Category-specific video summarization

Speaker:
Danila Potapov

Joint work with:
Matthijs Douze Zaid Harchaoui Cordelia Schmid

LEAR team, Inria Grenoble Rhône-Alpes

Christmas Colloquium on Computer Vision
Moscow, 28.12.2015

Introduction

- size of video data is growing
  - 300 hours of video uploaded on YouTube every minute
- types of video data: user-generated, sports, news, movies

User-generated Sports

News Movies

- common need for structuring video data

Video summarization

Detecting the most important part in a “Landing a fish” video

Goals

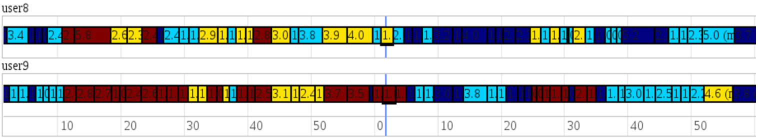
- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Quantitative evaluation of video analysis algorithms

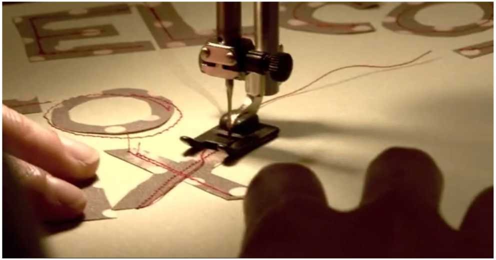
Goals

- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Quantitative evaluation of video analysis algorithms

Goals

- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Quantitative evaluation of video analysis algorithms





Contributions

- supervised approach to video summarization
- temporal localization at test time
- MED-Summaries dataset for evaluation of video summarization

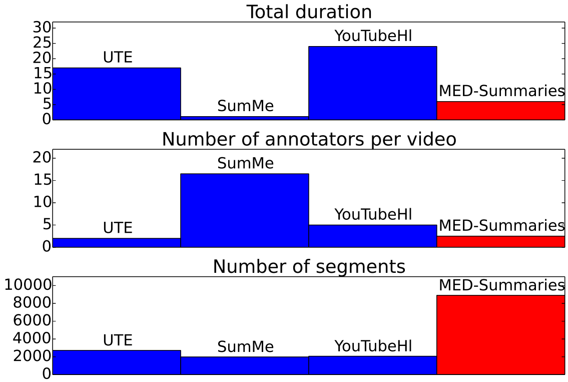
Publication

- D. Potapov, M. Douze, Z. Harchaoui, C. Schmid
  "Category-specific video summarization", ECCV 2014
- **MED-Summaries** dataset online

  http://lear.inrialpes.fr/people/potapov/med_summaries

MED-Summaries dataset

- evaluation benchmark for video summarization
- subset of TRECVID Multimedia Event Detection 2011 dataset
- 10 categories



Definition

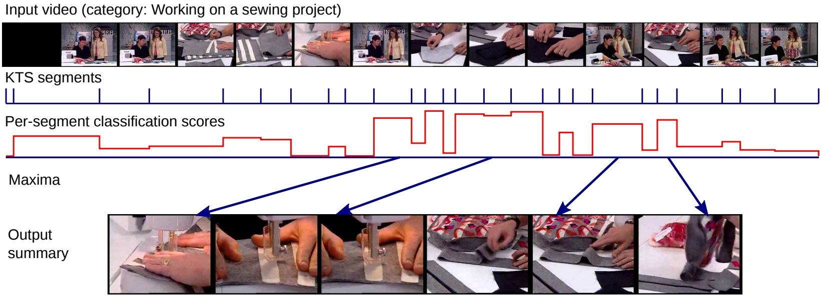
A *video summary*

- built from subset of temporal segments of original video
- conveys the most important details of the video

Original video, and its video summary for the category “Birthday party”

Overview of our approach

- produce *visually coherent* temporal segments
  - no shot boundaries, camera shake, etc. inside segments
- identify important parts
  - *category-specific importance*: a measure of relevance to the type of event

