Supervised Learning Approaches for Automatic Structuring of Videos

Danila Potapov

Supervisors:
Cordelia Schmid
Zaid Harchaoui
Matthijs Douze

Jury:
Patrick Pérez
Ivan Laptev
Florent Perronnin

LEAR team, Inria Grenoble Rhône-Alpes, 2015

Outline

Part I

Introduction

- size of video data is growing
  - 300 hours of video uploaded on YouTube every minute (June 2015)
- types of video data: user-generated, sports, news, movies

User-generated

Sports

News

Movies

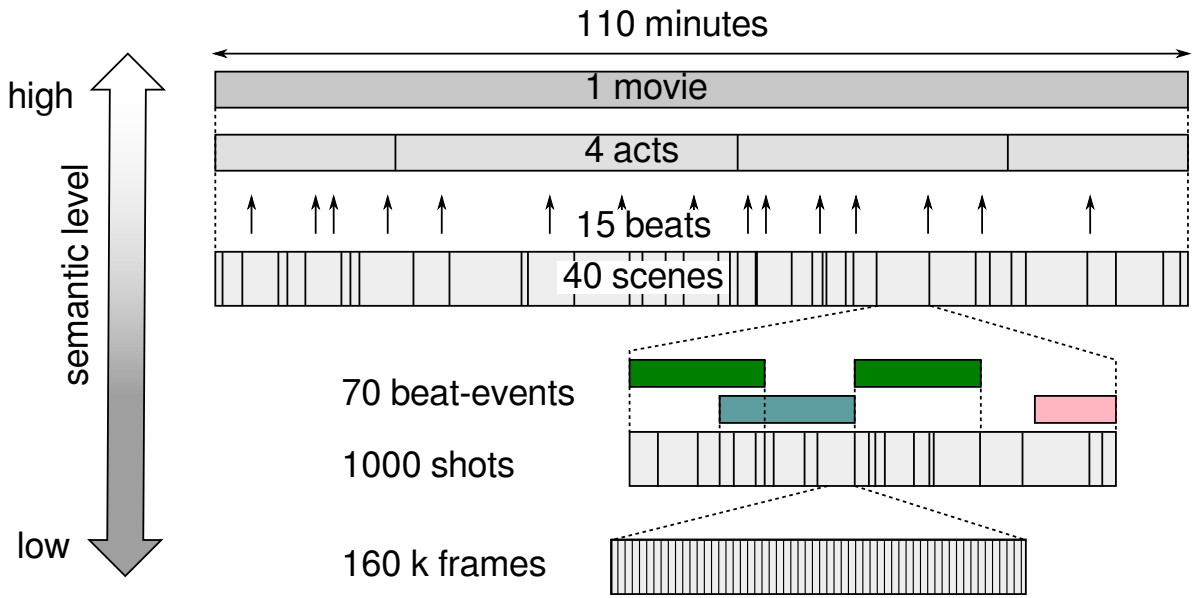
- common need for structuring video data

Video summarization

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

Temporal structuring of movies

- *movie* — a special type of video, with a complete storyline
- a typical Hollywood movie follows a set of scriptwriting rules



B. Snyder “Save the cat!: The Last Book on Screenwriting You’ll Ever Need”


pursuit
romance
victory bad
despair good
good argue bad
bad argue bad
battle
victory good
preparation
joy bad
good argue good
NULL
good argue good
ground truth
SVM output
CRF output

Current challenges

- complexity of the temporal structure
  - long events (battle preparation, birthday party, etc.)
    vs.
    short events (open door, victory good, joy bad, etc.)

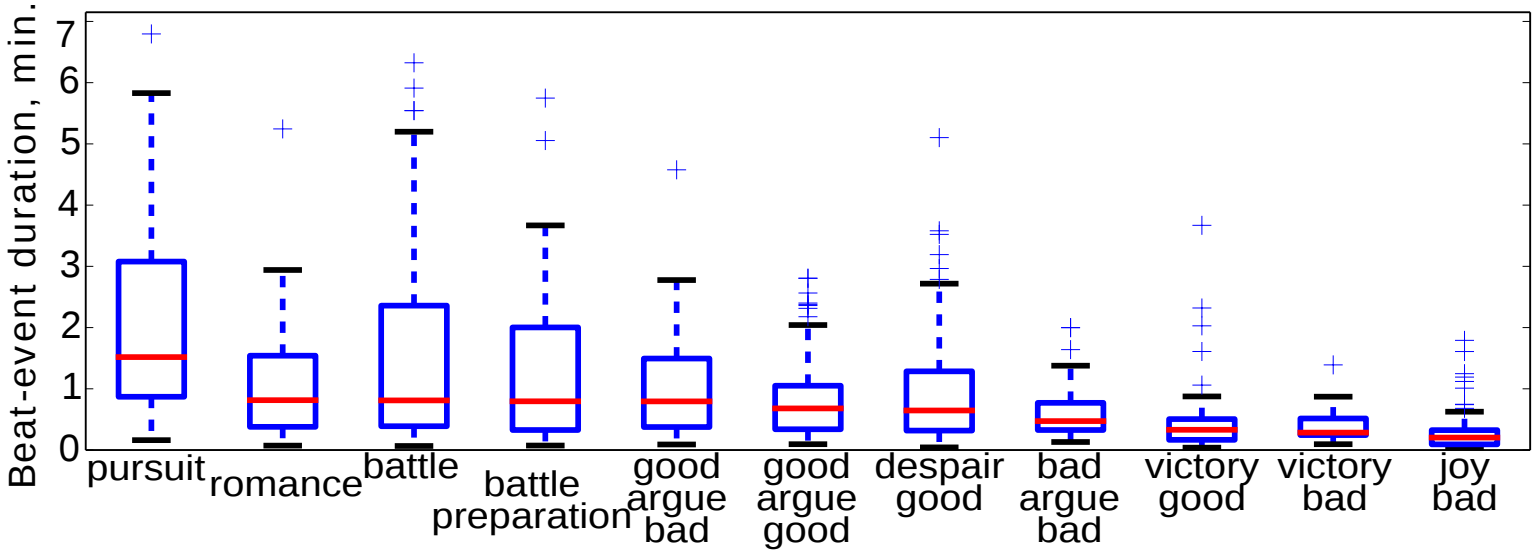
- high intra-class variation
  - highly-semantic events (romance, battle preparation, victory good, etc.)
    vs.
    low-semantic events (open door, stand up, etc.)



Current challenges

- complexity of the temporal structure
  - long events (battle preparation, birthday party, etc.)
    vs.
    short events (open door, victory good, joy bad, etc.)

- high intra-class variation
  - highly-semantic events (romance, battle preparation, victory good, etc.)
    vs.
    low-semantic events (open door, stand up, etc.)



Current challenges

- complexity of the temporal structure
  - long events (battle preparation, birthday party, etc.)
    vs.
    short events (open door, victory good, joy bad, etc.)

- high intra-class variation
  - highly-semantic events (romance, battle preparation, victory good, etc.)
    vs.
    low-semantic events (open door, stand up, etc.)



Goals

- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Adapt the classifiers to a specific dataset
- Quantitative evaluation of video analysis algorithms

Goals

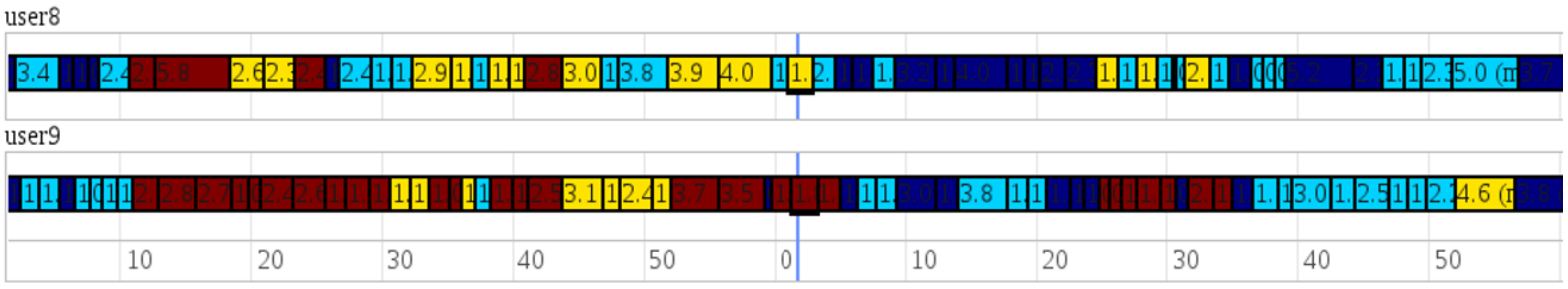
- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Adapt the classifiers to a specific dataset
- Quantitative evaluation of video analysis algorithms

