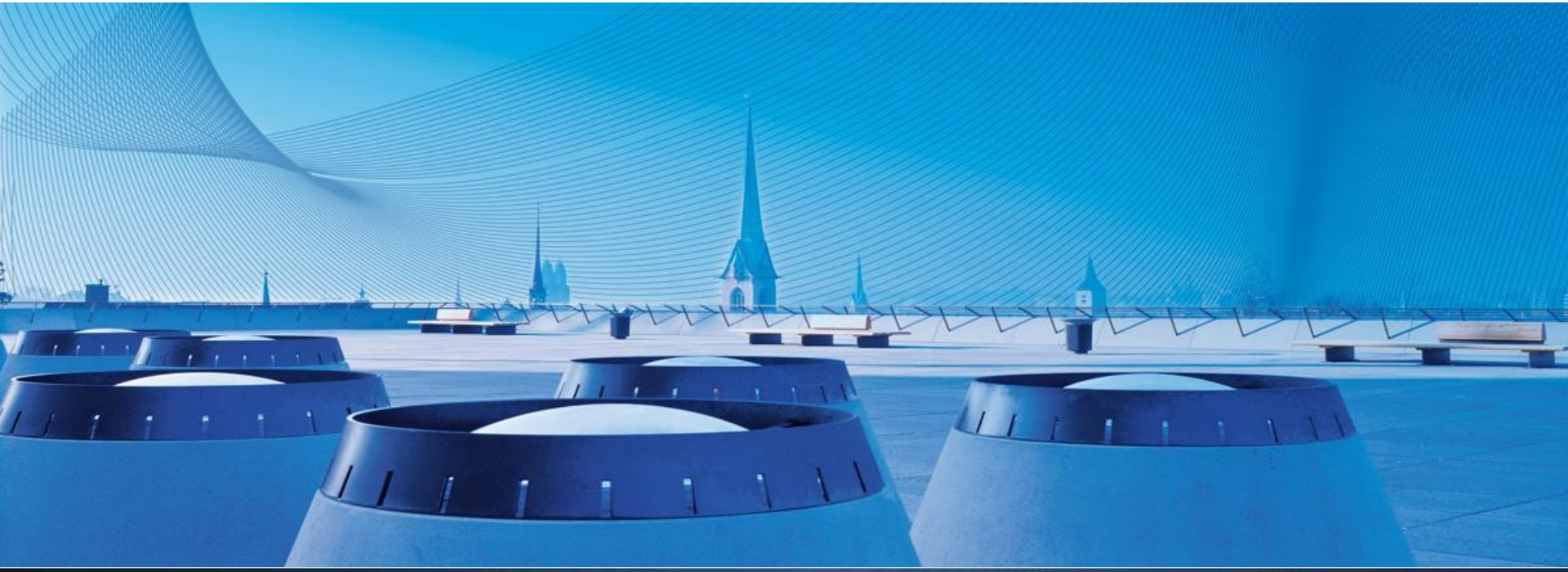
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CVL Computer Vision Lab

Video Summarization as Subset Selection

Tutorial @ Optimization Algorithms for Subset Selection and Summarization in Large Data Sets, CVPR 2016

Michael Gygli, PhD student @ Computer Vision Lab, ETH Zurich



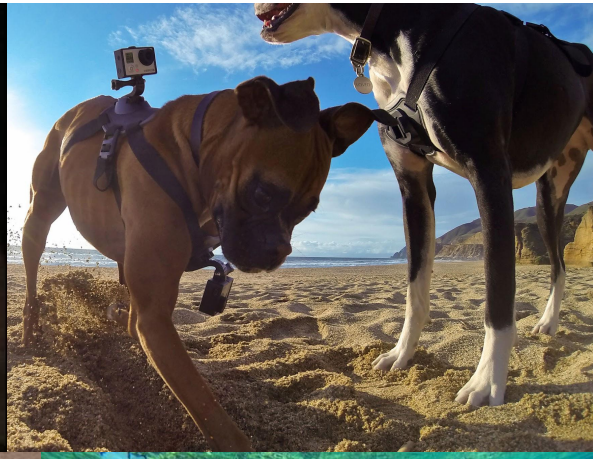
Overview

- Motivation
- Overview of Video Summarization and Related Work
- Finding the Most Interesting and Relevant Content
- Find a Good Subset of Frames / Segments
- Demo
- Conclusion and Outlook

Why do we need summarization?

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







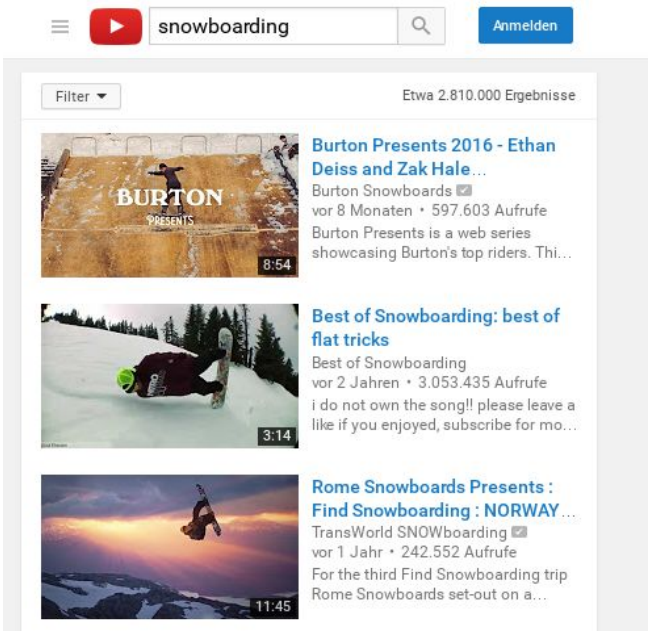

WAIT
FOR

Why do we need summarization?

- Video capture is omnipresent and vast
- 400 hours of video are uploaded to YouTube every minute
- Users have a "capture first, filter later" mentality
  - Later often means never

Why do we need summarization?

- Large amounts of video
  - Need to search for relevant content quickly
    - Based on text search
    - Based on visual inspection (Thumbnails [1])



[1] http://googleresearch.blogspot.ch/2015/10/improving-youtube-video-thumbnails-with.html

Why do we need summarization?

- Content quality degradation due to cheap acquisition
  - Need better presentation
    - Through frame preview
    - Through automatic editing and summarization

Source: YouTube

What does it mean to summarize?

- Definition of “summarize” (text)

  “Give a brief statement of the main points of (something)” [Oxford dictionary]

- Similarly for video summaries
  - *Brief*
    - The summary should be non redundant or diverse
  - Cover the main points
    - *Main:* Frequently occurring content
    - *Points:* Interesting and visually informative frames or segments

Why is it challenging?

