Beyond Supervised Learning:
Incidental Supervision for Video Summarization

Amit K. Roy-Chowdhury, UC Riverside

Why Video Summarization?

[The Verge, “We are all Glassholes now”]

Why Video Summarization?

UCR VideoWeb camera network- 37 AXIS-215 PTZ cameras

With a frame rate of 0.15 Mbps (10% of Netflix Standard) --- produces 1TB of data on an average 3 weeks of operation

Over an year, 14 disks are required -- One 1TB disk costs about $60 today and storing this information requires $840 per year for a small network of 37 cameras

It may be possible to store all the data for a small network of 37 cameras

How many CCTV Cameras are there globally?

According to IHS, there were **245 million** professionally installed video surveillance cameras active and operational globally in 2014.

Video Summarization

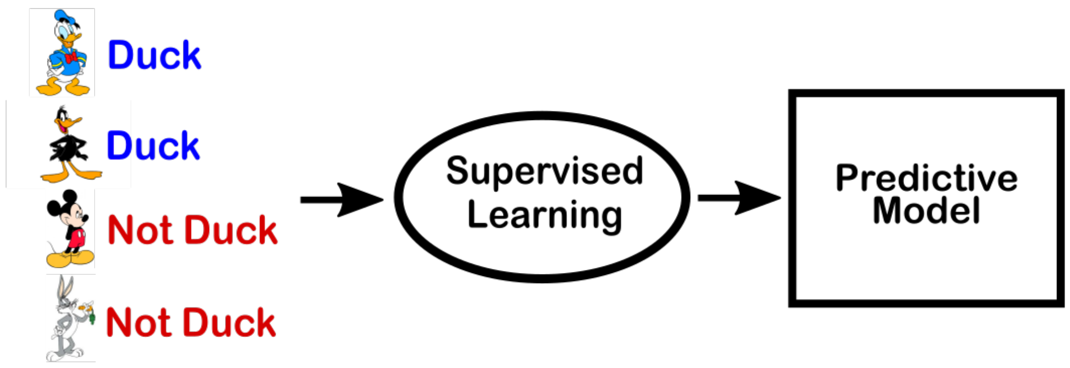
- Video capture is omnipresent and vast
- Users have a "capture first, filter later" mentality
- Large amounts of video – need to search for relevant content quickly

Definition of "summarize"

"Give a brief overview of the main parts of the video(s)"

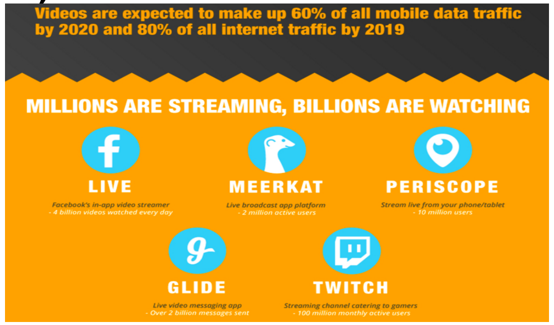
Supervised Learning

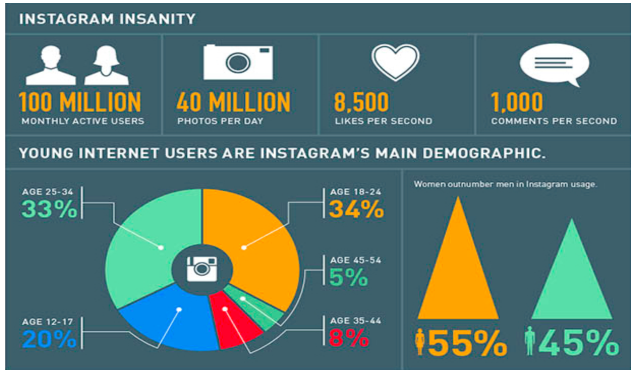
- Learning (training): Learn a model using labeled training data
- Testing: Test the model using unseen test data to assess the model performance



Beyond Supervised Learning

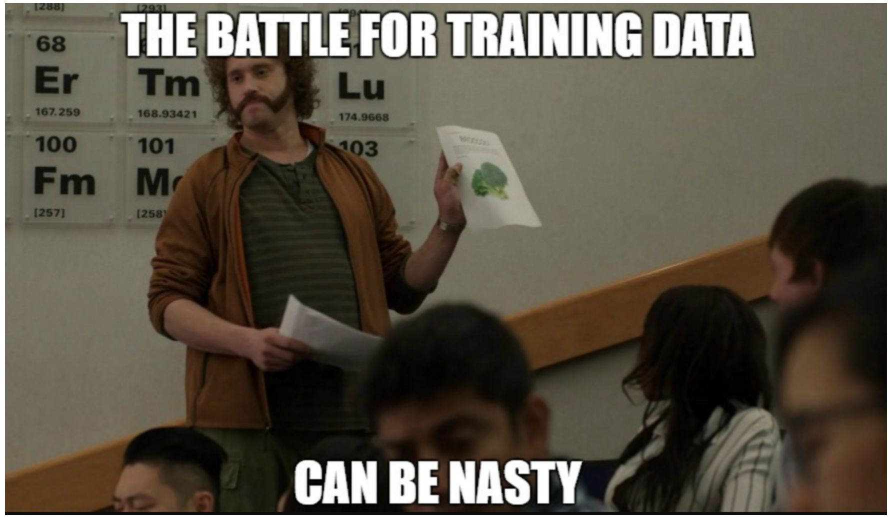
- Most of the existing methods follow supervised approaches where all data are labeled
- Unrealistic assumption: all data will be labeled and available beforehand to train a model
    - Infeasible: Labeling is expensive and time consuming
    - **5000 users** working **24 hours** will take **1 month** to label Google Image database (425,000,000)





Labeling (Big) Data is Infeasible

Incidental Supervision



Annotating data for complex tasks is difficult, costly, and sometimes impossible – summarization and Re-ID

Can we move beyond current annotation heavy approaches?

Learning should be driven by incidental signals

Incidental Signals refer to a collection of weak signals that exist in the data and the environment, independently of the tasks at hand. [Dan Roth, AAAI'17]

Incidental Supervision for Video Summarization

- Single-Video Summarization
    - Collaborative Video Summarization
    - Weakly Supervised Video Summarization
- Multi-Video Summarization
    - Diversity-aware Video Summarization
    - Multi-View Video Summarization in a Camera Network

Preliminaries - I

A way to find a dictionary/set of basis functions (Dictionary Learning) such that a signal (or a set of of signals) has a sparse representation (Sparse Coding) over the set of basis functions

$$\underset{D,C}{\text{minimize}} \quad ||Y - DC||_F^2$$
$$\text{subject to} \quad ||C_i||_0 \leq s, \; i = 1, \ldots, N.$$

$Y - R^{m \times N}$
$D - R^{m \times l}$
$C - R^{l \times N}$

L1 relaxation: Simultaneously learn D and C in an alternative fashion

"Reconstruction error" and "Sparsity" term naturally fits into the problem of summarization

Summaries are from the data itself, it can't be from outside (Self expressiveness property)

Goal is to find how many data points are in fact required to represent the whole data