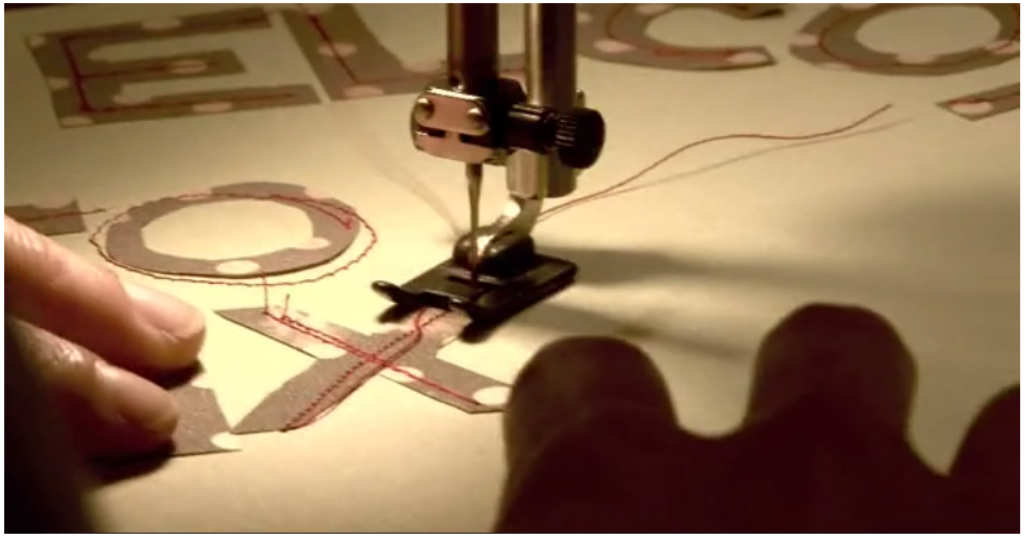
Goals

- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Adapt the classifiers to a specific dataset
- Quantitative evaluation of video analysis algorithms

Goals

- Recognize events accurately and efficiently
- Identify the most important moments in videos
- Adapt the classifiers to a specific dataset
- Quantitative evaluation of video analysis algorithms





Contributions: video summarization

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

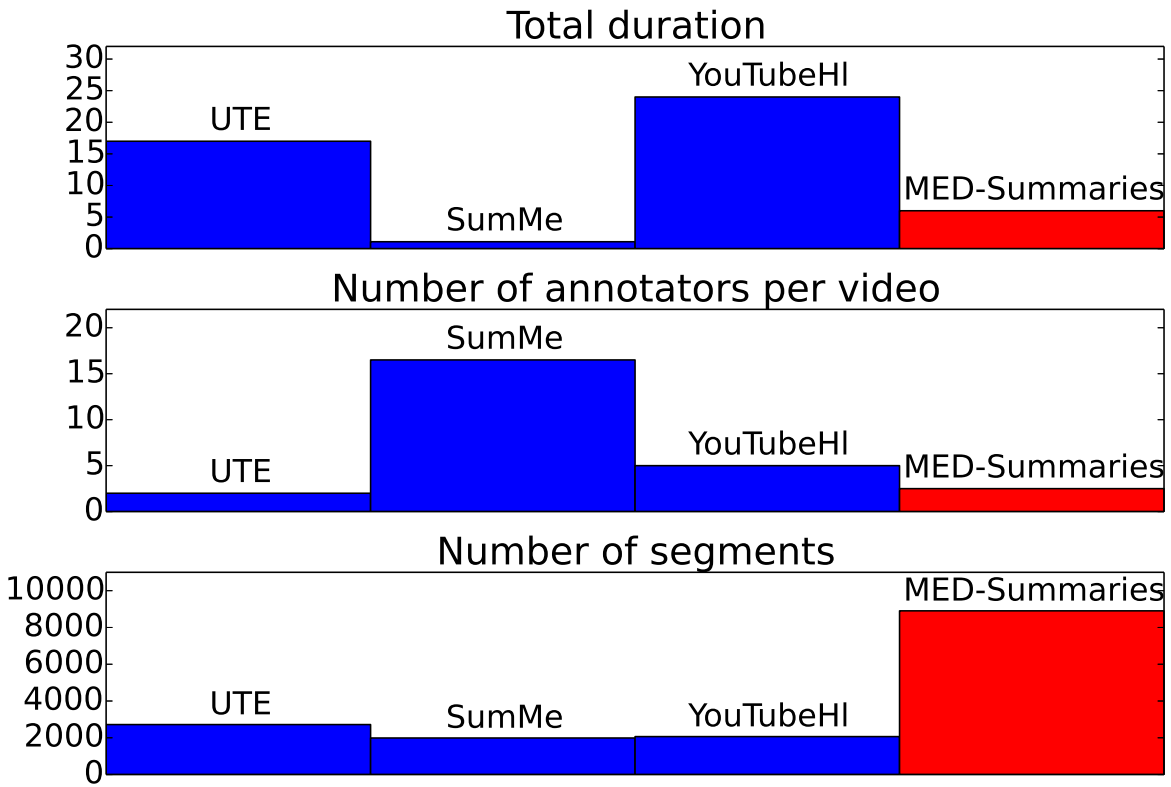
Publication

- 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



Contributions: temporal structuring of movies

- dataset of 20 movies with shot- and event-level annotations
- 11 beat-categories + NULL class, complete annotation of movies
- localization of low-salient beat-events
- temporal structure of shots
- movie-specific information

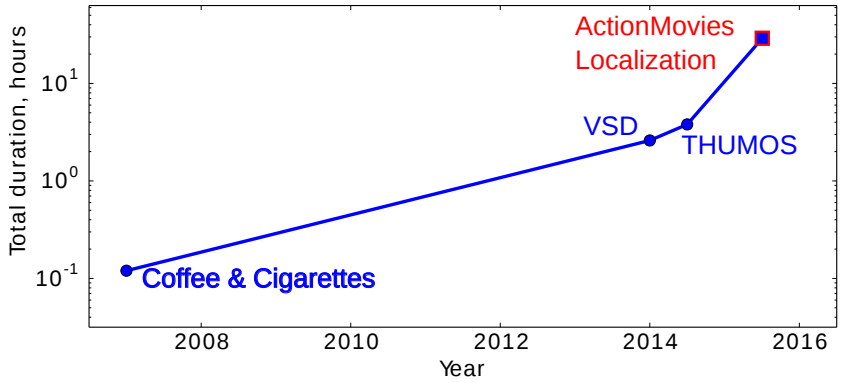
Publication

- D. Potapov, M. Douze, J. Revaud, Z. Harchaoui, C. Schmid "Beat-Event Detection in Action Movie Franchises", arXiv, 2015

