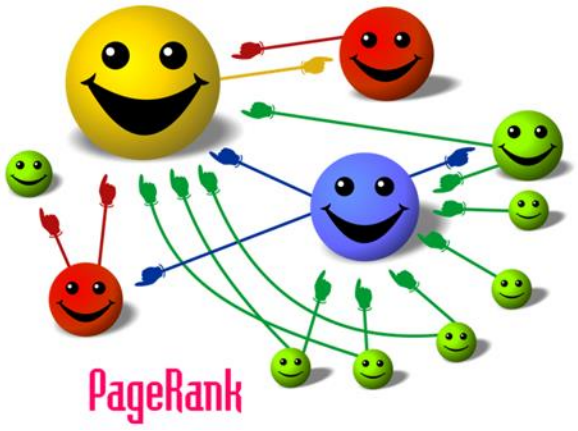
TextRank

- Identify important words or sentences
- Formulate the problem with graph-based solution
- Keyword extraction & sentence extraction

Rada Mihalcea and Paul Tarau, TextRank : Bringing Order into Texts, EMNLP 2004.

PageRank Revisit

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$



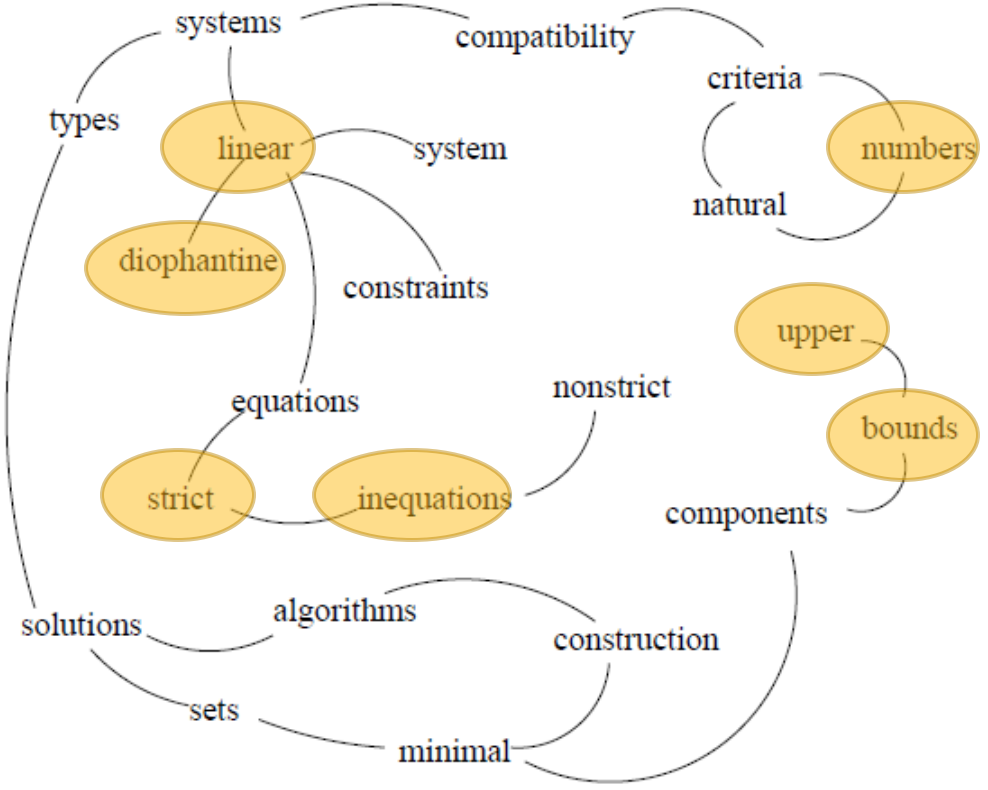
- $S(V_i)$: Score of the Vertex
- V_i : Vertex
- $In(V_i)$: the set of vertices that point to it (predecessors)
- $Out(V_i)$: the set of vertices that vertex points to (successors)
- d : damping factor
  - The probability of jumping from a given vertex to another vertex
    - Random surfer model
    - 0.85 (PageRank)

Graph Construction

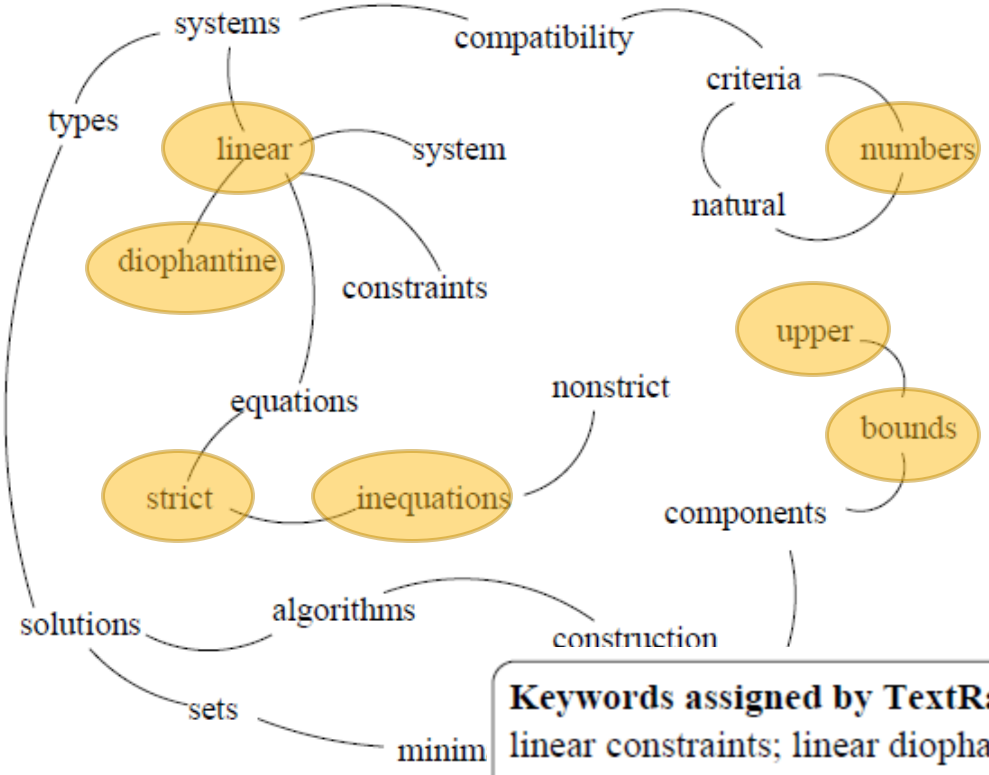
- Vertex
  - Smallest text units (e.g., keyword, sentence)
    - Different types of keywords (e.g., noun, verb)
- Edge
  - Keyword: co-occurrence in a sliding window
  - Sentence: similarity between sentences
    - Knowledge based: WordNet
    - Data driven: Google Distance
    - Empirical: overlap (over common tokens/words)

Sample on Keyword Extraction

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in sc systems and systems of mixed types.



Sample on Keyword Extraction



"importance" by the TextRank algorithm are (with the TextRank score indicated in parenthesis): numbers (1.46), inequations (1.45), linear (1.29), diophantine (1.28), upper (0.99), bounds (0.99), strict (0.77). Notice that this ranking is different than the one rendered by simple word frequencies. For the same text, a frequency approach provides the following top-ranked lexical units: systems (4), types (3), solutions (3), minimal (3), linear (2), inequations (2), algorithms (2). All other lexical units have a fre-

Keywords assigned by TextRank:
linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:
linear constraints; linear diophantine equations; minimal generating sets; nonstrict inequations; set of natural numbers; strict inequations; upper bounds

Quantitative Result

Method	Assigned		Correct		Precision	Recall	F-measure
	Total	Mean	Total	Mean			
TextRank							
Undirected, Co-occ.window=2	6,784	13.7	2,116	4.2	**31.2**	43.1	**36.2**
Undirected, Co-occ.window=3	6,715	13.4	1,897	3.8	28.2	38.6	32.6
Undirected, Co-occ.window=5	6,558	13.1	1,851	3.7	28.2	37.7	32.2
Undirected, Co-occ.window=10	6,570	13.1	1,846	3.7	28.1	37.6	32.2
Directed, forward, Co-occ.window=2	6,662	13.3	2,081	4.1	31.2	42.3	35.9
Directed, backward, Co-occ.window=2	6,636	13.3	2,082	4.1	31.2	42.3	35.9
Hulth (2003)							
Ngram with tag	7,815	15.6	1,973	3.9	25.2	**51.7**	33.9
NP-chunks with tag	4,788	9.6	1,421	2.8	29.7	37.2	33.0
Pattern with tag	7,012	14.0	1,523	3.1	21.7	39.9	28.1

Table 1: Results for automatic keyword extraction using TextRank or supervised learning (Hulth, 2003)

Sample on Sentence Extraction

3: BC−HurricaineGilbert, 09−11 339
4: BC−Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dom
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic
9: Hurricaine Gilbert Swept towrd the D
alerted its heavily populated south coa
10: The storm was approaching from the s
to 92 mph.
11: "There is no need for alarm," Civil De
alert shortly after midnight Saturday.
12: Cabral said residents of the province o
13: An estimated 100,000 people live in th
about 125 miles west of Santo Doming
14. Tropical storm Gilbert formed in the e
Saturday night.
15: The National Hurricaine Center in Mia
16.1 north, longitude 67.5 west, about
southeast of Santo Domingo.
16: The National Weather Service in San .
at 15 mph with a "broad area of cloudi
of the storm.
17. The weather service issued a flash floc
at least 6 p.m. Sunday.
18: Strong winds associated with the Gilb
and up to 12 feet to Puerto Rico's sout
19: There were no reports on casualties.
20: San Juan, on the north coast, had heav
the night.
21: On Saturday, Hurricane Florence was
pushed inland from the U.S. Gulf Coa
22: Residents returned home, happy to fin
23: Florence, the sixth named storm of the
24: The first, Debby, reached minimal hur
last month.

TextRank extractive summary

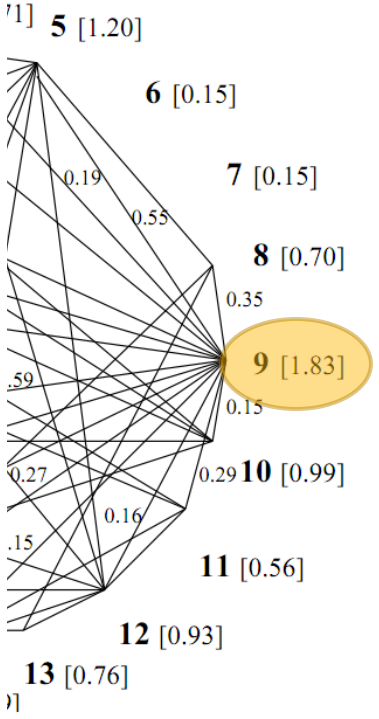
Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil De−fense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm. Strong winds associated with Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's south coast.

Manual abstract I

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high wind and seas. Tropical storm Gilbert formed in the eastern Carribean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and in the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and down−graded to a tropical storm.

Manual abstract II

Tropical storm Gilbert in the eastern Carribean strenghtened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15 mph with a broad area of cloudiness and heavy weather with sustained winds of 75 mph gusting to 92 mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.



Sample on Sentence Extraction

- TextRank goes beyond the sentence "connectivity" in a text
  - Sentence 15 would not identified as "important" based on the number of connection
  - But it is identified as "important" by TextRank
  - Human also identify the sentence as "important"

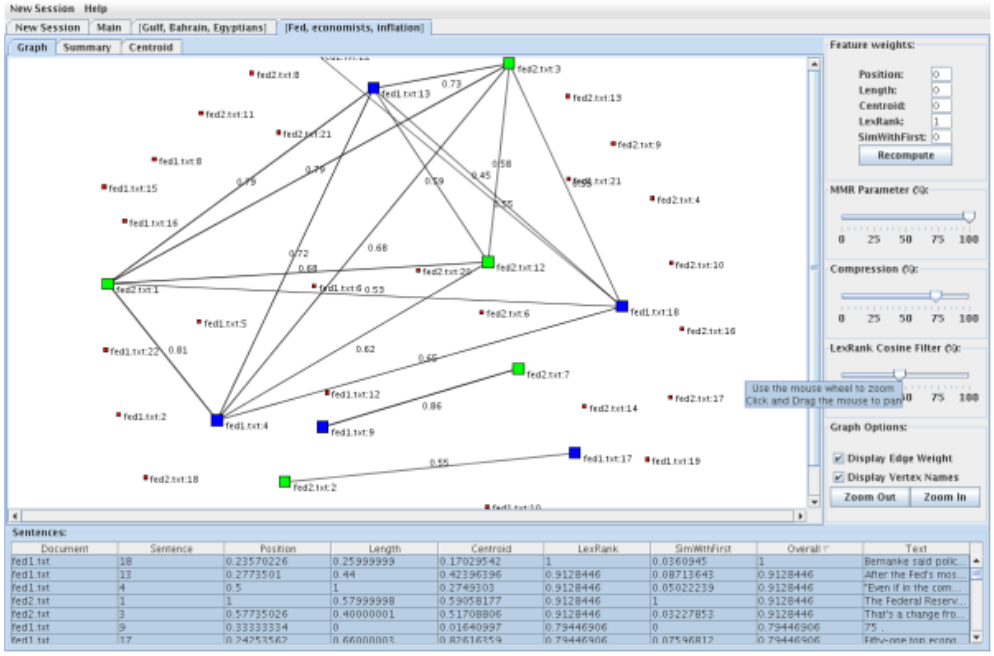
Quantitative Result

System	ROUGE score – Ngram(1,1)		
	basic (a)	stemmed (b)	stemmed no-stopwords (c)
S27	0.4814	0.5011	0.4405
S31	0.4715	0.4914	0.4160
TextRank	**0.4708**	**0.4904**	**0.4229**
S28	0.4703	0.4890	0.4346
S21	0.4683	0.4869	0.4222
Baseline	*0.4599*	*0.4779*	*0.4162*
S29	0.4502	0.4681	0.4019

Table 2: Results for single document summarization: TextRank, top 5 (out of 15) DUC 2002 systems, and baseline. Evaluation takes into account (a) all words; (b) stemmed words; (c) stemmed words, and no stop-words.

LexRank

- Sentence level
- Cosine similarity between sentences



http://141.211.245.18/demos/lexrank/lexrankmead.html

G. Erkan, D. Radev, LexRank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research, 2004.

http://clair.si.umich.edu/demos/lexrank/

Video Summarization Problem

- Representativeness
  - Maximized
- Redundancy
  - Minimized
- Presentation
  - Keyframe/Storyboard
  - Skim
  - Collage

The Problem

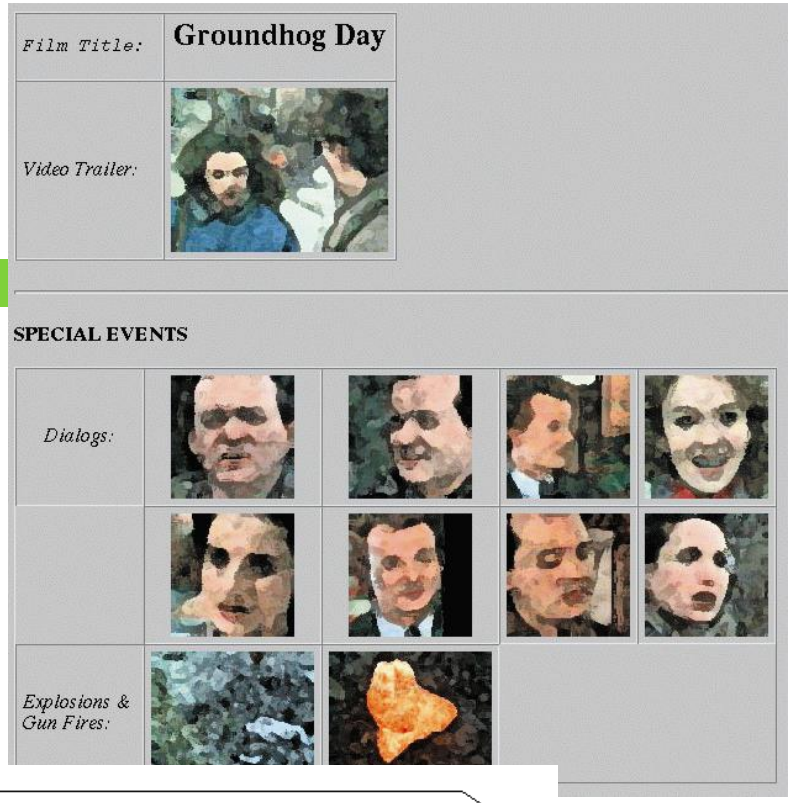

SIEMENS


#1 #21 #41 #61 #81 #101 #121 #141 #161 #181 #201 #221 #241 #261 #268(last)
#1 #21 #41 #61 #81 #101 #121 #141 #161 #181 #201 #221 #241 #261 #281
#1 #7 #13 #19 #25 #31 #37 #43 #49 #55 #61 #67 #73 #79 #85
#1 #2 #3 #4 #5 #6 #7 #9 #10 #11 #12 #13 #14 #15 #17

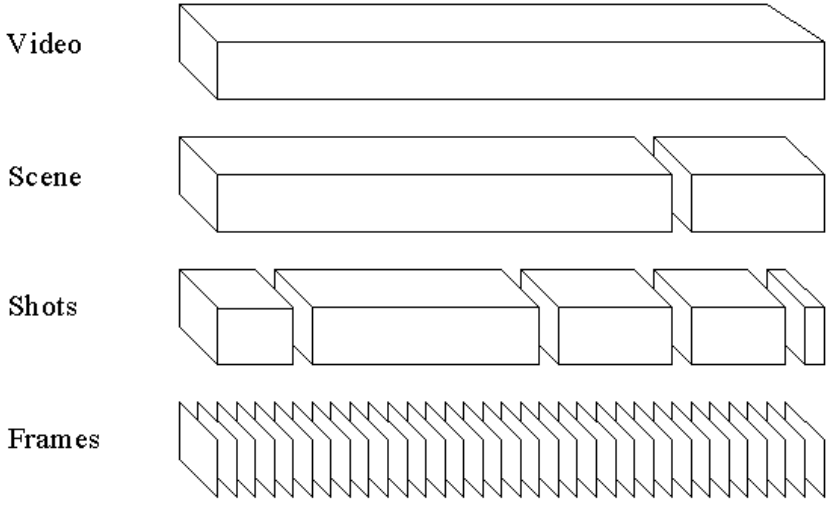
Related Work

- Clustering based
  - K-means, graph cuts, … …
- Learning based
  - Important vs unimportant
- Reconstruction based
  - Curve fitting
  - Data fitting
- Different features
  - Visual
  - Audio
  - Semantics such as who, what, where, when

Video Abstraction

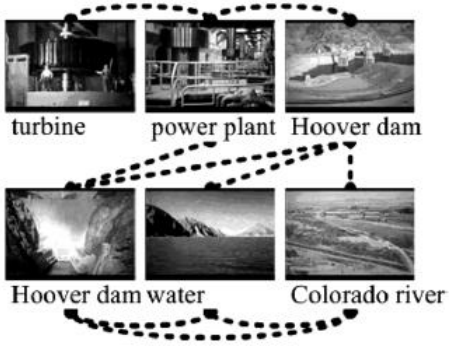


Step 1	Step 2	Step 3
Video Segmentation and Analysis	Clip Selection	Clip Assembly
Extracts and analyzes • Shots and scenes • clips containing special events (close-ups of main actors, gunfire, explosions, text)	Determines video clips to be part of the abstract based on • general considerations (balanced coverage of entire input video) • special events	Assembles the selected video clips into their final format

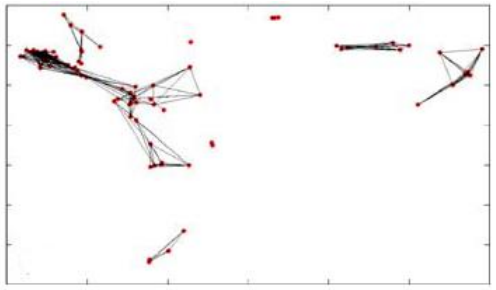


Rainer Lienhart, Silvia Pfeiffer, and W. Effelsberg, Video abstracting, Communications of the ACM 40(12): 54–62, 1997.