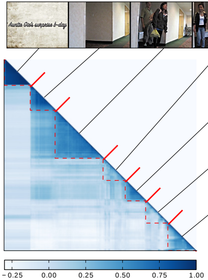


Related works

- specialized domains
  - Lu and Grauman [2013], Lee et al. [2012]: summarization of egocentric videos
  - Khosla et al. [2013]: keyframe summaries, canonical views for cars and trucks from web images
- Sun et al. [2014] “Ranking Domain-specific Highlights by Analyzing Edited Videos”
  - automatic approach for harvesting data
  - highlight detection vs. temporally coherent summarization
- Gygli et al. [2014] “Creating Summaries from User Videos”
  - cinematic rules for segmentation
  - small set of informative descriptors

Kernel temporal segmentation

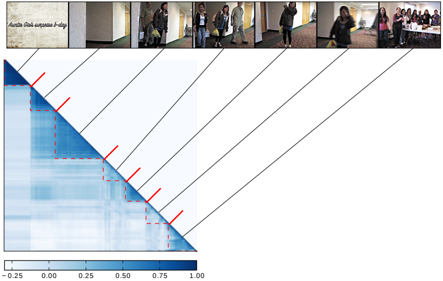
- goals: group similar frames such that semantic changes occur at the boundaries
- kernelized Multiple Change-Point Detection algorithm
  - change-points divide the video into temporal segments
- input: robust frame descriptor (SIFT + Fisher Vector)



Kernel matrix and temporal segmentation of a video

Kernel temporal segmentation

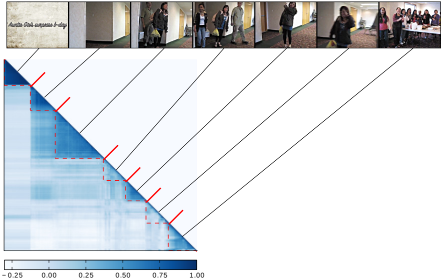
- goals: group similar frames such that semantic changes occur at the boundaries
- kernelized Multiple Change-Point Detection algorithm
  - change-points divide the video into temporal segments
- input: robust frame descriptor (SIFT + Fisher Vector)



Kernel matrix and temporal segmentation of a video

Kernel temporal segmentation

- goals: group similar frames such that semantic changes occur at the boundaries
- kernelized Multiple Change-Point Detection algorithm
  - change-points divide the video into temporal segments
- input: robust frame descriptor (SIFT + Fisher Vector)



Kernel matrix and temporal segmentation of a video

Kernel temporal segmentation algorithm

Input: temporal sequence of descriptors $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$

1. Compute the Gram matrix A : $\quad a_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$
2. Compute cumulative sums of A
3. Compute unnormalized variances

$$v_{t,t+d} = \sum_{i=t}^{t+d-1} a_{i,i} - \frac{1}{d} \sum_{i,j=t}^{t+d-1} a_{i,j}$$

$$t = 0, \ldots, n-1, \quad d = 1, \ldots, n-t$$

4. Do the forward pass of dynamic programming

$$L_{i,j} = \min_{t=i,\ldots,j-1} \left(L_{i-1,t} + v_{t,j} \right), \quad L_{0,j} = v_{0,j}$$

$$i = 1, \ldots, m_{\max}, \quad j = 1, \ldots, n$$

5. Select the optimal number of change points

$$m^\star = \arg\min_{m=0,\ldots,m_{\max}} L_{m,n} + C\, m \left(\log\left(n/m \right) + 1 \right)$$