Preliminaries - II

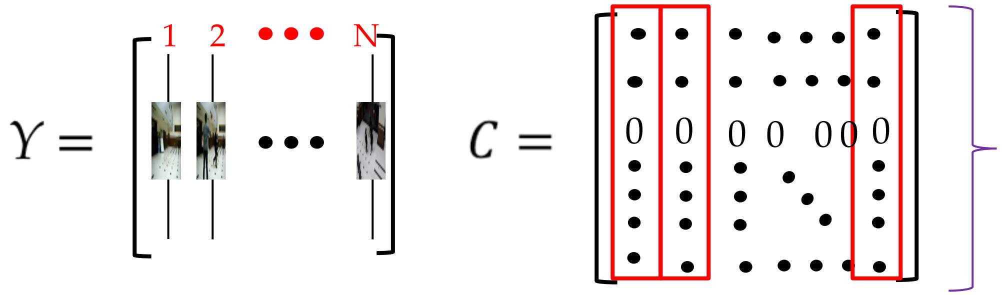
$$\underset{C}{\text{minimize}} \quad ||Y - YC||_F^2 + \lambda ||C||_{2,0}$$

$Y - R^{d \times N}$

$C - R^{N \times N}$

$||C||_{2,0}$ counts the number of nonzero rows of C

*NP-hard – Changed to $l_{2,1}$ norm (Sum of l_2 norm of rows)



Frame 3 does not take part in reconstruction of any frames in the video

Solution: indices of the nonzero rows of C correspond to the indices of the columns of Y are chosen as representative summaries

Assumptions

- Videos are given beforehand – no streaming/online setting has been considered (although it can be handled with little changes to the proposed solutions)
- Basic processing unit for summarization is a video shot (detected using any standard method)
- Each video shot is represented by a feature vector (C3D feature)
- User preferences are not considered – personalization; can be added with small changes

Incidental Supervision for Video Summarization

- **Single-Video Summarization**
  - **Collaborative Video Summarization**
  - Weakly Supervised Video Summarization
- Multi-Video Summarization
  - Diversity-aware Video Summarization
  - Multi-View Video Summarization in a Camera Network

Collaborative Video Summarization

Are these videos independent of each other or something common exists across them?

They all belongs to the same topic "Eiffel Tower"

Summaries of these videos will have significant common information

Incidental Supervision

Rameswar Panda, Amit K. Roy-Chowdhury, "Collaborative Summarization of Topic-Related Videos", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

Problem Statement

Goal: Finding a sparse set of **representative and diverse** shots that simultaneously capture both important particularities arising in the **given video**, as well as, generalities identified from the set of **topic-related videos**

Basic Idea: Exploit visual context from topic-related videos to identify important parts of a video

Builds upon the idea of collaborative techniques from IR and NLP

Use attributes of similar objects to predict attribute of a given object

Problem Formulation

select an unified set of video shots that simultaneously cover the important particularities arising in the target video, as well as the generalities arising in the video collection

Collaborative Sparse Representative Selection

Representative: summary should reconstruct the topic-related videos
Sparsity: summary length should be as small as possible
Diversity: summary should be collectively diverse

$$\min_{\mathbf{Z},\,\tilde{\mathbf{Z}}} \underbrace{\frac{1}{2}\left(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\right)}_{Representative}+\underbrace{\lambda_s\left(\|\mathbf{Z}\|_{2,1}+\|\tilde{\mathbf{Z}}\|_{2,1}\right)}_{Sparsity}$$

$$+\underbrace{\lambda_d\left(tr(\mathbf{D}^T\mathbf{Z})+tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\right)}_{Diversity}+\underbrace{\beta||\mathbf{Z}_c||_{2,1}}_{Consensus} \quad s.t.\;\; \mathbf{Z}_c=[\mathbf{Z}|\tilde{\mathbf{Z}}],\;\; \mathbf{Z}_c\in\mathbb{R}^{n\times(n+\tilde{n})}$$

Diversity Regularization Functions

$$f_d(\mathbf{Z})=\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}Z_{ij}=tr(\mathbf{D}^T\mathbf{Z}),$$

$$f_d(\tilde{\mathbf{Z}})=\sum_{i=1}^{n}\sum_{j=1}^{\tilde{n}}\tilde{d}_{ij}\tilde{Z}_{ij}=tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})$$

J. Yao, et. al. AAAI, 2015

Half-Quadratic Optimization

Original objective function

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{X}-\mathbf{XZ}\|_F^2 + \lambda_s\|\mathbf{Z}\|_{2,1} \quad \text{--------} \quad (1)$$

Augmented objective function

$$\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{X}-\mathbf{XZ}\|_F^2 + \lambda_s tr(\mathbf{Z}^T\mathbf{PZ})$$

$$\mathbf{P}_{i,i} = \frac{1}{2\sqrt{||\mathbf{Z}_i||_2^2+\epsilon}} \quad \text{--------} \quad (2)$$

$$(\mathbf{X}^T\mathbf{X} + 2\lambda_s\mathbf{P})\mathbf{Z} = \mathbf{X}^T\mathbf{X} \quad \text{--------} \quad (3)$$

Solve alternatively

Algorithm 2 Algorithm for Solving Eq. (1)

Input: Feature matrix $\mathbf{X}$, Parameters λ_s, set $t=0$;
Initialize $\mathbf{Z}$ randomly;
Output: Optimal sparse coefficient matrix $\mathbf{Z}$.
while *not converged* **do**
1. Compute $\mathbf{P}^t$ using Eq. (2);
2. Compute $\mathbf{Z}^{t+1}$ using Eq. (3);
4. $t=t+1$;
end while

[*] R. He. Half-quadratic based iterative minimization for robust sparse representation. In TPAMI, 2014

Optimization

Overall problem is non-smooth involving multiple-norms
Half-quadratic optimization [4] is effective in solving these sparse optimization problems

$$\min_{\mathbf{Z},\, \tilde{\mathbf{Z}}} \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big)+\lambda_s\big(tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z})+tr(\tilde{\mathbf{Z}}^T\mathbf{Q}\tilde{\mathbf{Z}})\big)$$
$$+\lambda_d\big(tr(\mathbf{D}^T\mathbf{Z})+tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big)+\beta\big(tr(\mathbf{Z}_c^T\mathbf{R}\mathbf{Z}_c)\big)$$

Augmented cost function according to half-quadratic theory

$$\mathbf{P}_{ii}=\frac{1}{2\sqrt{||\mathbf{Z}_i||_2^2+\epsilon}},\quad \mathbf{Q}_{ii}=\frac{1}{2\sqrt{||\tilde{\mathbf{Z}}_i||_2^2+\epsilon}},\quad \mathbf{R}_{ii}=\frac{1}{2\sqrt{||\mathbf{Z}_{ci}||_2^2+\epsilon}}$$

[*] R. He. Half-quadratic based iterative minimization for robust sparse representation. In TPAMI, 2014.

Algorithm

Algorithm 1 Collaborative Sparse Representative Selection

Input: Video feature matrices $\mathbf{X}$ and $\tilde{\mathbf{X}}$;
Parameters $\alpha, \lambda_s, \lambda_d, \beta$, set $t = 0$;
Construct $\mathbf{D}$ and $\hat{\mathbf{D}}$ using inner product similarity;
Initialize $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ randomly, set $\mathbf{Z}_c$= $[\mathbf{Z}, \tilde{\mathbf{Z}}]$;
Output: Optimal sparse coefficient matrix **Zc**.
while *not converged* **do**
1. Compute $\mathbf{P}^t$, $\mathbf{Q}^t$ and $\mathbf{R}^t$;
2. Compute $\mathbf{Z}^{t+1}$ and $\tilde{\mathbf{Z}}^{t+1}$;
3. Compute $\mathbf{Z}_c^{t+1}$ as: $\mathbf{Z}_c^{t+1}$= $[\mathbf{Z}^{t+1} \mid \tilde{\mathbf{Z}}^{t+1}]$;
4. $t = t + 1$;

end while

Results

Eiffel Tower

Attempting Bike Tricks

Role of topic-related visual context in summarizing videos.
Top: CVS w/o topic-related visual context, Bottom: CVS w/ topic-related visual context