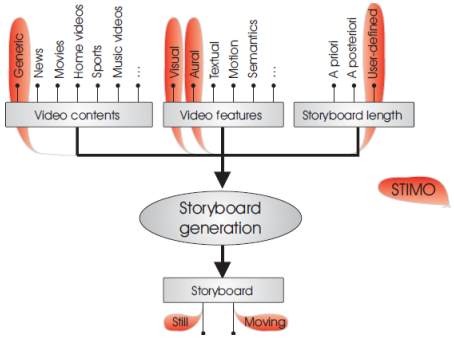
Static Video Summarization (Among many others)



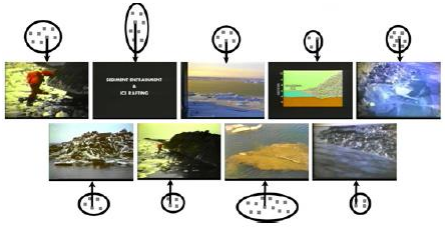
[Chen *et al.*, 09]
story-structure



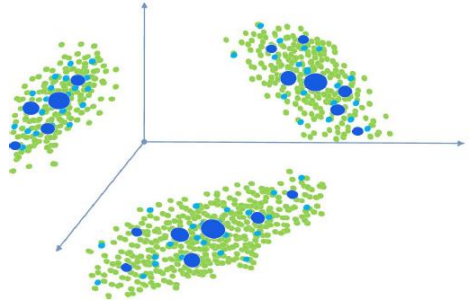
[Makedonas *et al.*, 09]
graph connectivity



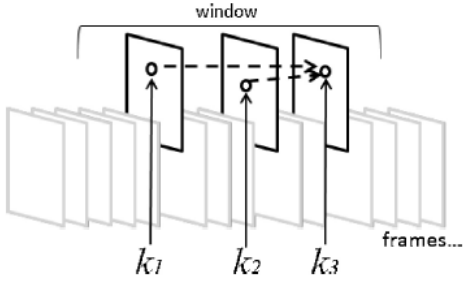
[Furini *et al.*, 10]
STIMO



[Avila *et al.*, 11]
Vsumm



[Cong *et al.*, 12]
sparse dictionary



[Guan *et al.*, 12]
keypoint-based

Limitation

- Utilizing global visual features
  - Color and texture computed over the entire frame

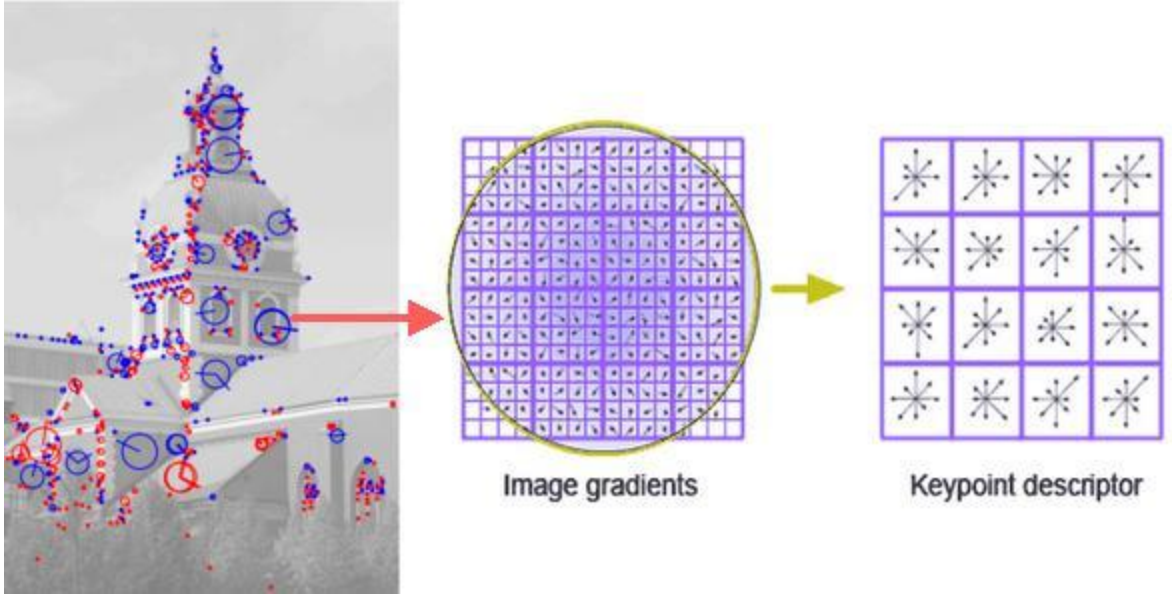
- Subtle yet important details could be swallowed by global features

Local Descriptors

- Local keypoint features
  - Distinctive representation capacity (e.g. invariant to location, scale and rotation, and robust to affine transformation).
  - Played a significant role in many application domains of visual content analysis
    - Object recognition
    - Landmark recognition
    - Image classification
    - … …

Local Descriptors

- SIFT
  - Scale Invariant Feature Transform
- SURF
  - Speeded Up Robust Features
- … ..



Problem Formulation

- What makes a video
  - Video frame vs video shot vs video story
  - A video shot depicts a scene
  - Object can be characterized with a number of keypoints
- What contributes to redundancy
  - Redundancy exists among adjacent frames
  - Removing overlapped objects could reduce redundancy
- Keyframe selection is to identify a number of frames which
  - Best cover the keypoints
  - Share minimal redundancy

Keyframe Selection

- The global pool is separated into two sets, $K_{covered}$ and $K_{uncovered}$. At the beginning, $K_{uncovered}$ contains all keypoints in K and $K_{covered}$ is empty

- For frame f_i, denote its keypoint set as FP_i,
- Coverage
  - the cardinality of the intersection between FP_i and $K_{uncovered}$

$$C(f_i) = |FP_i \cap K_{uncovered}|$$

- Redundancy
  - how many keypoints it contains in $K_{covered}$

$$R(f_i) = |FP_i \cap K_{covered}|$$

Keyframe Selection

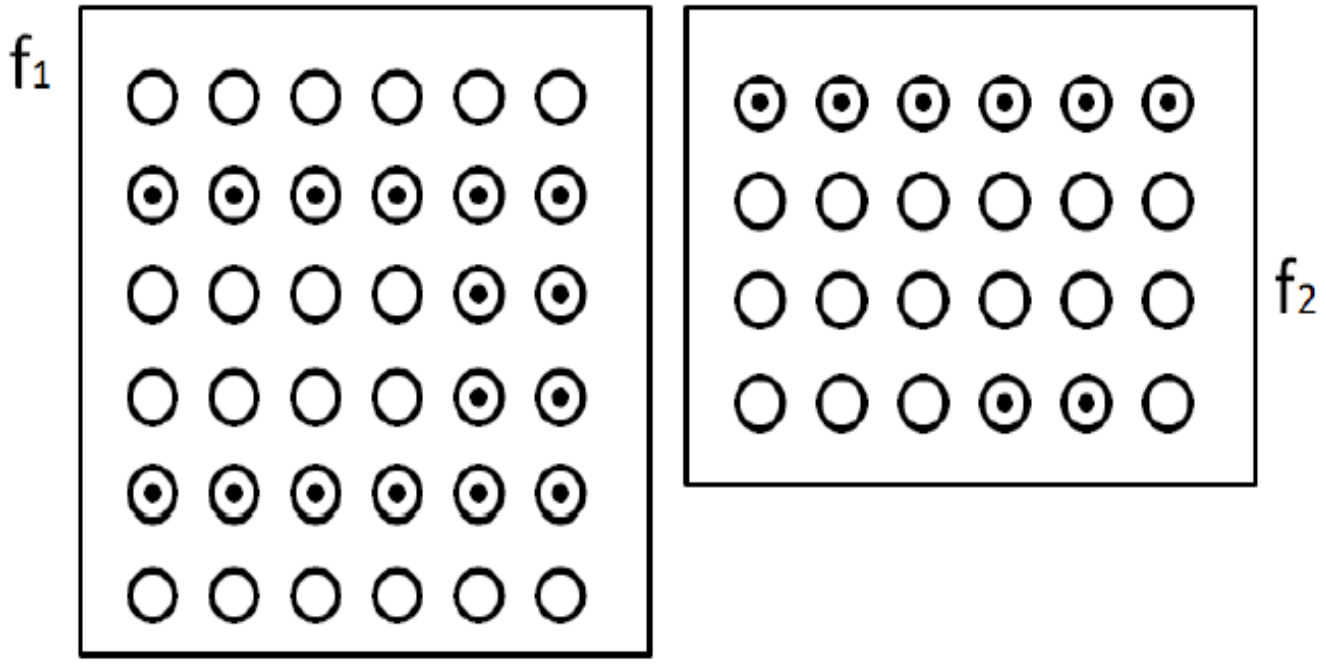
- The influence of frame f_i is calculated as a balance of $C(f_i)$ and $R(f_i)$ controlled by alpha (set to 1 empirically in the experiments)

$$Influence(f_i) = C(f_i) - \alpha R(f_i)$$

- At the end of each iteration, the frame with the highest influence value and positive coverage will be selected as a keyframe, and $K_{covered}$ and $K_{uncovered}$ will be updated

- The iteration repeats until the whole keypoint pool is covered, or a predefined percentage of coverage STOP of the pool K is reached.

Toy Example


f1
f2
⊙ - already-covered keypoint in the pool
○ - uncovered keypoint in the pool
C(f1) = 20, R(f1) = 16
Influence(f1) = 20 – 16 = 4
C(f2) = 16, R(f2) = 8
Influence(f1) = 16 – 8= 8

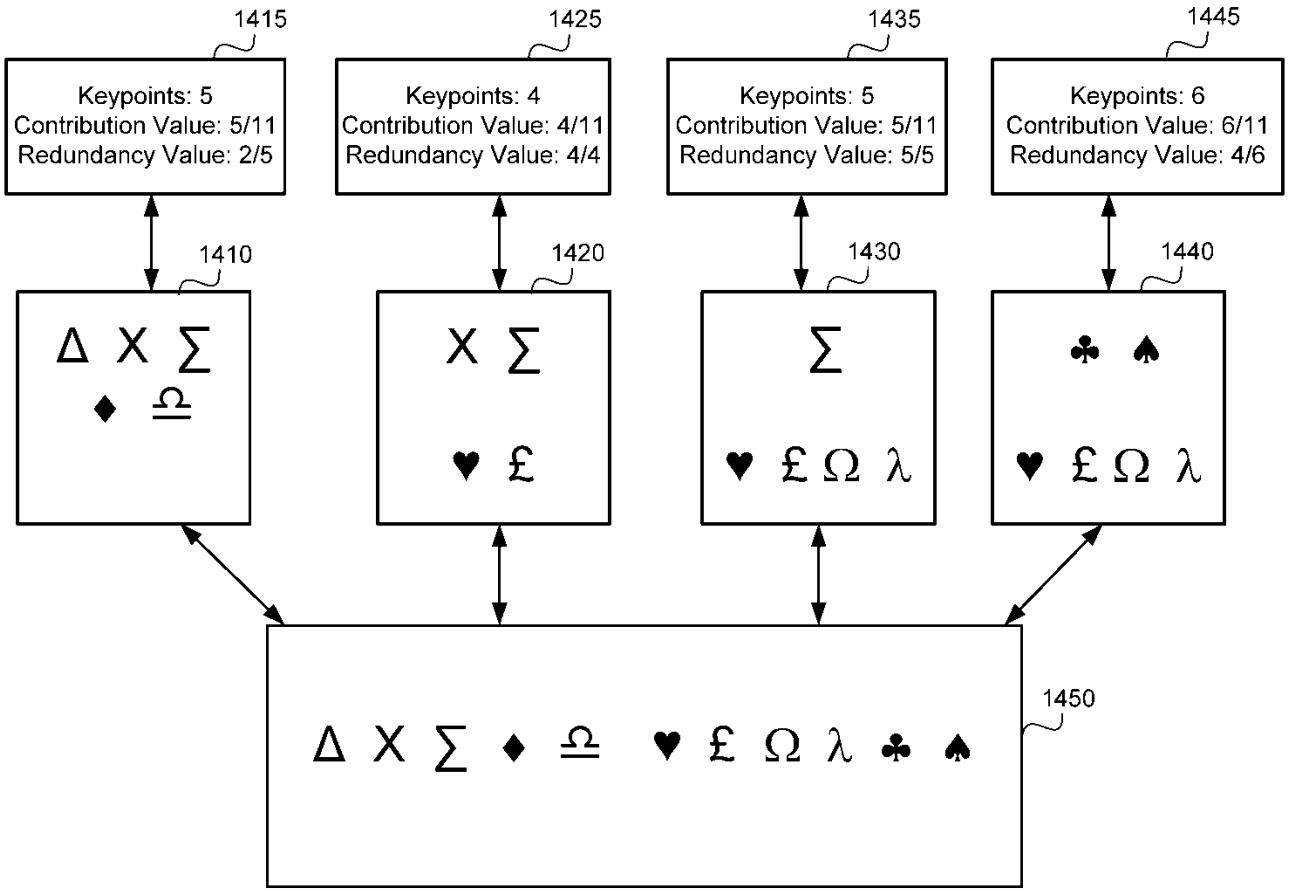

1415
Keypoints: 5
Contribution Value: 5/11
Redundancy Value: 2/5
1425
Keypoints: 4
Contribution Value: 4/11
Redundancy Value: 4/4
1435
Keypoints: 5
Contribution Value: 5/11
Redundancy Value: 5/5
1445
Keypoints: 6
Contribution Value: 6/11
Redundancy Value: 4/6
1410
Δ X ∑
♦ Ω
1420
X ∑
♥ £
1430
∑
♥ £ Ω λ
1440
♣ ♠
♥ £ Ω λ
1450
Δ X ∑ ♦ Ω ♥ £ Ω λ ♣ ♠

Fig. 14

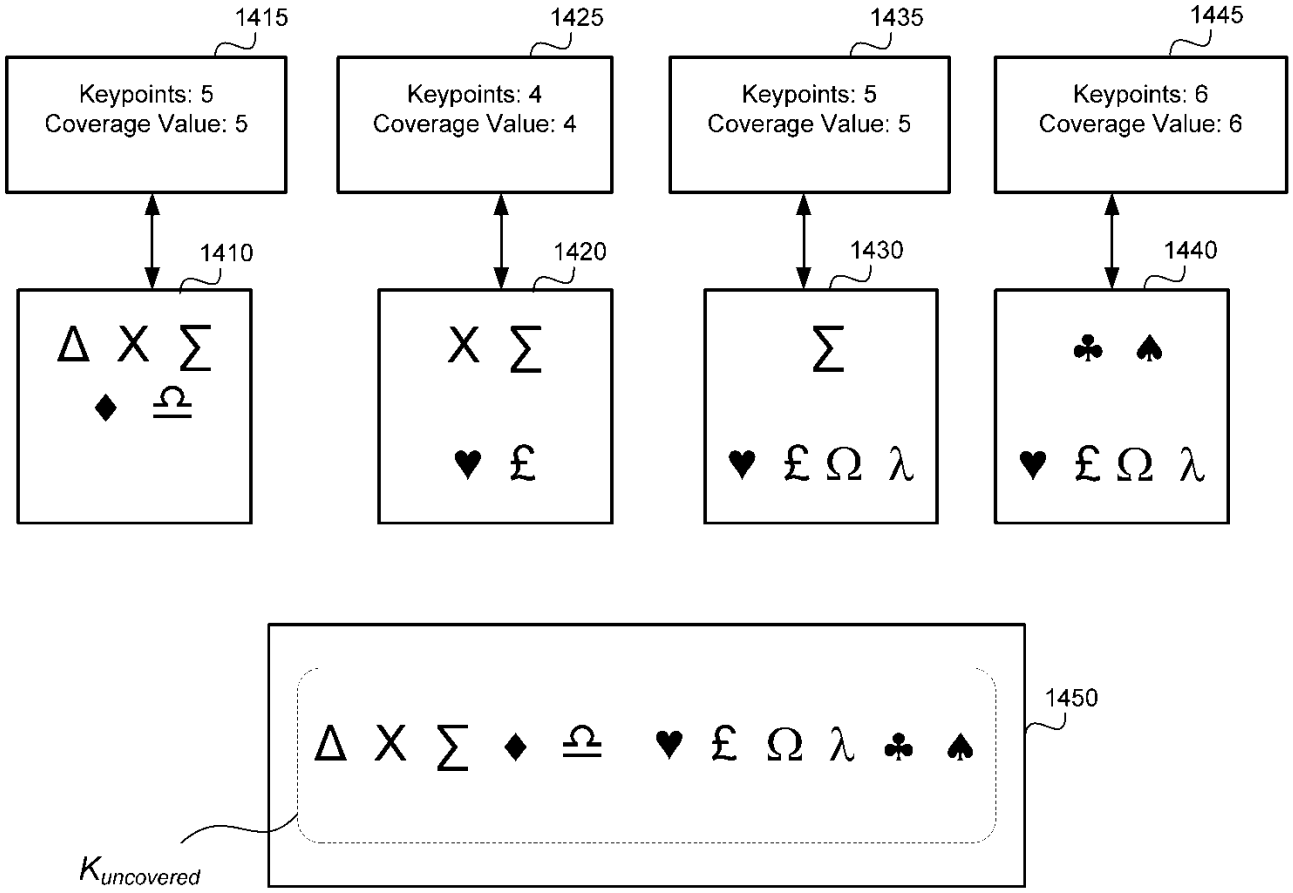

1415
Keypoints: 5
Coverage Value: 5
1425
Keypoints: 4
Coverage Value: 4
1435
Keypoints: 5
Coverage Value: 5
1445
Keypoints: 6
Coverage Value: 6
1410
1420
1430
1440
1450
$K_{uncovered}$

