Video Co-summarization: Summarizing Videos Using Visual Co-occurrence

Wen-Sheng Chu, Yale Song and Alejandro Jaimes

YAHOO! LABS

Video summarization

surfing

Travel Nicaragua: Surfing

Summaries attractive to users?

Summaries attractive to users?

Summaries attractive to users?

A desired method

- ☐ Generates adaptive summaries that fits user's interests

Statistics about videos

- On December 2012
  - 100 hours: # of hours of videos uploaded / minute
  - 82.5%: % of US audience that viewed videos online
  - 200B: # of videos viewed online / month
  - 4B: # of hours of video viewed / month

http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

A desired method

- ☐ Generates adaptive summaries that fits user's interests

- ☐ Scales to large datasets

Supervised video summarization

- Sports videos
  - Canonical views

E.g., Fleischman et al. [ACMMM'07]

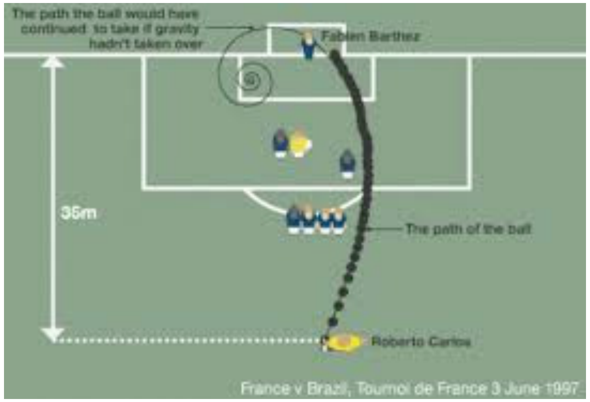
E.g., Chen & Vleechouwer [TCSVT'11]

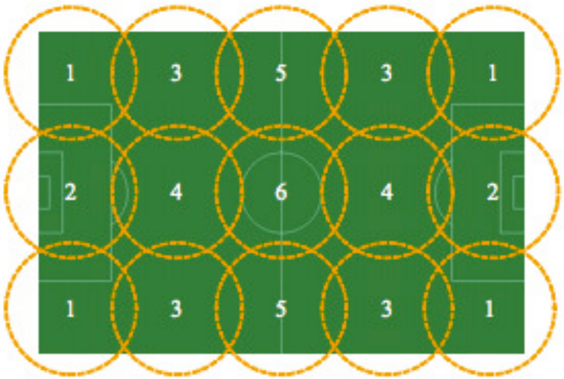
Supervised video summarization

- Sports videos
  - Canonical views

E.g., Fleischman et al. [ACMMM'07]

E.g., Chen & Vleechouwer [TCSVT'11]





E.g., Zhu et al. [ACMMM'07]

Supervised video summarization

- News videos
  - Topic themes
  - Rich texts/transcripts

E.g. Wu et al. [SPM'06], Liu et al. [ACMMM'12]

Supervised video summarization

- Surveillance videos
  – Stationary background

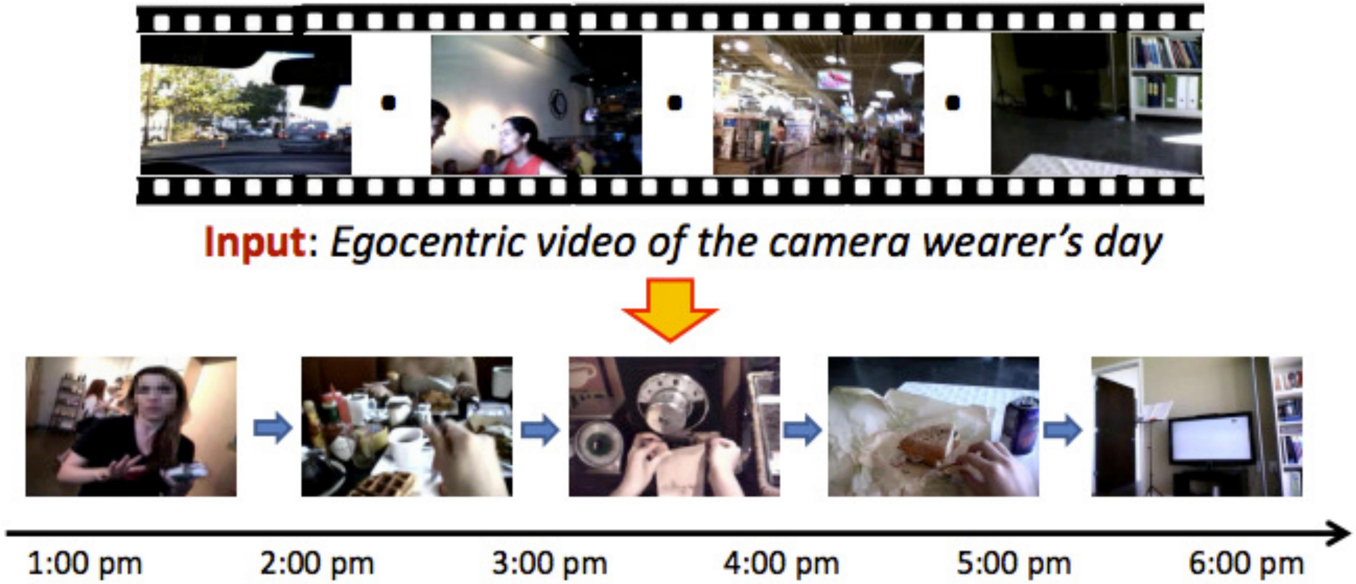
(a)

Synopsis: Pritch et al. [TPAMI'08]

Online video condensation: Feng et al. [CVPR'12]

Supervised video summarization

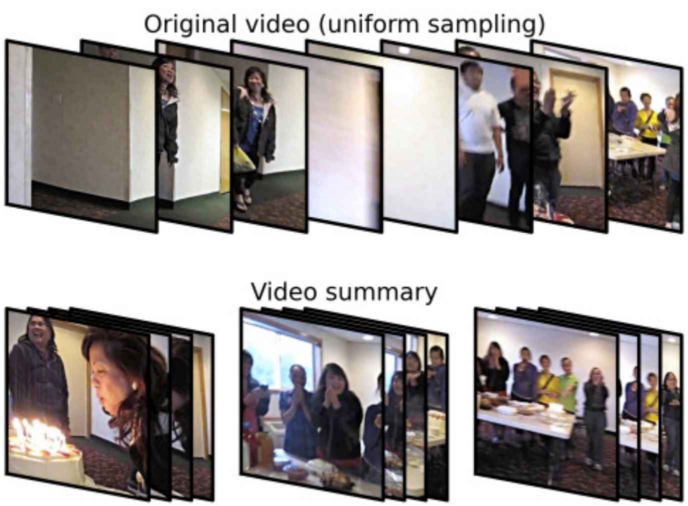
- Learn to summarize videos
  - Egocentric videos: use clues from faces, hands, interesting objects



E.g., Lee et al. [CVPR'12], Lu and Grauman [CVPR'13]

Supervised video summarization

- Learn to summarize videos
  - Consumer videos: learn to estimate per-frame interestingness from annotated data



Potapov et al. [ECCV'14]

Sun et al. [ECCV'14]

A desired method

- ☐ Generates adaptive summaries that fits user's interests
- ☐ Scales to large datasets
- ☐ Requires limited/no human supervision