Incidental Supervision for Video Summarization

- **Single-Video Summarization**

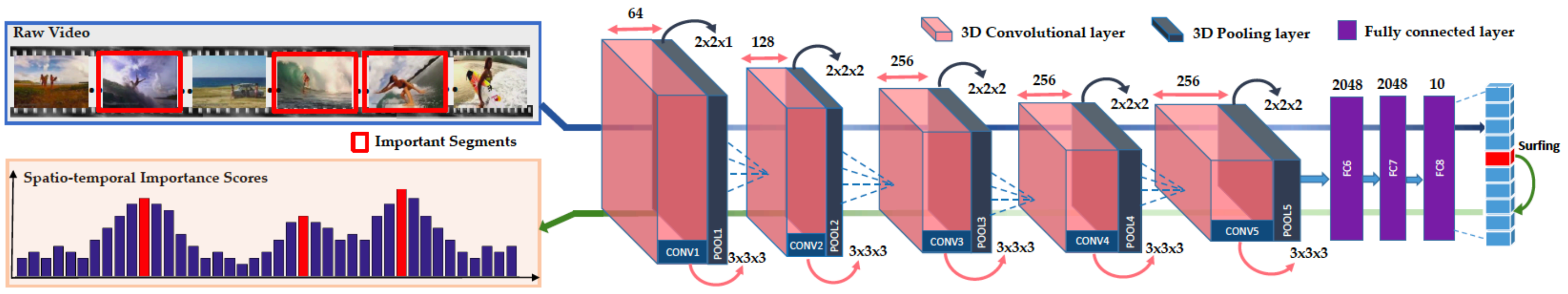
  - Collaborative Video Summarization
  - **Weakly Supervised Video Summarization**

- Multi-Video Summarization

  - Diversity-aware Video Summarization
  - Multi-View Video Summarization in a Camera Network

Weakly-Supervised Video Summarization

- Incidental Supervision: video level annotations (easy to obtain)



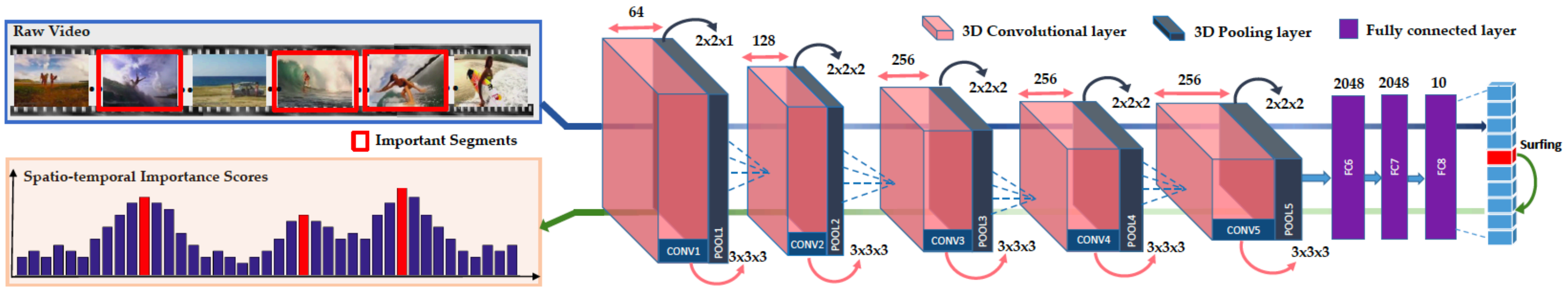
Deep Summarization Network (DeSumNet)

- Training: Given a set of videos, learn what aspects are important within a category (e.g., surfing)

- Testing: Compute importance score via back-propagation guided by category with highest score

Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, Amit K. Roy-Chowdhury, “Weakly Supervised Summarization of Web Videos", IEEE International Conference on Computer Vision (ICCV), 2017

Gradient-based Importance Computation



Deep Summarization Network (DeSumNet)

- Leverage multiple videos belonging to a specific category to automatically learn a parametric model for categorizing videos
- Adopt the learned model to find important segments from a given video as the ones which have the maximum influence to the model

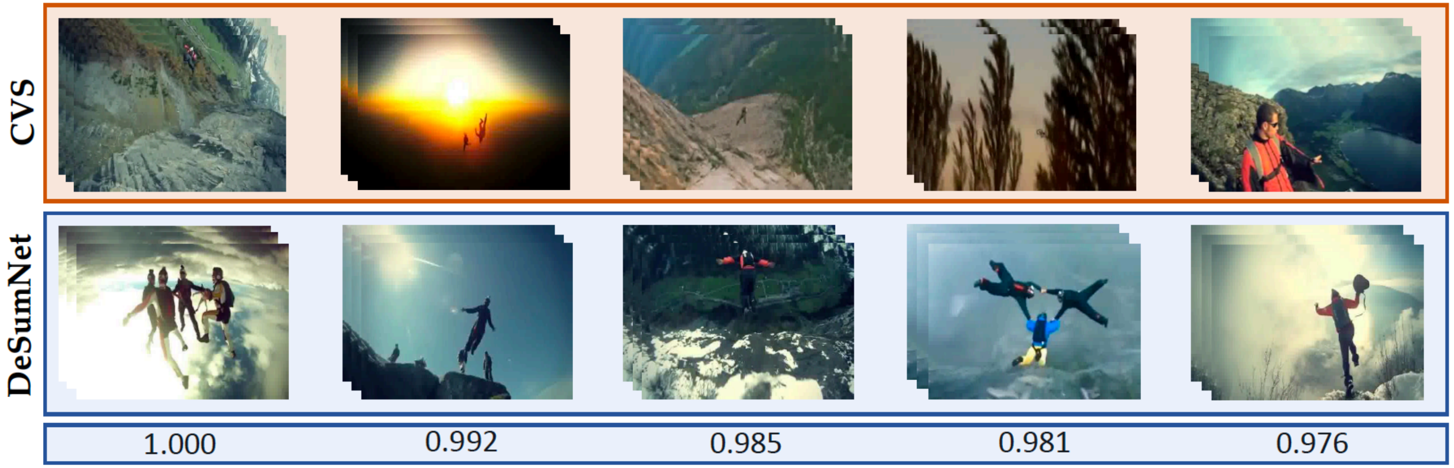
Input Video $$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$$

Spatio-temporal importance map $$\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h}) = \frac{\partial}{\partial \mathbf{x}} \langle \mathbf{h}, \phi(\mathbf{x}) \rangle \Big|_{\mathbf{x}=\mathbf{x}_i}$$