- Human-centric task
  - Context dependent
  - Subject dependent
- What is summary-worthy depends on high-level semantics
  - Objects, actions
  - Motion, emotion




- Highly diverse inputs
  - Summarization should work in any setting
- Collecting ground truth is demanding
  - Prevents training large-scale models
  - Objectively comparing methods is difficult

- Motivation
- **Overview of Video Summarization and Related Work**
- Finding the Most Interesting and Relevant Content
- Find a Good Subset of Frames / Segments
- Demo
- Conclusion and Outlook

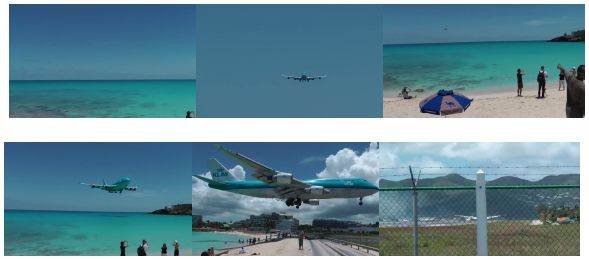
Overview of Video Summarization and Related Work

Focus: Select the “right” visual content:

- Skims (short videos)

[Gygli et al. ECCV 14], [Song et al. CVPR 15], [Arev et al. Siggraph 2014] [Potapov et al. ECCV 2014], [Sun et al. ECCV 2014], [Lu & Grauman CVPR 13], [Zhang et al. CVPR 16], [Huan et al. ICCV 2015], [Lin et al. ICCV Workshop 15], [Zhang et al. arXiv 16], [Xu et al. CVPR 2015], [Guoa et al. Neurocomputing 16]

- Keyframes



[Wolf, ICASSP 1996], [Huang & Mehrotra 98], [Smith & Kanade 98], [Liu et al. IJCAI 2009], [Liu et al. PAMI 10], [Lee et al. CVPR 12], [Ejaz et al. 13], [Khosla et al. CVPR 13], [Kim et al. CVPR 14], [Gong et al. NIPS 2015], [Liu et al. CVPR 15], [Elhamifar et al PAMI 16], [Zen et al ICMR16]

Overview of Video Summarization and Related Work

Focus: Present content in a condensed and appealing way

- Video Synopses

[Rav-Acha et al. CVPR 06], [Pritch et al. ICCV 07 PAMI 08, AVSS 09]

- Montages

[Aner & Kender ECCV 02], [Kang et al. CVPR 06], [Liu et al. ACM MM 08], [Sun et al. ECCV 14],

- Storyboards

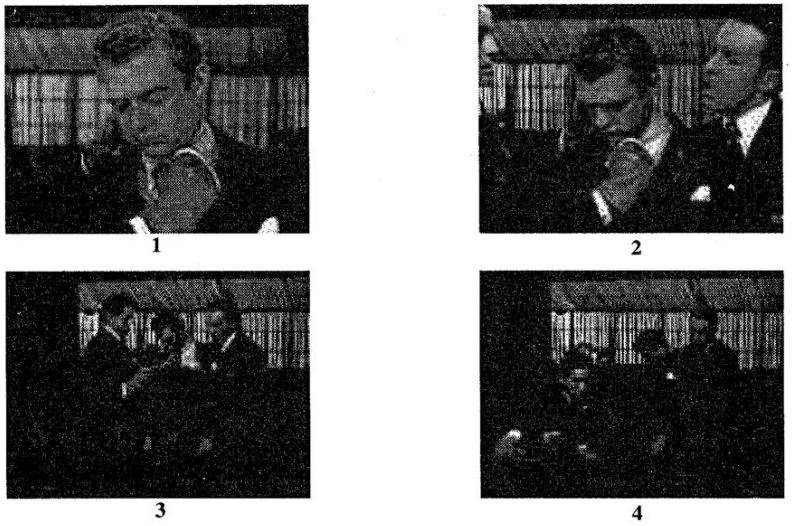
[Goldman et al. Siggraph 06], [Lee et al. CVPR 12]

- Hyperlapes

[Kopf et al. SIGGRAPH 14], [Joshi et al. Siggraph 15], [Poleg et al. CVPR 15], [Halperin arXiv 16]

Early Works

- Tackling edited videos
- Low-level cues for keyframe selection
  - Color histograms for shot detection
  - Optical flow for motion; select still frames [Wolf, ICASSP 96] or frames before/after a translation [Guironnet et al. EURASIP 07]
  - Based on heuristics, not explicitly optimizing any objectives



[Wolf, ICASSP 1996]

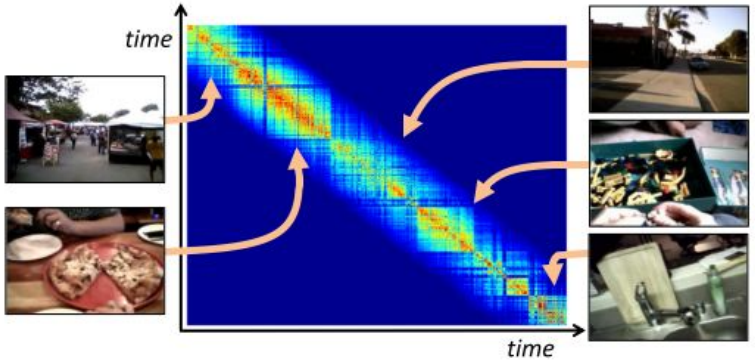
Discovering important people and objects for egocentric video summarization [Lee et al. CVPR 2012]

- Transition to high-level analysis
- Cluster frames into events using visual and temporal proximity
  - Per event, generate region proposals
  - Find high importance regions using a learnt model

GT: Important

GT: Unimportant

  - Cluster them and select highest scoring people/objects per cluster

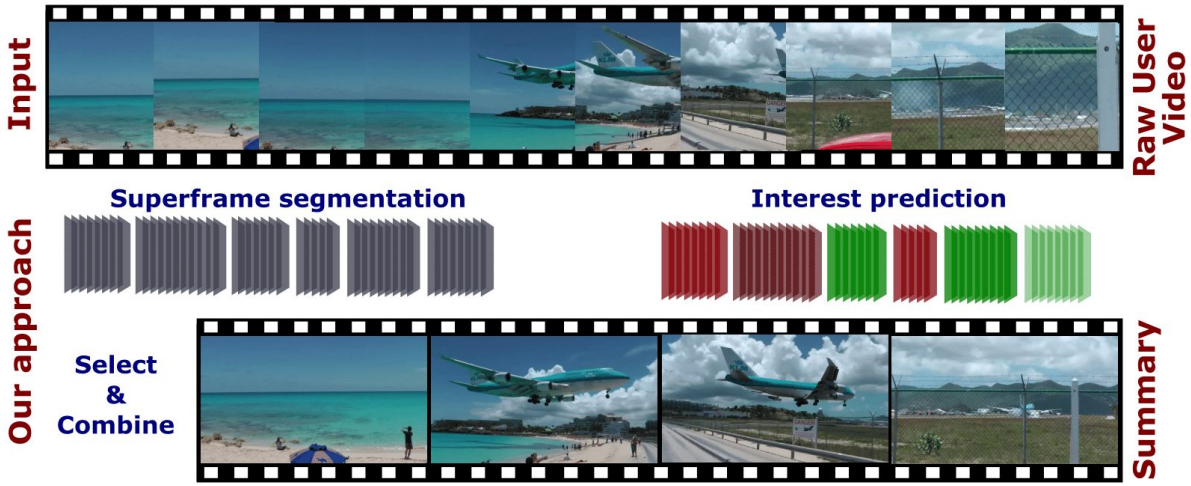


[Lee et al. CVPR 12]

Thus: importance prediction plus clustering for diversity

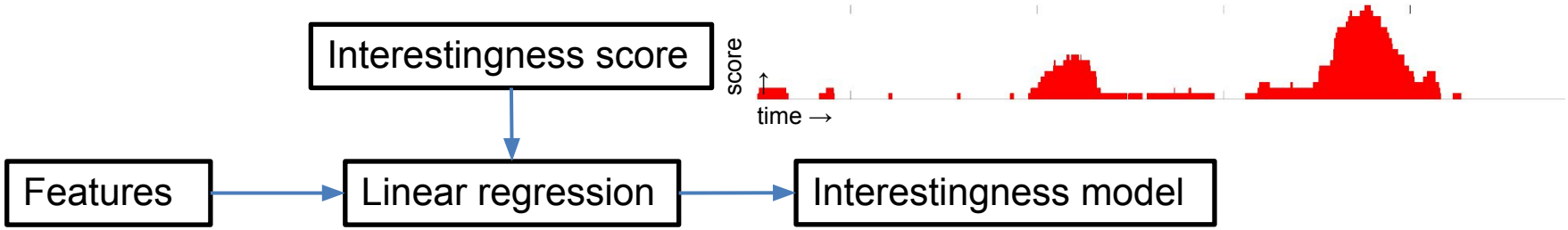
Creating Summaries from User Videos [Gygli et al. ECCV 14]