- **Action Movie Franchises** dataset online

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

Comparison to existing video localization datasets



Coffee&Cigarettes Laptev and Pérez [2007],

VSD = MediaEval Violent Scene Detection Demarty et al. [2014],

THUMOS Challenge 2014

http://crcv.ucf.edu/THUMOS14/

Outline

Definition

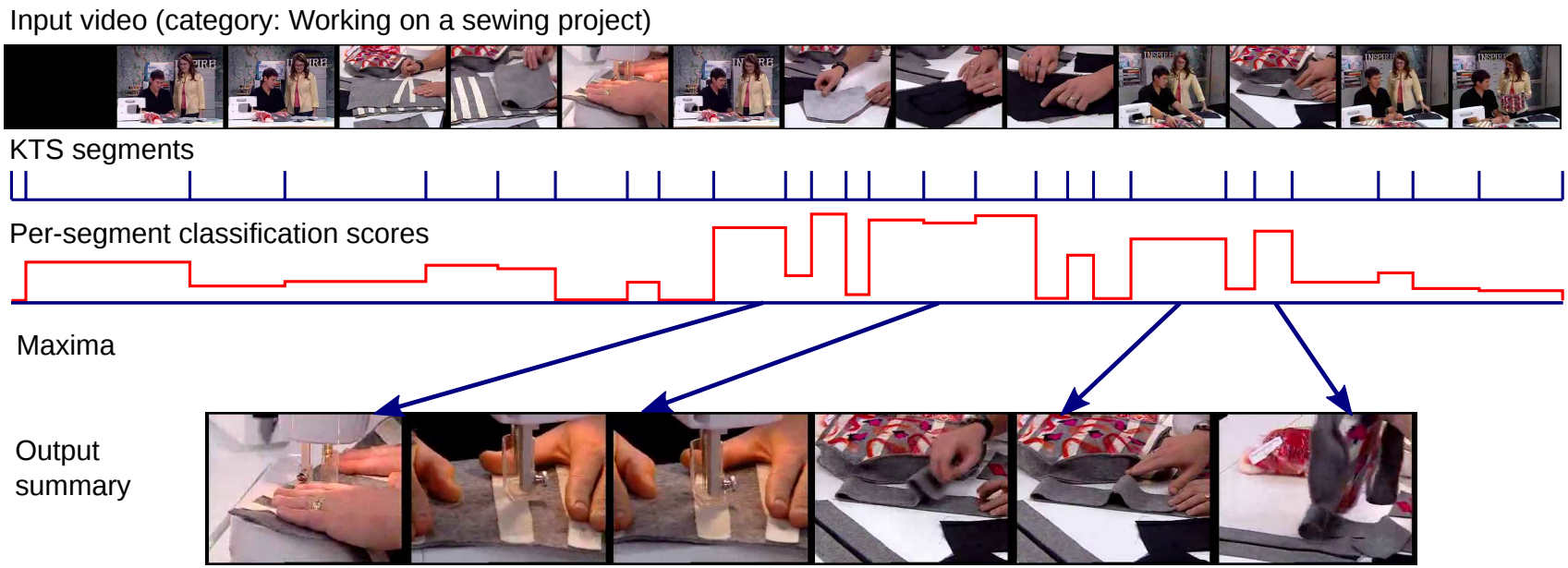
A *video summary*

- 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



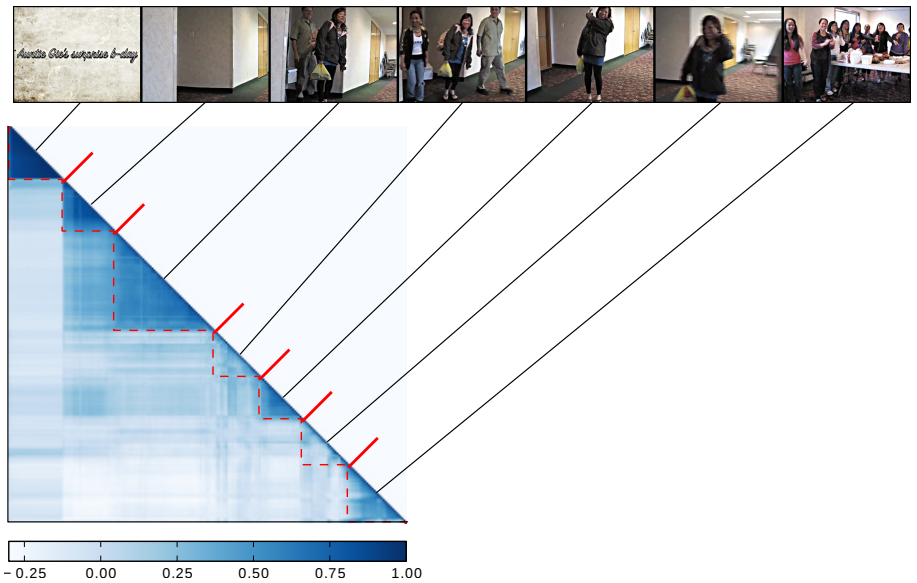
Related works

- 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

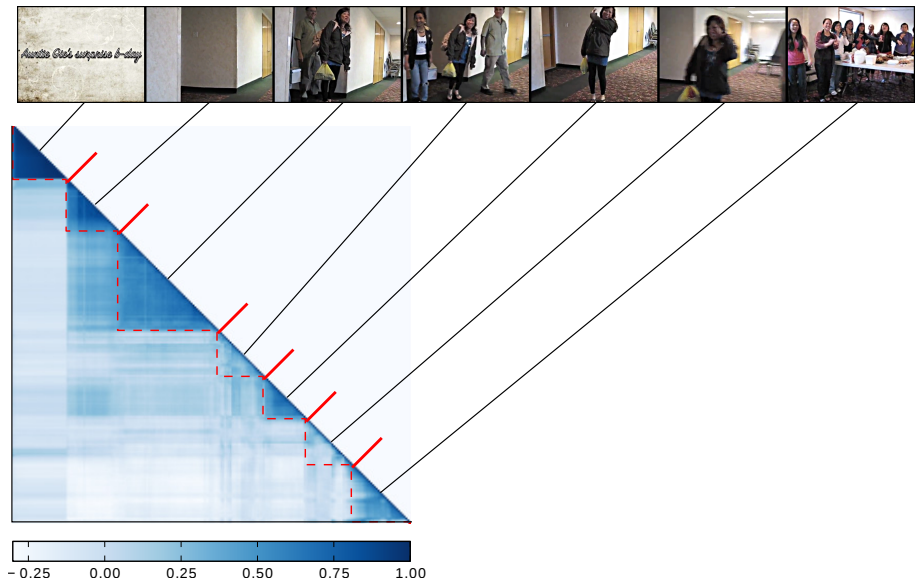
- goals: group similar frames such that semantic changes occur at the boundaries
- kernelized Multiple Change-Point Detection algorithm
  - change-points divide the video into temporal segments
- input: robust frame descriptor (SIFT + Fisher Vector)



Kernel matrix and temporal segmentation of a video

Kernel temporal segmentation

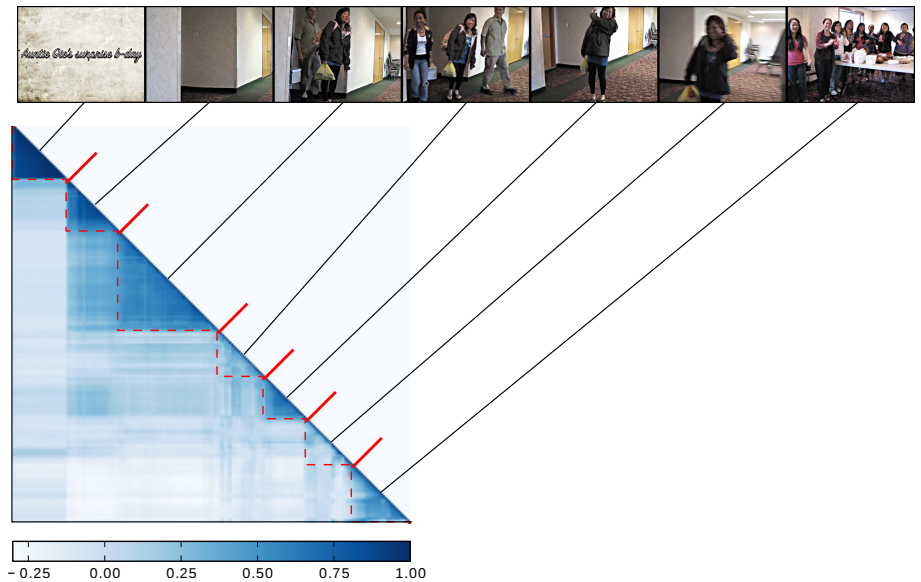
- goals: group similar frames such that semantic changes occur at the boundaries
- kernelized Multiple Change-Point Detection algorithm
  - change-points divide the video into temporal segments
- input: robust frame descriptor (SIFT + Fisher Vector)



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 : $a_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$
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