A desired method

- ☐ Generates adaptive summaries that fits user's interests

- ☐ Scales to large datasets

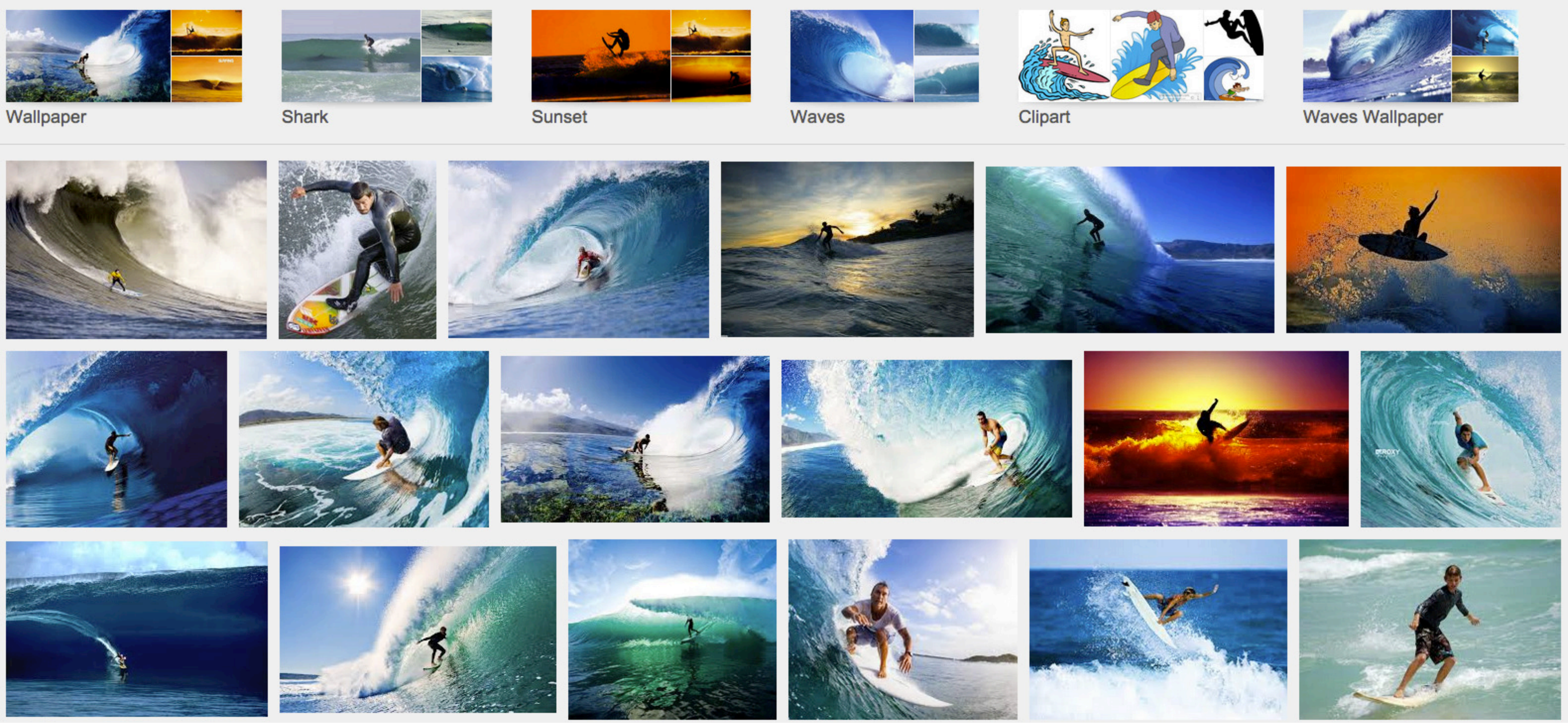
- ☐ Requires limited/no human supervision

Unsupervised video summarization

- No prior knowledge and annotated data
  - Sparse dictionary learning: Cong et al. [TMM'12], Zhao and Xing [CVPR'14]
  - Hierarchical clustering: Mahmoud et al. [ICMLA'13]
- Additional resources
  - Human attention during video watching: Ngo et al. [TCSVT'05]
  - Web image priors: Khosla et al. [CVPR'13], Kim et al. [CVPR'14]