Fig. 15A

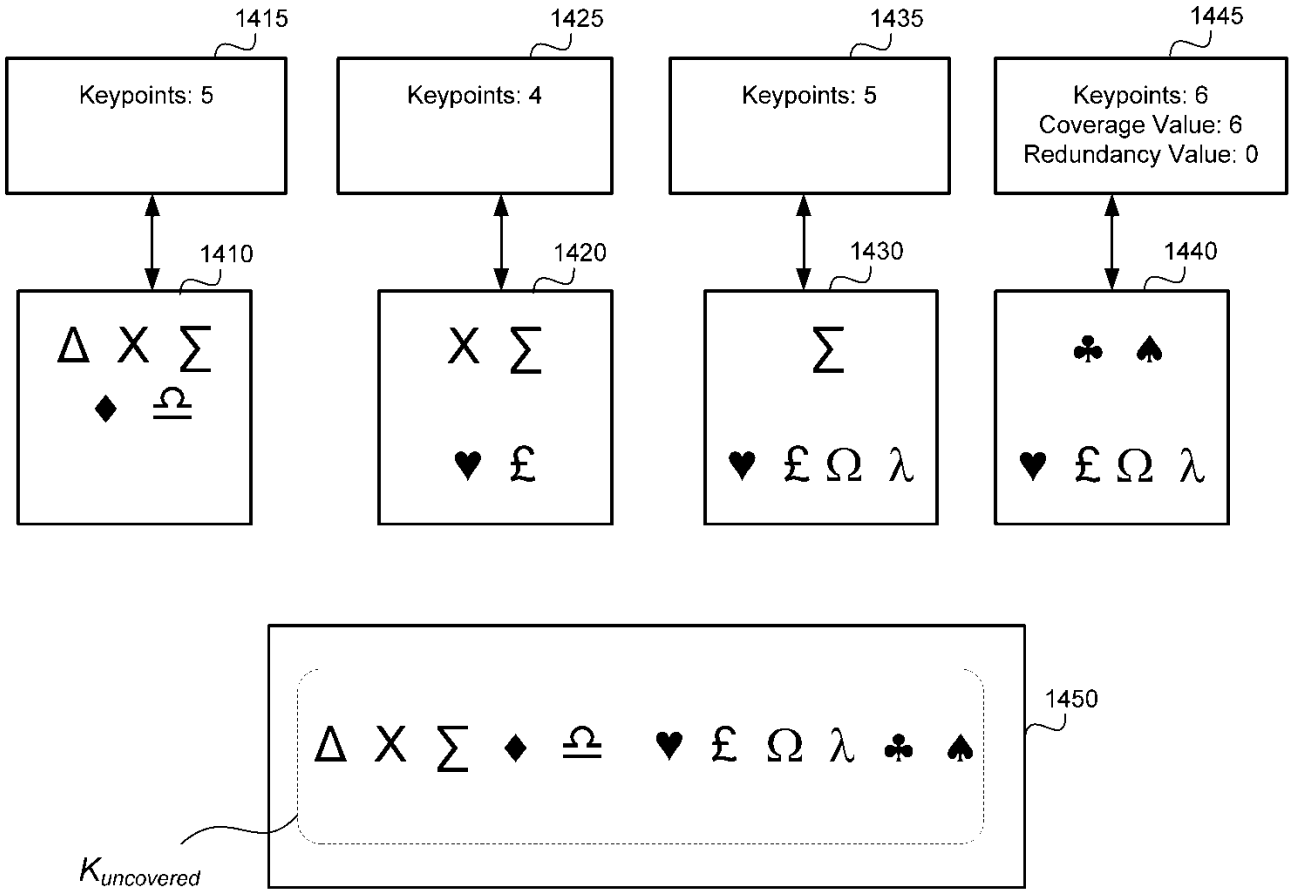

1415
Keypoints: 5
1425
Keypoints: 4
1435
Keypoints: 5
1445
Keypoints: 6
Coverage Value: 6
Redundancy Value: 0
1410
1420
1430
1440
Δ X ∑ ♦ Ω
X ∑ ♥ £
∑ ♥ £ Ω λ
♣ ♠ ♥ £ Ω λ
1450
Δ X ∑ ♦ Ω ♥ £ Ω λ ♣ ♠
$K_{uncovered}$

Fig. 15B

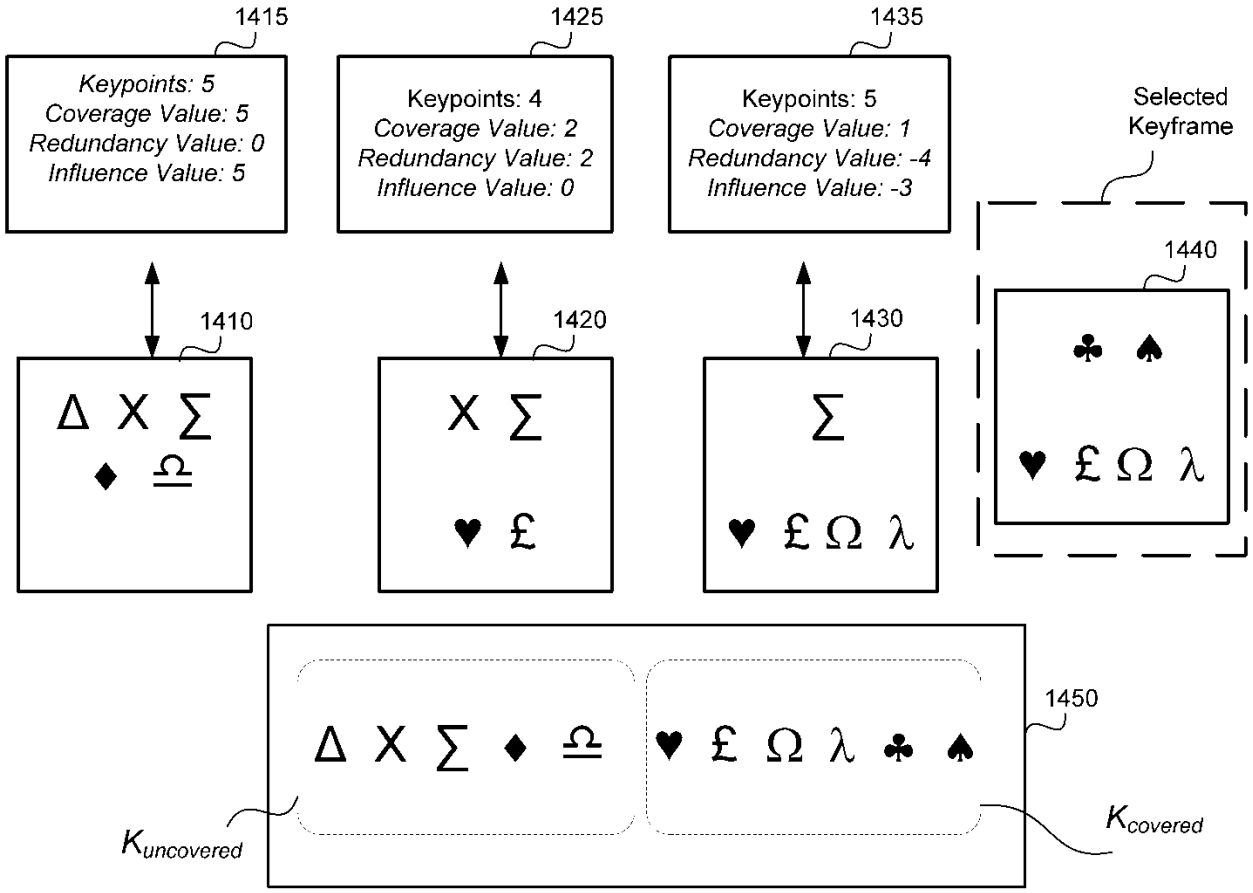

1415
Keypoints: 5
Coverage Value: 5
Redundancy Value: 0
Influence Value: 5
1425
Keypoints: 4
Coverage Value: 2
Redundancy Value: 2
Influence Value: 0
1435
Keypoints: 5
Coverage Value: 1
Redundancy Value: -4
Influence Value: -3
Selected Keyframe
1440
1410
1420
1430
1450
$K_{uncovered}$
$K_{covered}$

Fig. 15C

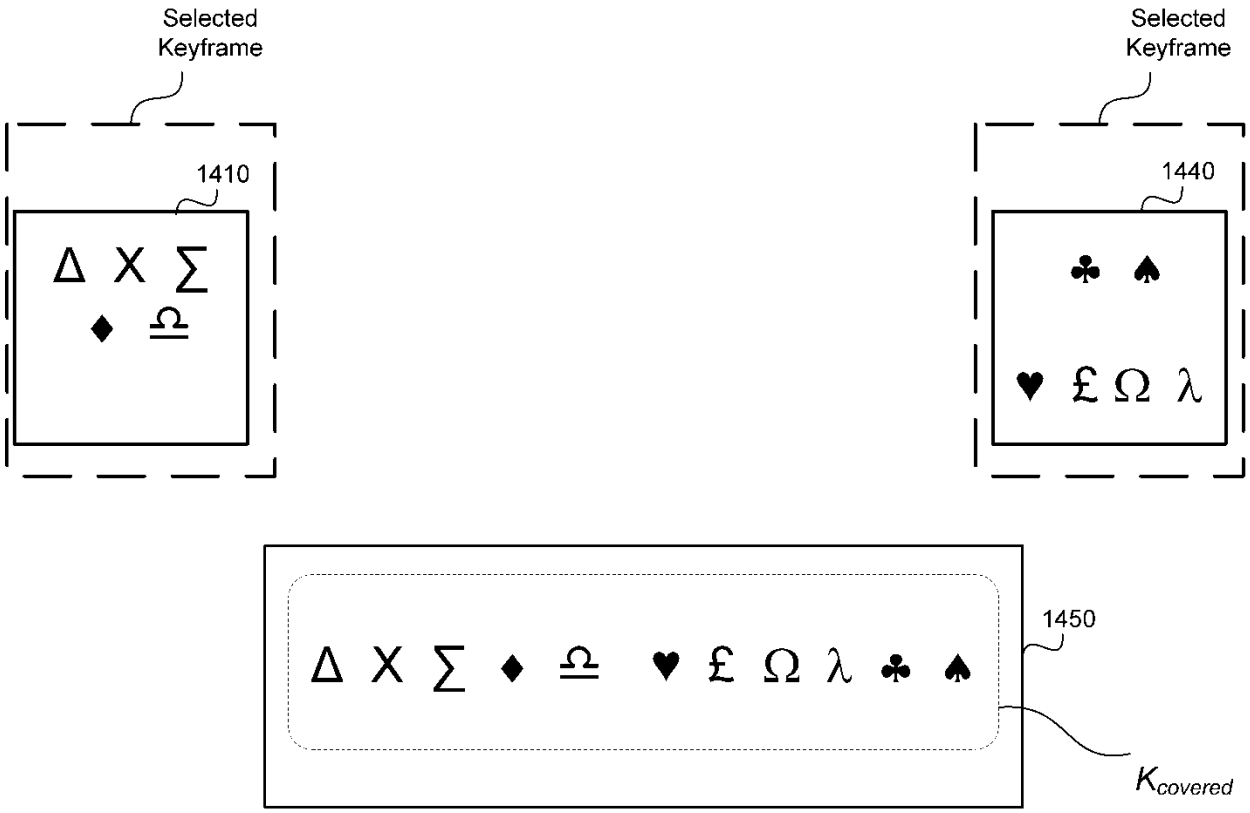
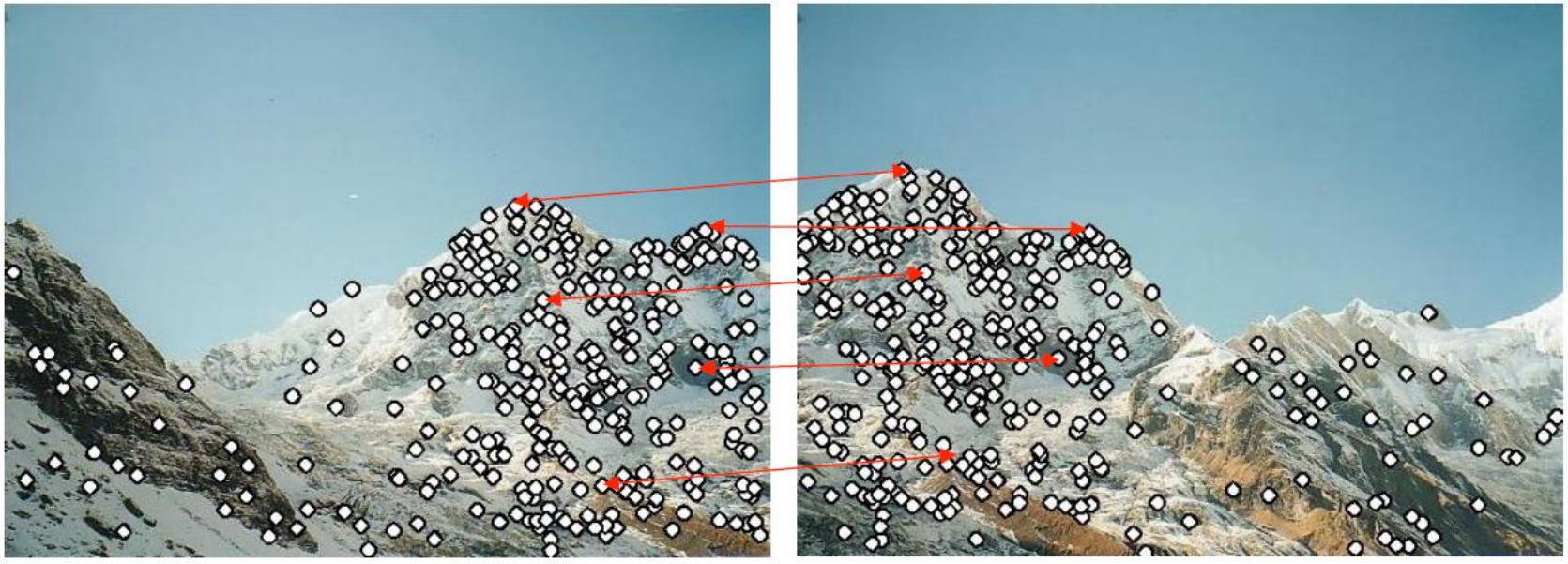

Selected Keyframe
1410
Δ X ∑
♦ Ω
Selected Keyframe
1440
♣ ♠
♥ £ Ω λ
Δ X ∑ ♦ Ω ♥ £ Ω λ ♣ ♠
1450
$K_{covered}$

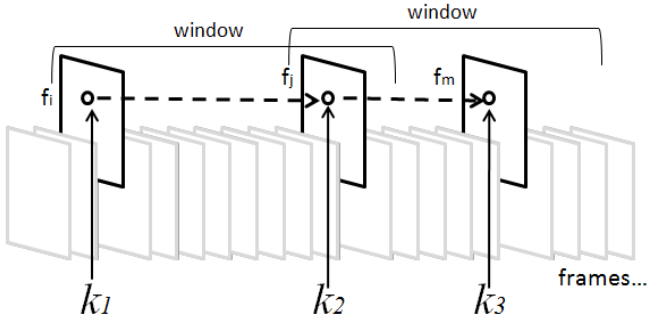
Fig. 15D

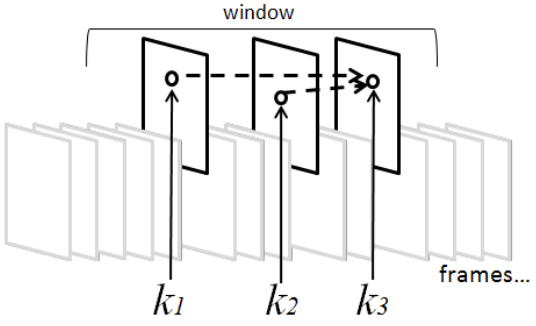
Keypoint Matching



Keyframe Selection

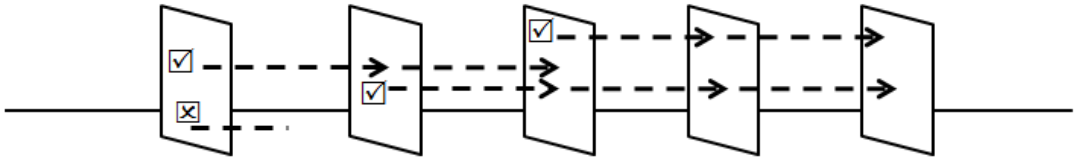
- Keypoint Pool Construction
  - Inter-window Keypoint Chaining
    - Constrain the pairing within a temporal window of size W without losing the discriminative power of keypoint matching
  - Intra- Window Keypoint Chaining
    - make the matching more reliable





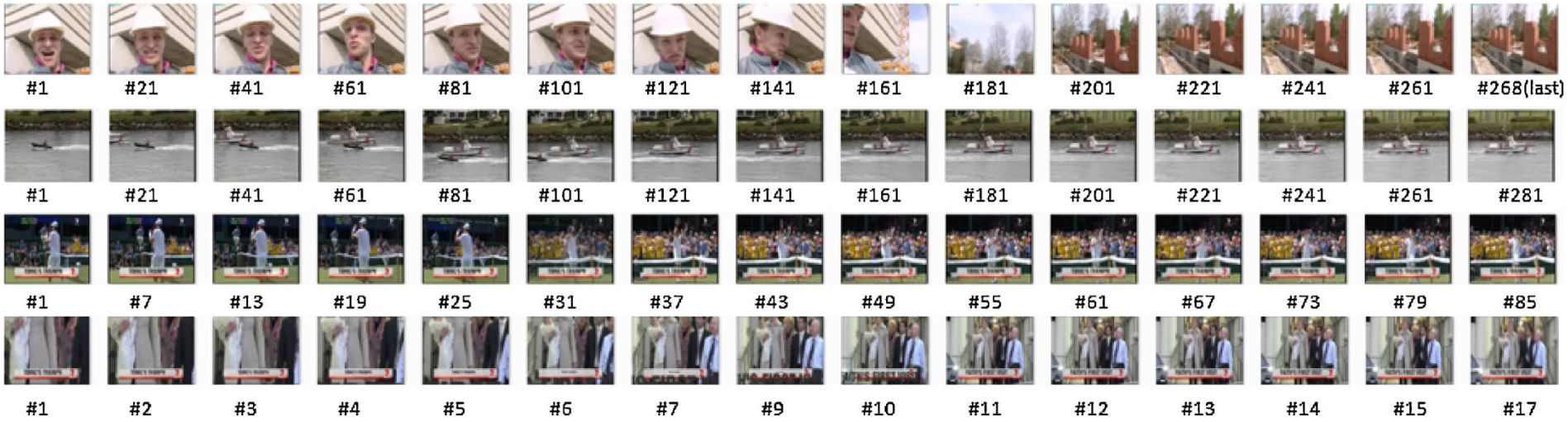
Keyframe Selection

- Keypoint Pool Construction
  - Each keypoint either belongs to a chain of matched keypoints or becomes an singleton without any connection
    - Each chain is represented by its HEAD keypoint
    - Chains with the number of keypoints greater than T (set to 10) are kept



☑ are HEAD keypoints in each chain, which are included in the global keypoint pool;
☒ are singleton keypoints, which are excluded from the pool.