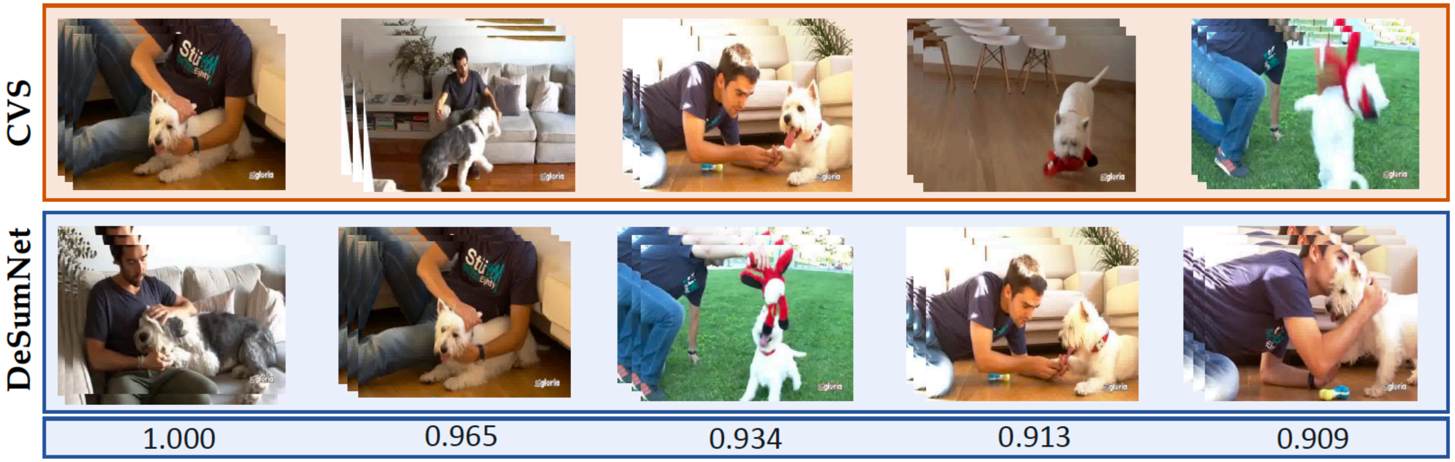
$$\mathtt{vec}[\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h})] = \mathbf{h}^\top \times \frac{\partial\ \mathtt{vec}[\phi_l]}{\partial\ \mathtt{vec}[\mathbf{x}_l]^\top} \times \cdots \times \frac{\partial\ \mathtt{vec}[\phi_1]}{\partial\ \mathtt{vec}[\mathbf{x}_i]^\top}$$

Chain rule (vector notation)

Example Summaries



Base Jumping



Grooming An Animal

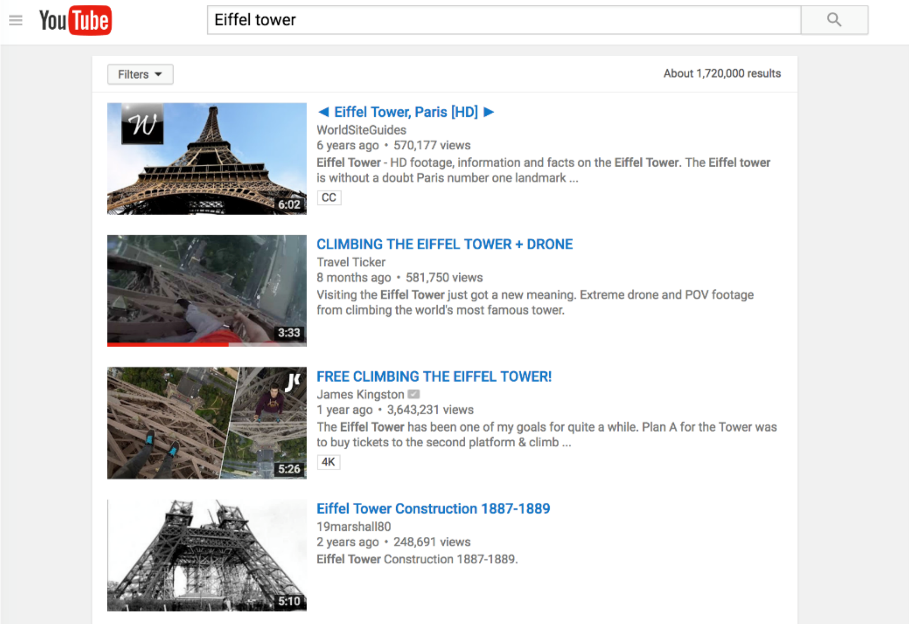
Incidental Supervision for Video Summarization

- Single-Video Summarization
    - Collaborative Video Summarization
    - Weakly Supervised Video Summarization
- **Multi-Video Summarization**
    - **Diversity-aware Video Summarization**
    - Multi-View Video Summarization in a Camera Network

Diversity-aware Multi-Video Summarization

Questions Asked: Can we generate a single summary from all the videos without any manual supervision?

Incidental Supervision: each video in the set may contain some information that other videos do not have



Can we get an idea of the video content without watching all the videos entirely?

Rameswar Panda, Niluthpol C. Mithun, Amit K. Roy-Chowdhury, “Diversity-aware Multi-Video Summarization", IEEE Transactions on Image Processing (TIP), vol. 26, no. 10, pp. 4712-4724, Oct. 2017.

Problem Statement

- Input: a set of m relevant web videos given a video search

$$X^v = \{X^v_{\cdot,i} \in \mathbb{R}^d, i = 1, \cdots, n_v\}, v = 1, \cdots, m$$

$X^v_{\cdot,i}$: Feature descriptor of a video shot in d-dimensional space

C3D features computed using a 3D CNN architecture

- Output: find a summary that conveys the most important details of the video collection

Problem Formulation

$$\min_{Z^v} \|X^v - X^v Z^v\|_F^2 + \lambda_s^v \|Z^v\|_{2,1} \;\; s.t. \;\; {Z^v}^T 1 = 1$$

Sparse Optimization for summarizing a single video

All shots are treated equally in selecting representatives

Introducing Prior Knowledge via Weighted norm:

$$\min_{Z^v} \|X^v - X^v Z^v\|_F^2 + \lambda_s^v \|Q^v Z^v\|_{2,1} \;\; s.t. \;\; {Z^v}^T 1 = 1$$

$Q^v = [diag(q^v)]^{-1}$ $q^v \in \mathbb{R}^{n_v}$: interestingness score of each video shot

favors selection of interesting shots

Problem Formulation

Introducing Diversity of Multiple Videos – Incidental Supervision

$$\min_{Z^1,Z^2,\cdots,Z^m} \sum_{v=1}^{m} \|X^v - X^v Z^v\|_F^2 + \lambda_s \sum_{v=1}^{m} \|Q^v Z^v\|_{2,1} + \lambda_d \sum_{\substack{1\le v,w\le m\\ v\ne w}} f_d(Z^v, Z^w)$$

$$s.t.\ \ {Z^v}^T 1 = 1,\ Z^v \in \mathbb{R}^{n_v \times n_v},\ \forall\ 1 \le v \le m$$

favors selection of interesting and diverse shots

$$f_d(Z^v, Z^w) = \sum_{i=1}^{n_v} \sum_{j=1}^{n_w} ||Z^v_{i,.}||_2 C_{i,j} ||Z^w_{j,.}||_2 = ||W^{vw} Z^v||_{2,1}$$

$C_{i,j}$: measure the correlation between i-th shot from v-th video and the j-th shot in w-th video

$$W^{vw}_{i,i} = \sum_{j=1}^{n_w} C_{i,j} \|Z^w_{j,.}\|_{2,1}$$

Optimization

- Alternating minimization: minimizing the function with respect to one video at a time while fixing the other videos

$$\min_{Z^v} \|X^v - X^v Z^v\|_F^2 + \lambda_s \|Q^v Z^v\|_{2,1} + \lambda_d \sum_{w=1, v \neq w}^{m} \|W^{vw} Z^v\|_{2,1} \quad s.t. \quad {Z^v}^T 1 = 1$$

- Convex weighted norm minimization problem – Optimization via Alternating Direction Method of Multipliers (ADMM)

- Alternate over multiple videos until convergence – in practice, convergence less than 10 iterations

Experiments

Dataset Statistics:

- No publicly available dataset for evaluation
- Selected 20 tourist attractions from the Tripadvisor travelers choice landmarks 2015 list
- Collected 140 videos from YouTube under the CC-BY 3.0 license

Performance Measures :

Precision: Ratio of correctly detected shots to the number of shots in system-generated summary
Recall: Ratio of correctly detected shots to the number of detected shots in ground truth summary
F1-measure: Harmonic mean of Precision and Recall