- Segment the video according to cinematographic rules (motion cues)
- Score segments according to estimated interestingness
- Select best segments under duration constraints (knapsack)



Creating Summaries from User Videos [Gygli et al. ECCV 14]

Interestingness Model



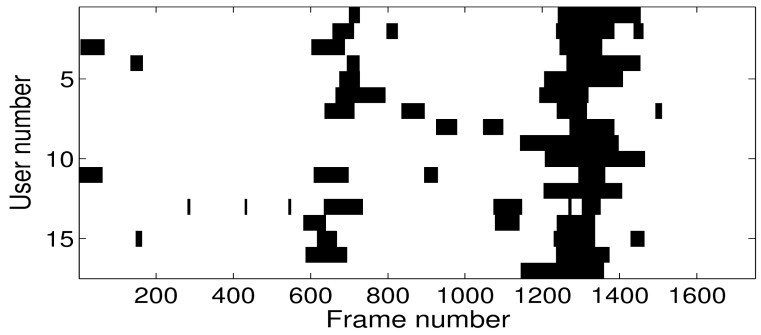
Spatio-temporal attention model
Aesthetics/Quality
Presence of landmarks
Face/Person area
«Follow object»: Predict, if the camera tracks a moving object

Creating Summaries from User Videos

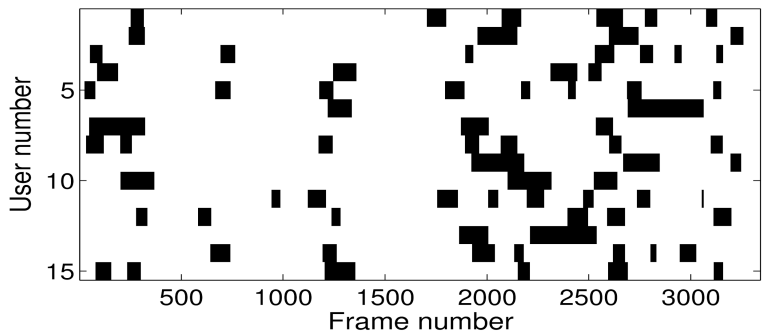
Dataset for training and testing: SumMe

- 25 videos
- 390 ground truth summaries

St Maarten Landing



Bearpark climbing

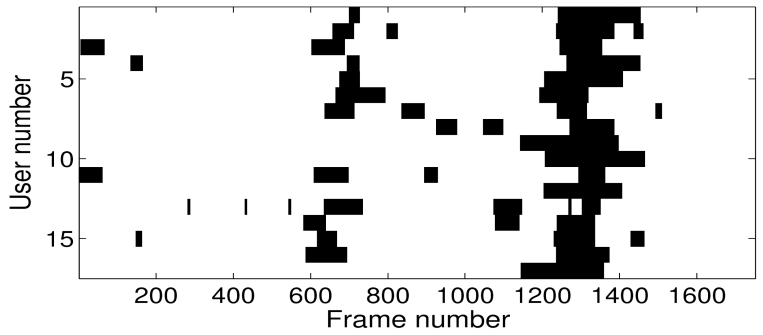


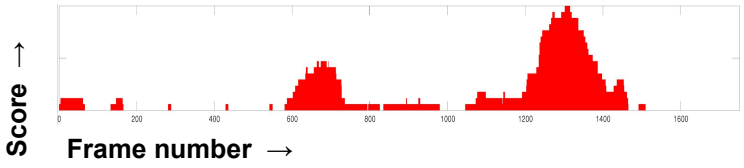
Creating Summaries from User Videos

Dataset for training and testing: SumMe

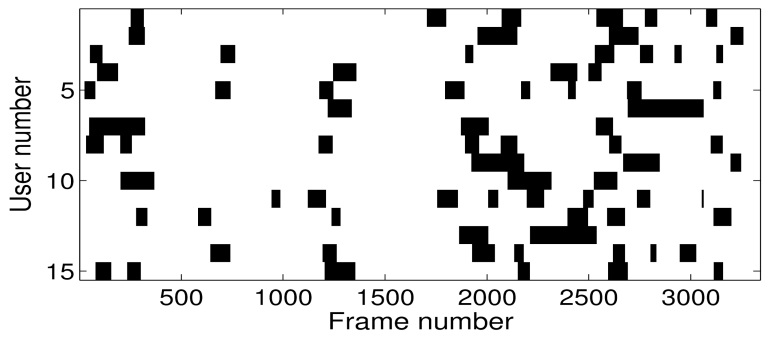
- 25 videos
- 390 ground truth summaries

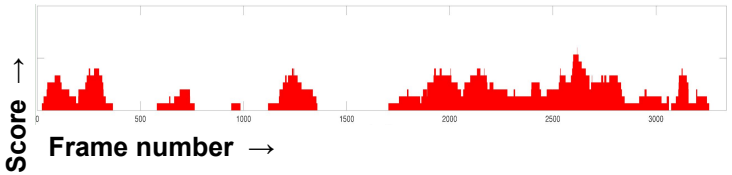
St Maarten Landing



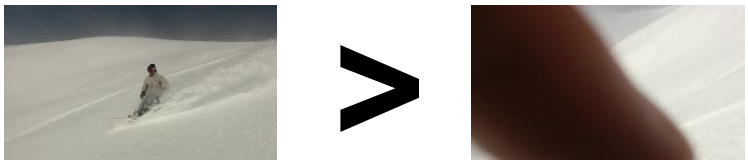


Bearpark climbing





- Motivation
- Overview of Video Summarization and Related Work
- **Finding the Most Interesting and Relevant Content**



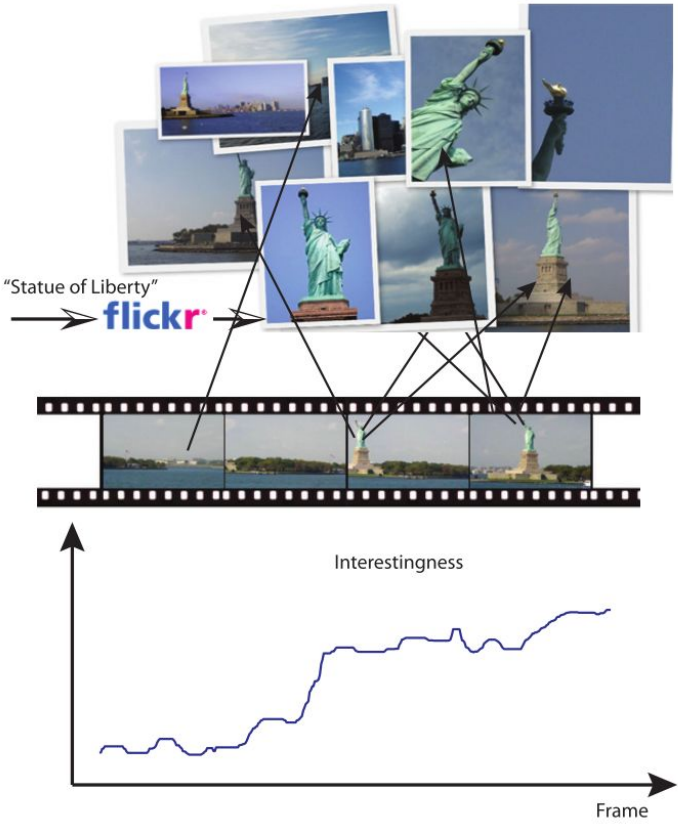
- Find a Good Subset of Frames / Segments
- Demo
- Conclusion and Outlook