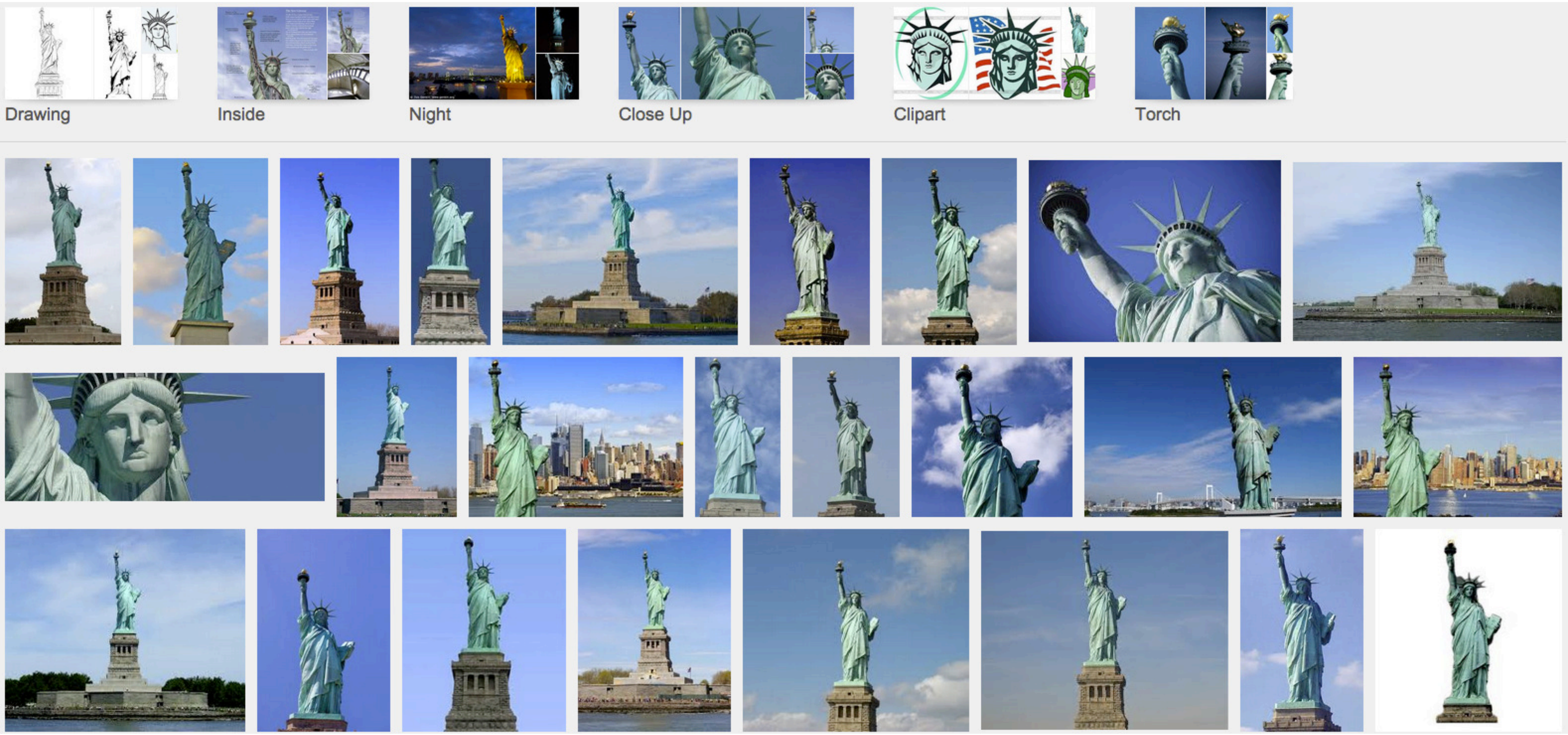
Important concepts repeat visually

- Surfing



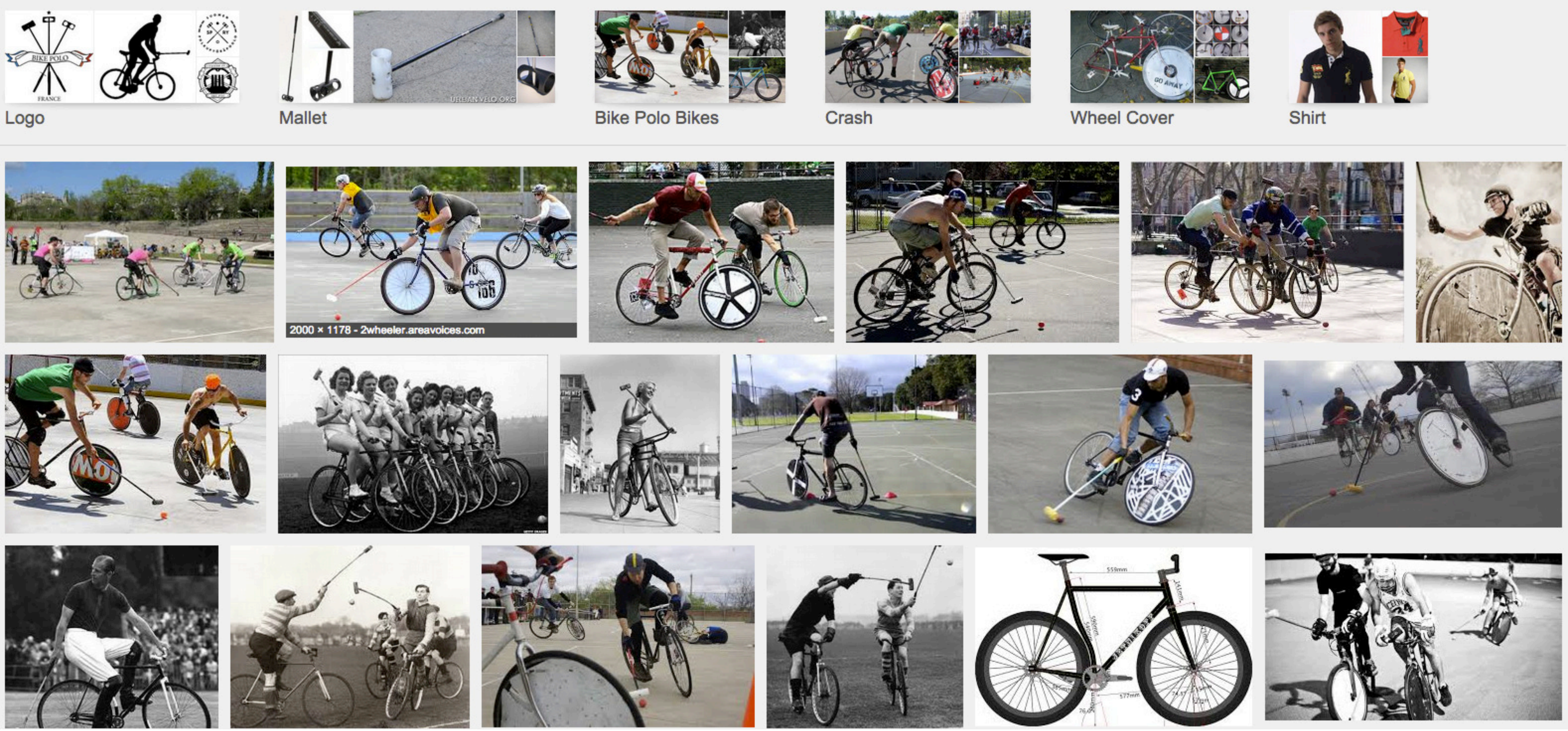
Important concepts repeat visually

- Statue of Liberty

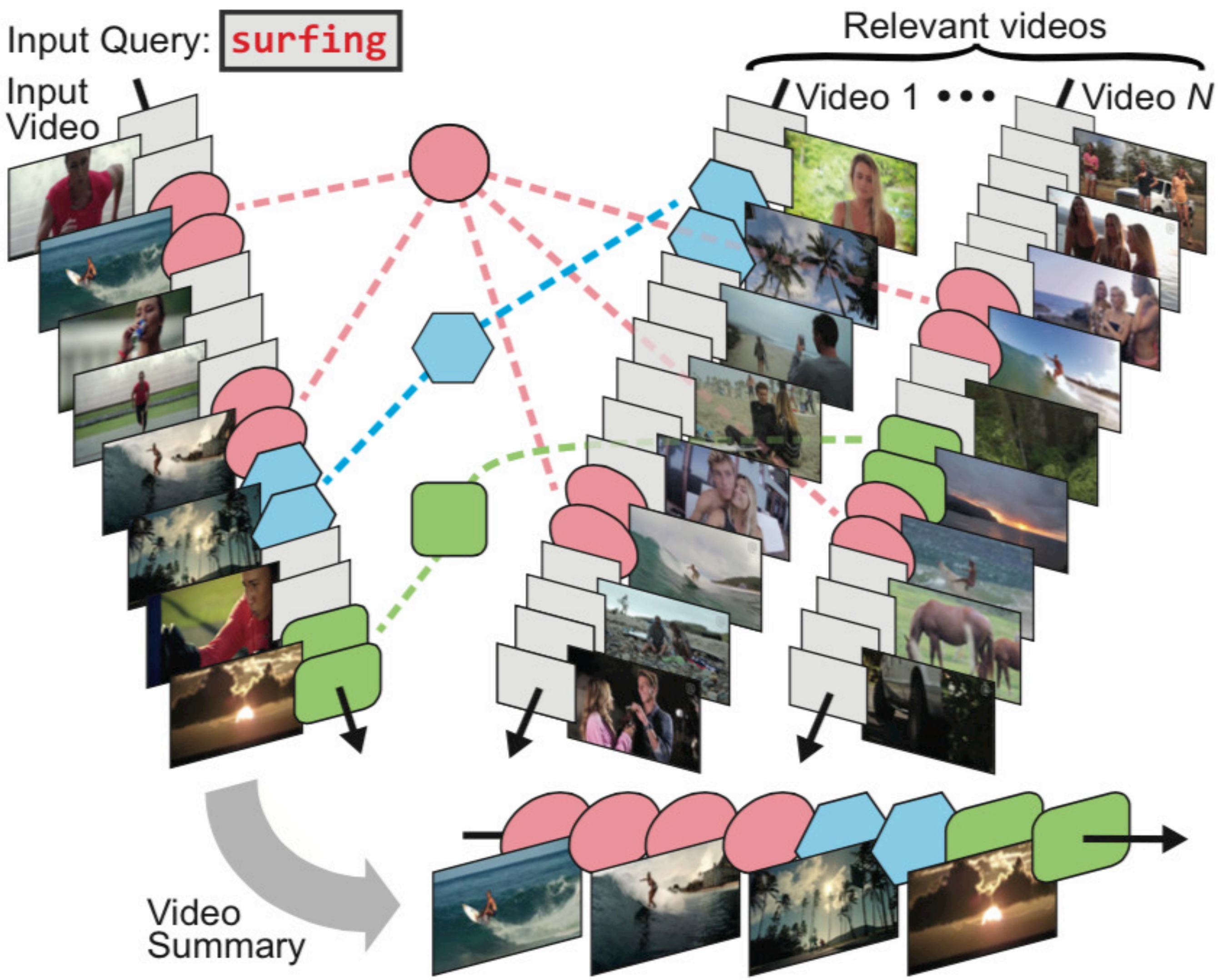


Important concepts repeat visually

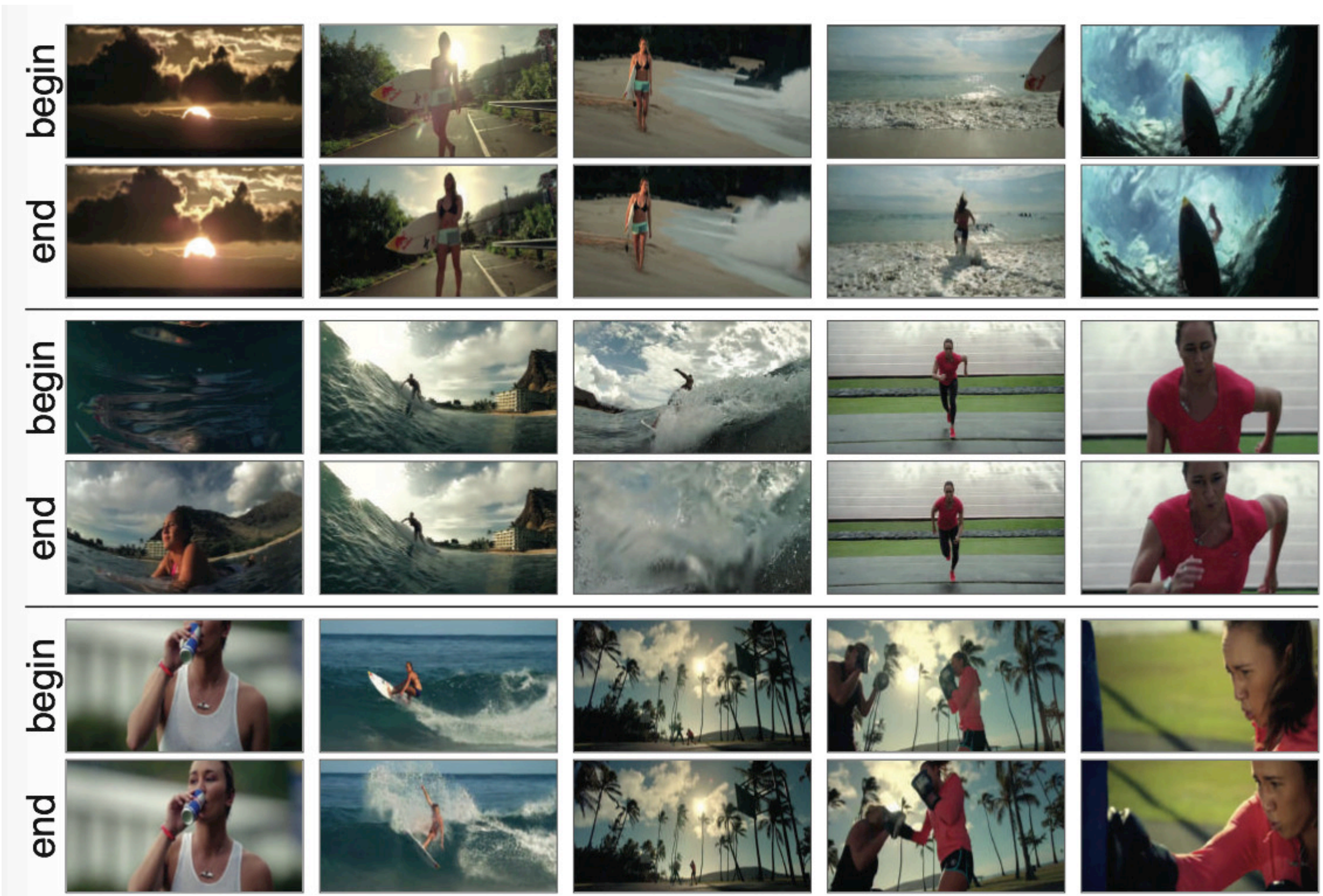
- Bike polo



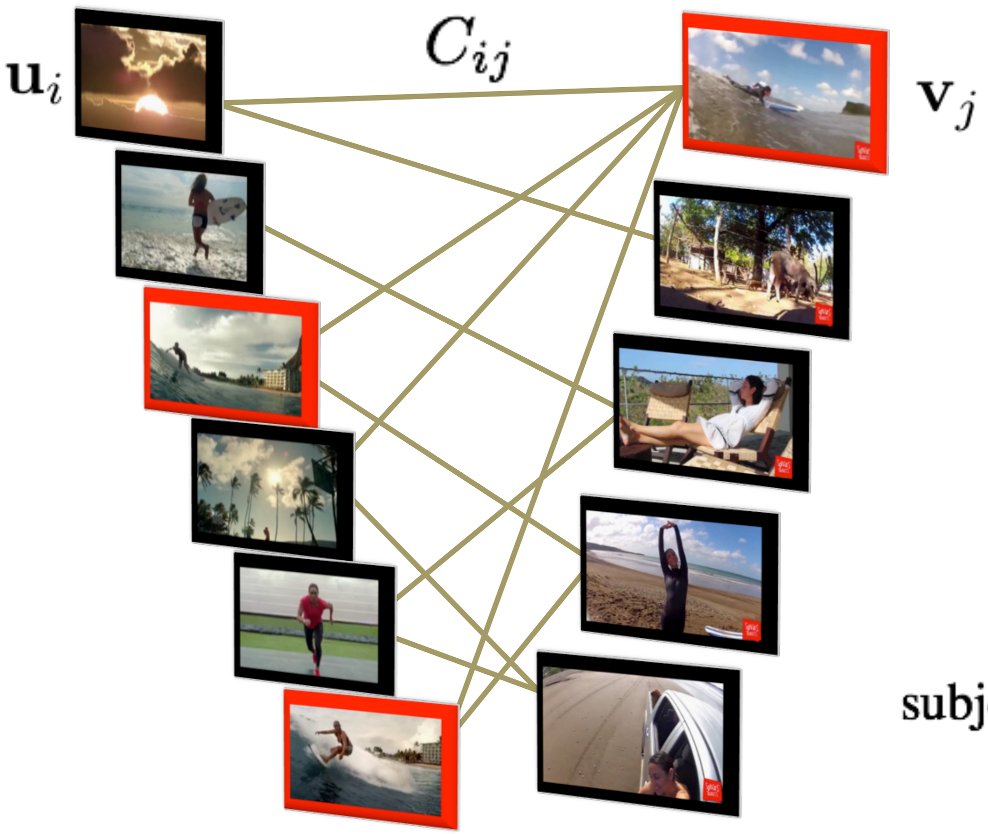
Video Co-Summarization



Video segmentation



Formulation



Discovering visual co-occurrence as "maximal bi-cliques"

$$\max_{\mathbf{u},\mathbf{v}} \quad \sum_{ij} C_{ij} u_i v_j - \lambda_u \|\mathbf{u}\|_1 - \lambda_v \|\mathbf{v}\|_1$$

$$\text{subject to} \quad u_i + v_j \leq 1 + I(C_{ij} \geq \epsilon), \forall i, j$$

$$\mathbf{u} \in [0,1]^m, \mathbf{v} \in [0,1]^n,$$

Algorithm

Input : Bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathbf{W}$ is described by the co-occurrence matrix $\mathbf{C}$; parameters $\lambda_u \geq 0$, $\lambda_v \geq 0$, and ϵ.

Output: Maximal biclique indicated by $\mathbf{u}$ and $\mathbf{v}$

Initialize $\mathbf{v} \leftarrow \texttt{rand}(n) \in [0, 1]^n$;

while *not converged* **do**

  Compute $\widehat{u}_i = \min\{I(\mathbf{C}_{ij} \geq \epsilon) - v_j\}_{j=1}^{n}$;

  Update $u_i = \min(I(\mathbf{C}_{i:}\mathbf{v} \geq \lambda_u), 1 + (\widehat{u}_i)_-)$;

  Compute $\widehat{v}_j = \min\{I(\mathbf{C}_{ij} \geq \epsilon) - u_i\}_{i=1}^{m}$;

  Update $v_j = \min(I(\mathbf{u}^\top \mathbf{C}_{:j} \geq \lambda_v), 1 + (\widehat{v}_j)_-)$;