Samples Results



Sample Result 1

Coverage = 94%

KBKS, Coverage = 73%

KBKS, Coverage = 84%

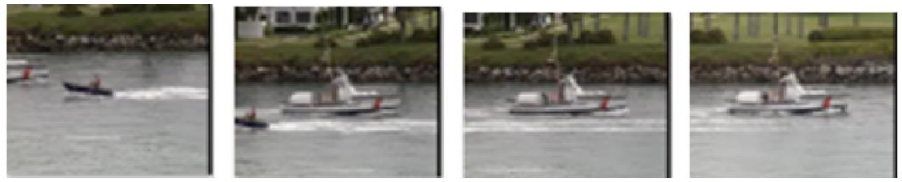
KBKS, Coverage = 95%

Clustering (5 clusters), Coverage = 88%

Iso-Content Distance, Coverage = 85%

Iso-Content Distortion, Coverage = 87%

Sample



#19 #107 #161 #264

KBKS, Coverage = 95%

#58 #118 #197 #271

Clustering (4 clusters), Coverage = 89%

#1 #70 #179 #300

Iso-Content Distance, Coverage = 90%

#1 #68 #175 #300

Iso-Content Distortion, Coverage = 91%

Sample Result 3

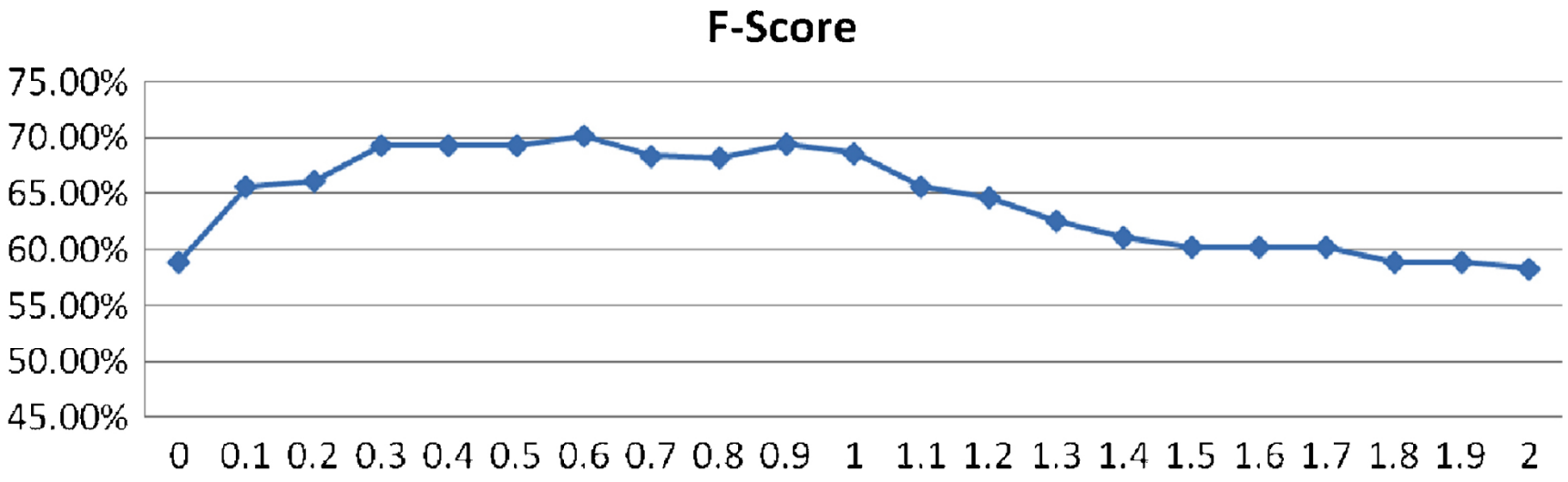
Sample Frames of the Zooming Shot

Our Selected Keyframes (Coverage=86%)

Clustering-based Result

Equidistance-based Result

Impact of α

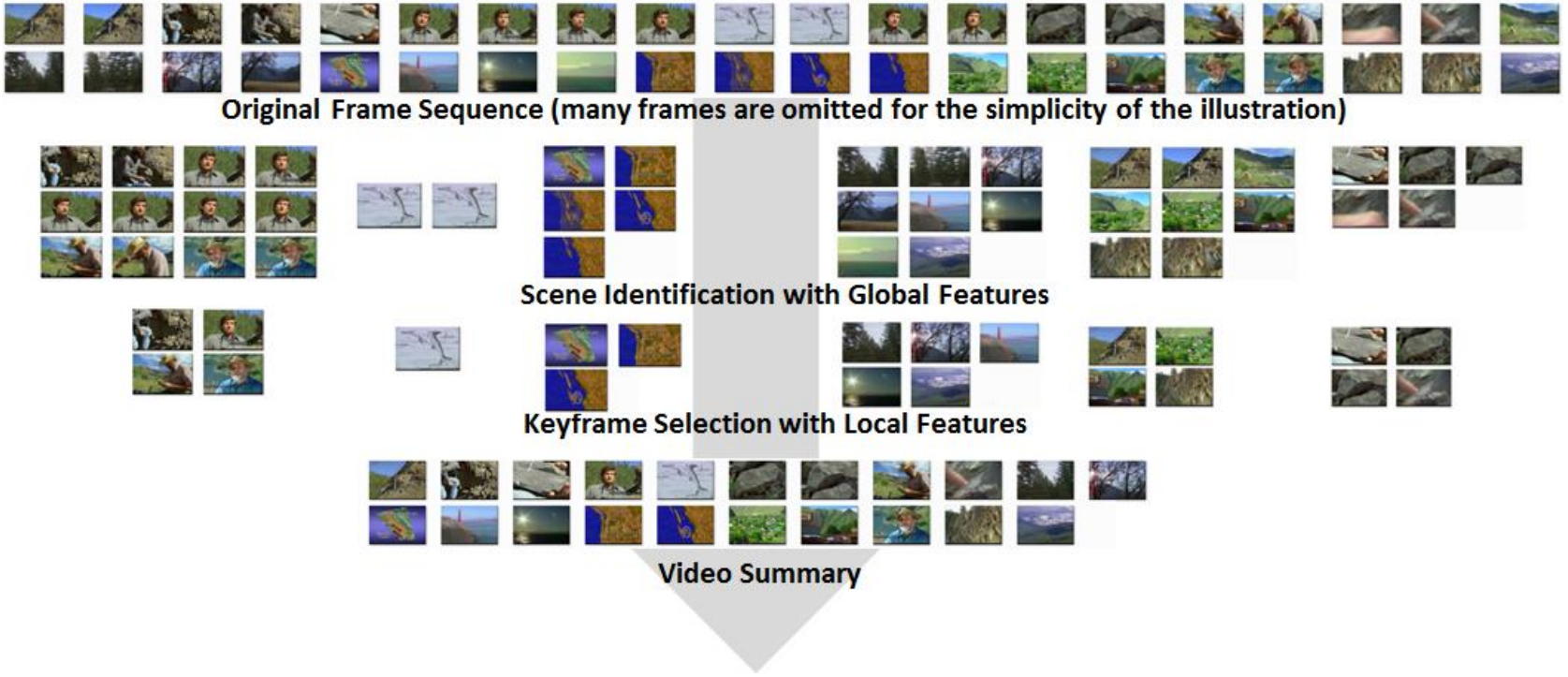

F-Score
75.00%
70.00%
65.00%
60.00%
55.00%
50.00%
45.00%
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2

Keypoint Matching

- Computationally expensive
  - Thousands of keypoints per frame
  - Matching candidate keypoints within a certain radius R (set to 100)
  - RANdom Sample Consensus algorithm (RANSAC) is iteratively invoked to enforce geometrical consistence among keypoint matches

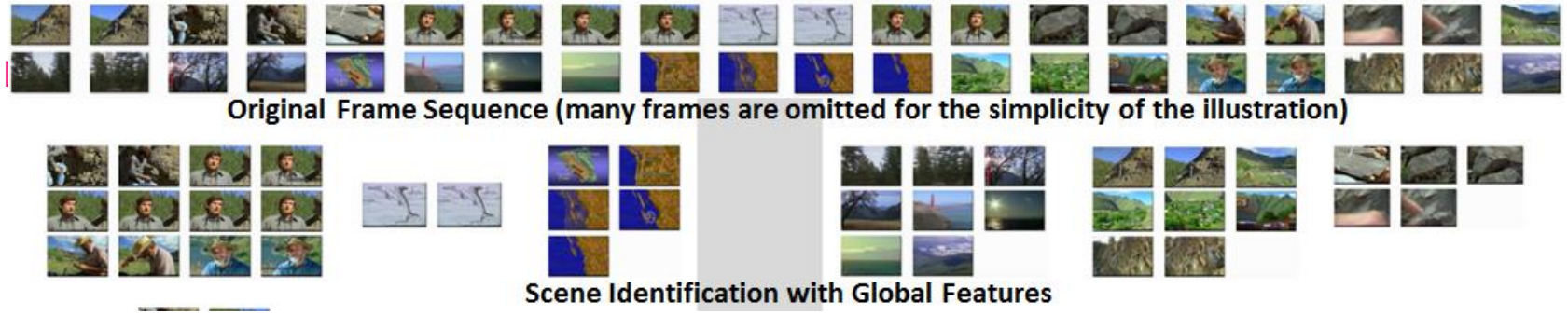
Video Summarization Framework

- Utilizing both global and local visual features



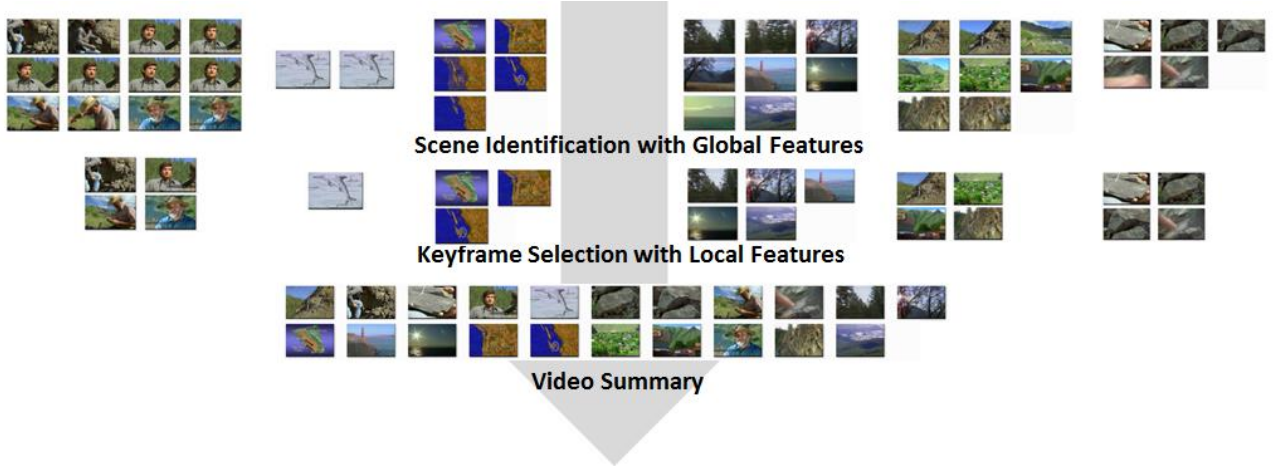
Scene Identification

- A video consists of multiple scenes and the frames of each scene are visually similar, though the frames of the same scene may scatter in the video
- Represent each video frame with the CEDD feature which is a histogram characterizing both color and texture features
- Perform frame clustering with K-Means



Keyframe Selection

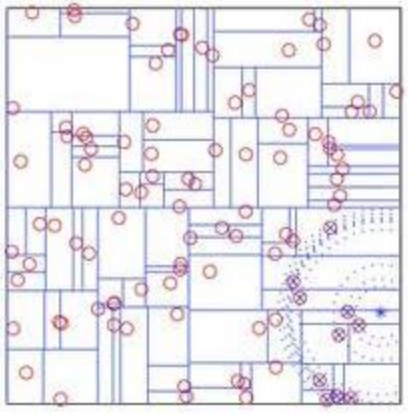
- Within each cluster (i.e. scene)
  1. Represent each frame with local keypoints
  2. Generate a keypoint pool
  3. Select the frames that covers the pool best (maximum coverage and minimum redundancy)
  4. Combine keyframes from each scene as a summary

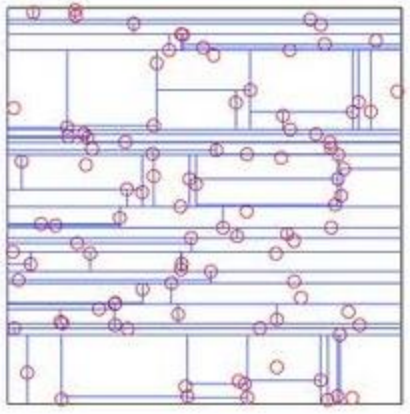


Keypoint Filtering with Saliency

Fast Solution – Keypoint Forest

- Randomized kd-tree
  1. Gather all keypoints from all frames
  2. Split the data along different features that have the greatest variance to generate a few trees
  3. Matching a keypoint against the trees to find the best match





More details

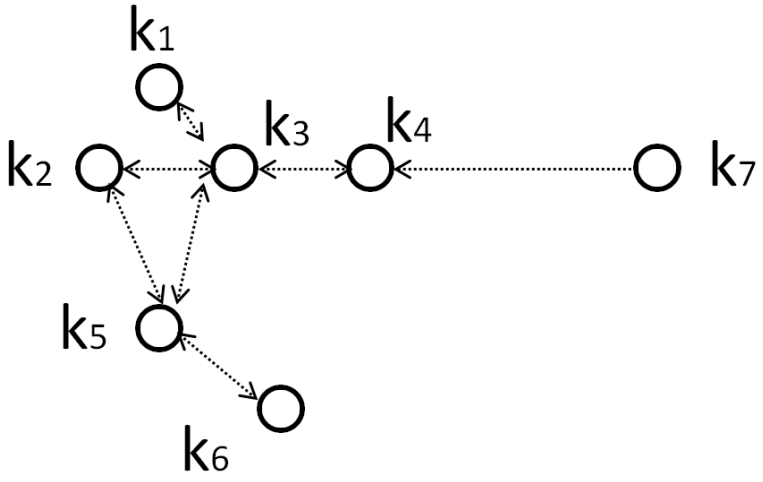
Fast Solution – Keypoint Forest

- Randomized kd-tree performance
  - Appropriate number of trees (e.g. 5)
  - Keypoint Matching accuracy can be above 90%
  - Keypoint Matching can be 100 times faster
  - Previously 0.5 second /frame --> now 0.01 second / frame
  - Do not have noticeable impact on the keyframe selection

Local Visual Word Model

- grouping neighbouring keypoints into local visual words to accommodate variance of the same keypoint appearing in different frames.

- Simple mutual neighbourhood relationship



Calculate Influence

$$Influence(f_i) = \frac{\alpha * C(f_i) - (1-\alpha)R(f_i)}{GlobalSim(f_i)},$$

$$GlobalSim(f_i) = \sum_j Similarity(f_i, f_j).$$

For GlobalSim(), is j from the whole sequence or the selected keyframes.

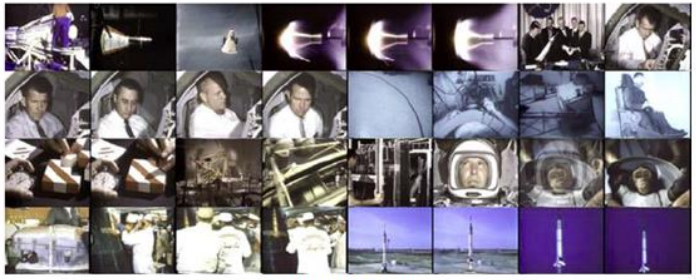
Experiments

- Dataset 1
  - 50 videos from Open Video Project (OVP)
    - http://www.open-video.org/
  - 1 to 4 minutes
- Dataset 2
  - 50 Youtube videos
  - 1 to 10 minutes

Sample Result

- NASA 25th Anniversary Show Segment 03
  - There are 8 frames of pilot shots in our result, covering 6 out of 7 pilots mentioned in the story.
  - This indicates that our approach focuses more on local details compared to other global-feature based approaches