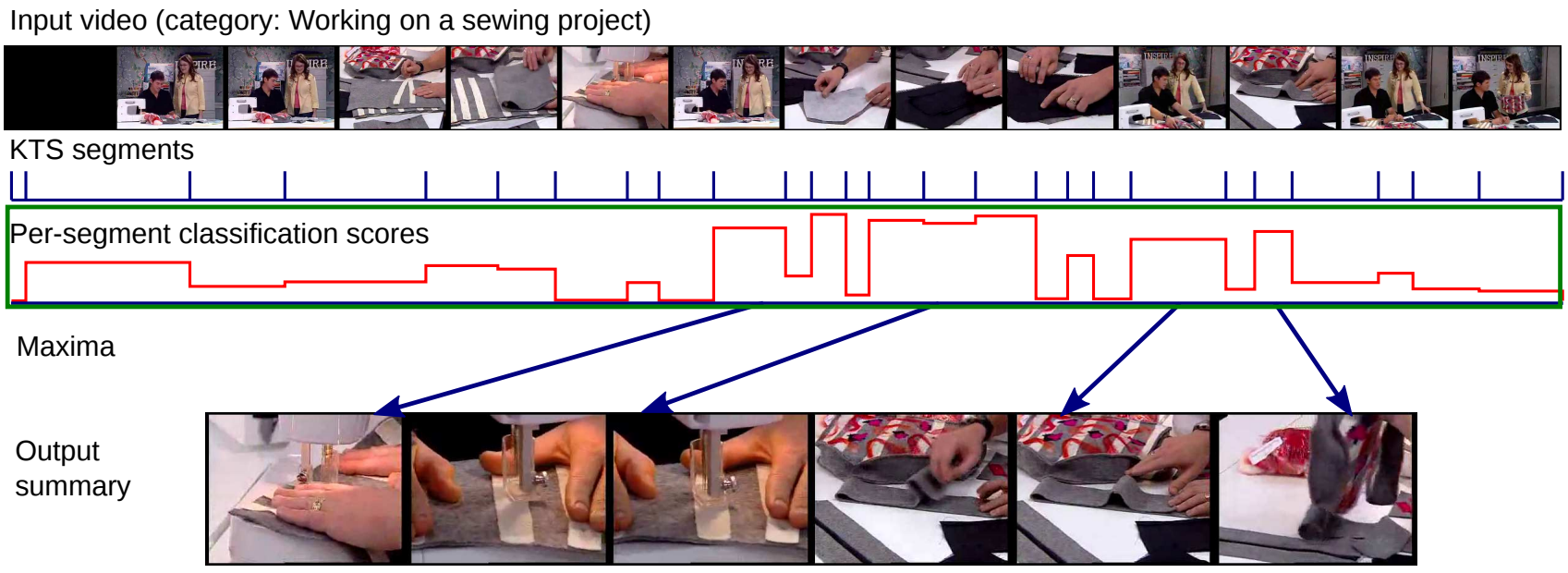
$$t_{m^\star} = n, \quad t_{i-1} = \arg\min_t \left(L_{i-1,t} + v_{t,t_i} \right)$$

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

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

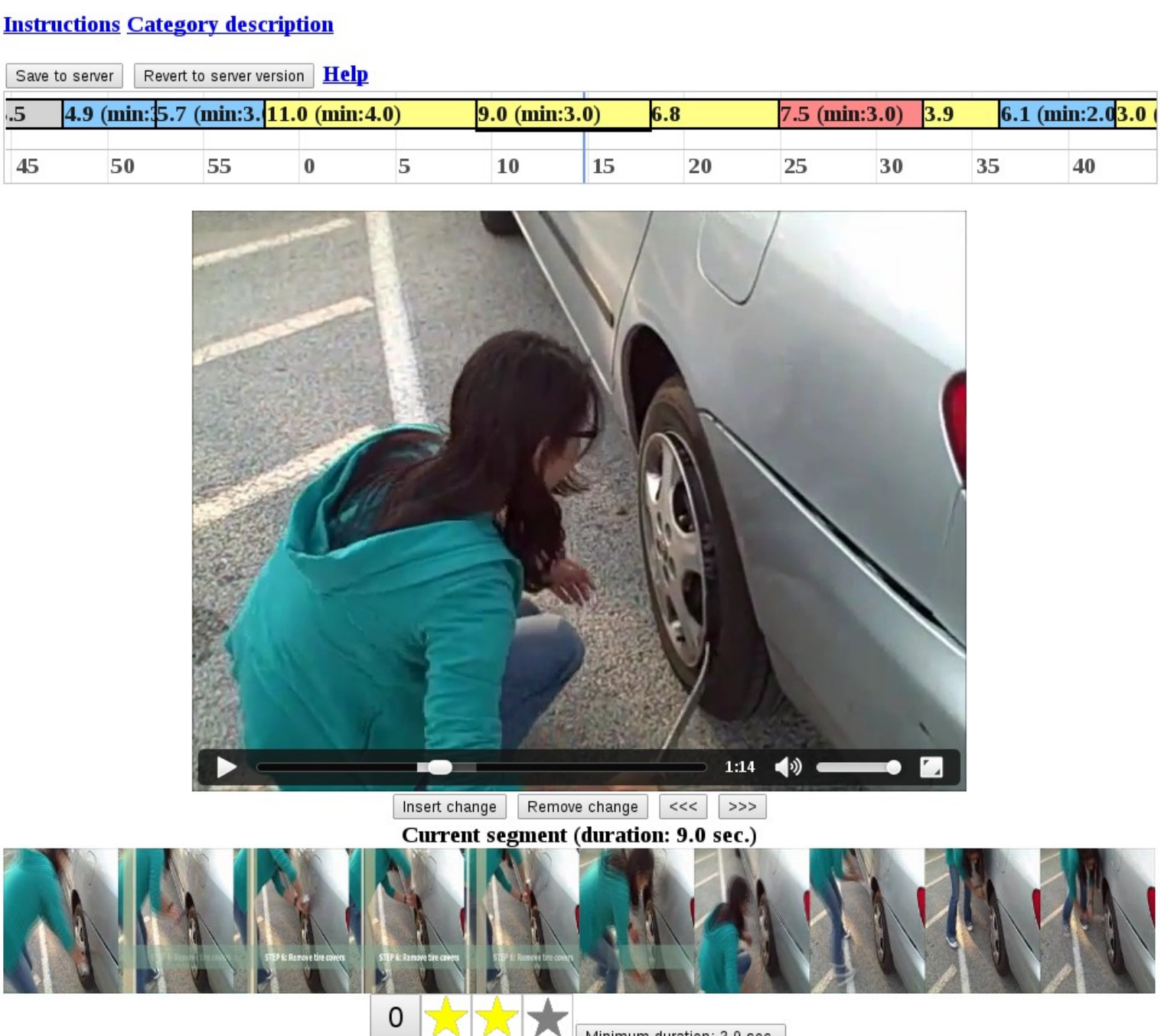


MED-Summaries dataset

- 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

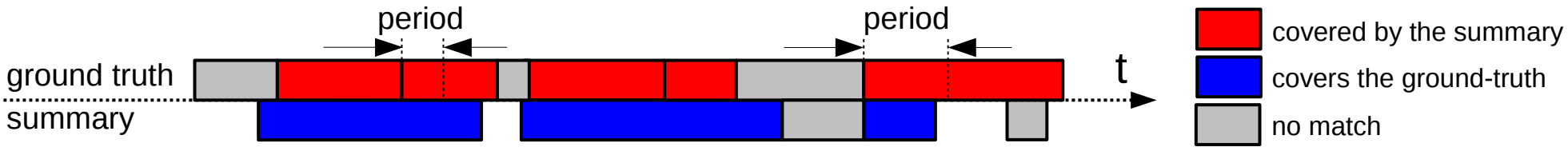


Dataset statistics

	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: $\text{duration}\left(S_i \cap \widetilde{\mathbf{S}}_j\right) > \alpha P_i$



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

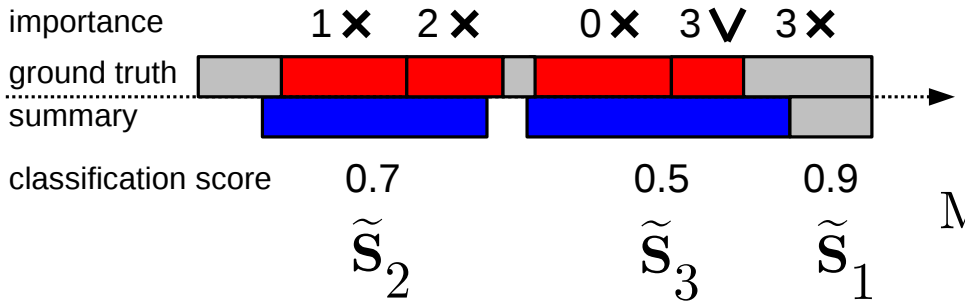
$$\mathcal{I}^*(\widetilde{\mathbf{S}}) = \frac{\mathcal{I}(\widetilde{\mathbf{S}})}{\mathcal{I}^{\max}(\mathbf{T})}$$