Exp (1/3): Sanity check

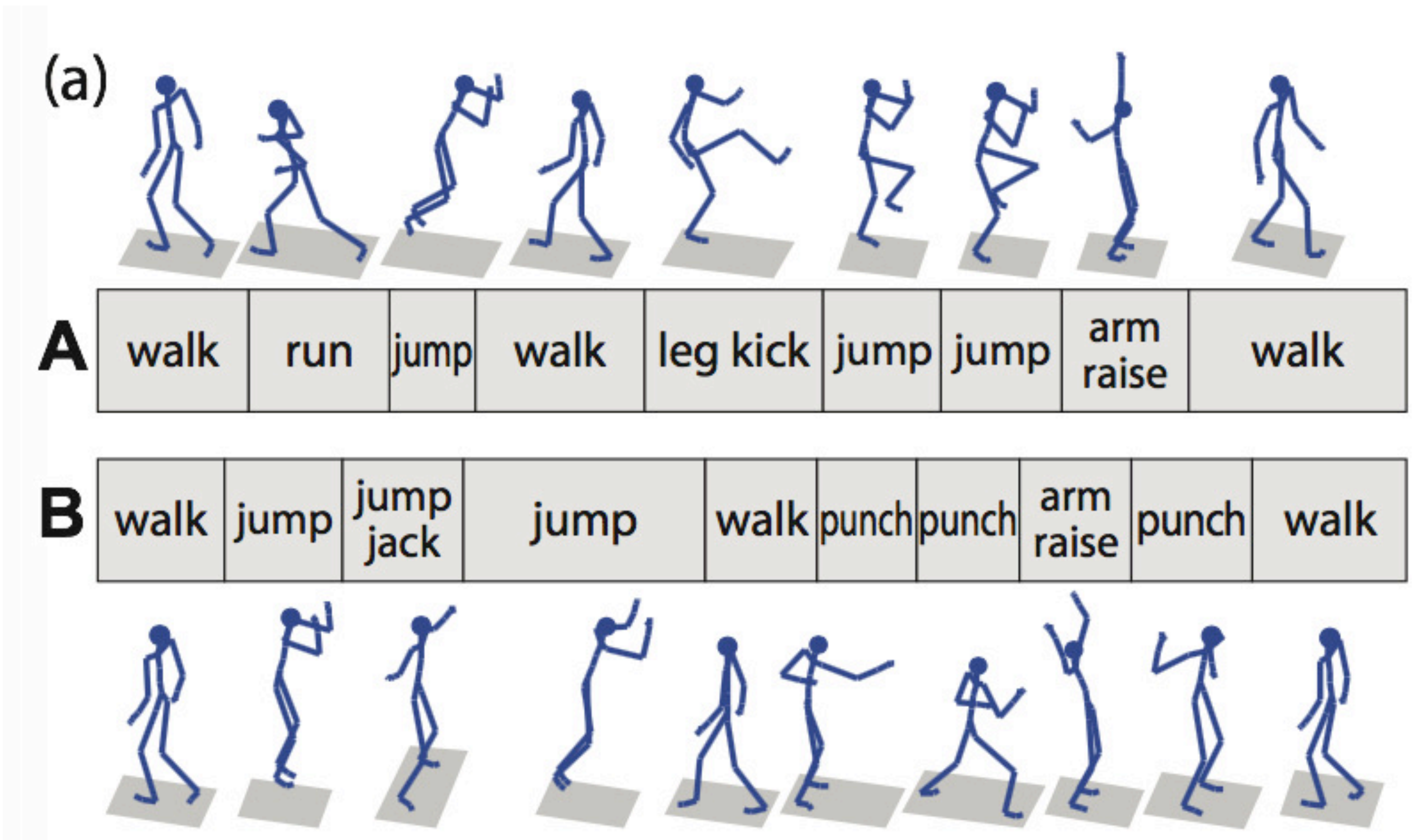
- CMU-Mocap dataset
  - We used the Subject 86 that contains 14 long sequences labeled with segment boundaries [3]
  - Thousands of frames / sequence
  - Up to 10 human actions / sequence (out of a total of 48 pre-defined actions)
- Representation
  - Each frame is represented a 30-D feature vector from 10 joints

Competitive methods

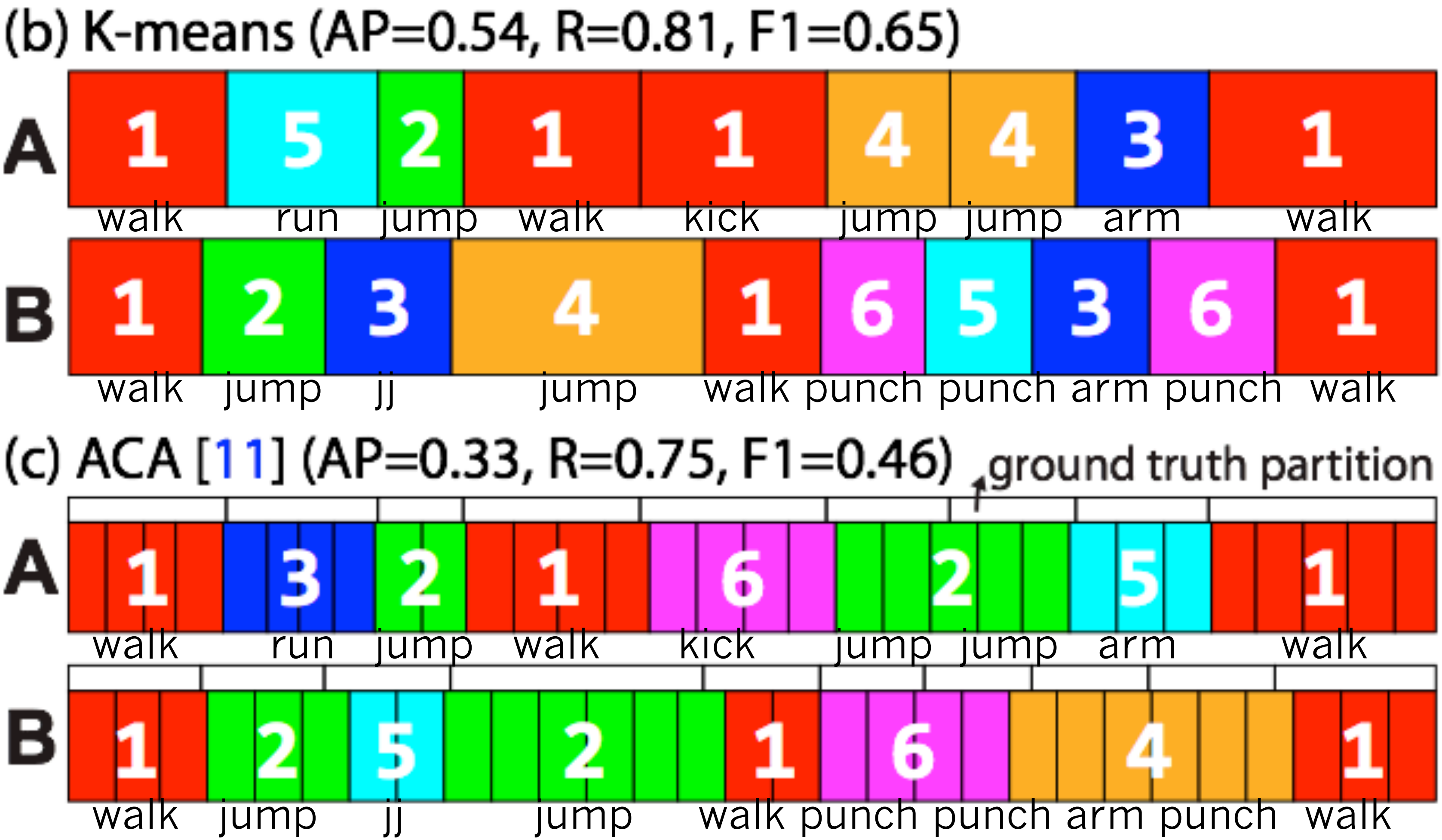
1. Baseline k-means
   – k=#groundtruth actions
2. Co-clustering (Dhillon [SIGKDD’01])
3. ACA (Zhou et al. [TPAMI’13])
4. MBF (our method)