Tour20 Dataset Statistics

Tourist Attractions	# Videos	Length	# Frames	# Shots
Angkor Wat, Cambodia	7	26m57s	44,410	803
Machu Picchu, Peru	7	26m15s	43,125	914
Taj Mahal, India	7	22m21s	36,554	705
Basilica of the Sagrada Familia, Spain	6	23m30s	22,641	400
St. Peter's Basilica, Italy	5	14m39s	23,777	406
Milan Cathedral, Italy	10	24m18s	37,749	768
Alcatraz, United States	6	05m22s	09,733	223
Golden Gate Bridge, United States	6	19m21s	33,063	521
Eiffel Tower, Paris	8	106m10s	26,071	495
Notre Dame Cathedral, Paris	8	26m49s	44,583	862
The Alhambra, Spain	6	21m20s	38,087	779
Hagia Sophia Museum, Turkey	6	24m27s	38,608	853
Charles Bridge, Prague	6	27m33s	48,395	769
Great Wall at Mutiantu, Beijing	5	13m16s	22,117	477
Burj Khalifa, Dubai	9	23m21s	40,557	809
Wat Pho, Bangkok	5	11m48s	20,461	382
Chichen Itza, Mexico	8	16m51s	28,737	545
Sydney Opera House, Sydney	10	25m55s	49,735	695
Petronas Twin Towers, Malaysia	9	18m32s	30,009	470
Panama Canal, Panama	6	17m33s	31,625	623
Total	**140**	**6h46m18s**	**669,497**	**12,499**

[Publicly available at: http://vcg.engr.ucr.edu/datasets](http://vcg.engr.ucr.edu/datasets)

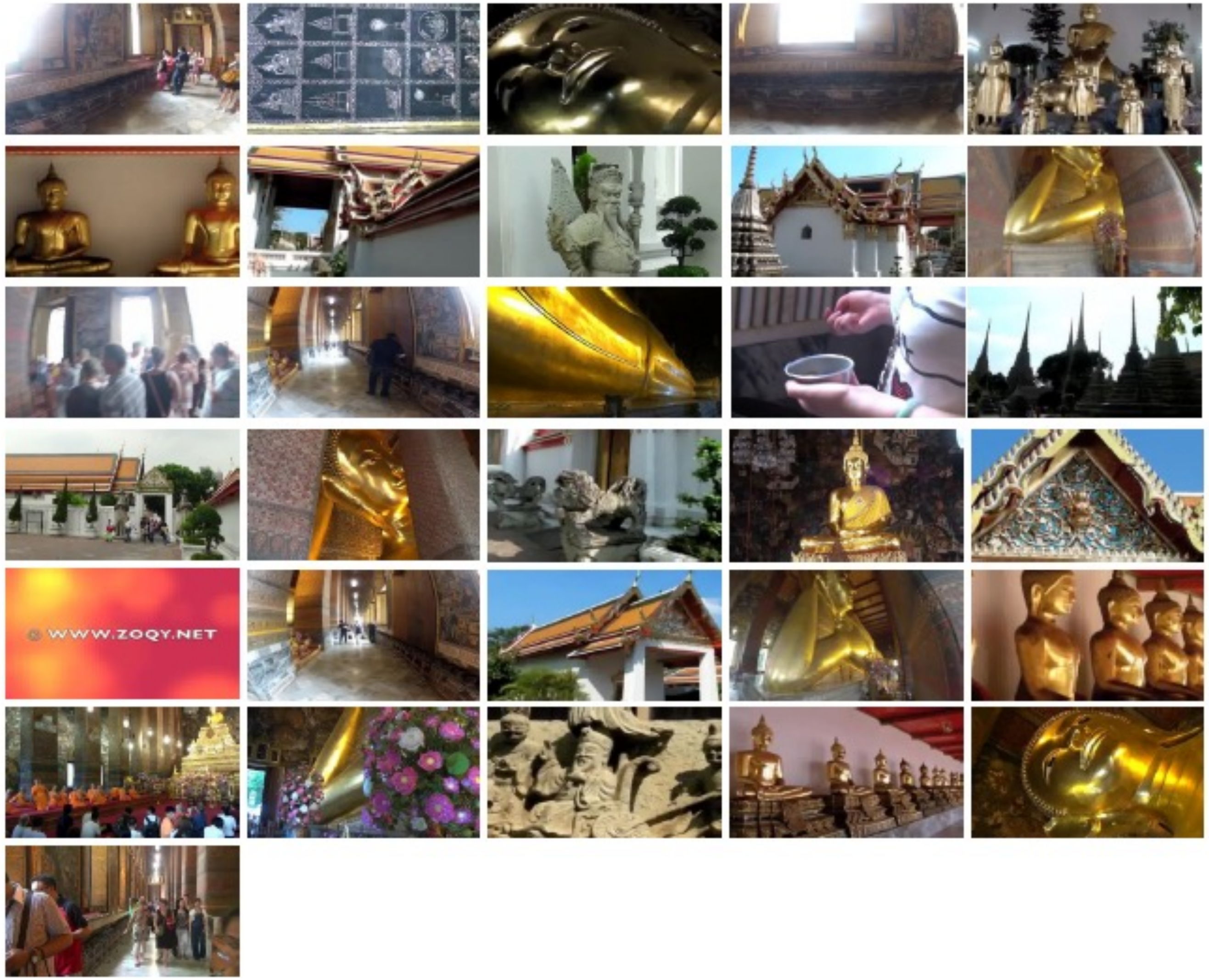
Ground Truth Summaries

Pairwise F-measure - 0.756

Exemplar Summaries: Alcatraz

- Summaries at 10% length (i.e., 22 shots out of total 223 shots)
- F-measure achieved by our approach for this topic is the highest (0.755) in our experimented dataset