6. Find change-point positions by backtracking

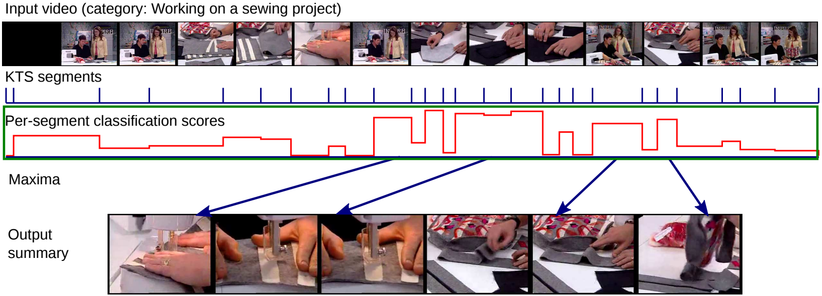
$$t_{m^\star} = n, \quad t_{i-1} = \arg\min_t \left(L_{i-1,t} + v_{t,t_i} \right)$$

$$i = m^\star, \ldots, 1$$

Output: Change-point positions $t_0, \ldots, t_{m^\star - 1}$

Supervised summarization

- **Training:** train a linear SVM from a set of videos with just video-level class labels
- **Testing:** score segment descriptors with the classifiers trained on full videos; build a summary by concatenating the most important segments of the video

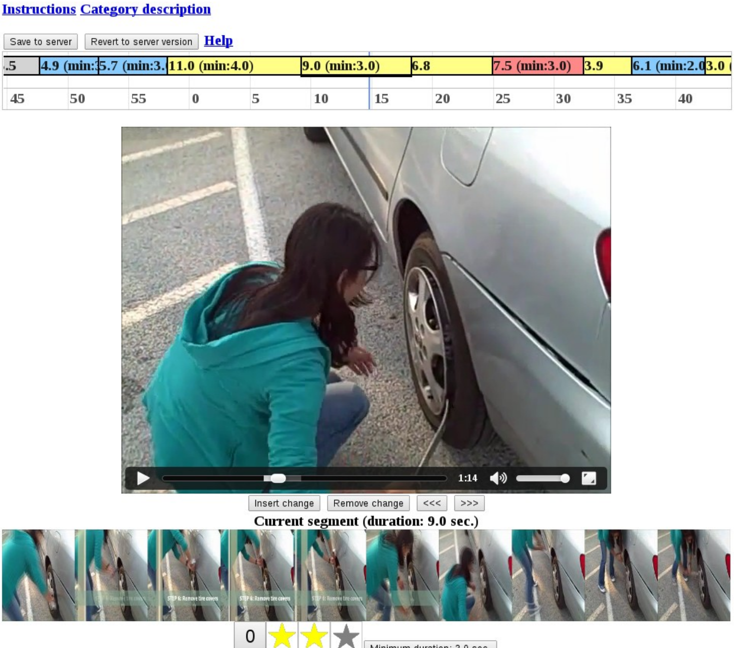


MED-Summaries dataset

- 100 test videos (= 4 hours) from TRECVID MED 2011
- multiple annotators
- 2 annotation tasks:
  - segment boundaries (median duration: 3.5 sec.)
  - segment importance (grades from 0 to 3)
    - 0 = not relevant to the category
    - 3 = highest relevance

Central frame for each segment with importance annotation for category “Changing a vehicle tyre”.

Annotation interface

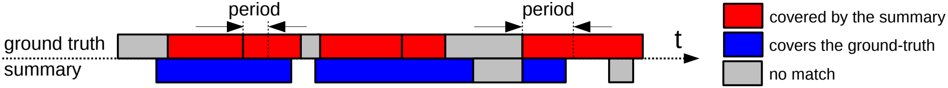


Dataset statistics

	Training	Validation	Test
MED dataset			
Total videos	10938	1311	31820
Total duration, hours	468	57	980
MED-Summaries			
Annotated videos	—	**60**	**100**
Total duration, hours	—	3	4
Annotators per video	—	1	2-4
Total annotated segments (units)	—	**1680**	**8904**

Evaluation metrics for summarization (1)

- often based on user studies
  - time-consuming, costly and hard to reproduce
- **Our approach:** rely on the annotation of test videos
- ground truth segments $\{S_i\}_{i=1}^{m}$
- computed summary $\{\widetilde{\mathbf{S}}_j\}_{j=1}^{\tilde{m}}$
- coverage criterion: $\text{duration}\big(S_i \cap \widetilde{\mathbf{S}}_j\big) > \alpha P_i$