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

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

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

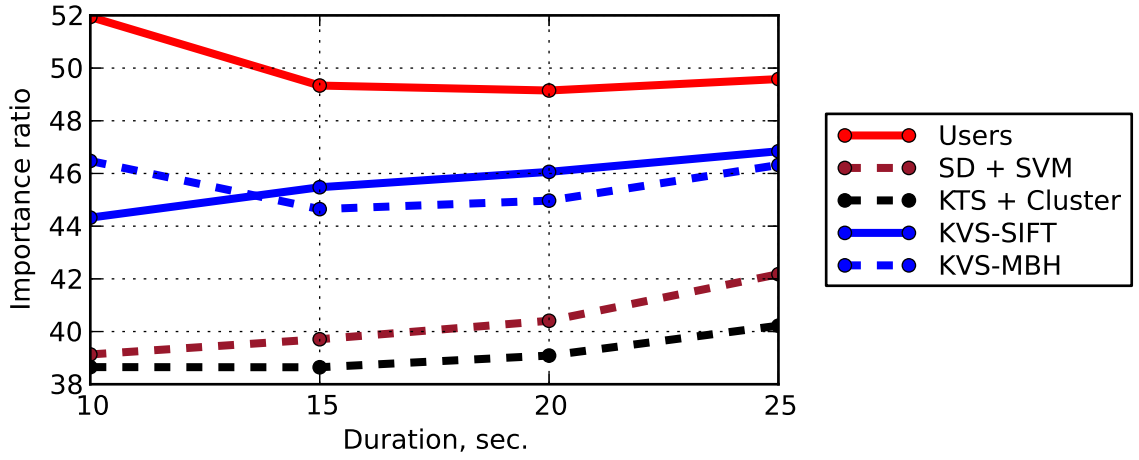
Our approach

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

Results

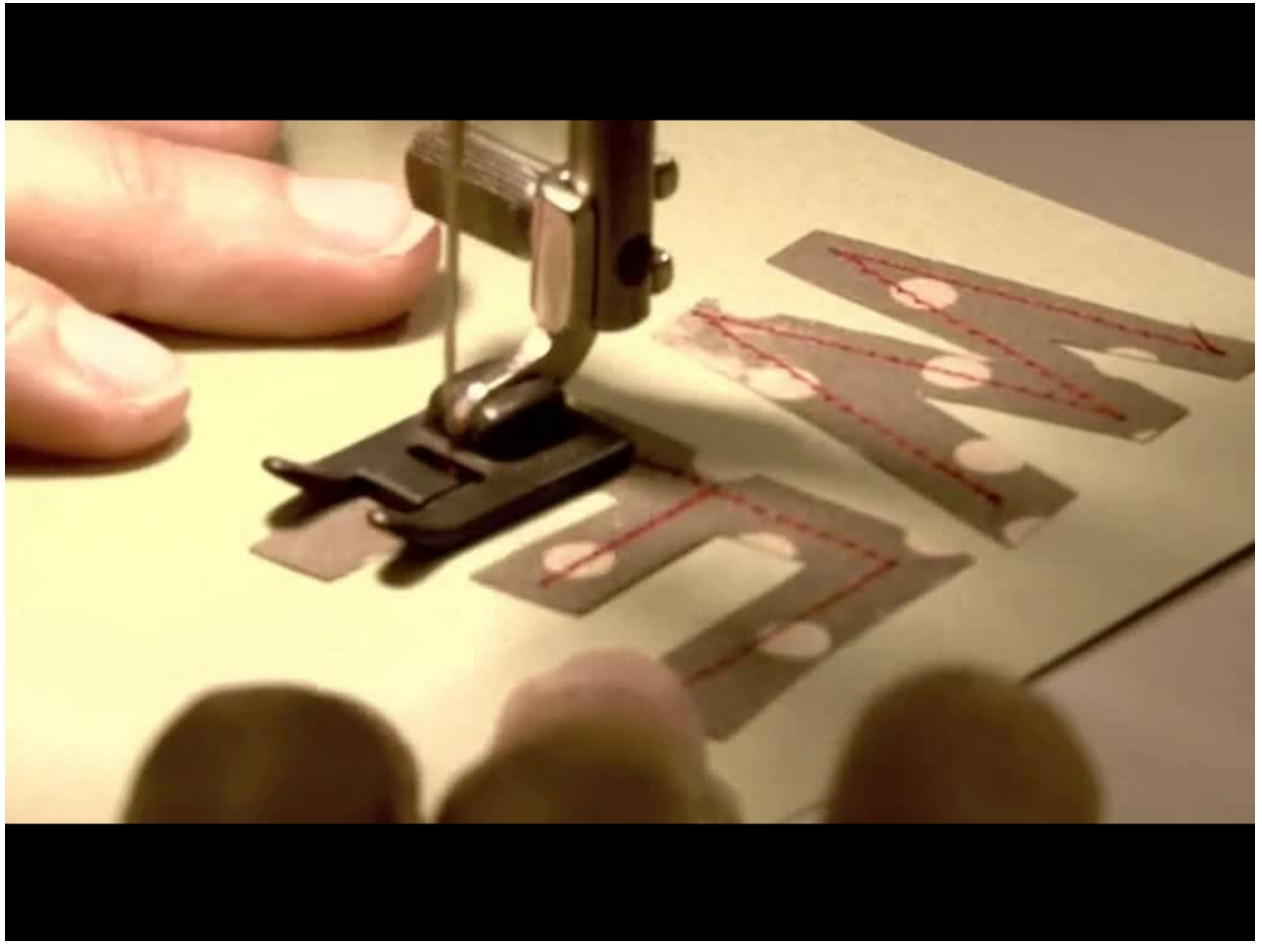
Method	Segmentation Avg. f-score higher better	Summarization Med. MSD (s) lower better
Users	49.1	10.6
SD + SVM	30.9	16.7
KTS + Cluster	**41.0**	13.8
KVS	**41.0**	**12.5**

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

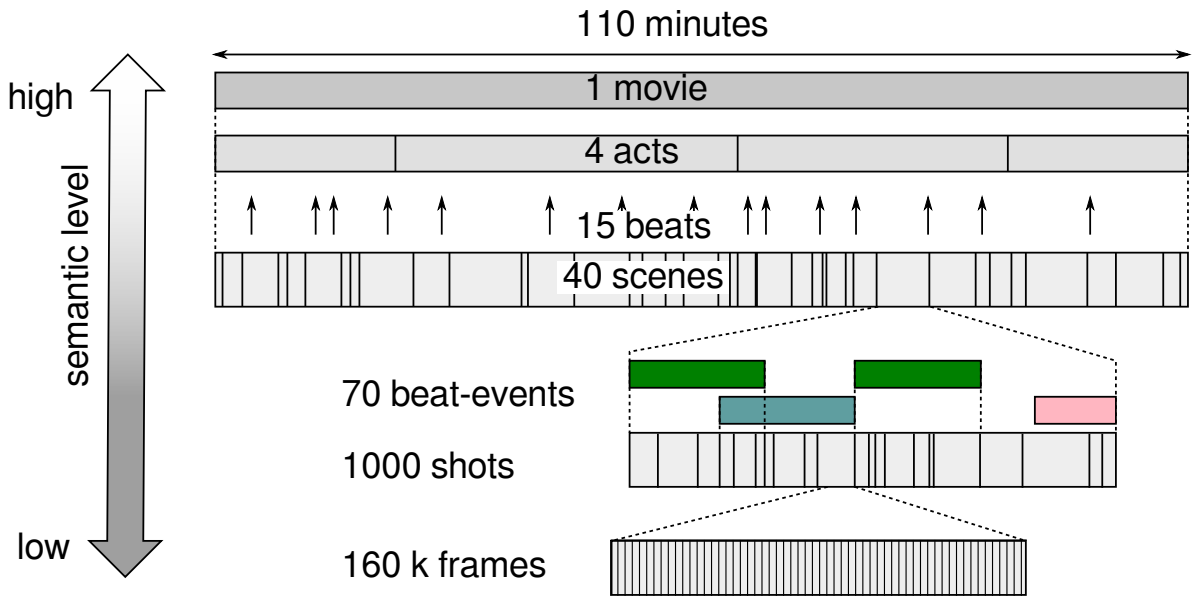
- D. Potapov, M. Douze, Z. Harchaoui, C. Schmid "Category-specific video summarization", ECCV 2014
- **MED-Summaries** dataset online