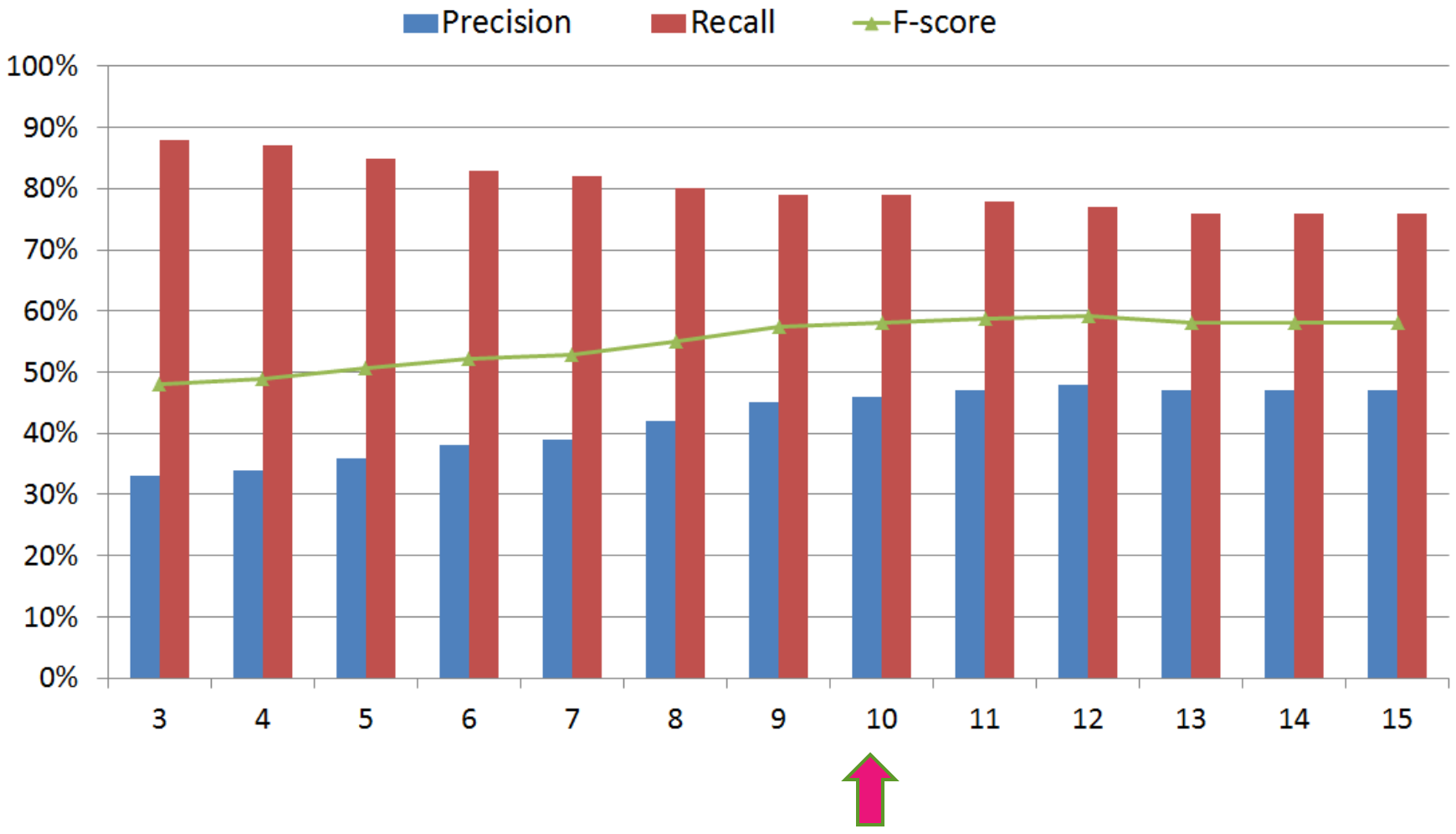
Our Result



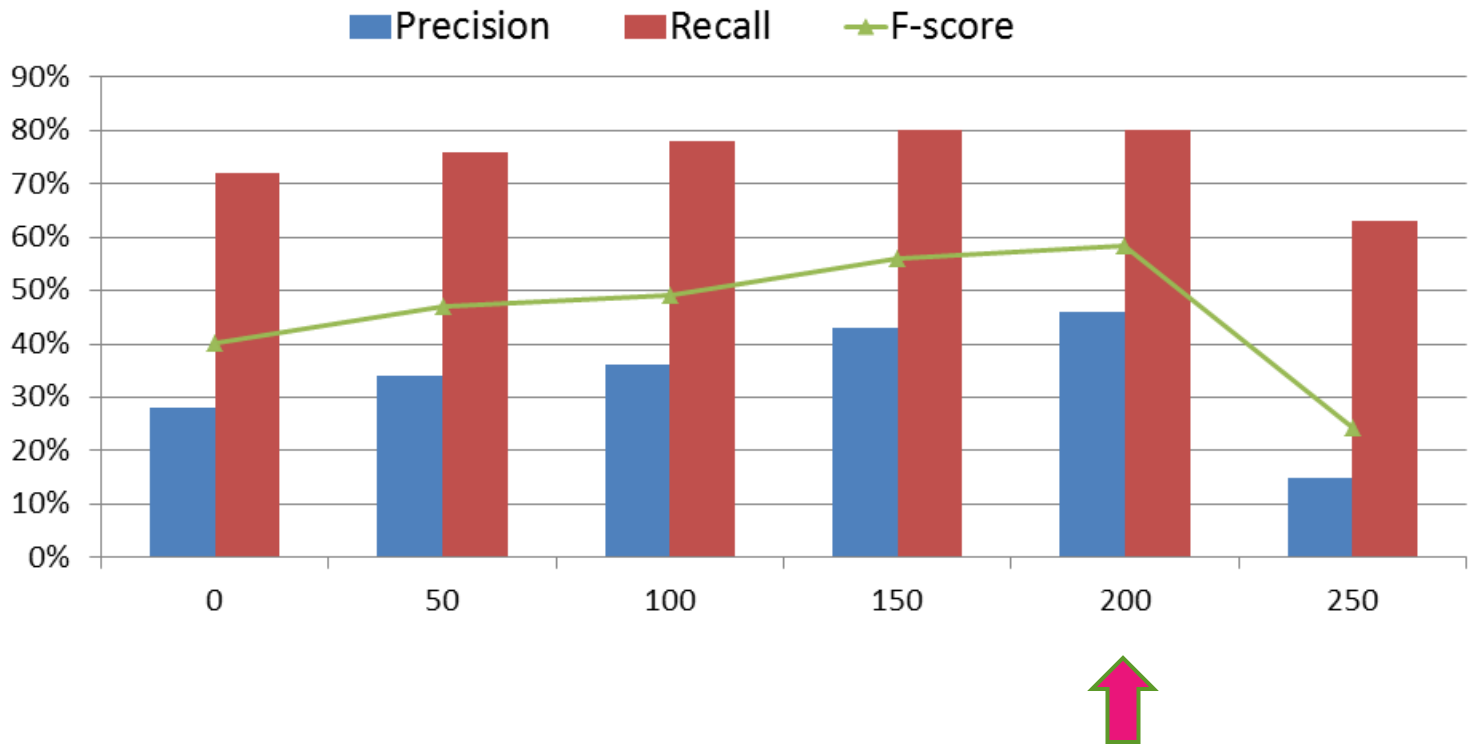
Besiris' Result

OVP Storyboard

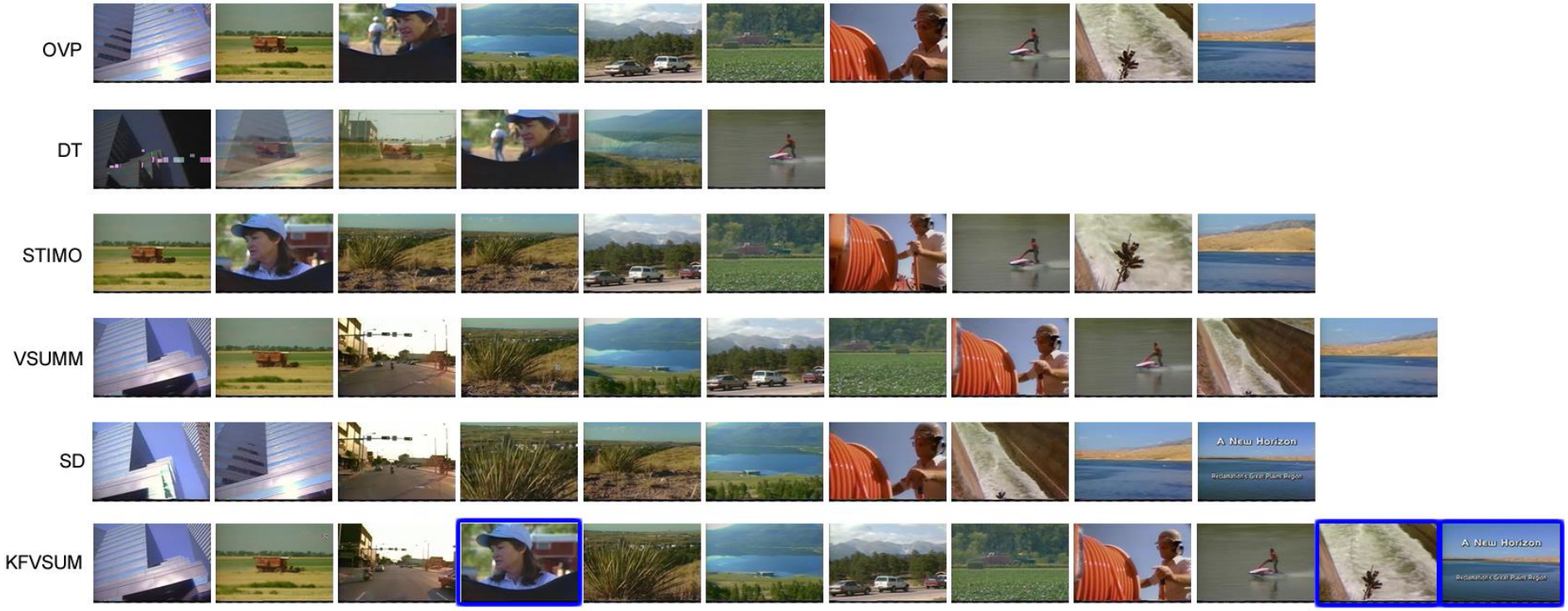
Impact of Clustering



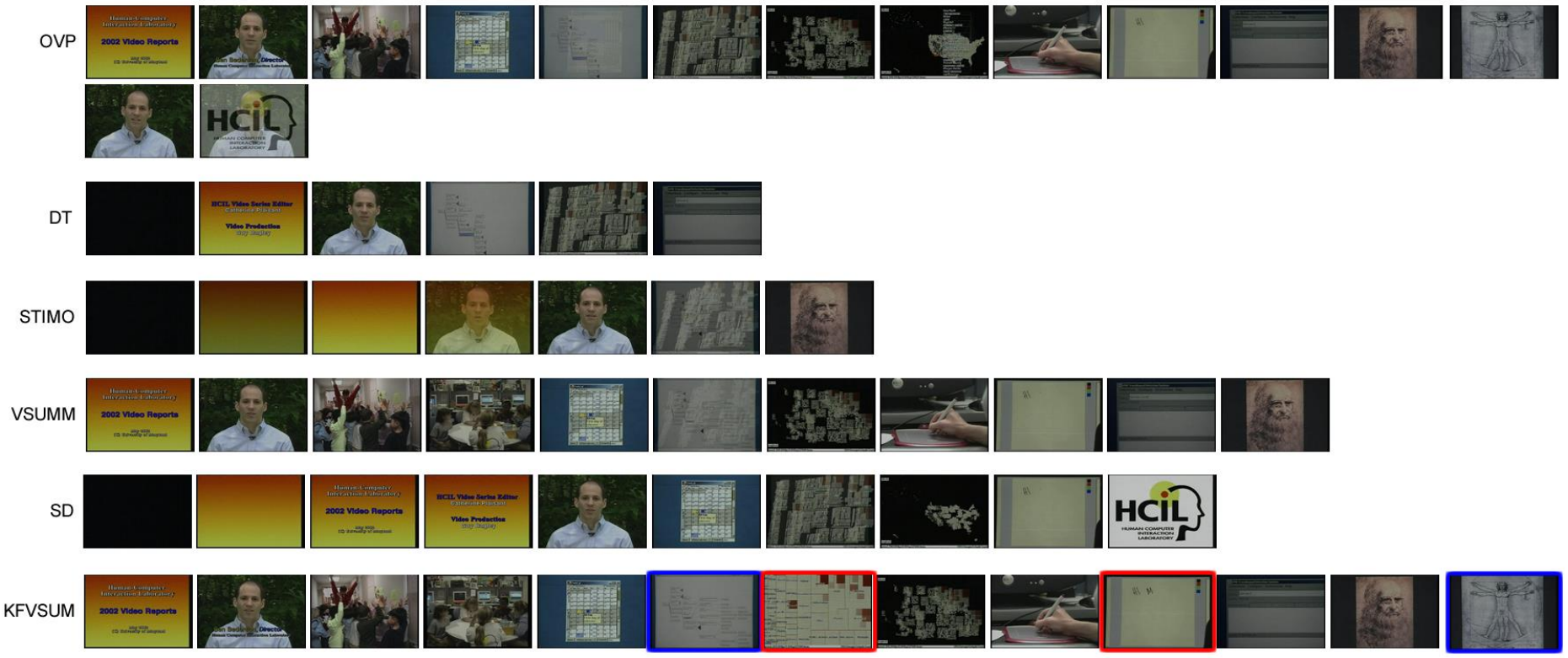
Impact of Saliency Map



Sample Result 1



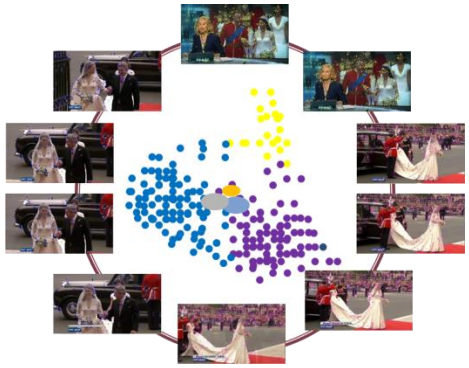
Sample Result 2



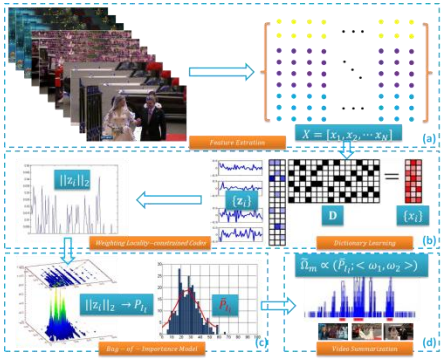
Sample Result 3

- Summarization with Different Lengths

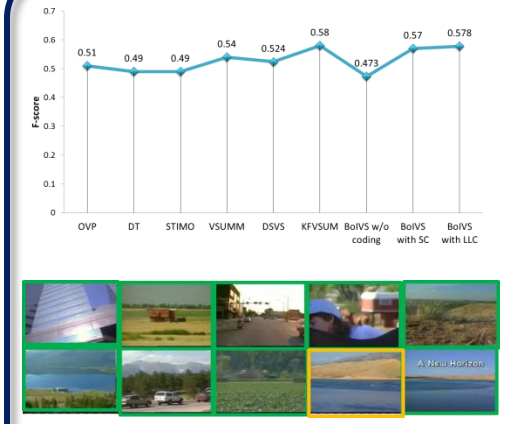
Bag-of-Importance (BoI) Model



Part I:
Motivations



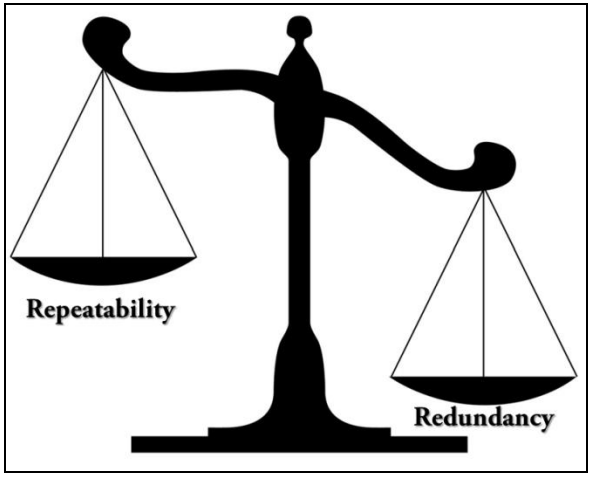
Part II:
Methodology

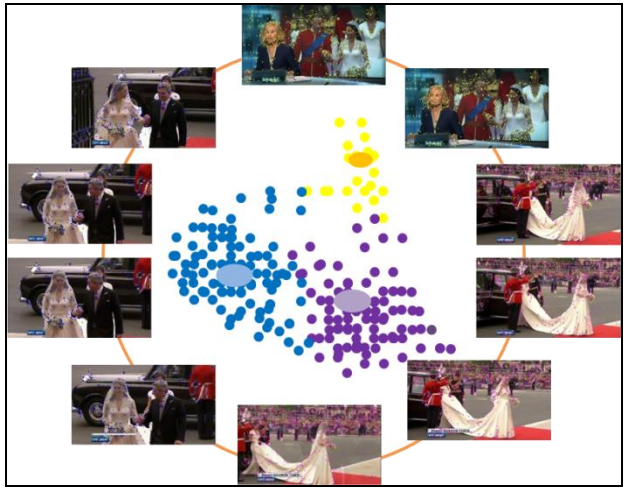


Part III:
Evaluations

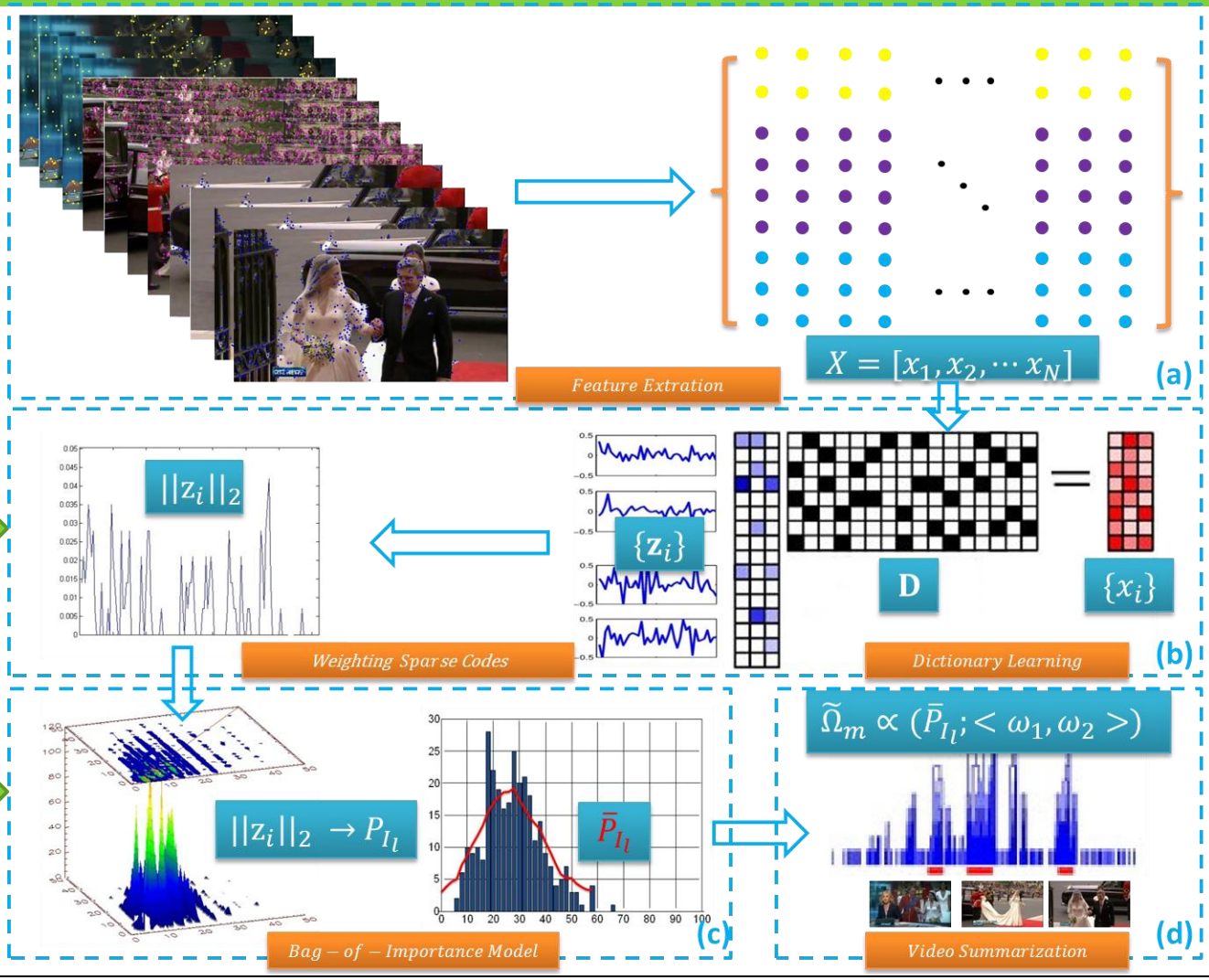
Motivations

- Propose a paradigm for video summarization
  - Identify the invariant and repeatable patterns
    - Capture the essence of the visual patterns
  - Eliminate the redundancy
    - Capture the discriminative details
- Characterize individual features for video summarization





Framework


Identify Repetition
Eliminate Redundancy
$X = [x_1, x_2, \cdots x_N]$
Feature Extration
(a)
$||z_i||_2$
$\{z_i\}$
D
$\{x_i\}$
Weighting Sparse Codes
Dictionary Learning
(b)
$||z_i||_2 \rightarrow P_{I_l}$
$\bar{P}_{I_l}$
Bag – of – Importance Model
(c)
$\widetilde{\Omega}_m \propto (\bar{P}_{I_l}; < \omega_1, \omega_2 >)$
Video Summarization
(d)

Feature Learning

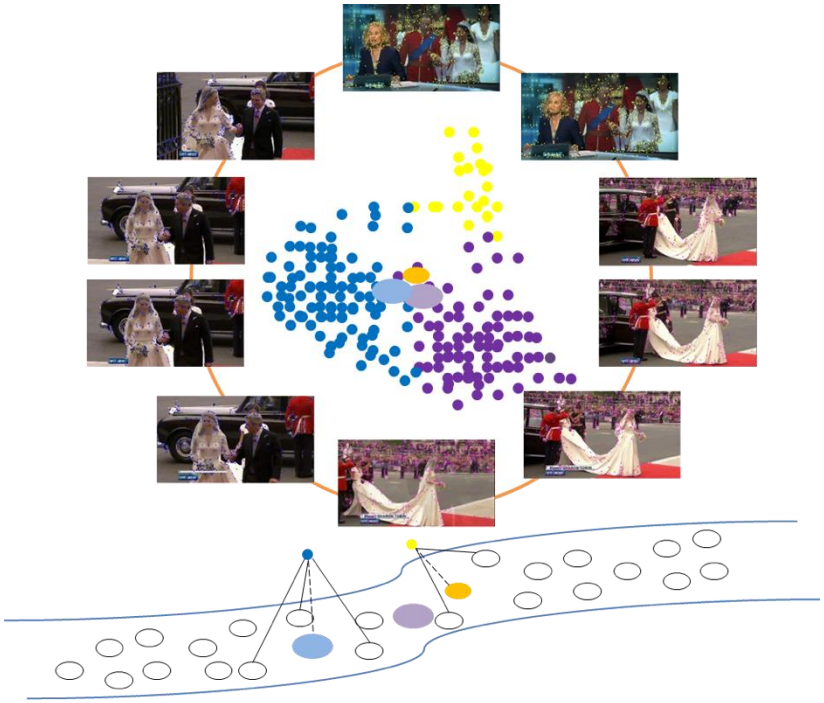
- Learn the Dictionary by Sparse Pursuit

$$\min_{\mathbf{D},\{\mathbf{z}_i\}_{i=1,\ldots,N}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\mathbf{z}_i\|_{l_2}^2 + \lambda\|\mathbf{z}_i\|_{l_1},$$

- Transform the local features into sparse space

$$\min_{\mathbf{d}^j,\{\mathbf{z}_i^j\}} \sum_{i=1}^{N} \|\mathbf{x}_i - \sum_j \mathbf{d}^j * \mathbf{z}_i^j\|_{l_2}^2 + \lambda \sum_j \|\mathbf{z}_i^j\|_{l_1},$$

- Weight the learned feature $\|\mathbf{z}_i\|_2$
  - Project the raw features to an anchor point the transformed space
  - Anchor points – assemble the repetition

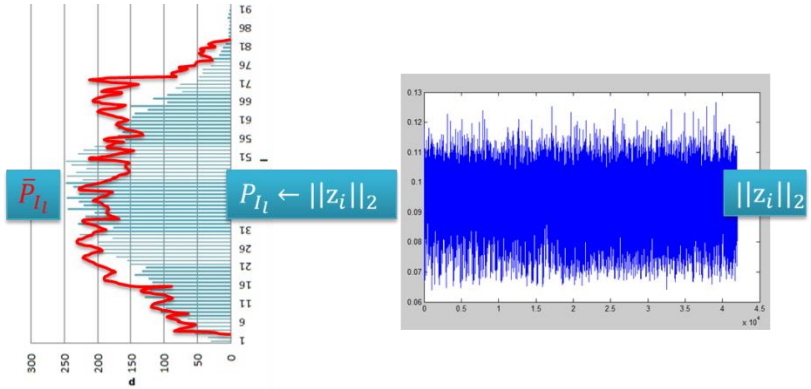


Identify the Bag-of-Importance

- Derive the distribution of the weight coefficients

$$P(||\boldsymbol{z}_i||_2)$$

- The most repeatable learned features are with the highest *P* Value.
- We further borrow TF-IDF concept to reweight $\bar{P}$
  - The “common words” are stopped
  - The discriminative words may be weighted a higher value