Finding the most interesting and relevant content

- Two dominant approaches for relevance prediction
  - Supervised learning with generic features and large training set
    E.g. [Potapov et al. ECCV 2014], [Sun et al. ECCV 2014], [Gygli et al. CVPR 16]
  - Use of textual information (title, category) to obtain a video specific model
    E.g. [Khosla et al. CVPR 13], [Liu et al. IJCAI 09], Song et al. CVPR 15]
  - Some methods use both
    E.g. [Liu et al. CVPR 15]

Web image priors for relevance prediction

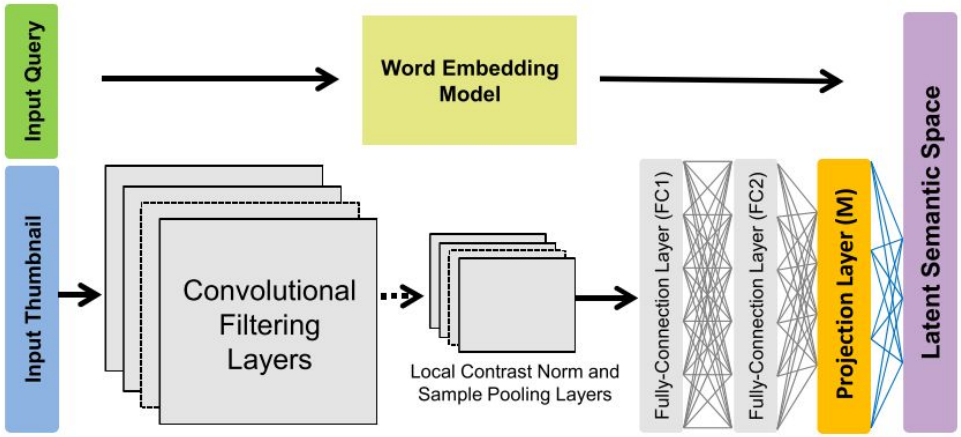
- Central idea: The average image on the web is more relevant and aesthetic than a typical video frame of the same topic
- Typically build a model per video category [Khosla et al. CVPR 13] or video title [Liu et al. IJCAI 09, Song et al. CVPR 15]



[Liu et al. IJCAI 2009]

Multi-task deep visual-semantic embedding for video thumbnail selection [Liu et al. CVPR 15]

- Use Bing image search data (query, image, # of clicks) to learn a joint embedding space for images and text
- Compute frame relevance as cosine similarity between the query or title embedding and the frame embedding

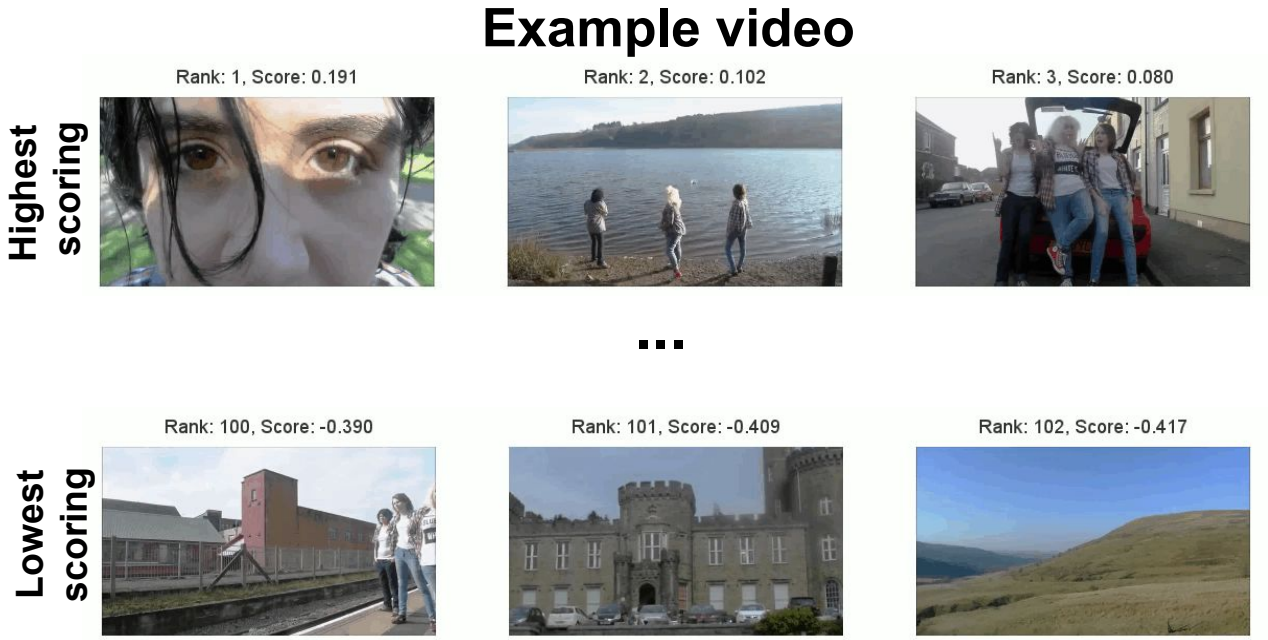


[Liu et al. CVPR 2015]

Video2GIF: Automatic Generation of Animated GIFs from Video [Gygli et al. CVPR 16]

Approach

- Work with segments as units
  - obtained through change-point detection [Song et al. CVPR 15]
- Train a deep neural network for ranking segments



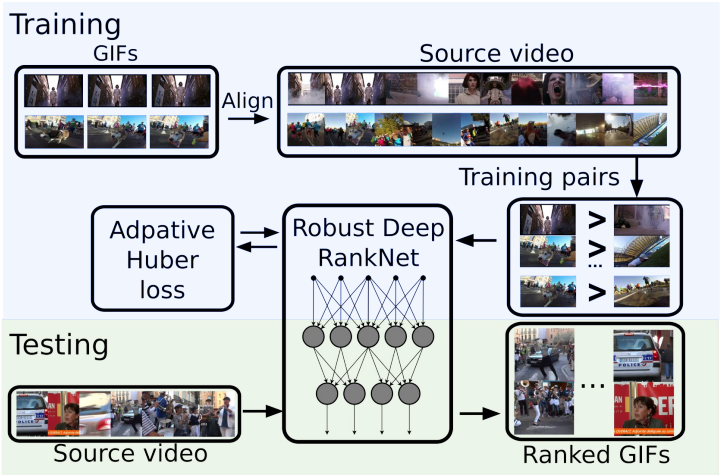
Video2GIF: Automatic Generation of Animated GIFs from Video