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

Outline

Temporal structure of the movie

- *movie* — a special type of video, with a complete storyline
- a typical Hollywood movie follows a set of scriptwriting rules[1]



[1] B. Snyder “Save the cat!: The Last Book on Screenwriting You’ll Ever Need”

Beat-event definition

- *beat-event* — a sequence of consecutive shots, tied by a single “action” in the movie
  - typically lasts a few minutes
- semantically higher level than the actions in most current benchmarks, but lower than the beats

Examples:

actions	beat-events	beats
fighting	battle	“all is lost”
hug person	romance	B-story
open door	good-argue-bad	debate
eating	despair-good	“dark night of the soul”

Action Movie Franchises dataset

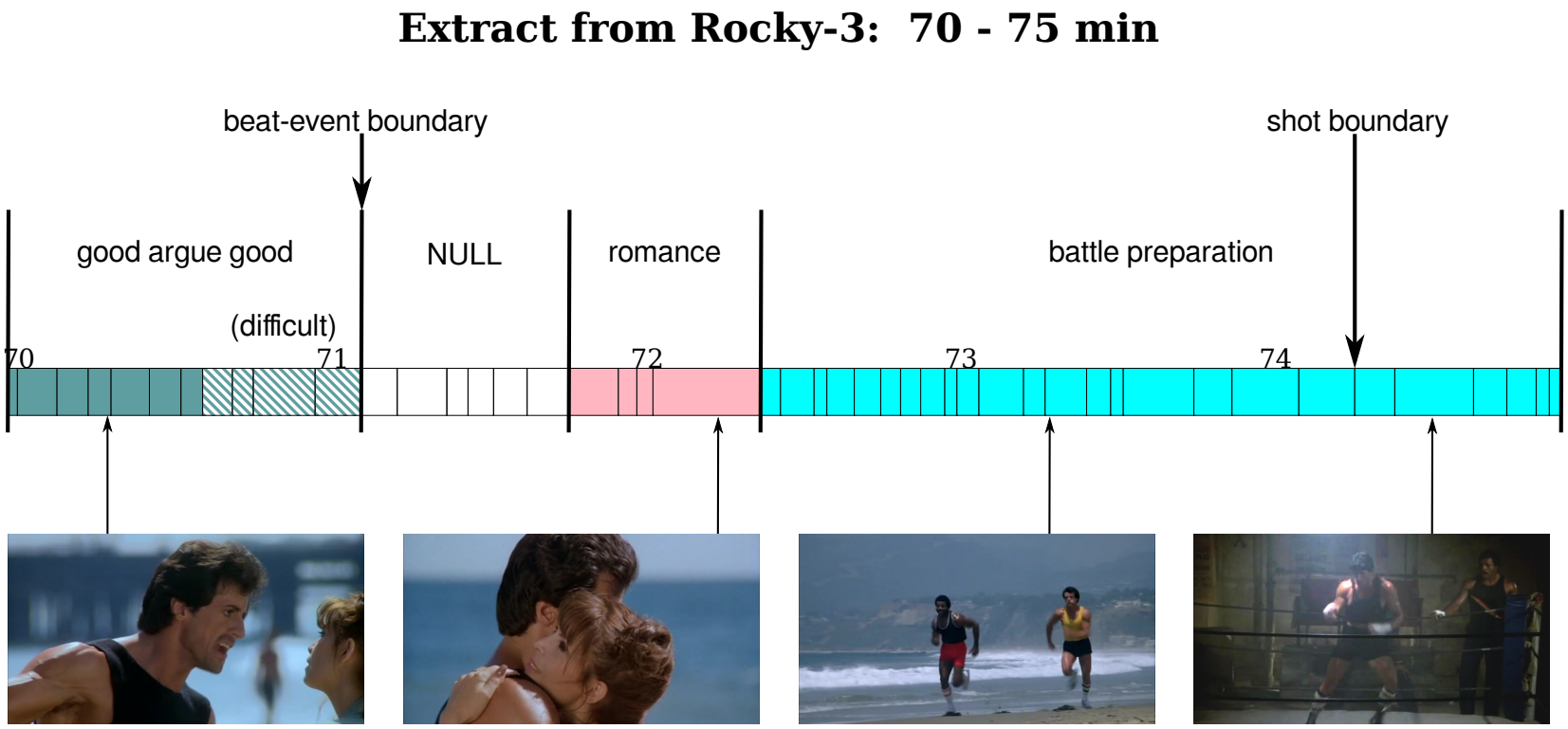
- 20 action movies of 5 franchises (22811 shots, 36.5 hours)
  - *franchise* — series of movies with similar characters and topic
- dense annotations of 11 beat-categories
- both shot and event levels
- franchises minimize intra-class variance

Franchise information example

Battle in Indiana Jones

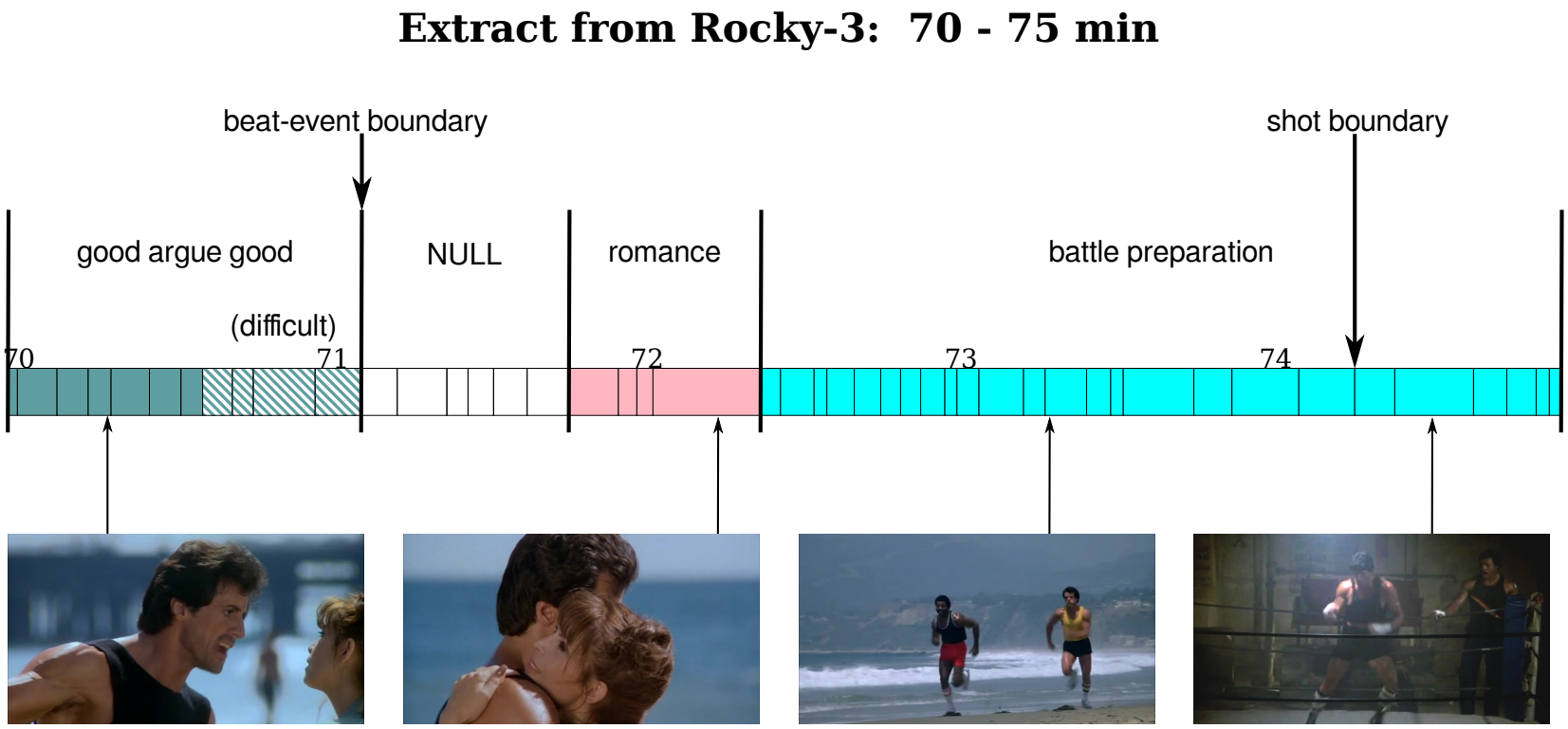
Annotation example

- shots detected automatically
- minimal annotation unit = shot
- “difficult” tag for ambiguous cases