On a sequence pair

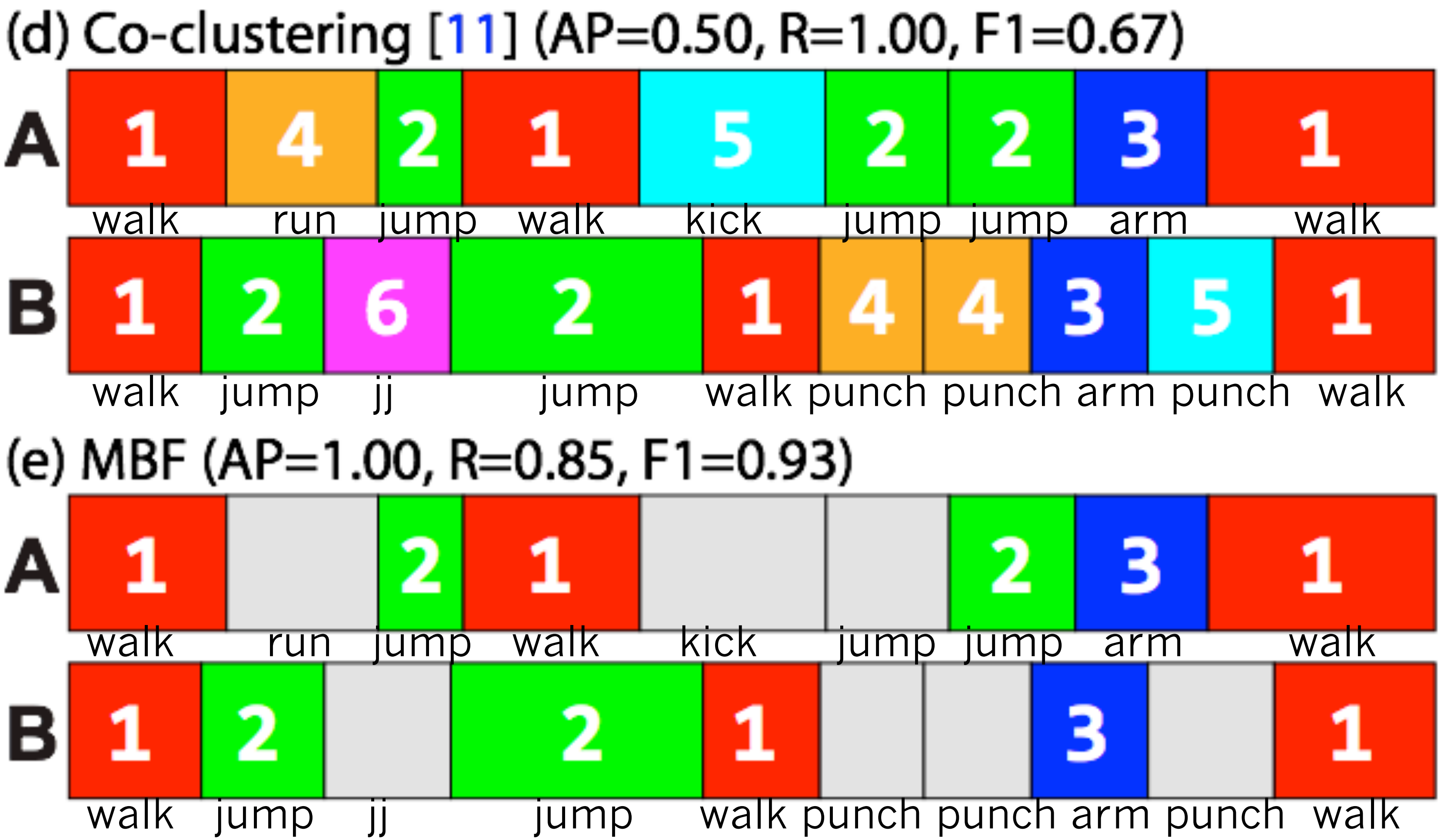
- Sequences 86_03 and 86_05



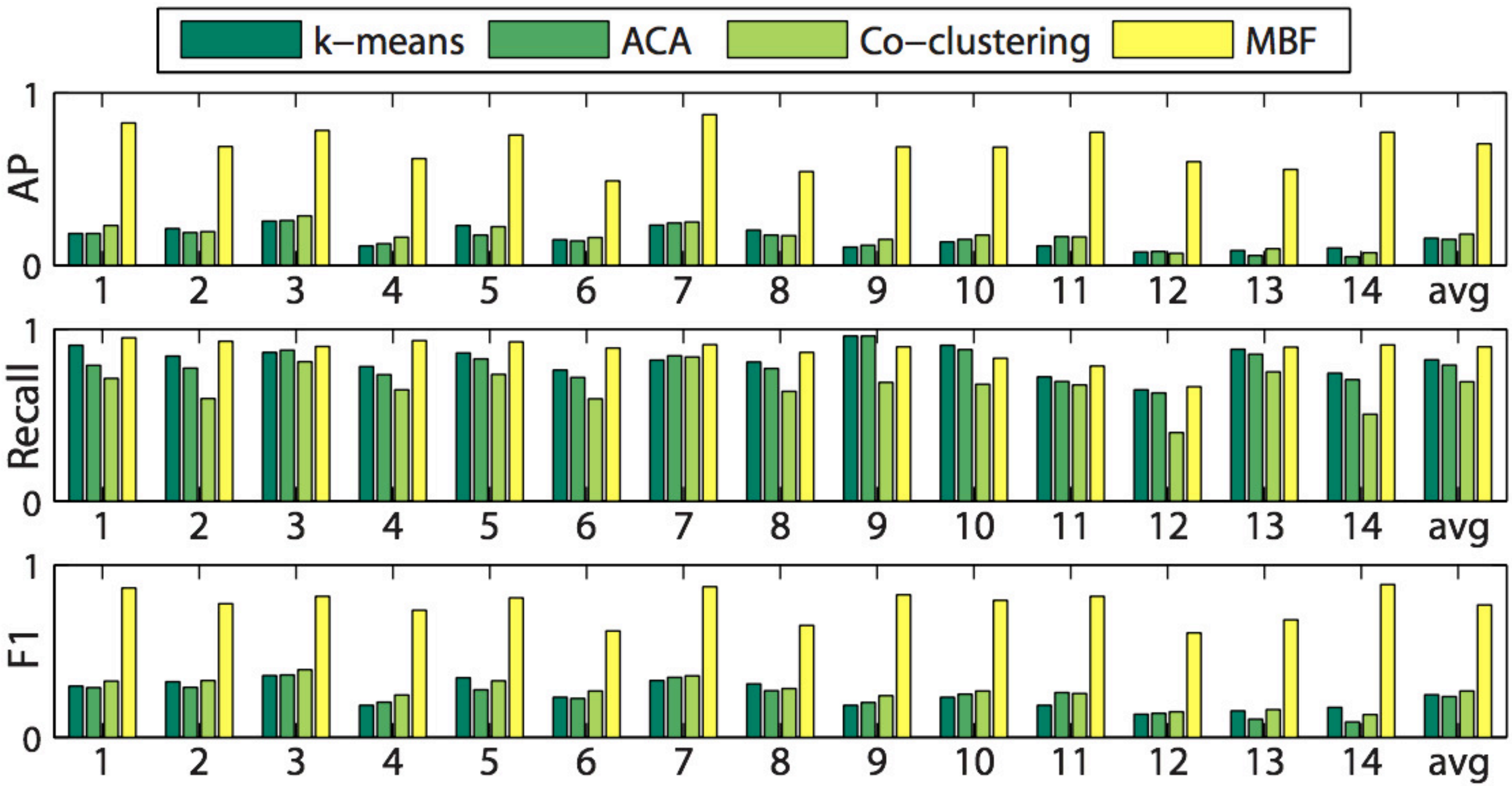
On a sequence pair



On a sequence pair

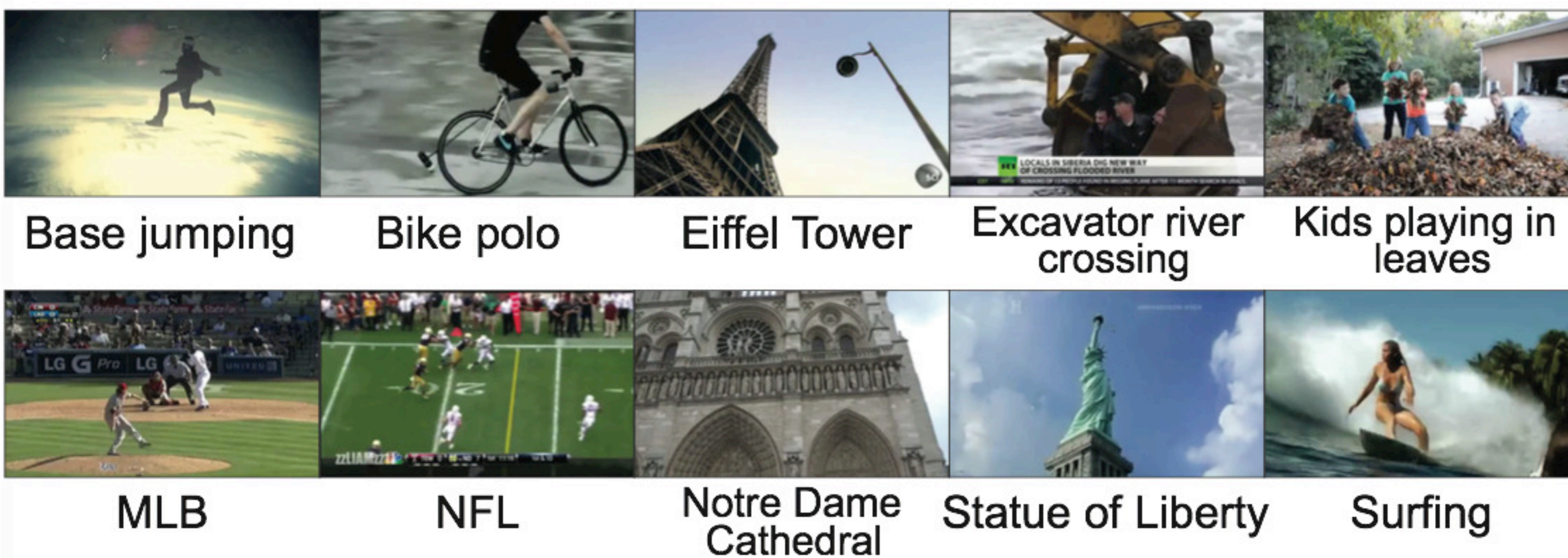


On all sequence pairs



Exp (2/3): Query-specific video summarization

- We compiled a dataset using 10 queries following SumMe [ECCV'14]
- 10 categories, 51 videos, 150 minutes.
- 246k frames, 2.8k segments.



Features

- CENTRIST (Wu and Rehg [TPAMI'11])
  - 254-D
- Dense-SIFT
  - Resize each frame to 620x420
  - 3840-D
- HSV color moments (Cong et al. [TMM'12])
  - 108-D
- Concatenated features and reduced to 400-D using PCA
- Use 200-entry BoTW to represent each segment

Competitive methods

- ACA [TPAMI'13] is not directly comparable
  - The assumption of repetitive temporal patterns barely occur in real-world videos
  - Building a kernel matrix for >15k frames is computationally prohibitive.

1. Baseline k-means (different values of k)
2. Co-clustering (Dhillon [SIGKDD'01])
3. LiveLight (Zhao and Xing [CVPR'14])