Approach

- Train a deep neural network for ranking segments
  - Built on C3D network [Tran et al. ICCV 2015]
- Objective: score positives higher than negatives

$$h(s^+) > h(s^-), \quad \forall \left(s^+, s^-\right)$$

h: scoring function
s^+: positive segment
s^-: negative segment



Video2GIF - Method

Loss function

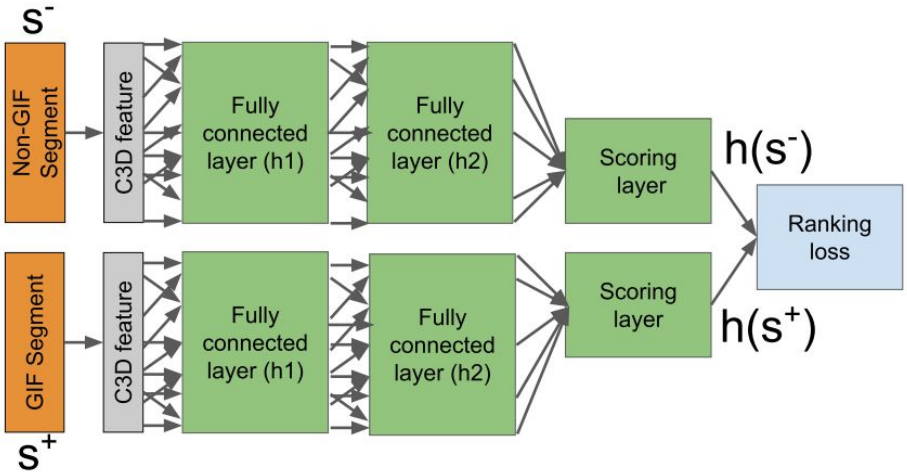
$$l_p(s^+, s^-) = \max\left(0, 1 - h(s^+) + h(s^-)\right)^p$$

p = 1:linear loss
p = 2: squared loss

- Use Huber loss: squared of small margin violations, linear for large violations (outliers)

Architecture

- Siamese network
- 2 hidden layers
- Dropout



Video2GIF: Dataset

- Large-scale training data: GIFs created from YouTube videos
- Align GIF back to video
  - This part defines a positive example
  - Assume non-selected parts are less interesting than selected part

Crawling idea inspired by [Sun et al. ECCV 2014]

Video2GIF: Dataset

Property	Quantity
Total number of animated GIFs	121,647
Mean GIF duration	5.8 sec
Total number of videos	84,754
Total video duration	7,379 hr
Mean video duration	5.2 min
Total number of videos (CC-BY)	432
GIFs used in experiment	100,699
Videos used in experiment	70,456

Most frequent tags

Available on github.com/gyglim/video2gif_dataset

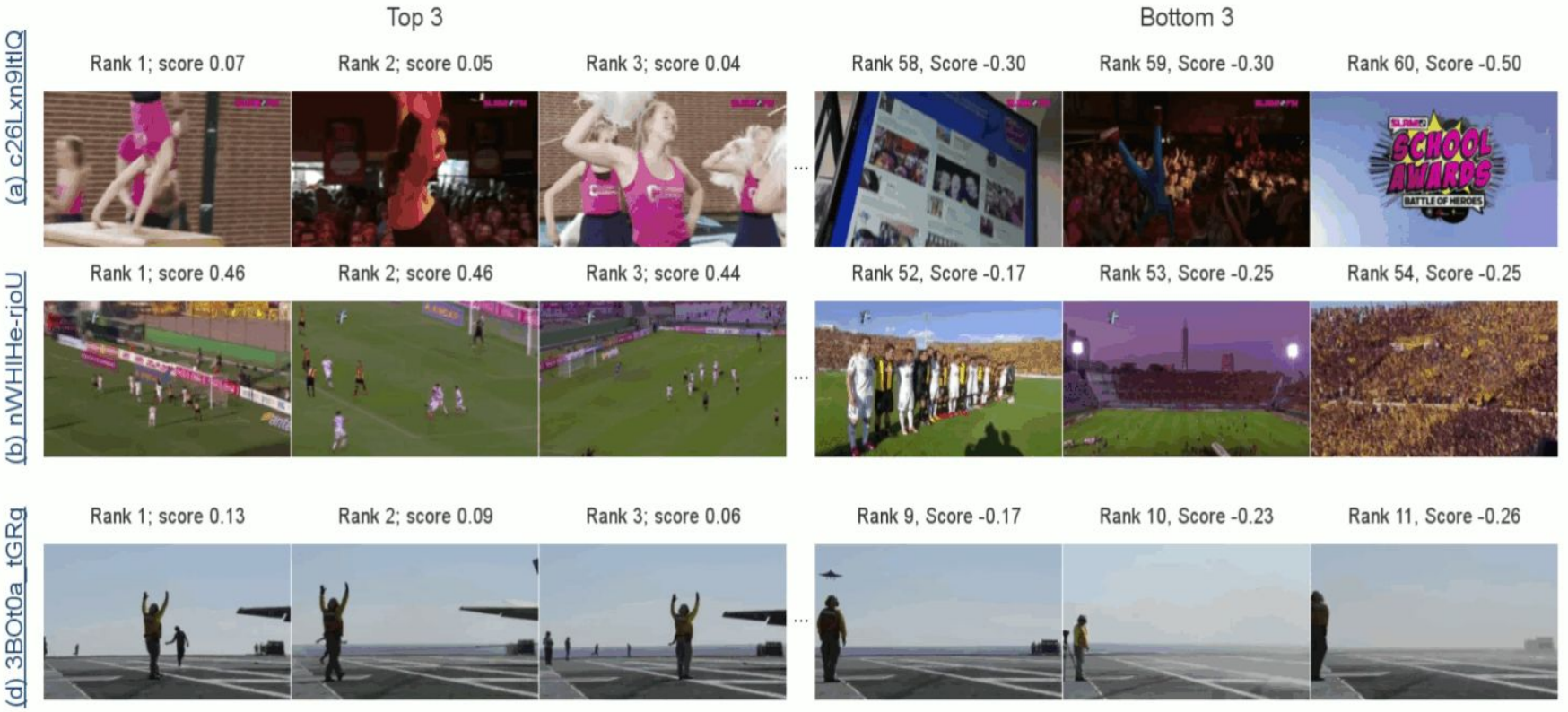
Video2GIF results

Method	nMSD (lower is better)	mAP (higher is better)
Category-specific SVM [Potapov et al. ECCV14]	52.98%	13.46%
Domain-specific rankSVM [Sun et al. ECCV 14]	46.40%	16.08%
Classification	61.37%	11.78%
Ranking across videos	53.71%	13.25%
Ranking within videos	44.60%	16.09%
Ours full	**44.08%**	**16.21%**
Approx. bounds	38.77%	21.30%

Observations:

- Ranking within video better than classification or ranking across videos
- Non-linear model improves performance (despite using deep features as input to SVM)

Video2GIF results



Results on the full test set, see: video2gif.info

- Motivation
- Overview of Video Summarization and Related Work
- Finding the Most Interesting and Relevant Content
- **Find a Good Subset of Frames / Segments**

- Demo
- Conclusion and Outlook

Video summarization and submodularity

- Imaging going through your camera feed, cleaning out videos and categorizing them
  - Should you delete this one? From what event is it?

Probably not from a beach holiday, but what then? Is it worth keeping?

Video summarization and submodularity

- And if we observe some (bad) frame?

Event: Some scene in the snow. Is the video bad or just this part?

Video summarization and submodularity

- And if we observe some (bad) frame?

Event: Some scene in the snow. Is the video bad or just this part?

- A good frame?

Event: My brother snowboarding of off-piste

Video summarization and submodularity

- And if we observe some (bad) frame?

Event: Some scene in the snow. Is the video bad or just this part?

- A good frame?

Event: My brother snowboarding of off-piste

- Or two frames?


Second frame doesn't give new information

Video summarization and submodularity

- And if we observe some (bad) frame?

Event: Some scene in the snow. Is the video bad or just this part?

- A good frame?

Event: My brother snowboarding of off-piste

- How about these two frames?