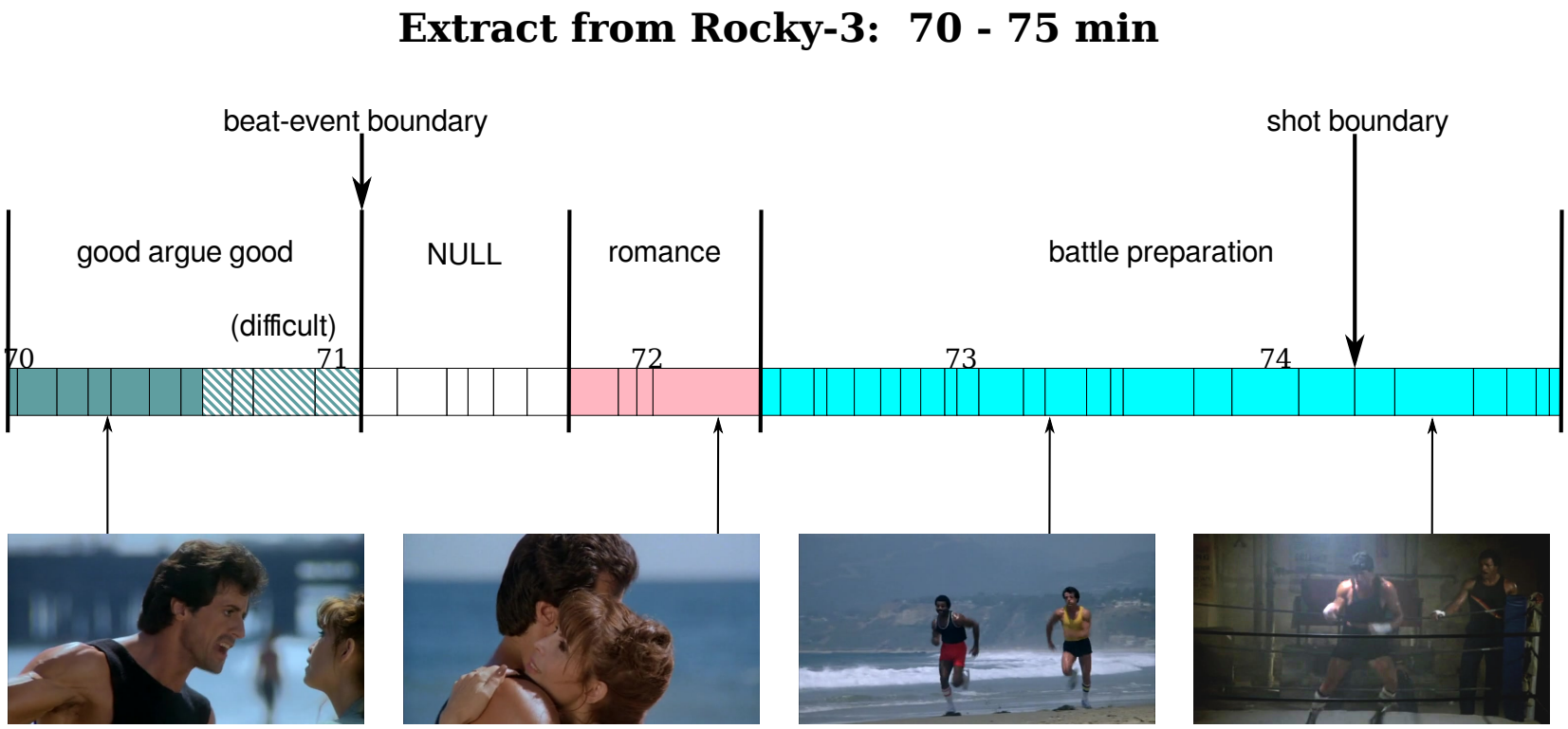
Annotation example

- shots detected automatically
- minimal annotation unit = shot
- “difficult” tag for ambiguous cases



Annotation example

- shots detected automatically
- minimal annotation unit = shot
- “difficult” tag for ambiguous cases



Contributions

- **Action Movie Franchises** dataset
  - multiple evaluation protocols (same/different franchise, classification/localization)
  - multiple train-test splits
- classification of video shots based on multimodal feature extraction, classification and fusion.
- localizing beat-events using a temporal structure inferred by a Conditional Random Field (CRF) model

Shot classification

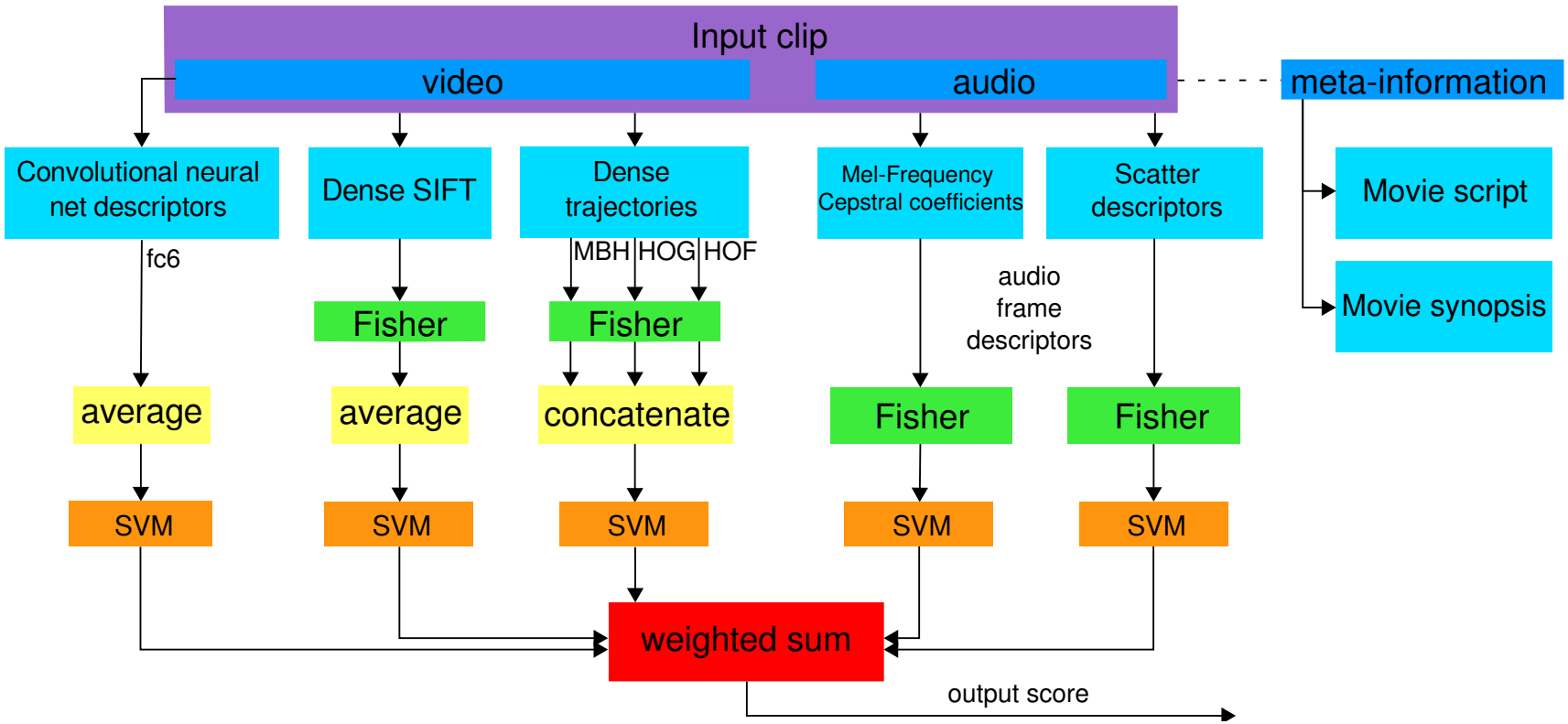
Descriptors: Fisher Vector descriptors of shots