Exemplar Summaries: Wat Pho



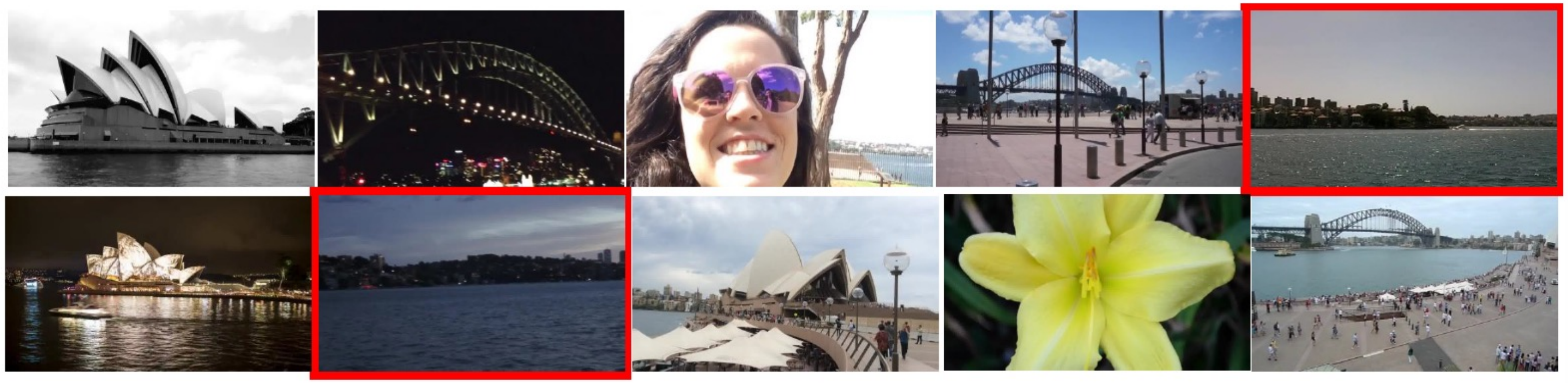
F-measure - 0.722

Qualitative Example

Summary w/o Diversity Constraint

Summary w/ Diversity
Constraint

MultiVideoMMR vs Our Approach



Summary by MultiVideoMMR

Summary by Our Approach

[*] Yingbo Li. Multi-video summarization based on Video-MMR. In WIAMIS,

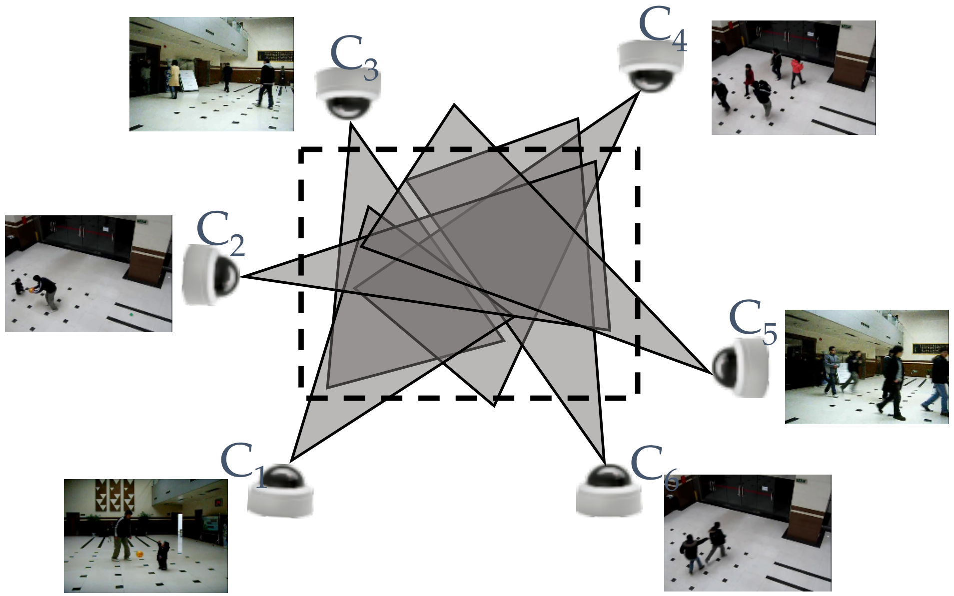
Incidental Supervision for Video Summarization

- Single-Video Summarization
  - Collaborative Video Summarization
  - Weakly Supervised Video Summarization
- **Multi-Video Summarization**
  - Diversity-aware Video Summarization
  - **Multi-View Video Summarization in a Camera Network**

Multi-View Video Summarization

Questions Asked: Can we generate a single summary from all the videos without any manual supervision?

Incidental Supervision: large amount of correlations (both intra-view as well as inter-view)



Rameswar Panda, Amit K. Roy-Chowdhury, “Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization", IEEE Transactions on Multimedia (TMM), vol. 19, no. 9, pp. 2010-2021, Sept. 2017.

Basic Idea

Split the problem into 2 sub-problems:

Capturing the multi-view content correlations via an embedded representation

Apply sparse representative selection over the embedding space to generate the summaries

Top row: SC applied to each view separately and then the results are combined to produce a single summary

Middle row: SC applied by simply concatenating all three videos into a single long video,

Bottom row: SC applied on a embedded representation that takes into account multi-view correlations

Joint Embedding and Sparse Representative Selection

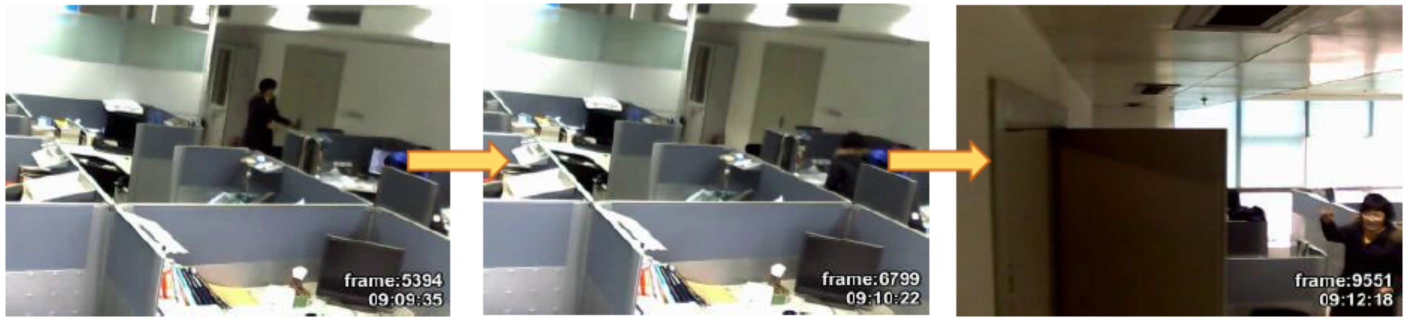
$$\min_{Y,Z,YY^T=I} tr(YLY^T) + \alpha\left(||Y - YZ||_F^2 + \lambda||Z||_{2,1}\right)$$

Optimization: Alternating minimization with Half-quadratic optimization

$$\min_{Y,Z,YY^T=I} tr(YLY^T) + \alpha\left(||Y - YZ||_F^2 + \lambda tr(Z^T PZ)\right)$$

$$P_{i,i} = \frac{1}{2\sqrt{||z^i||_2^2 + \epsilon}}$$

More Informative Summary



(a)

(b)

Sequence of Events detected related to the activities of a member **(A0)** inside the Office dataset.

(a): Summary produced by RandomWalk (TMM'10), and

(b): Summary produced by Our Proposed Framework.

3rd: **A0** is looking for a thick book to read (as per the ground truth) – not detected in (a)

Exemplar Summary

Summarized events for the Office dataset

Scalability (Analyze once, Generate many)

Summary for different user length requests

Video Summary (Office Dataset)

Total Video Duration: 46:19 mins
Summary Duration: 02:01 mins
(only 4.4% of total data)

Summary

Physics/Structure

Unsupervised *Supervised*

Incidental Supervision

Summarization Under Resource Constraints

Thank You

Additional Slides

Optimization

$$\min_{\mathbf{Z}} \ \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \lambda_d tr(\mathbf{D}^T\mathbf{Z}) + \lambda_s tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z}) + \beta tr(\mathbf{Z}^T\mathbf{R}\mathbf{Z})$$

$$\min_{\tilde{\mathbf{Z}}} \ \frac{\alpha}{2}\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2 + \lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}) + \lambda_s tr(\tilde{\mathbf{Z}}^T\mathbf{Q}\tilde{\mathbf{Z}}) + \beta tr(\tilde{\mathbf{Z}}^T\mathbf{R}\tilde{\mathbf{Z}})$$

$$(\mathbf{X}^T\mathbf{X} + 2\lambda_s\mathbf{P} + 2\beta\mathbf{R})\mathbf{Z} = (\mathbf{X}^T\mathbf{X} - \lambda_d\mathbf{D})$$

$$(\alpha\mathbf{X}^T\mathbf{X} + 2\lambda_s\mathbf{Q} + 2\beta\mathbf{R})\tilde{\mathbf{Z}} = (\alpha\mathbf{X}^T\tilde{\mathbf{X}} - \lambda_d\tilde{\mathbf{D}})$$

Solve the above two linear systems to obtain sparse coefficient matrices

Summary Generation: Sort shots according to ℓ_2 norms of the rows in $\mathbf{Z}_c$;
Construct summary from top-ranked shots

Results

- Goal: summarize each video by exploiting visual context from others
- Human Evaluation: mean Average Precision (mAP)