Here, the second frame does give new information

Video summarization and submodularity

- Amount of new information obtained from an additional frame depends on what other frames we observe

  Let $f(S)$ be a function scoring how informative a set of frames S is.

- Given f_1=  f_2 = 

  A ={} , B = {f_2}, or, more generally A$\subseteq B$

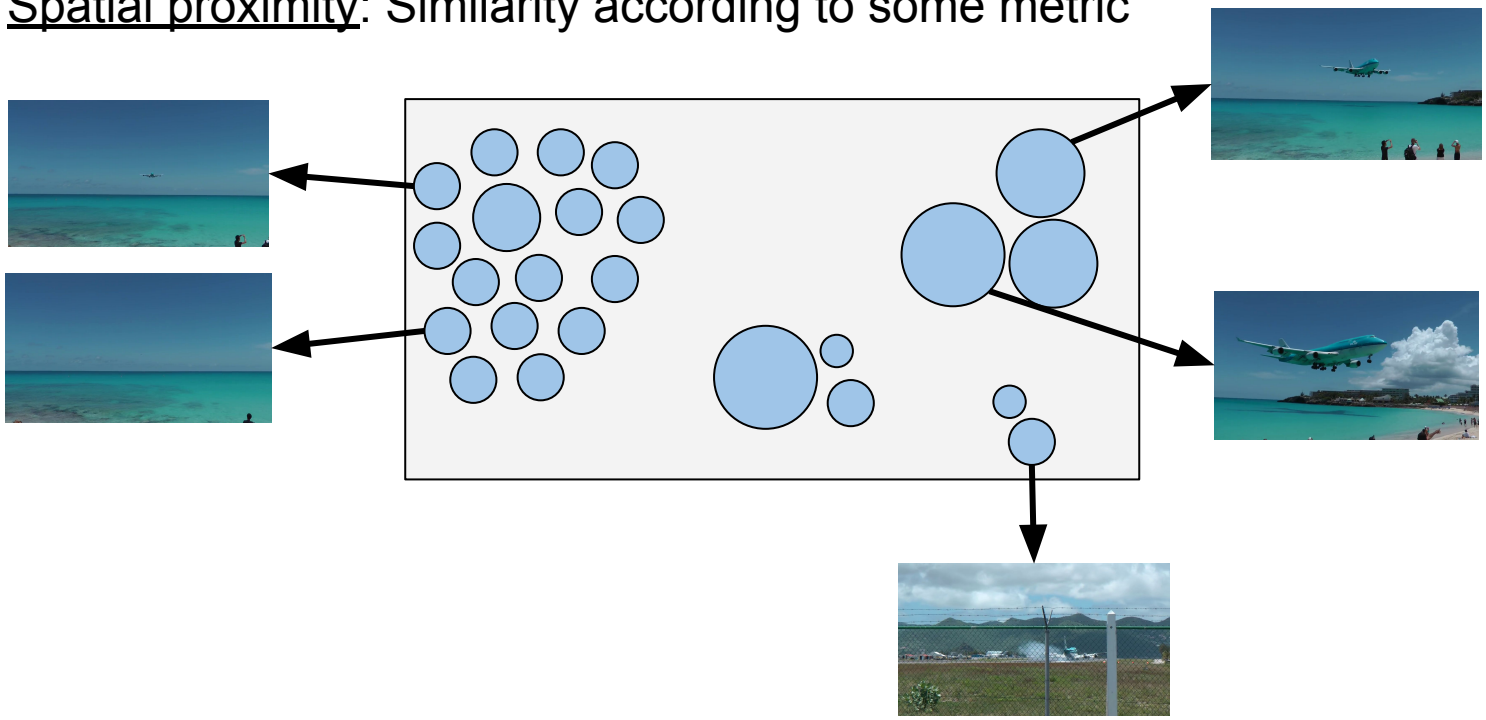
$$f(A \cup f_1) - f(A) \geq f(B \cup f_1) - f(B)$$

Definition of submodularity

Video summarization as submodular maximization

Visualize a video as segments (Circles)

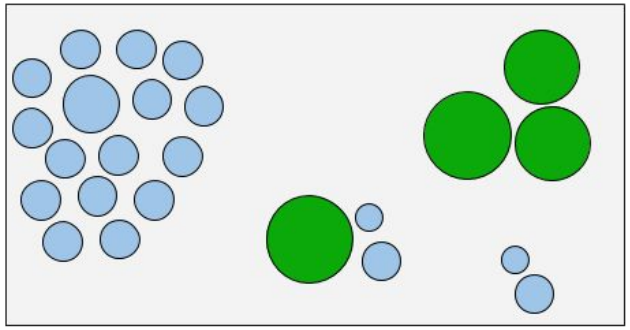
- Circle size: An interestingness estimate
- Spatial proximity: Similarity according to some metric



Video summarization as submodular maximization

Green: Selected segment

Only optimize for interestingness



[Liu et al. IJCAI 09, Gygli et al., Sun et al., Patapov et al. ECCV 14]

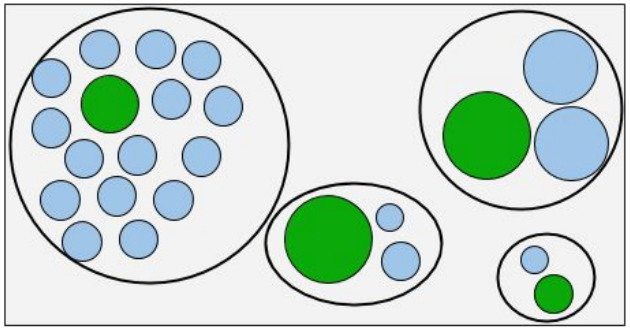
Video summarization as submodular maximization

Green: Selected segment

Only optimize for interestingness

[Liu et al. IJCAI 09, Gygli et al., Sun et al., Patapov et al. ECCV 14]

Interest+diversity through clustering



[Lee et al. CVPR 12, De Avila et al. PR letters 11, etc.]

Video summarization as submodular maximization

Green: Selected segment

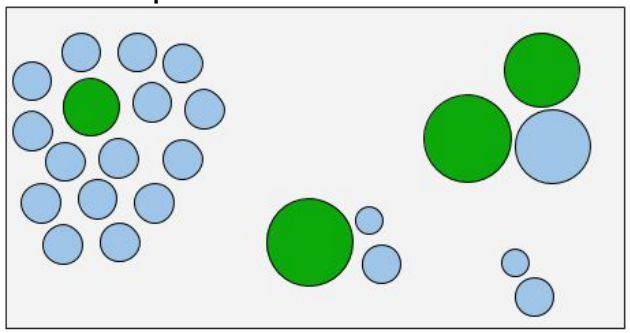
Only optimize for interestingness

[Liu et al. IJCAI 09, Gygli et al., Sun et al., Patapov et al. ECCV 14]

Interest+diversity through clustering

[Lee et al. CVPR 12, De Avila et al. PR letters 11, etc.]