4. MBF (our method)

User study

- 3 judges label relevant segments in each video (#segments is >10% and <50%)
- Groundtruth is compiled by pooling those segments selected by >1 judges.
- Mean average precision (mAP) is computed for evaluation.

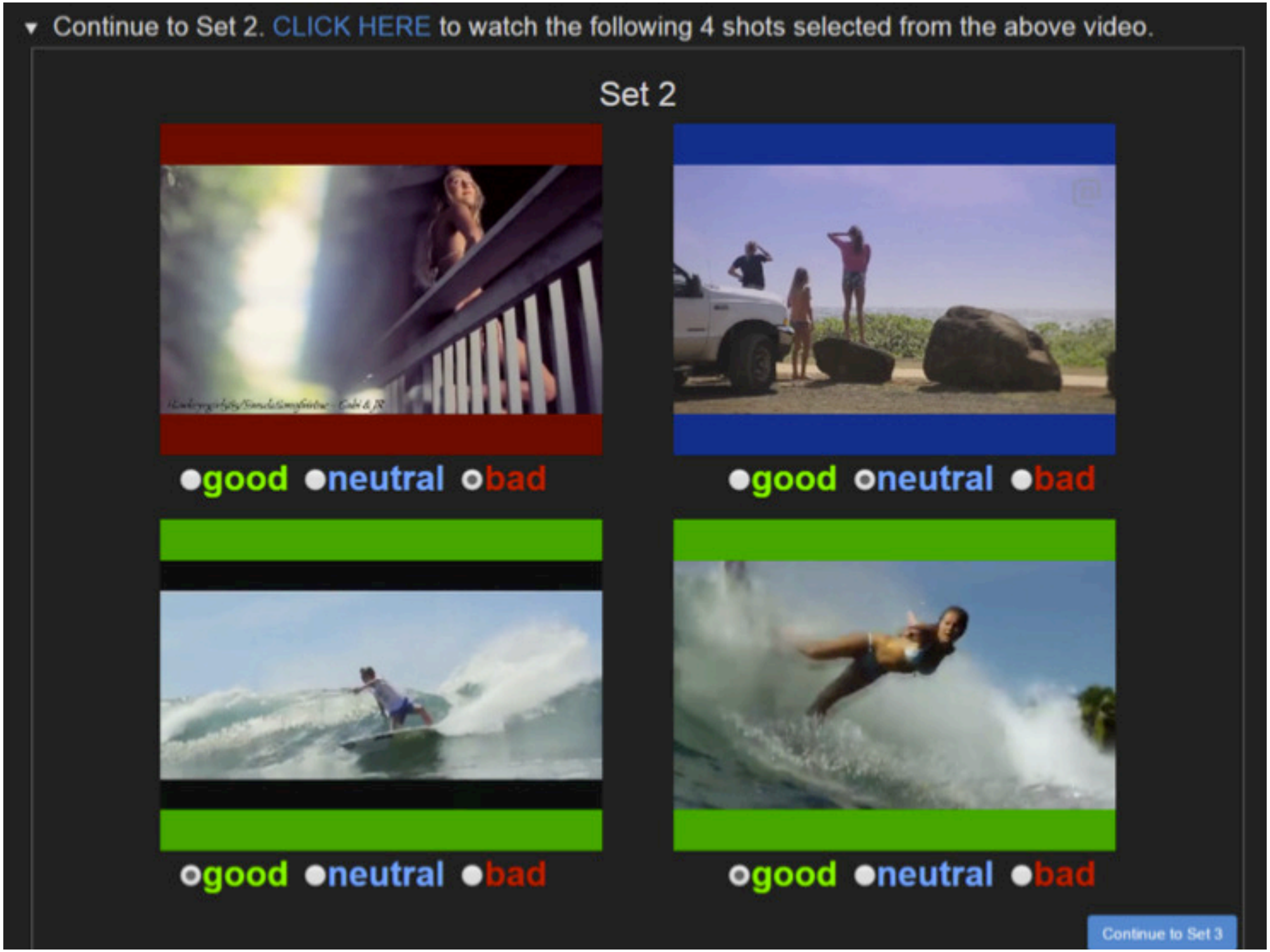
	Methods	Base*	Bike*	Eiffel*	Excavators*	Kids*	MLB	NFL	Notre Dame*	Statue*	Surfing	*Avg.*
$k=5$	k-means	0.432	0.427	0.422	0.289	0.791	0.556	0.663	0.392	0.543	0.550	0.507
	LL	0.226	0.305	0.413	0.667	0.744	0.508	0.710	0.568	0.763	0.334	0.524
	COC	0.495	**0.802**	0.580	**0.713**	**0.859**	0.561	**0.762**	**0.803**	0.378	0.668	0.662
	MBF	**0.680**	0.788	**0.596**	0.690	0.798	**0.638**	0.680	0.715	**0.810**	**0.684**	**0.707**
$k=15$	k-means	0.397	0.369	0.422	0.338	0.772	0.485	0.562	0.442	0.597	0.481	0.487
	LL	0.318	0.459	0.468	0.671	0.710	0.499	0.737	0.592	0.653	0.337	0.545
	COC	0.496	**0.795**	**0.561**	0.656	0.852	0.503	**0.823**	0.676	0.458	0.586	0.641
	MBF	**0.747**	0.663	**0.562**	**0.674**	**0.859**	**0.755**	0.760	**0.680**	**0.661**	**0.652**	**0.701**