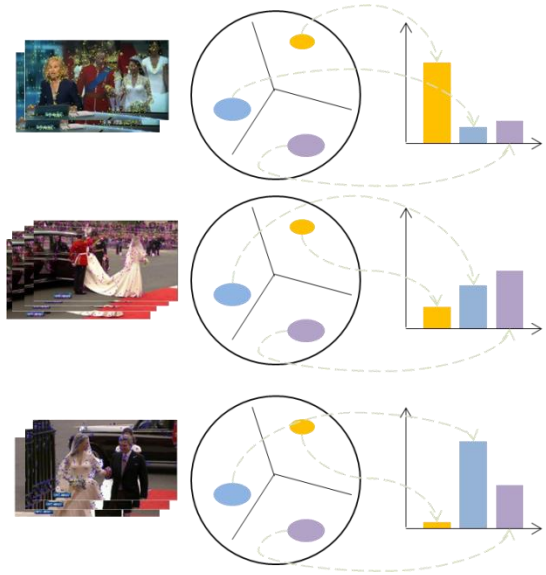


$$\bar{P}_{I_l} = [\frac{f(1|I_l)}{\sum_{t=1}^{L}(f(t|I_l))} log(\frac{M}{f(I_l|1)}), \cdots, \frac{f(t|I_l)}{\sum_{t=1}^{L}(f(t|I_l))} log(\frac{M}{f(I_l|t)}), \cdots, \frac{f(L|I_l)}{\sum_{t=1}^{L}(f(t|I_l))} log(\frac{M}{f(I_l|L)})], \quad (7)$$

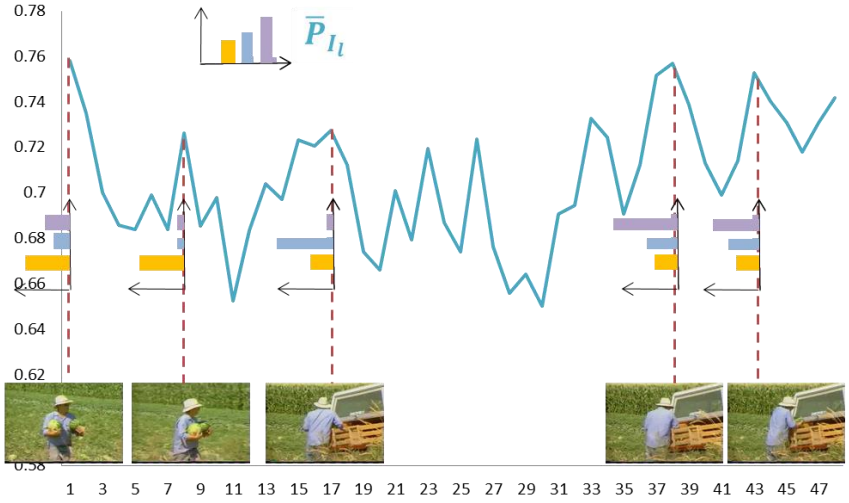
Video summarization by BoI

- We calculate the representativeness score for each frame, by aggregating the important codes inside the frame

$$\Omega_m = \frac{\sum_{i=1}^{N_m}\{\sum_{j=1}^{L} sgn_I(\Delta_i^{X-P(I_L)}, j)\cdot \bar{P}_{I_l}\}}{\sum_{m=1}^{M} N_m},$$



- We generate the representativeness curve, representative frames are detected by identifying the top K local maximum.



Evaluations

Datasets

- Annotated Videos from Open Video Project
  - www.openvideo.org
- Youtube videos

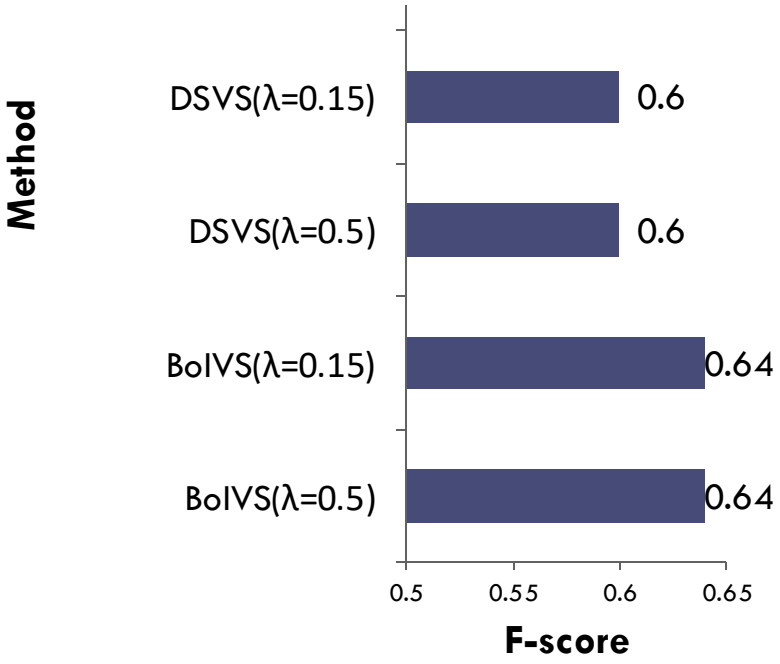
Metric

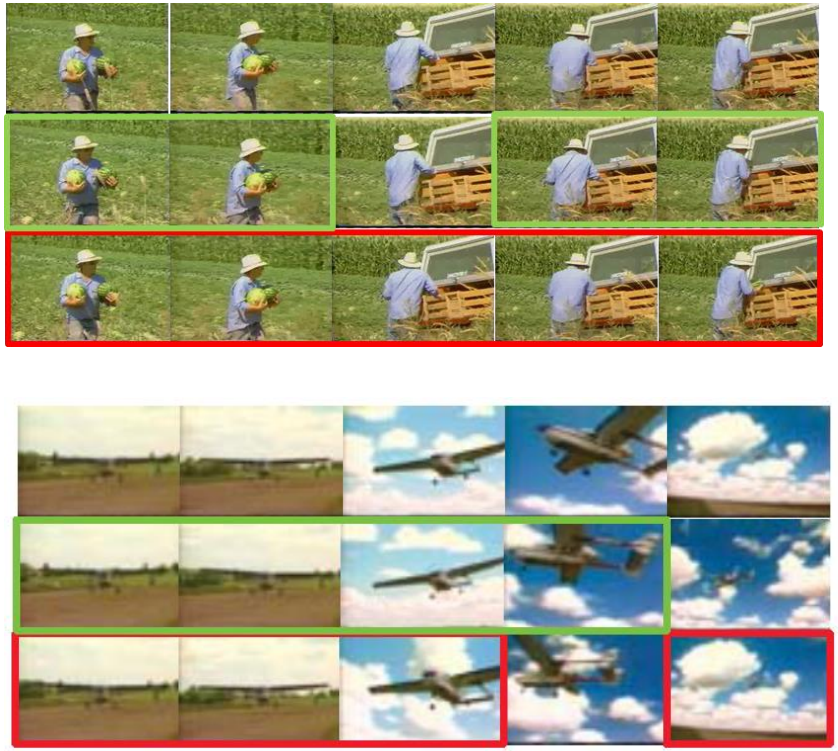
- F-score:

$$F_\beta = \frac{(\beta^2 + 1) Precision \cdot Recall}{\beta^2 Precision + Recall},$$

- β controls the balance between precision and recall.
- The *F-score* can be interpreted as a weighted average of precision and recall, where a score reaches its best value at 1 and worst at 0.

Evaluations at a short length level

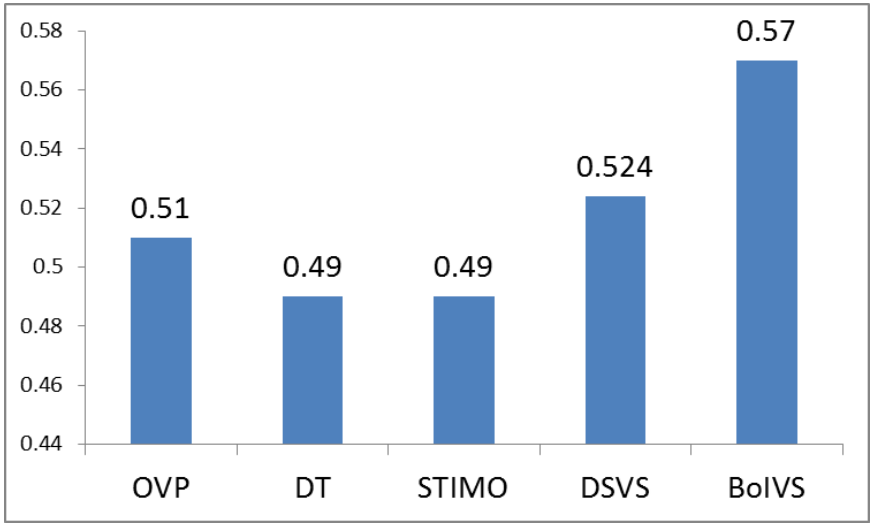




Dsvs-: [Cong *et al.*, 12] **sparse dictionary**

BoIVS: our proposed method

Evaluations at a long length level



OVP: service provider

DT: [Mundur, 2006]

STIMO: [Avila et al., 2011]

DSVS: [Cong, 2012]

BoIVS: proposed by us

Impact of various factors

TABLE IV
PERFORMANCE COMPARISON BETWEEN P_{I_l} VALUE AND $\bar{P}_{I_l}$ VALUE.

Dataset	VSUMM (P_{I_l})	VSUMM ($\bar{P}_{I_l}$)	YouTube (P_{I_l})	YouTube ($\bar{P}_{I_l}$)
F-score	54%	**56.5%**	50%	**52%**

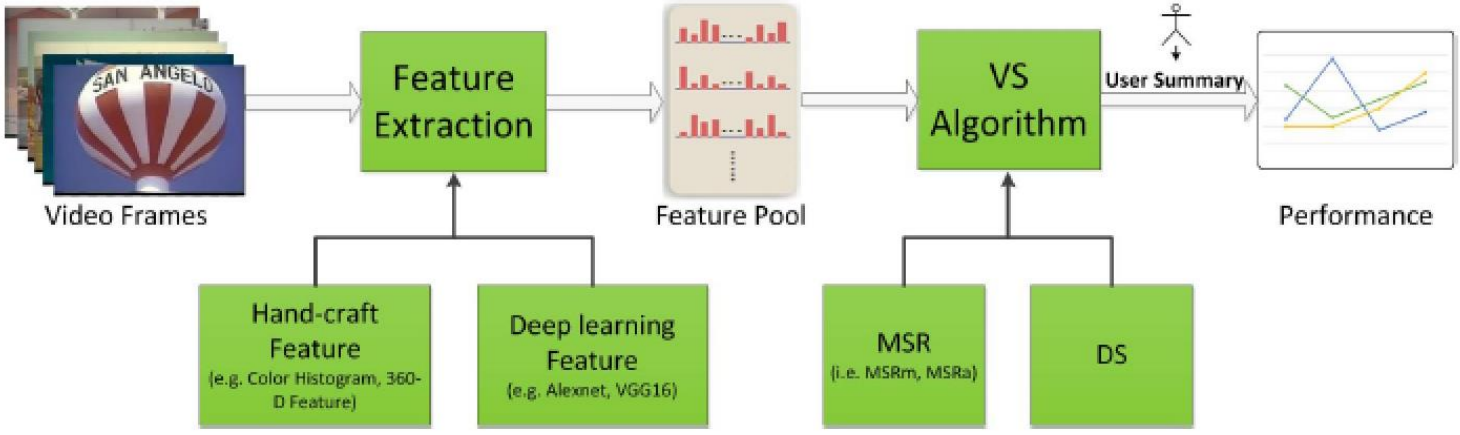
TABLE V
PERFORMANCE COMPARISON BETWEEN THE 1_1 NORM AND THE l_2 NORM.

Dataset	VSUMM (l_1 norm)	VSUMM (l_2 norm)	YouTube (l_1 norm)	YouTube (l_2 norm)
F-score	51.3%	**56.5%**	49.4%	**52%**

Summary

- Introduce a new perspective into video summarization
- Utilize local features for video summarization at finer level
- Introduce a new BoI framework for video summarization
- Promising future for exploiting the value of local features

Deep Features



M. Ma, et al., Exploring the Influence of Feature Representation for Dictionary Selection based Video Summarization, ICIP 2017.

Deep Features

Table 1. Details of the CNNs for feature extraction of video frames.

Network	Input size	Layer	Dimension
Alexnet	$224 \times 224 \times 3$	'fc6'	4096-D
BN-Inception	$224 \times 224 \times 3$	global_pool	1024-D
Inception-v3	$229 \times 299 \times 3$	global_pool	2048-D
VGG16	$229 \times 299 \times 3$	fc6	4096-D

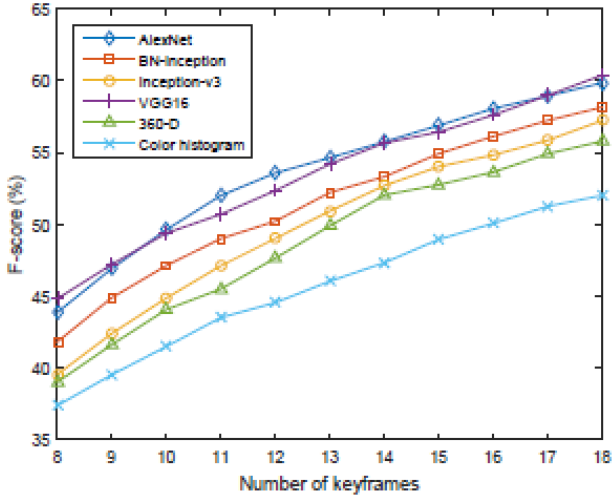
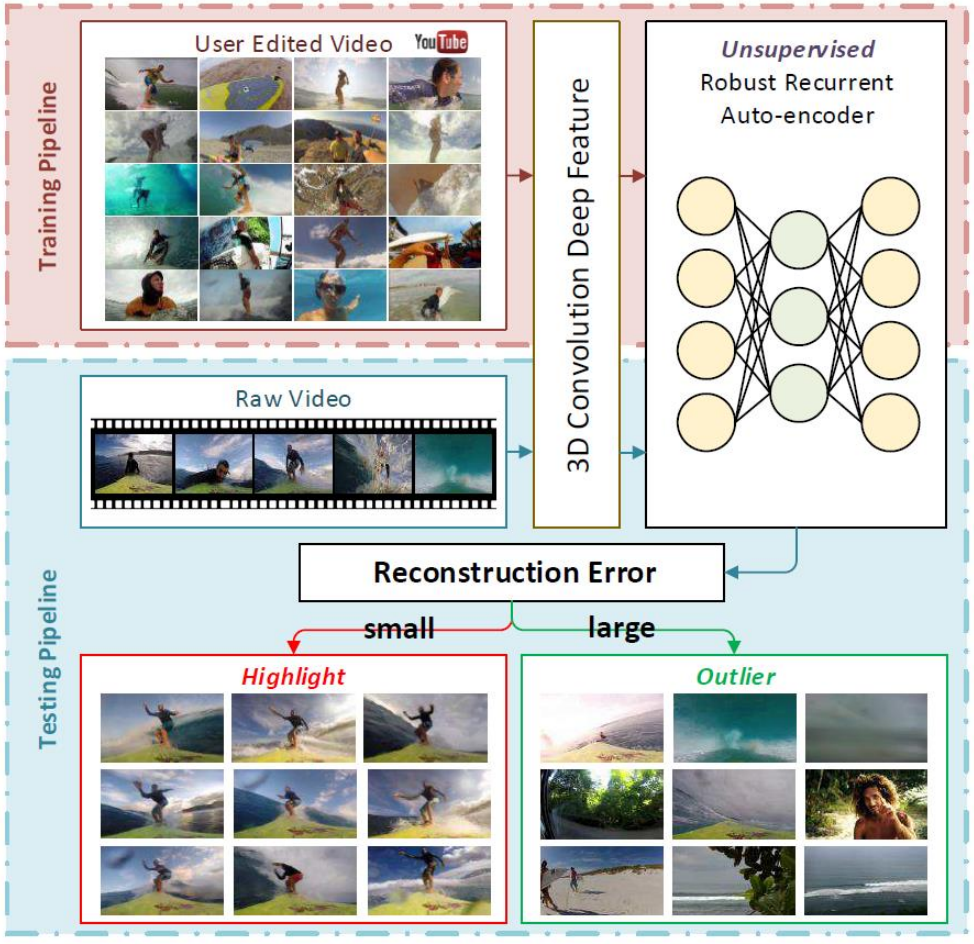


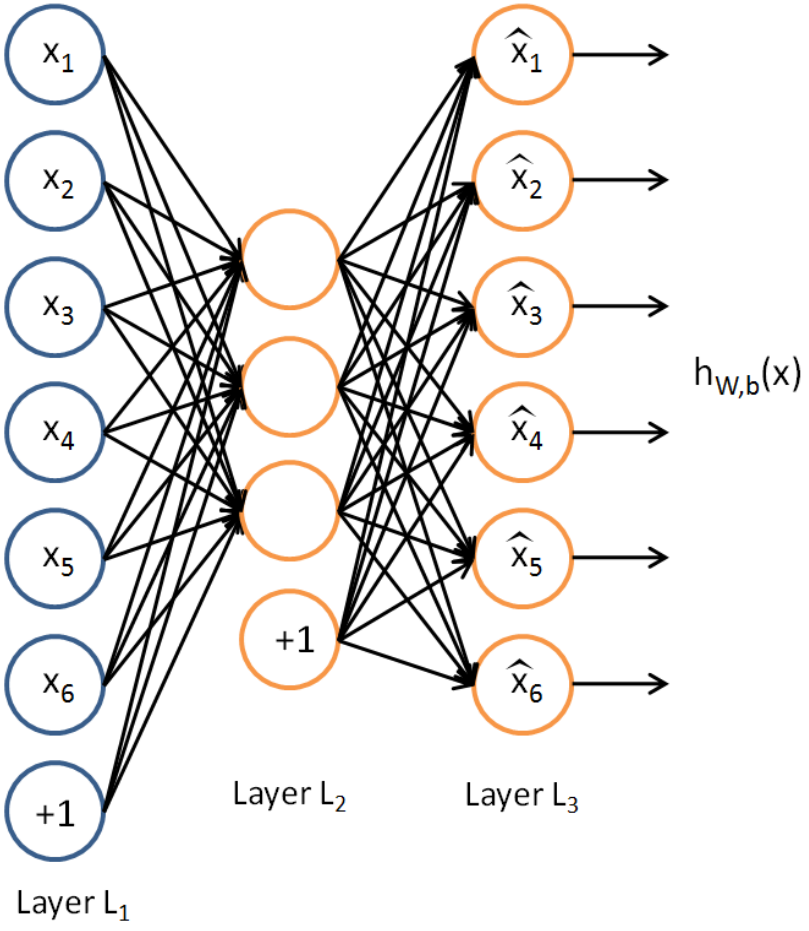
Fig. 2. Experimental results of MSRm with different features.