Joint optimization for both

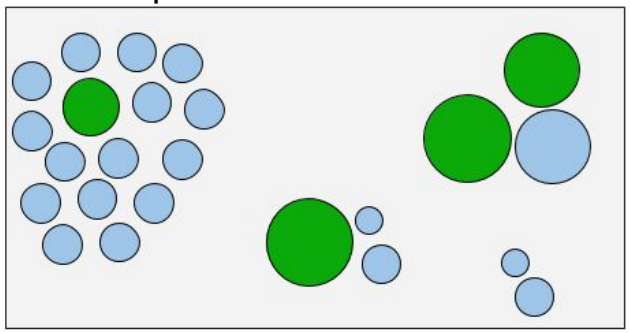


[Gygli et al., Xu et al. CVPR 15]

- Ideally, a method avoids hard decisions and optimizes for multiple objectives simultaneously (right)
- Submodular mixtures are ideal for this and extensively used for video summarization and image collection summarization

Video Summarization by Learning Submodular Mixtures of Objectives [Gygli et al. CVPR 15]

Joint optimization for both



Global summarization model

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathbf{Y_V}} \mathbf{w}^{\mathrm{T}} \mathbf{f_x}(\mathbf{y})$$

Restricting $\mathbf{f_x}(\mathbf{y})$ to be submodular and learning the weights **w.**

Y_v: candidate segments (ground set)
y: a possible subset of **Y_v** (candidate solution)
$f_x(y)$: vector of scores each objective f assigns to set **y**

Video Summarization by Learning Submodular Mixtures of Objectives

Objectives

- Interestingness / relevance
  - $\sum_{s \in S} I(x_s)$: simple sum over the scores of the elements in S
- Representativeness (diversity)
  - K-medoid objective: minimize distance of each segment in the video to closest segment in summary

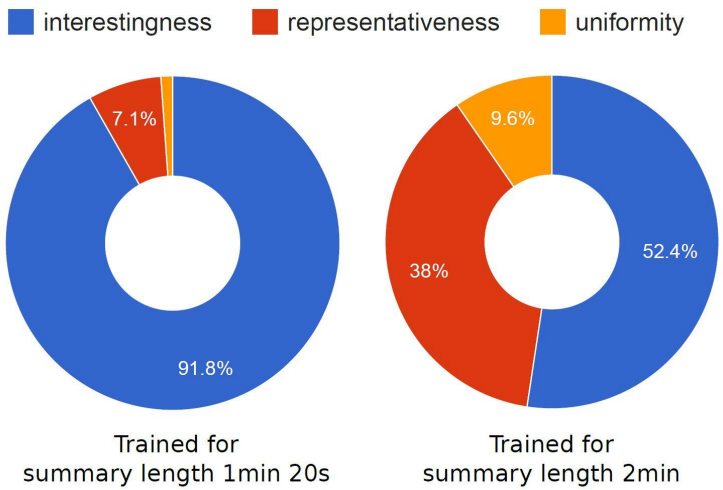
$$L(S) = \sum_{i \in V} min_{s \in S} \quad ||x_s - x_i||^2$$

- Temporal representativeness (uniformity)

Video Summarization by Learning Submodular Mixtures of Objectives

Weight learning

- Structured learning
- Loss: $\hat{L}_t(\mathbf{w}) = \max_{\mathbf{y} \subseteq \mathcal{Y}_{\mathcal{V}^{(t)}}} (\mathbf{w}^{\mathrm{T}} f(\mathbf{x}_{\mathcal{V}}^{(t)}, \mathbf{y}) + l_t(\mathbf{y})) - \mathbf{w}^{\mathrm{T}} f(\mathbf{x}_{\mathcal{V}}^{(t)}, \mathbf{y}_{\mathrm{gt}}^{(t)})$
- Optimized using Subgradient Descent



On life-logging dataset [Lee et al. CVPR 12]

Results

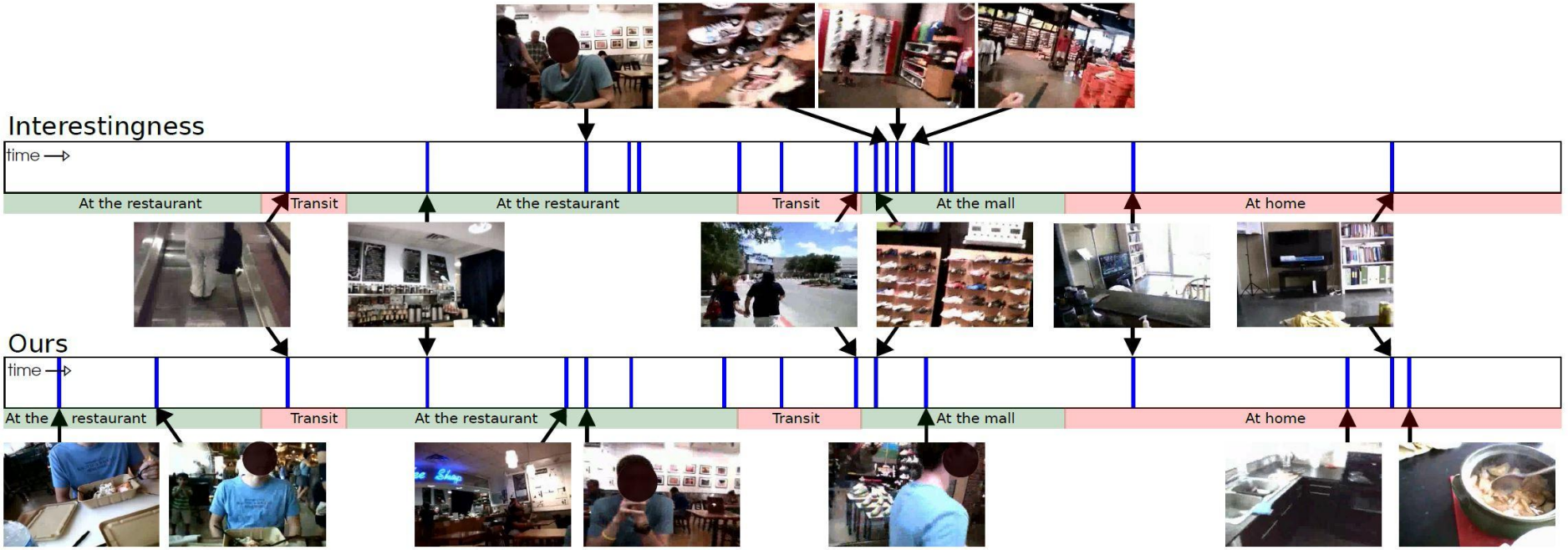
Egocentric life-logging dataset [Lee et al. CVPR 12]

Method	F-measure	Recall
Video MMR	25.57	23.10
Uniformity	25.41	22.27
Interestingness	27.07	24.78
Representative	27.02	23.51
Combined	29.01	26.61

- Performance metric: Based on the text associated with the selected segments [Yeung et al. arXiv 14]
- Uniform selection is very competitive
  - Hours long
  - No clear highlights
- Combination of objectives works best