Video Topics	Humans			Computational methods						
	Worst	Mean	Best	CK	CS	SMRS	LL	CoC	CoSum	CVS
Base Jumping	0.652	0.831	0.896	0.415	0.463	0.487	0.504	0.561	0.631	**0.658**
Bike Polo	0.661	0.792	0.890	0.391	0.457	0.511	0.492	0.625	0.592	**0.675**
Eiffel Tower	0.697	0.758	0.881	0.398	0.445	0.532	0.556	0.575	0.618	**0.722**
Excavators River Xing	0.705	0.814	0.912	0.432	0.395	0.516	0.525	0.563	0.575	**0.693**
Kids Playing in Leaves	0.679	0.746	0.863	0.408	0.442	0.534	0.521	0.557	0.594	**0.707**
MLB	0.698	0.861	0.914	0.417	0.458	0.518	0.543	0.563	0.624	**0.679**
NFL	0.660	0.775	0.865	0.389	0.425	0.513	0.558	0.587	0.603	**0.674**
Notre Dame Cathedral	0.683	0.825	0.904	0.399	0.397	0.475	0.496	0.617	0.595	**0.702**
Statue of Liberty	0.687	0.874	0.921	0.420	0.464	0.538	0.525	0.551	0.602	**0.715**
Surfing	0.676	0.837	0.879	0.401	0.415	0.501	0.533	0.562	0.594	**0.647**
mean	**0.679**	**0.812**	**0.893**	**0.407**	**0.436**	**0.511**	**0.525**	**0.576**	**0.602**	**0.687**
relative to average human	83%	100%	110%	51%	54%	62%	64%	70%	74%	85%

CoSum Dataset (Top-5 mAP)

Ablation analysis on CoSum dataset: CVS w/ VGG features -> 0.643 mAP

CVS – Neighborhood -> 0.538 mAP, CVS – Diversity -

Results

Video Topics	Humans			Computational methods						
	Worst	Mean	Best	CK	CS	SMRS	LL	CoC	CoSum	CVS
Changing Vehicle Tire	0.285	0.461	0.589	0.225	0.235	0.287	0.272	0.336	0.295	0.328
Getting Vehicle Unstuck	0.392	0.505	0.634	0.248	0.241	0.305	0.324	0.369	0.357	0.413
Grooming an Animal	0.402	0.521	0.627	0.206	0.249	0.329	0.331	0.342	0.325	0.379
Making Sandwich	0.365	0.507	0.618	0.228	0.302	0.366	0.362	0.375	0.412	0.398
ParKour	0.372	0.503	0.622	0.196	0.223	0.311	0.289	0.324	0.318	0.354
PaRade	0.359	0.534	0.635	0.179	0.216	0.247	0.276	0.301	0.334	0.381
Flash Mob Gathering	0.337	0.484	0.606	0.218	0.252	0.294	0.302	0.318	0.365	0.365
Bee Keeping	0.298	0.515	0.591	0.203	0.247	0.278	0.297	0.295	0.313	0.326
Attempting Bike Tricks	0.365	0.498	0.602	0.226	0.295	0.318	0.314	0.327	0.365	0.402
Dog Show	0.386	0.529	0.614	0.187	0.232	0.284	0.295	0.309	0.357	0.378
mean	**0.356**	**0.505**	**0.613**	**0.211**	**0.249**	**0.301**	**0.306**	**0.329**	**0.345**	**0.372**
relative to average human	71%	100%	121%	42%	49%	60%	61%	65%	68%	74%

TVSum50 Dataset (Top-5 mAP)

Eiffel Tower

Attempting Bike Tricks

Role of topic-related visual context in summarizing videos.
Top: CVS w/o topic-related visual context, Bottom: CVS w/ topic-

Training DeSumNet

- Training the network is very difficult:
  - Challenges: video summarization datasets are very small (~ 50 videos)
  - Training 3D CNN with limited amount training data

- Our Solution:
  - Cross-Dataset Pre-training – UCF 101
  - Progressive Model Adaptation with Web Data – Webly Supervised Learning
  - Enhanced Data Augmentation – Horizontal flipping, Multi-scale jittering, Corner cropping

Experiments

- Datasets
  - CoSum and TVSum
- Compared Methods
  - Unsupervised: SMRS [CVPR'12], Quasi [CVPR'14], MBF [CVPR'15], CVS [CVPR'17]
  - Supervised: KVS [ECCV'14], seqDPP [NIPS'14], SubMod [CVPR'15]
- Settings
  - Network input: a segment of size 128 X 171 X 16, output: a video category label
  - Training: SGD with minibatch size of 50, momentum – 0.9, weight decay – 0.005
  - Learning rate – 0.003, decreased by 1/10 after 4 epochs
  - Training/Testing split: 80%/20%, dropout probability - 0.5
  - Video prediction: average over 10 random segments (88% in CoSum, 72% in TVSum)

Generating Video Skims

- Goal: generate video skim of user-defined summary length
- Human Evaluation: mean Average Precision

Table 1. Experimental results on CoSum dataset.

Mean Average Precision	Humans			Unsupervised Methods				Supervised Methods			Proposed
	Worst	Mean	Best	SMRS	Quasi	MBF	CVS	KVS	seqDPP	SubMod	DeSumNet
Top-5	0.668	0.814	0.887	0.491	0.507	0.588	0.676	0.684	0.692	0.735	0.721
Relative to average human	82.1%	100%	109.1%	60.4%	62.6%	72.3%	83.2%	84.1%	85.2%	90.3%	88.5%
Top-15	0.682	0.821	0.916	0.506	0.527	0.579	0.677	0.686	0.709	0.745	0.736
Relative to average human	83.0%	100%	111.5%	61.7%	64.3%	70.6%	82.5%	83.6%	86.5%	90.8%	89.7%

Table 2. Experimental results on TVSum dataset.

Mean Average Precision	Humans			Unsupervised Methods				Supervised Methods			Proposed
	Worst	Mean	Best	SMRS	Quasi	MBF	CVS	KVS	seqDPP	SubMod	DeSumNet
Top-5	0.382	0.516	0.608	0.322	0.334	0.353	0.388	0.398	0.447	0.461	0.424
Relative to average human	74.2%	100%	117.8%	62.5%	64.8%	68.5%	75.3%	77.3%	86.7%	89.6%	82.2%
Top-15	0.372	0.507	0.589	0.320	0.325	0.342	0.371	0.387	0.435	0.443	0.415
Relative to average human	73.5%	100%	116.3%	63.2%	64.1%	67.4%	73.2%	76.5%	85.8%	87.4%	81.8%

Effect of Training Strategies

Methods	CoSum	TVSum
Scratch	71.4	66.7
Scratch+NoisyWebData	76.3	69.5
Pre-train	83.5	75.2
Pre-train+NoisyWebData	84.4	77.3
Pre-train+ModelAdaptationwithRefinedWebData	87.7	80.8
Pre-train+ModelAdaptation+EnhancedDataAugmentation	88.5	82.2

Exploration study on training strategies. Numbers show top-5 mAP scores, relative to the average human score (in %)

Generating Video Time-lapse

- Goal: generate time-lapse videos by controlling the frame rate based on importance scores
- Segments with high importance score are played at a smaller rate and vice versa
- Compared Methods: CVS [CVPR'17], KVS [ECCV'14]
- Subjective Evaluation: 10 experts – rate overall quality from 1 (worst) to 5 (best)