Recurrent Auto-Encoder for Unsupervised Highlight Extraction



H. Yang, et al., Unsupervised Extraction of Video Highlights via Robust Recurrent Auto-encoders, ICCV 2015.

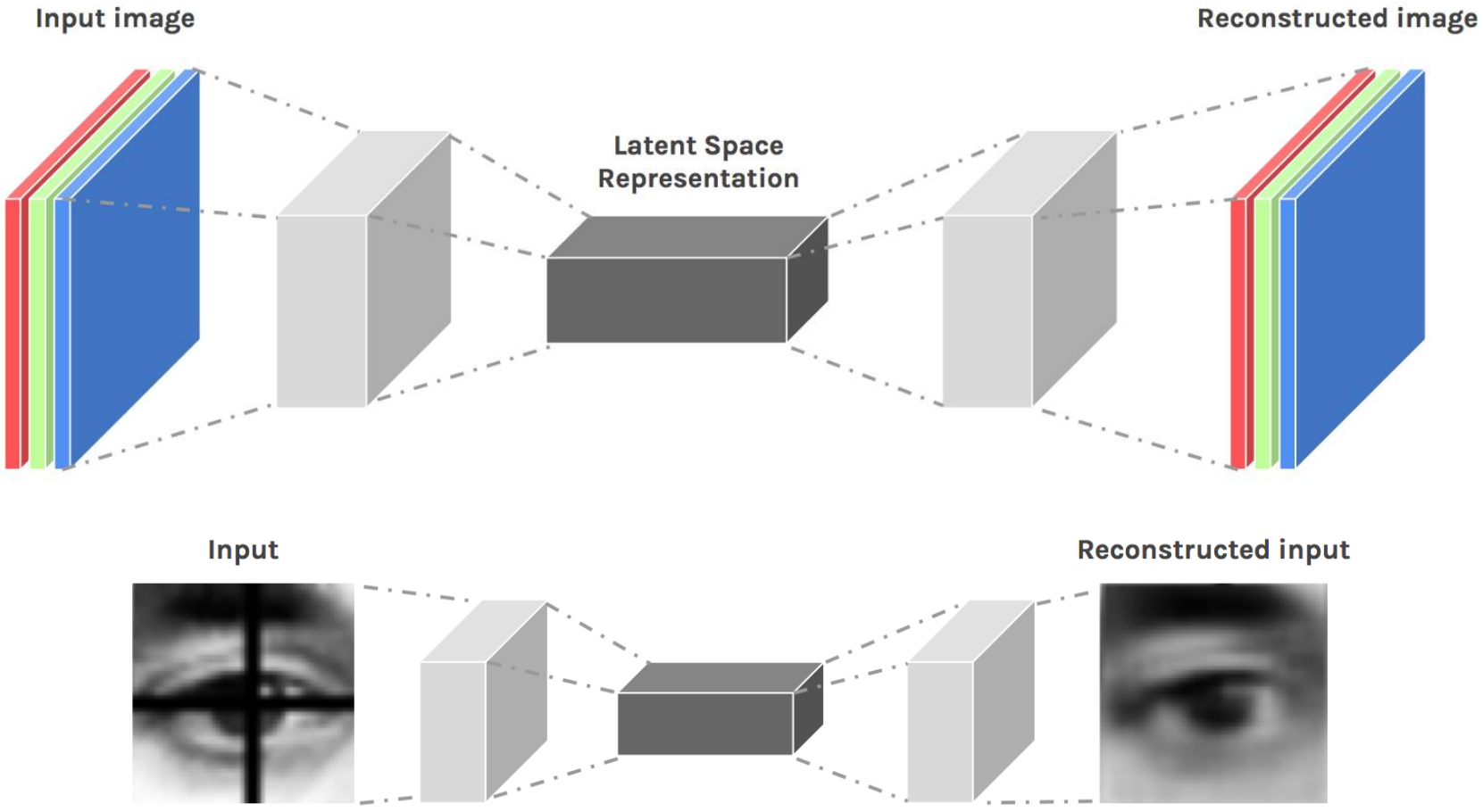
Auto-encoder



The autoencoder tries to learn a function $h_{W,b}(x) \approx x$. In other words, it is trying to learn an approximation to the identity function, so as to output $\hat{x}$ that is similar to x.

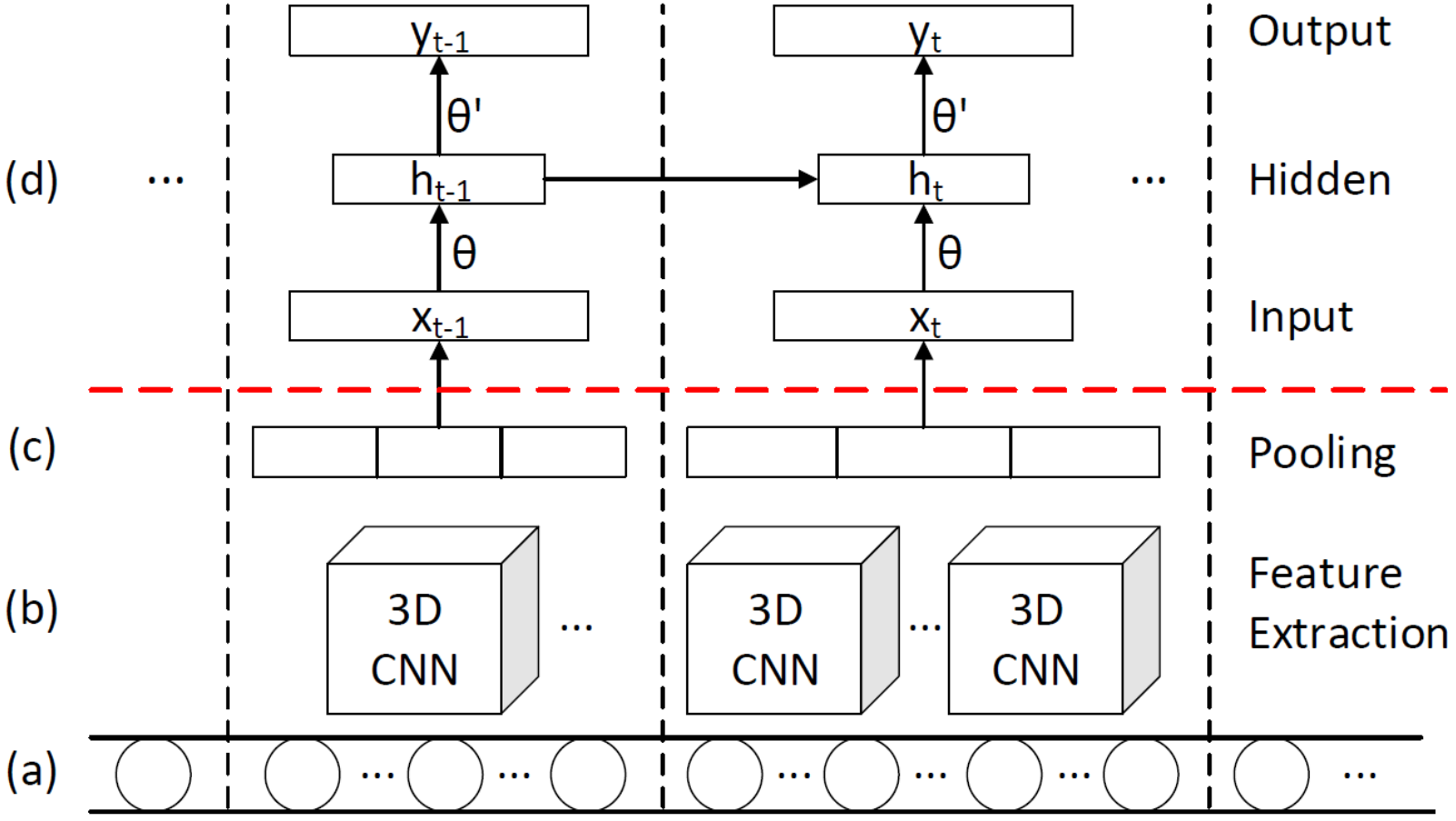
http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

Auto-encoder



https://towardsdatascience.com/autoencoders-are-essential-in-deep-neural-nets-f0365b2d1d7c

Recurrent Auto-Encoder for Unsupervised Highlight Extraction



Recurrent Auto-Encoder for Unsupervised Highlight Extraction

Figure 5. Highlight detection results in different video domains. The blue bar represents reconstruction error, where highlights tend to have smaller errors than non-highlight snippets. The red borders indicate snippets detected as highlights.

Recurrent Auto-Encoder for Unsupervised Highlight Extraction

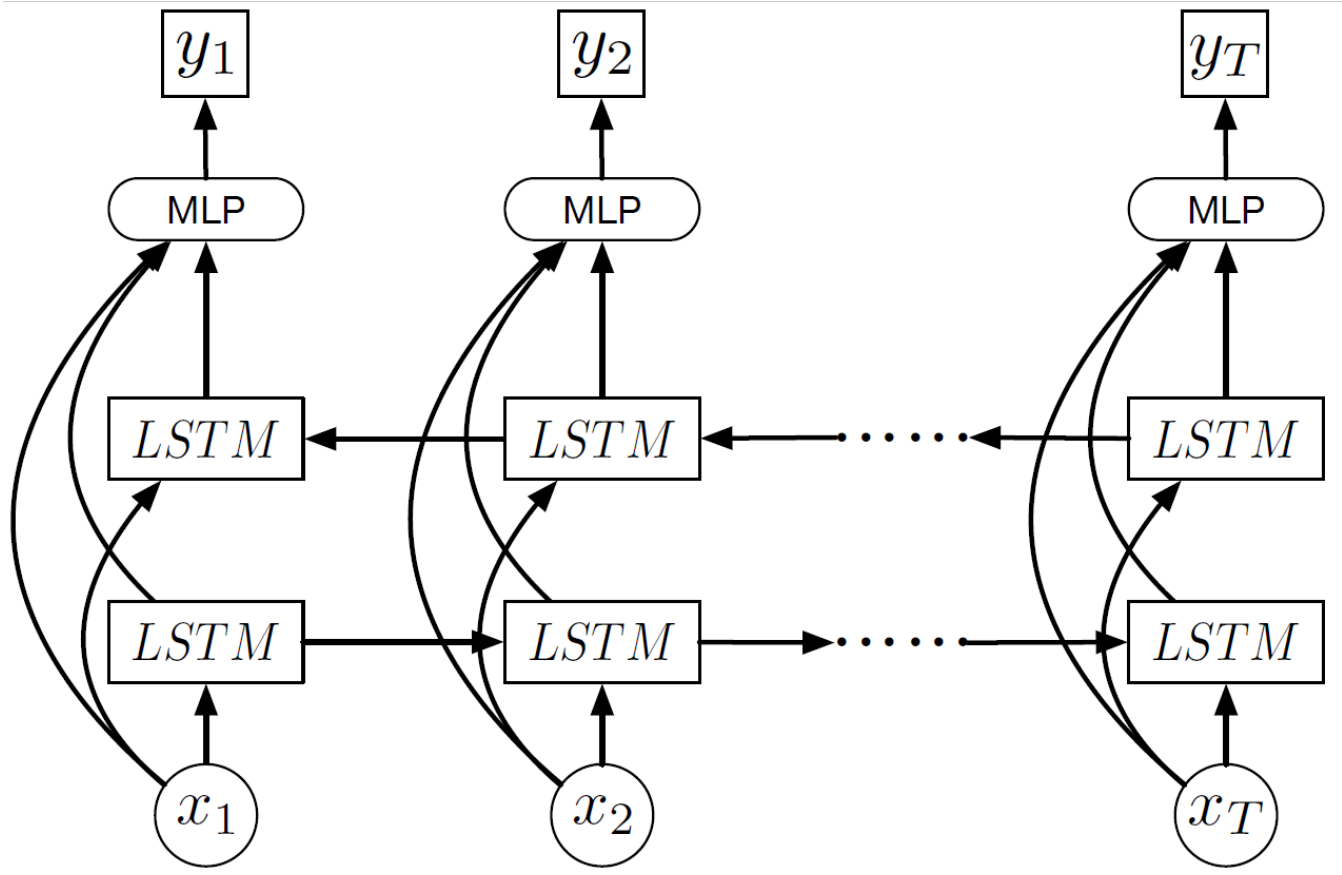
	LRSVM [24]	PCA	OCSVM	AE	Robust AE	Recurrent AE	RRAE
freeride	*	0.235	0.258	0.268	0.277	0.277	**0.288**
parkour	0.246	0.377	0.445	0.507	0.508	0.618	**0.675**
skating	0.330	0.251	0.297	0.308	0.306	0.322	**0.332**
skiing	0.337	0.388	0.412	0.428	0.472	0.478	**0.485**
skydiving	*	0.376	0.332	0.335	0.364	0.338	**0.390**
surfing	0.564	0.525	0.484	0.494	0.534	0.565	**0.582**
swimming	*	0.274	0.238	0.255	0.277	0.275	**0.283**
mAP		0.347	0.352	0.371	0.391	0.410	**0.434**

Table 4. Performance results of our methods and several baseline methods, all using C3D features [26]. The dimensionality of C3D features is reduced from 4096 by a domain specific PCA that keeps 90% of the total energy.

	Supervised[24]	*RRAE*
dog	0.60	0.49
gymnastics	0.41	0.35
parkour	0.61	0.50
skating	0.62	0.25
skiing	0.36	0.22
surfing	0.61	0.49

Table 5. mAP comparison to [24] on the YouTube dataset.

LSTM for VS



K. Zhang, et al., Video Summarization with Long Short-term Memory, ECCV 2016.

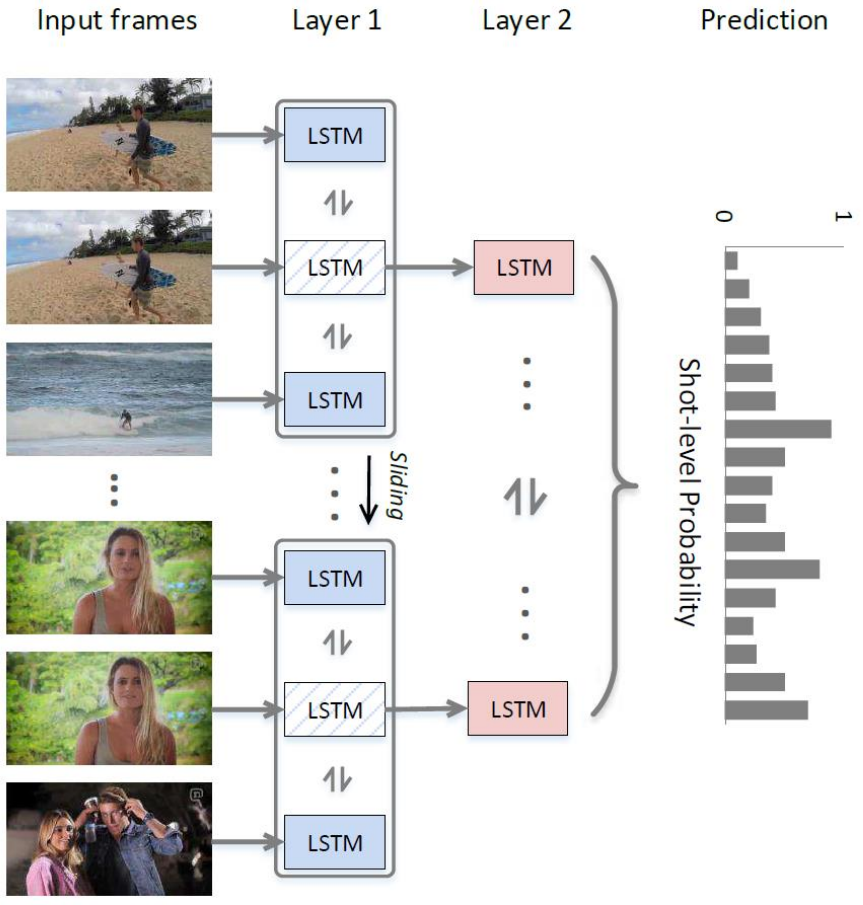
LSTM for VS

Dataset	Settings	Training & Validation	Testing
SumMe	Canonical	80% SumMe	20% SumMe
	Augmented	OVP + Youtube + TVSum + 80% SumMe	20% SumMe
	Transfer	OVP + Youtube + TVSum	SumMe
TVSum	Canonical	80% TVSum	20% TVSum
	Augmented	OVP + Youtube + SumMe + 80% TVSum	20% TVSum
	Transfer	OVP + Youtube + SumMe	TVSum

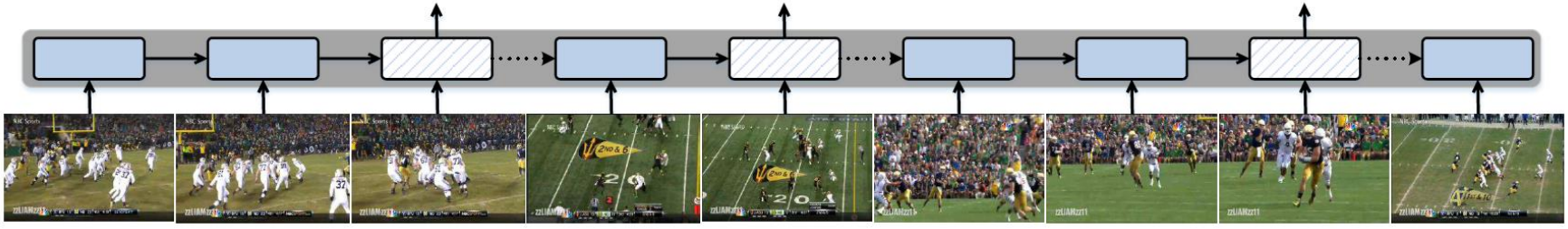
LSTM for VS

Dataset	Method	unsupervised	Canonical	Augmented	Transfer
SumMe	[30]	*26.6*			
	[17]		*39.4*		
	[15]		*39.7*		
	[16]		*40.9*†	41.3	38.5
	vsLSTM (ours)		37.6±0.8	41.6±0.5	40.7±0.6
	dppLSTM (ours)		38.6±0.8	42.9±0.5	41.8±0.5
TVSum	[34]	*46.0*			
	[11]‡	*36.0*			
	[35]‡	*50.0*			
	vsLSTM (ours)		54.2±0.7	57.9±0.5	56.9±0.5
	dppLSTM (ours)		54.7±0.7	59.6±0.4	58.7±0.4

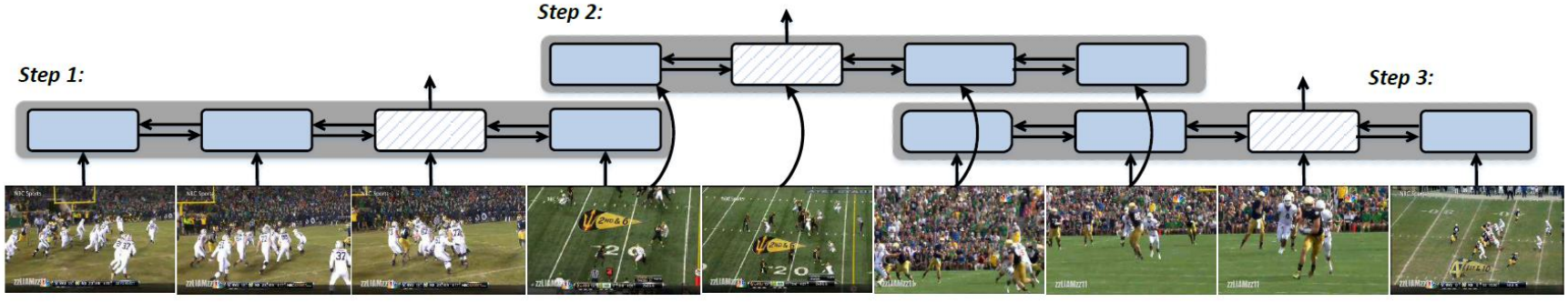
Hierarchical Structure-Adaptive RNN for Video Summarization



B. Zhao, et al., Hierarchical Structure-Adaptive RNN for Video Summarization, CVPR 2018.



(a) Shot boundary detection with long single LSTM



(b) Shot boundary detection with sliding bidirectional LSTM

HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization.(cvpr18)

Table 3. The summarization results of various approaches on the VTW dataset. (The scores in bold indicate the best values.)