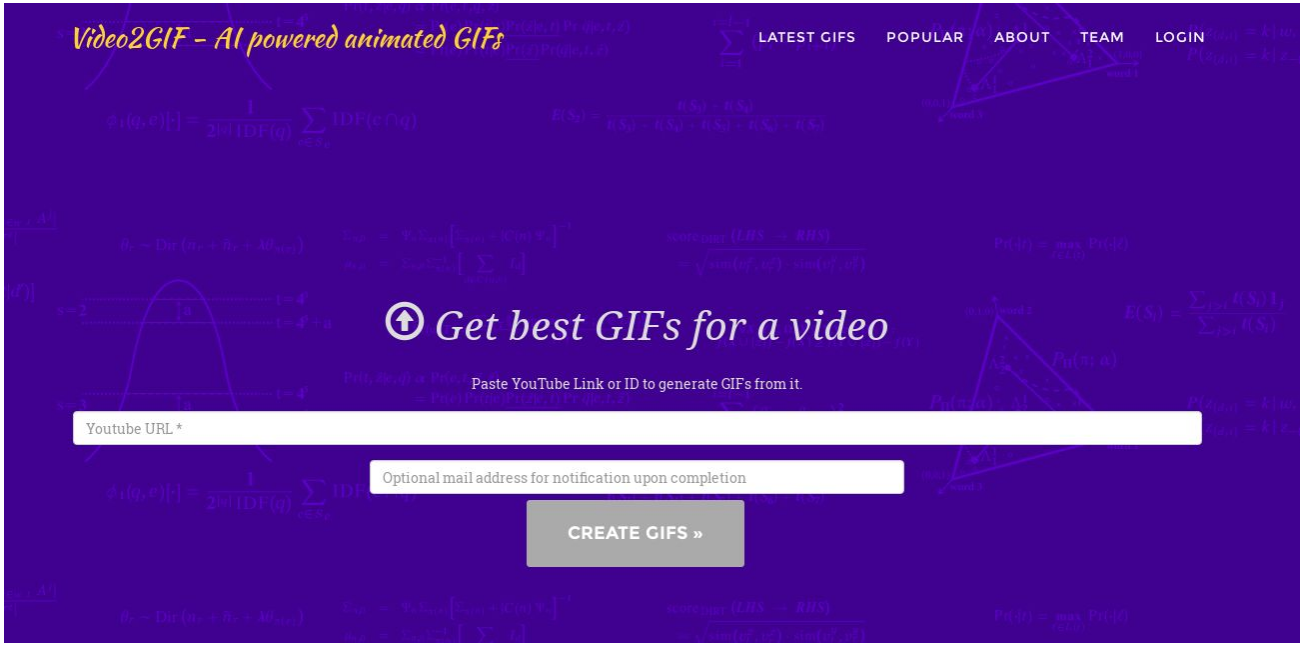
Results

Example result



- Motivation
- Overview of Video Summarization and Related Work
- Finding the Most Interesting and Relevant Content
- Find a Good Subset of Frames / Segments
- **Demo**
- Conclusion and Outlook

Online demo: Video2GIF + diversification



URL: http://video2gif.info/autogif

WITHOUT diversification


Rank: 1 (♥ 0)


Rank: 2 (♥ 1)


Rank: 3 (♥ 0)


Rank: 4 (♥ 0)


Rank: 5 (♥ 0)


Rank: 6 (♥ 0)

WITH diversification


Rank: 1 (♥ 0)


Rank: 2 (♥ 1)


Rank: 3 (♥ 0)


Rank: 4 (♥ 0)


Rank: 5 (♥ 0)


Rank: 6 (♥ 0)

- Motivation
- Overview of Video Summarization and Related Work
- Finding the Most Interesting and Relevant Content
- Find a Good Subset of Frames / Segments
- Demo
- **Conclusion and Outlook**

What does the perfect summary look like?

- Contains the highlights
- Is representative for the video and diverse
- Visually pleasing
- Is coherent / has a sense of story
- Adaptive to user / contexts

What does the perfect summary look like?

Underlined: little explored areas that allow for interesting research

- **Contains the highlights**
  - Two dominant approaches
    - Generic features and large training set e.g. Video2GIF
    - Use of textual information (title, category) to obtain a video specific model e.g. [Liu et al. IJCAI 09], [Liu et al. CVPR 15]
- Is representative for the video and diverse
- Visually pleasing
- Is coherent / has a sense of story
- Adaptive to user / contexts

What does the perfect summary look like?

- Contains the highlights

- **Is representative for the video and diverse**
  - Several subset selection methods
    - Submodular mixtures [Gygli et al. CVPR 15]
    - Video MMR [Li & Merialdo, WIAMIS 10]
    - Determinantal Point Processes [Gong & Grauman NIPS 14]

- Visually pleasing

- Is coherent / has a sense of story

- Adaptive to user / contexts

What does the perfect summary look like?

Underlined: little explored areas that allow for interesting research

- Contains the highlights
- Is representative for the video and diverse
- **Visually pleasing**
  - Improve shot quality, e.g. through path smoothing (Hyperlapse)
  - Select good parts (highlights)
  - Choose sensible shot boundaries
    - Typically done as preprocessing using hand-defined methods, e.g. [Potapov et al. ECCV 14]
- Is coherent / has a sense of story
- Adaptive to user / contexts

What does the perfect summary look like?

Underlined: little explored areas that allow for interesting research

- Contains the highlights
- Is representative for the video and diverse
- Visually pleasing
- **Is coherent / has a sense of story**
  - Typically not considered / researched
    - Notable exception “Story-driven summarization for egocentric video” [Lu & Grauman CVPR 13]
  - Difficult to optimize for / evaluate
- Adaptive to user / contexts

What does the perfect summary look like?

Underlined: little explored areas that allow for interesting research

- Contains the highlights
- Is representative for the video and diverse
- Visually pleasing
- Is coherent / has a sense of story

- **Adaptive to user / context**
  - Learn user-specific models
  - Use, e.g. a search query as context [Liu et al. CVPR 15]

Thanks to my collaborators!

Helmut Grabner

Hayko Riemenschneider

Yale Song

Prof. Luc Van Gool

ETH zürich

logitech

YAHOO!

Thank you for your attention.

Questions? Suggestions?

Backup slides

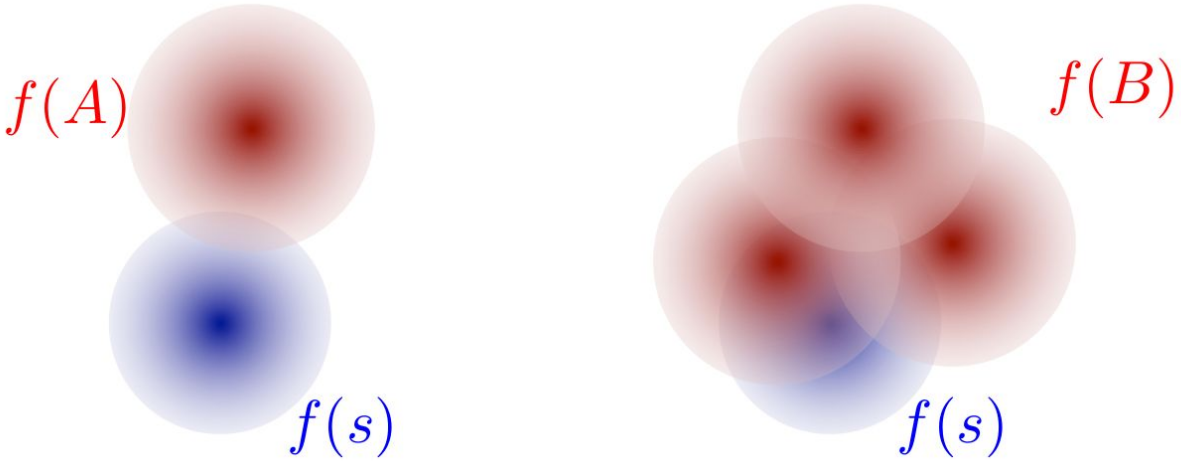
Video2GIF on highlight dataset

Category	Ours	rankSVM	Yang [40]	Sun [35]
skating	55.4%	26.2%	25%	61%
gymnastics	33.5%	25.5%	35%	41%
surfing	54.1%	45.0%	49%	61%
dog	30.8%	47.3%	49%	60%
parkour	54.0%	44.7%	50%	61%
skiing	32.8%	35.6%	22%	36%
Total	46.4%	37.9%	41.2%	53.6%

Submodularity – "Diminishing returns"

A set function f is called submodular, if for $A \subseteq B$

$$f(A \cup s) - f(A) \geq f(B \cup s) - f(B)$$



→ Very natural property for summarization problems