- *importance ratio* for summary $\widetilde{\mathbf{S}}$ of duration $\mathbf{T}$

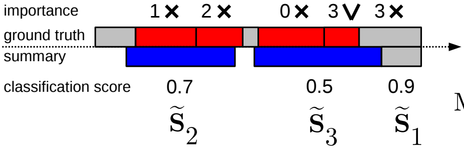
$$\mathcal{I}^*(\widetilde{\mathbf{S}}) = \frac{\mathcal{I}(\widetilde{\mathbf{S}})}{\mathcal{I}^{\max}(\mathbf{T})}$$

total importance
covered by the summary

max. possible total importance
for a summary of duration **T**

Evaluation metrics for summarization (2)

- a *meaningful summary* covers a ground-truth segment of importance 3



3 segments are required to see an importance-3 segment

$$\mathrm{MSD}(\widetilde{\mathbf{S}}) = \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{3}} \mathrm{duration}(\widetilde{\mathbf{S}}_{\mathbf{j}})$$

Meaningful summary duration (MSD): minimum length for a meaningful summary

Evaluation metric for temporal segmentation

- segmentation *f-score*: match when overlap/union $> \beta$

Experiments

Baselines

- **Users**: keep 1 user in turn as a ground truth for evaluation of the others
- **SD** + **SVM**: shot detector Massoudi et al. [2006] for segmentation + SVM-based importance scoring
- **KTS** + **Cluster**: Kernel Temporal Segmentation + k-means clustering for summarization
  - sort segments by increasing distance to centroid

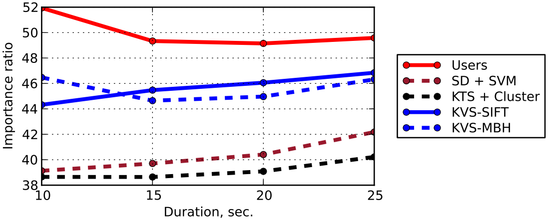
Our approach

Kernel Video Summarization =
Kernel Temporal Segmentation + **SVM-based importance scoring**

Results

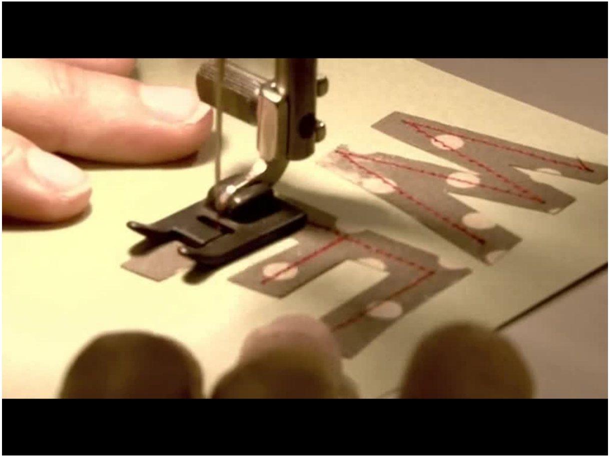
Method	Segmentation Avg. f-score higher better	Summarization Med. MSD (s) lower better
Users	49.1	10.6
SD + SVM	30.9	16.7
KTS + Cluster	**41.0**	13.8
KVS	**41.0**	**12.5**

Segmentation and summarization performance



Importance ratio for different summary durations

Example summaries



Conclusion

- KVS delivers short and highly-informative summaries, with the most important segments for a given category
- temporal segmentation algorithm produces visually coherent segments
- KVS is trained in a weakly-supervised way
  - does not require segment annotations in the training set
- MED-Summaries — dataset for evaluation of video summarization
  - annotations and evaluation code available online

Publication

- D. Potapov, M. Douze, Z. Harchaoui, C. Schmid "Category-specific video summarization", ECCV 2014
- **MED-Summaries** dataset online

  http://lear.inrialpes.fr/people/potapov/med_summaries

Thank you for your attention!