Feature	shallow feature			deep feature		
Metrics	Precision	Recall	F-measure	Precision	Recall	F-measure
CSUV [12]	0.367	0.423	0.393	–	–	–
HD-VS [33]	–	–	–	0.392	0.483	0.433
vsLSTM [37]	0.388	0.490	0.433	0.397	0.495	0.441
Hierarchical RNN [38]	0.408	0.516	0.456	0.417	0.525	0.465
HSA-RNN	**0.434**	**0.537**	**0.480**	**0.443**	**0.548**	**0.491**

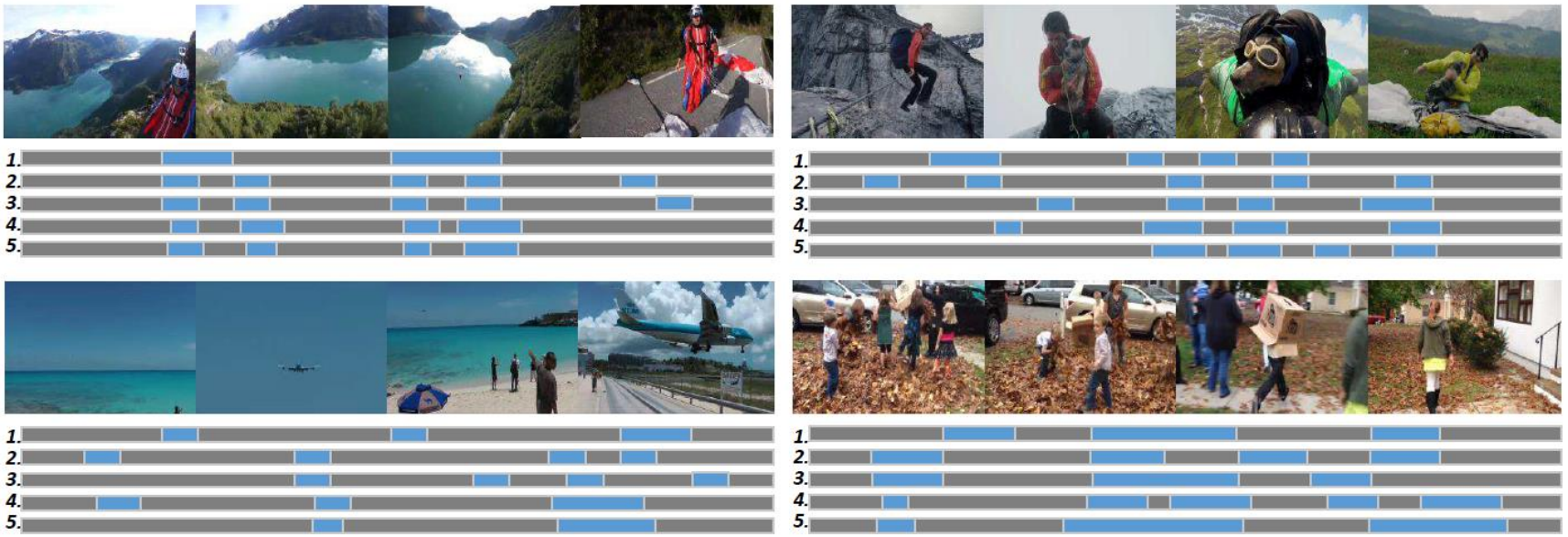
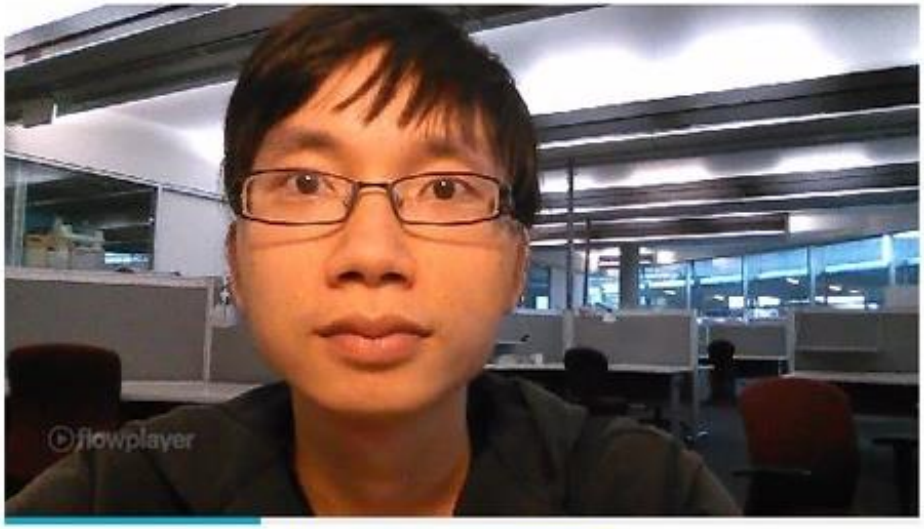


Figure 4. Four exemplar results from the Combined dataset. Each video is depicted by four frames. The five bars below each video represent the summaries generated by *vsLSTM*, *dppLSTM*, *Hierarchical RNN*, *HSA-RNN* and human beings, respectively. Specifically, the long gray bar stands for the whole video stream, and the short blue bar denotes the selected key shot.

HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization.(cvpr18)

Presentation

http://rp-www.cs.usyd.edu.au/~ggua5470/keyframe-demo/



(a) 2D BS-Video Collage (left: 2D collage, right: 2D collage with blending edges).

(b) 1D BS-Video Collage (top: 1D collage, down: 1D collage with blending edges).

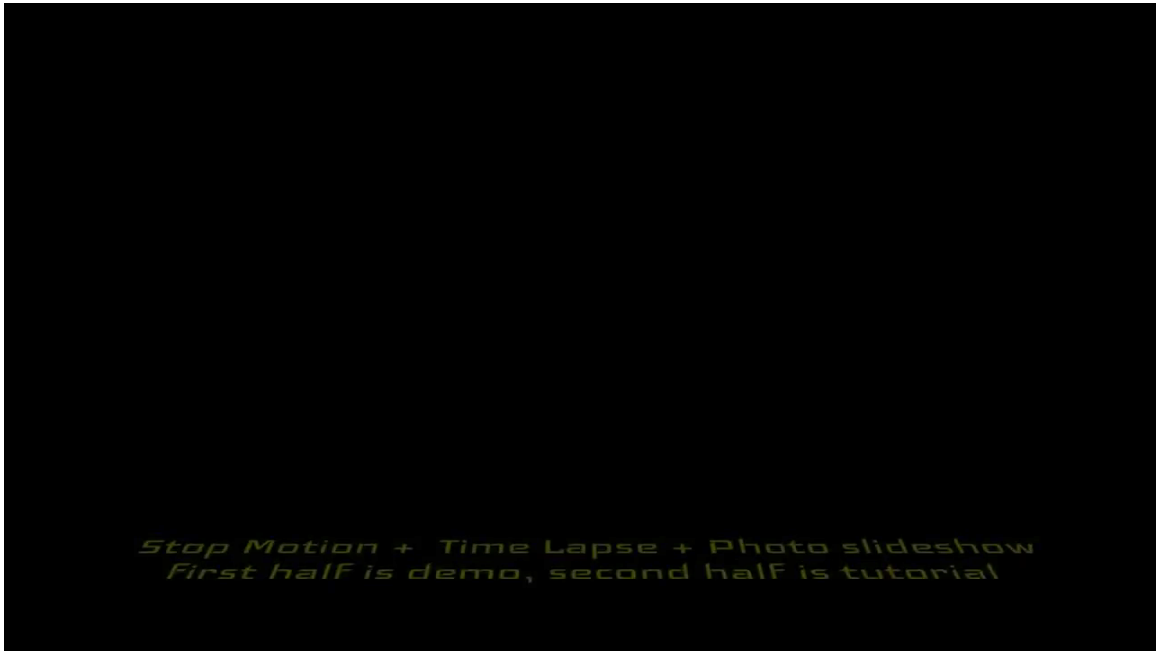
(c) FS-Video Collage (left: collage with "heart" template and *book* layout, right: collage with "fan" template and *spiral* layout).

T. Mei, et al., Video collage: presenting a video sequence using a single image, The Visual Computer 25(1): 39-51 (2009)

Presentation

- Video Synopsis of Brief Cam

PicPac Stop Motion



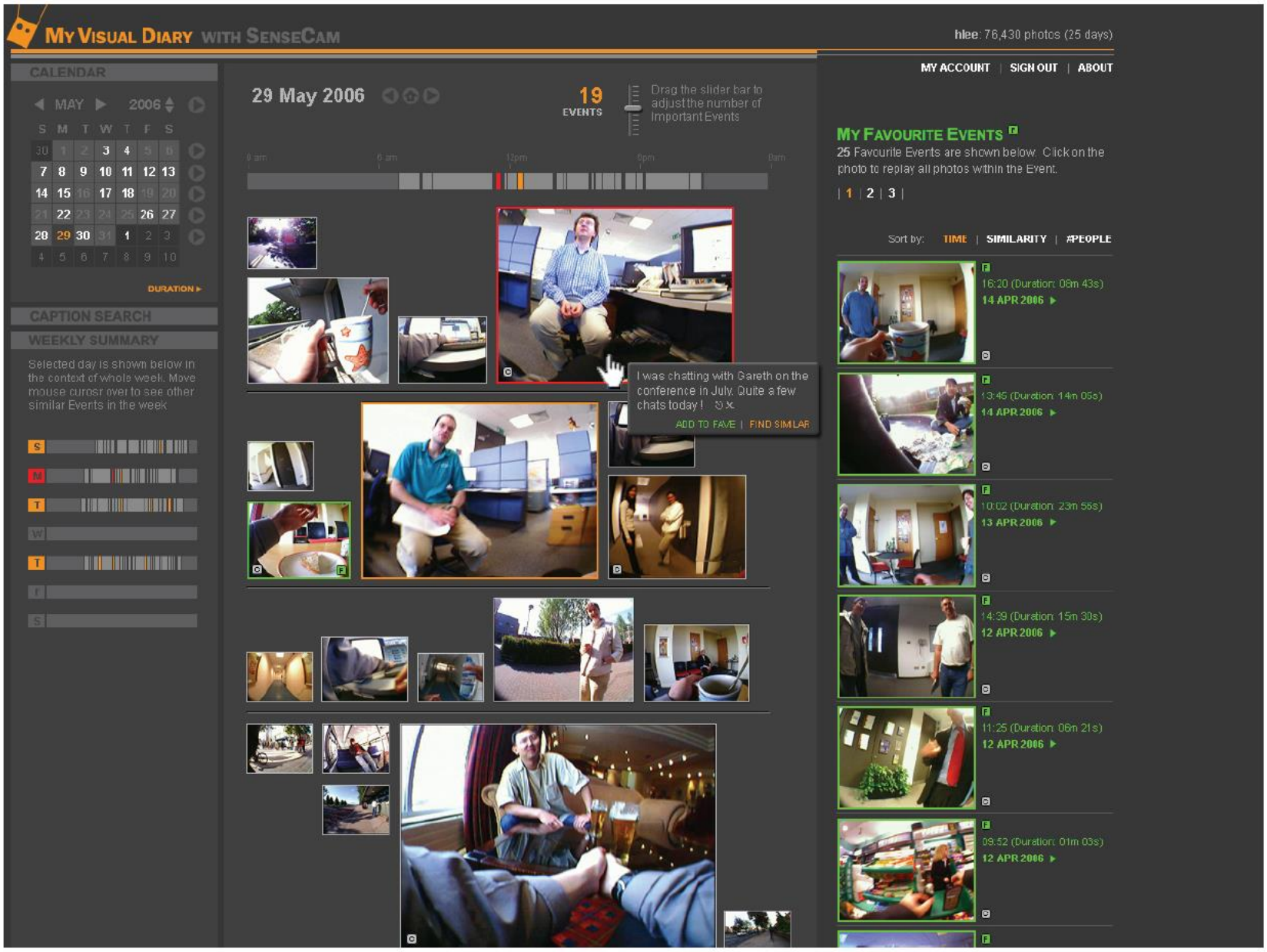
http://picpac.tv/

Applicaitons

- Summarizing LifeLog

Microsoft SenseCam

Hyowon Lee, Alan F. Smeaton, Noel E. O'Connor and Gareth J.F. Jones, Adaptive Visual Summary of LifeLog Photos for Personal Information, International Workshop on Adaptive Information Retrieval, 2006.


MY VISUAL DIARY WITH SENSECAM
hlee: 76,430 photos (25 days)
MY ACCOUNT | SIGN OUT | ABOUT
CALENDAR
MAY 2006
S M T W T F S
30 1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30 31 1 2 3
4 5 6 7 8 9 10
DURATION
CAPTION SEARCH
WEEKLY SUMMARY
Selected day is shown below in the context of whole week. Move mouse cursor over to see other similar Events in the week
S
M
T
W
T
F
S
29 May 2006
19
EVENTS
Drag the slider bar to adjust the number of important Events
6 am
6 am
12pm
6pm
0am
I was chatting with Gareth on the conference in July. Quite a few chats today !
ADD TO FAVE | FIND SIMILAR
MY FAVOURITE EVENTS
25 Favourite Events are shown below. Click on the photo to replay all photos within the Event.
| 1 | 2 | 3 |
Sort by: TIME | SIMILARITY | #PEOPLE
16:20 (Duration: 08m 43s)
14 APR 2006
13:45 (Duration: 14m 05s)
14 APR 2006
10:02 (Duration: 23m 56s)
13 APR 2006
14:39 (Duration: 15m 30s)
12 APR 2006
11:25 (Duration: 06m 21s)
12 APR 2006
09:52 (Duration: 01m 03s)
12 APR 2006

Applications

- PhotoSynth

Photo Tourism

Exploring photo collections in 3D

Microsoft

(a)

(b)

(c)

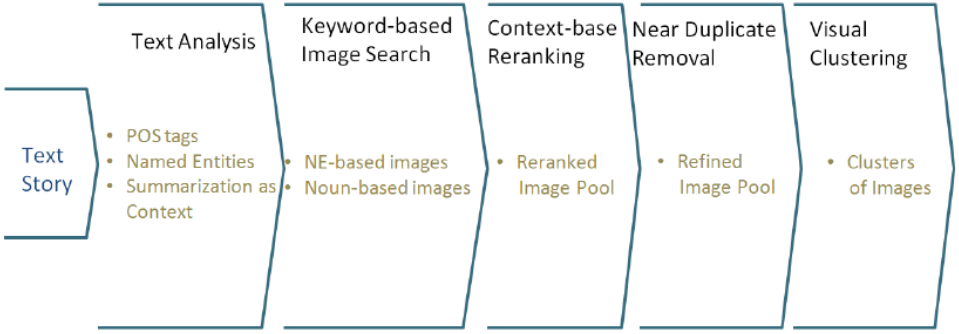
http://phototour.cs.washington.edu/

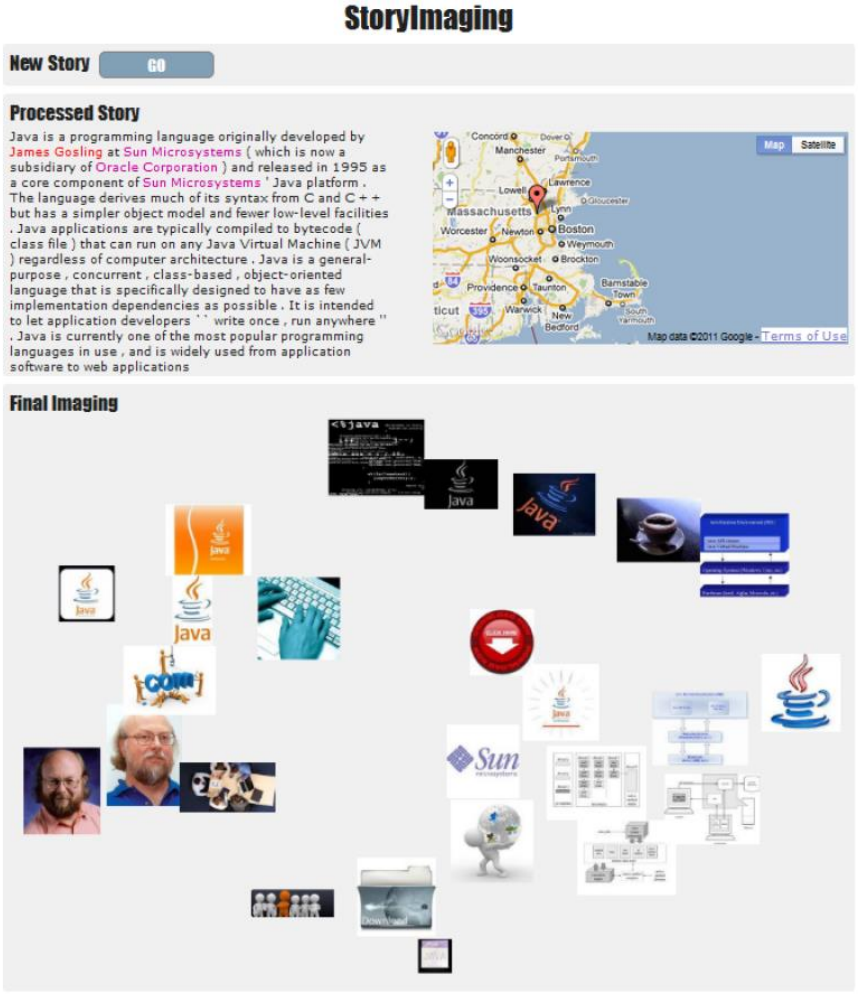
Applications

- PhotoSynth
  - http://photosynth.net
- How PhotoSynth can connect the world's images
  - http://www.ted.com/talks/blaise_aguera_y_arcas_demos_photosynth
- Photo Tourism
  - http://phototour.cs.washington.edu/

Applications

- StoryImaging





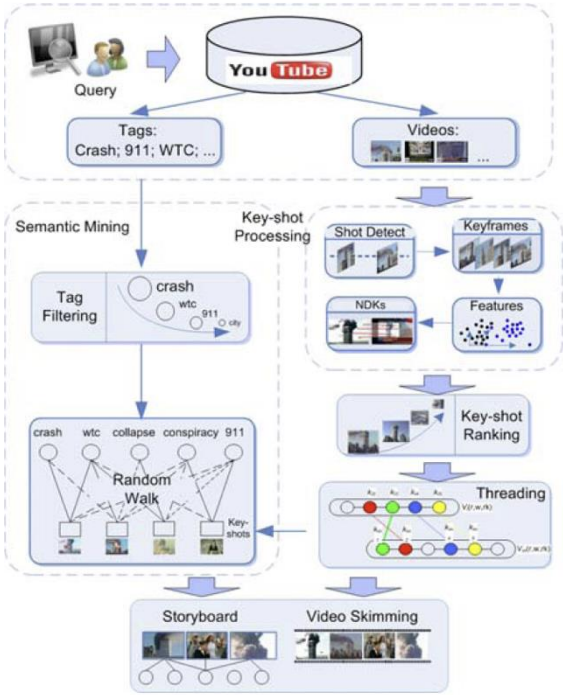
G. Guan, Z. Wang, X.-S. Hua, and D. Feng, StoryImaging: a media-rich presentation system for textual stories, ACM MM 2011.

Beyond Search: Event Driven Summarization for Web Videos TOMCCAP 2011 NGO

Undirected Graph

- NDK -> key-shots->graph
- Rank the key-shots
  - Informative scores
  - the chronological order
- Key-shot tagging
  - Tag filtering
  - Tag propagation
    - Random walk
- Summarization
  - Trade-off between the sum of relevance and time interval

$$D = \arg\max_{D}\Big(\sum_{l\in D} ifo(s_l) + \beta\frac{1}{|D|}\sum_{l,m\in D}\|\lambda_l - \lambda_m\|^2\Big)$$
$$s.t.\ \sum_{l\in D} length(s_l) < T$$



More on Summarisation

- Multi-document summarisation
- Multi-video summarisation
- Multi-modal summarisation
- Query based summarisation
- eXtreme summarisation
- Domain-specific summarisation
- … …

Need to Know

- Text summarization
- Video summarization problem
- Categories of existing solutions
- A new perspective into video summarization with local features
- Applications