Exp (3/3): Concept visualization

- Can a robot watch Youtube to learn about human's concepts?
- A natural extension of co-sum: visualize a concept as the most frequently co-occurring video clips

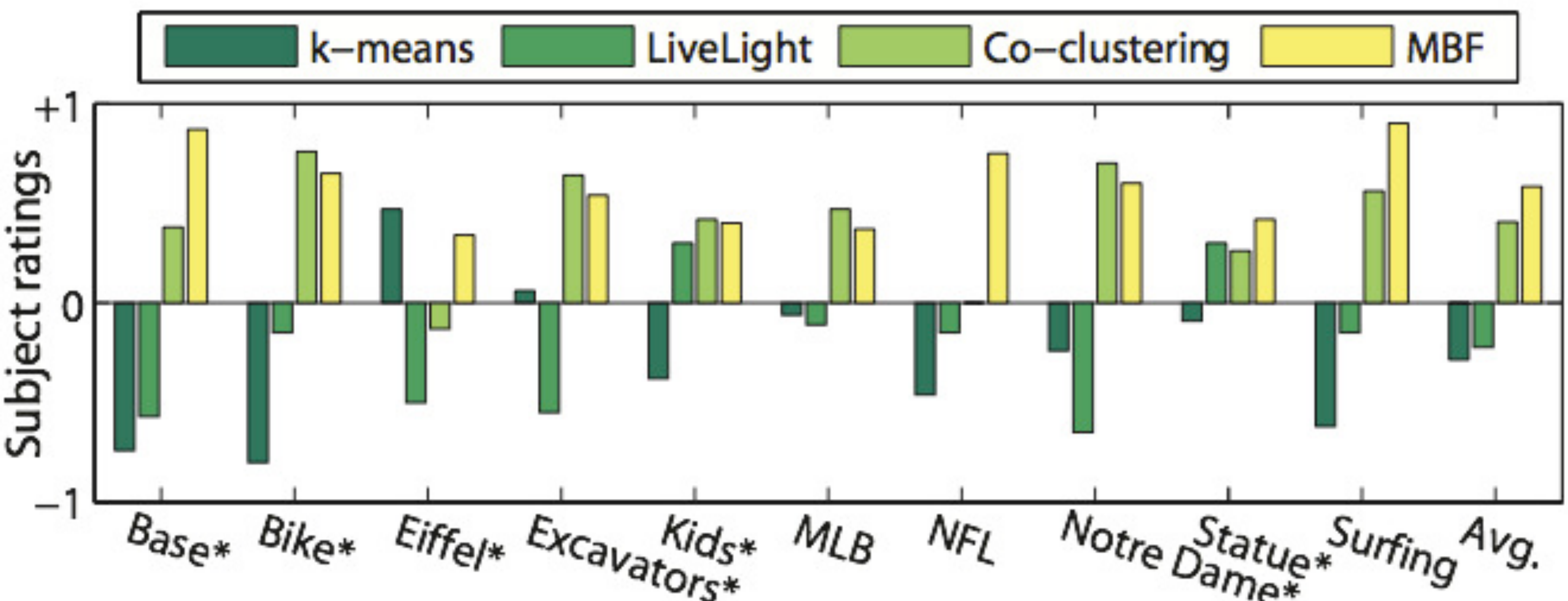
Surfing example

AMT-like user study

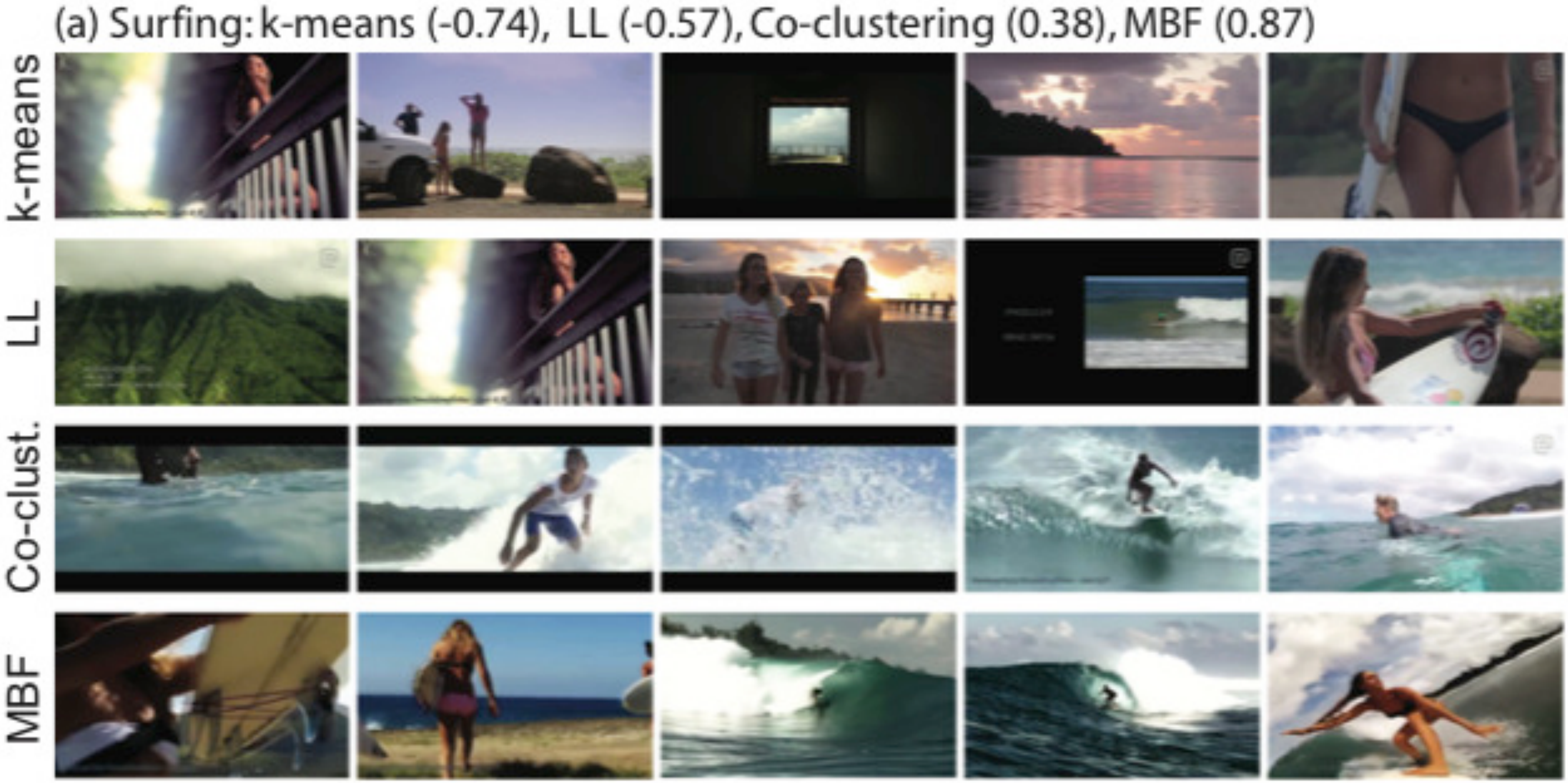


Subject ratings

- 20 subjects, ages ranging from 23-33
- 15 males, 5 females



Most winning case



Most losing case

Summary

- We propose video co-summarization that assumes important concepts are likely to visually repeat.
- We propose a maximal biclique finding algorithm that can be parallelized with closed-form updates
- Experiments suggest visually co-occurring clips are close to human summaries.

A desired method

☑ Generates adaptive summaries that fits user's interests

☑ Scales to large datasets

☑ Requires limited/no human supervision

Thank you!


Base jumping
MLB
Bike polo
NFL
Excavator river crossing
Notre Dame
Kids playing in leaves
Statue of Liberty