Datasets	CVS	KVS	DeSumNet
CoSum	3.23	3.15	4.03
TVSum	2.34	2.56	3.18

User Study: Average human ratings in evaluating video time-lapse

Performance Comparison

F-measure comparison at 10% summary length

Topic Names	ConcateKmeans	ConcateSpectral	ConcateSparse	KmeansConcate	SpectralConcate	SparseConcate	MultiVideoContent	MultiVideoMMR	Ours
Angkor Wat (7)	0.428	[illegible]	[illegible]	0.418	0.418	0.391	0.431	0.452	**0.567**
Machu Picchu (7)	0.336	0.367	0.379	0.373	0.394	0.427	0.438	0.507	**0.582**
Taj Mahal (7)	0.428	0.484	0.465	0.518	0.522	0.588	0.593	0.533	**0.679**
Basilica of Sagrada Familia (6)	0.423	0.415	0.461	0.382	0.427	0.478	0.488	0.492	**0.597**
St. Peter's Basilica (5)	0.437	0.458	0.497	0.533	0.526	0.575	0.586	0.602	**0.699**
Milan Cathedral (10)	0.475	0.430	0.451	0.449	0.442	0.489	0.481	0.473	**0.571**
Alcatraz (6)	0.601	0.550	0.638	0.631	0.651	0.729	0.652	0.668	**0.755**
Golden Gate Bridge (6)	0.447	0.443	0.508	0.504	0.475	0.509	0.527	0.515	**0.618**
Eiffel Tower (8)	0.408	0.390	0.460	0.401	0.427	0.448	0.436	0.446	**0.562**
Notre Dame Cathedral (8)	0.315	0.350	0.235	0.413	0.451	0.461	0.463	0.473	**0.550**
The Alhambra (6)	0.485	0.570	0.543	0.551	0.551	0.567	0.553	0.582	**0.662**
Hagia Sophia Museum (6)	0.305	0.346	0.315	0.433	0.384	0.523	0.473	0.536	**0.585**
Charles Bridge (6)	0.400	0.379	0.414	0.409	0.444	0.451	0.453	**0.534**	0.525
Great Wall at Mutiantu (5)	0.390	0.410	0.484	0.500	0.474	0.488	0.493	0.507	**0.673**
Burj Khalifa (9)	0.284	0.362	0.350	0.301	0.355	0.352	**0.450**	0.392	0.441
Wat Pho (5)	0.342	0.414	0.564	0.501	0.575	0.633	0.625	0.603	**0.722**
Chichen Itza (8)	0.337	0.361	0.430	0.413	0.426	0.507	0.514	0.492	**0.582**
Sydney Opera House (10)	0.400	0.391	0.497	0.409	0.458	0.474	0.503	0.512	**0.614**
Petronas Twin Towers (9)	0.302	0.326	0.421	0.418	0.376	0.445	0.453	0.486	**0.643**
Panama Canal (6)	0.377	0.410	0.492	0.539	0.523	0.528	0.512	0.544	**0.639**
mean	**0.396**	**0.413**	**0.450**	**0.455**	**0.465**	**0.503**	**0.506**	**0.517**	**0.613**

- Our approach statistically significantly outperforms all other compared methods ($p < 0.01$)
- Our method achieves the highest overall score of 0.613, while the strongest baseline reaches 0.517 (MultiVideoMMR)

Multi-View Video Embedding

Input: a set of K different videos $X^k = \{x_i^k \in R^D, i = 1, \cdots, N_k\}, k = 1, \cdots, K$

Output: a set of embedded coordinates $Y^k = \{y_i^k \in R^d, i = 1, \cdots, N_k\}, k = 1, \cdots, K$

x_i: D-dimensional feature descriptor of a shot

$d \ll D$

Constraints:

Intra-view correlations: shots with high feature similarity in a video should be close to each other

Inter-view correlations: shots from different videos with high feature similarity should also be close to each other

Objective Function

Aim: correctly match the proximity score between two shots x_i and x_j to the score between y_i and y_j respectively

$$\mathcal{F}_{\text{intra}}(Y^{(k)}) = \sum_{i,j} ||y_i^{(k)} - y_j^{(k)}||^2 C_{intra}^{(k)}(i,j)$$

$$\mathcal{F}_{\text{inter}}(Y^{(m)}, Y^{(n)}) = \sum_{i,j} ||y_i^{(m)} - y_j^{(n)}||^2 C_{inter}^{(m,n)}(i,j)$$

$$\sum_{k} \sum_{i,j} ||y_i^{(k)} - y_j^{(k)}||^2 C_{intra}^{(k)}(i,j) + \sum_{\substack{m,n \\ m \neq n}} \sum_{i,j} ||y_i^{(m)} - y_j^{(n)}||^2 C_{inter}^{(m,n)}(i,j)$$

$$\mathcal{F}(Y) = \sum_{m,n} \sum_{i,j} ||y_i^{(m)} - y_j^{(m)}||^2 C_{total}^{(m,n)}(i,j)$$

Objective is to minimize the function

$$C_{total}^{(m,n)}(i,j) = \begin{cases} C_{intra}^{(k)}(i,j) & \text{if } m = n = k \\ C_{inter}^{(m,n)}(i,j) & \text{otherwise} \end{cases}$$

Objective Function

$$\mathcal{F}(Y) = \sum_{m,n}\sum_{i,j} ||y_i^{(m)} - y_j^{(m)}||^2 W^{(m,n)}(i,j)$$

$$W = C_{total} + C_{total}^{T}$$

Equivalent to Laplacian embedding:

$$Y^* = \underset{Y, YY^T = I}{\operatorname{argmin}}\ tr\left(YLY^T\right)$$

Solution: Generalized Eigen vector problem

$$Ly = \lambda Dy$$

Bottom d non-zero Eigen vectors

Multi-Camera Video Summarization (TMM'17)

Experiments

Datasets	# Views	Total Durations (Mins.)	Settings	Camera Type
Office	4	46:19	Indoor	Fixed
Campus	4	56:43	Outdoor	Non-fixed
Lobby	3	24:42	Indoor	Fixed
Road	3	22:46	Outdoor	Non-fixed
Badminton	3	15:07	Indoor	Fixed
BL-7F	19	136:10	Indoor	Fixed

Performance Measures: Precision, Recall, F1-measure

Ground Truths: Events reported in Fu et. al. TMM'10

Dataset Source: http://cs.nju.edu.cn/ywguo/summarization.html

- Same dataset has been used by all previous works

Multi-Camera Video Summarization (TMM'17)

Results

Methods	Office			Campus			Lobby			Reference
	P	R	F	P	R	F	P	R	F	
Attention-Concate	100	46	63.01	40	28	32.66	100	70	82.21	TMM2005 [37]
Sparse-Concate	100	50	66.67	56	55	55.70	91	70	78.95	TMM2012 [8]
Concate-Attention	100	38	55.07	56	48	51.86	95	72	81.98	TMM2005 [37]
Concate-Sparse	93	58	71.30	56	62	58.63	86	70	77.18	TMM2012 [8]
Graph	100	26	41.26	50	48	49.13	100	58	73.41	TCSVT2006 [51]
RandomWalk	100	61	75.77	70	55	61.56	100	77	86.81	TMM2010 [14]
RoughSets	100	61	75.77	69	57	62.14	97	74	84.17	ICIP2011 [33]
BipartiteOPF	100	69	81.79	75	69	71.82	100	79	88.26	TMM2015 [28]
Ours	**100**	**81**	**89.36**	**84**	**72**	**77.78**	**100**	**86**	**92.52**	Proposed

Methods	Precision(%)	Recall(%)	F-measure(%)	Reference
GMM	58	61	60.00	JSTSP2015 [43]
Ours	73	70	71.29	Proposed

BL-7F

Methods	Office	Campus	Lobby	Reference
[45]	84.48	75.42	88.26	ICPR2016 [45]
Ours	89.36	77.78	92.52	Proposed

Advantage of Joint Optimization