Modality	Descriptor	Dimension	number of Gaussians
Static	Dense SIFT	34559	256
Static	CNN	4096	-
Motion	Dense Trajectories	108544	256
Audio	MFCC	20223	256
Face	Dense SIFT	16384	128

Classification

- SVM classifier trained on shot level for each channel
- weighted late fusion of classifier scores
  - random search over 5D space of coefficients
  - maximize mAP over the sub-folds

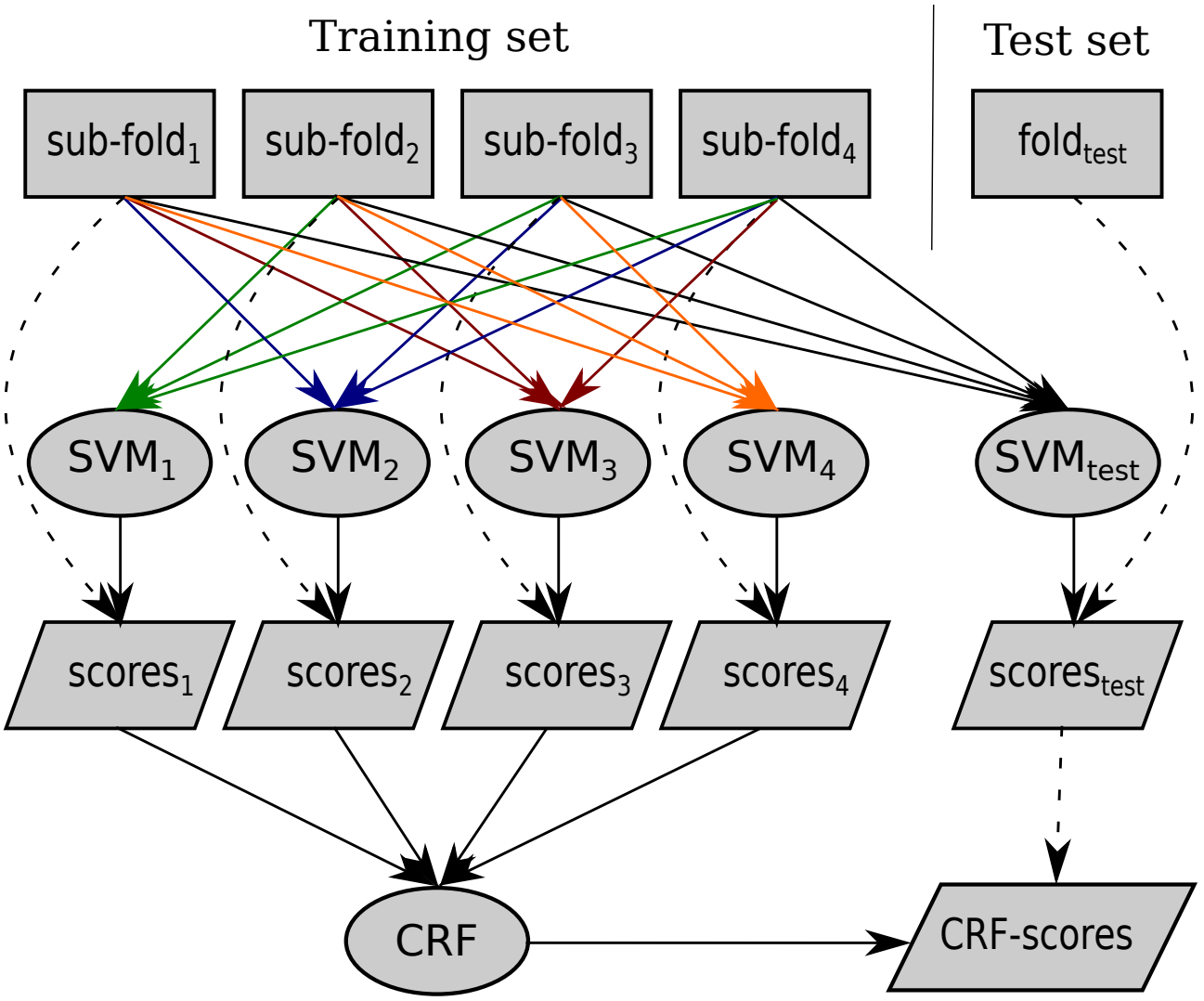
Feature representation of a video



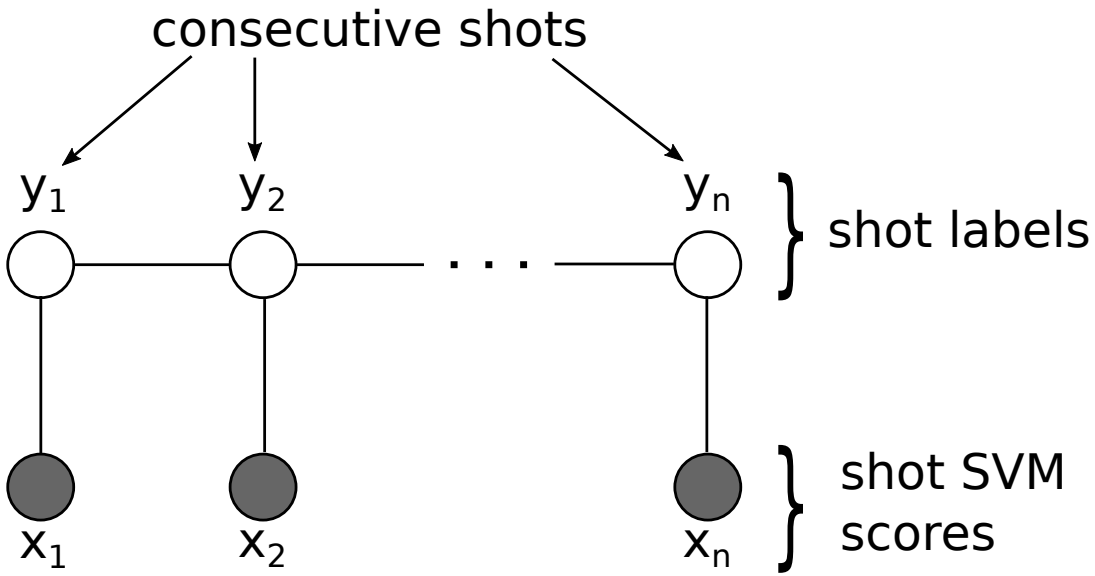
Face descriptor

- detect and track faces Everingham et al. [2006]
- encode face regions using SIFT and Fisher Vectors Simonyan et al. [2013]

Cross-validation scheme



Linear chain CRF



$$\log p(Y|X;\boldsymbol{\lambda},\boldsymbol{\mu}) = \sum_{k\in\mathcal{Y}} \lambda_k \sum_{i=1}^{n} p(y_i = k \mid x_i)\delta(y_i,k) + \sum_{\substack{k',k''\in\mathcal{Y} \\ (k',k'')\neq(c,c)}} \mu_{k',k''} \sum_{i=1}^{n-1} \delta(y_i,k')\delta(y_{i+1},k'')$$

Validation of the pipeline

- Coffee & Cigarettes dataset
- scoring fixed-size segments + non-maximum suppression of [Oneata et al., 2013]
- 65.5% mAP for “drinking” and 45.4% for “smoking”
- close to the result of [Oneata et al., 2013]:
  63.9% and 50.5% respectively

Evaluation metrics

- mean accuracy for classification
  - percentage of examples correctly detected for each class
- average precision for localization
  - 20% overlap over the union

Two types of evaluation splits

Classification results (accuracy)

	pursuit	battle	romance	victory good	victory bad	battle preparation	despair good	joy bad	good argue bad	good argue good	bad argue bad	mean accuracy
Leave 4 movies out												
SIFT	53.8	76.4	23.9	11.7	4.4	22.1	15.0	9.5	15.1	25.5	4.0	23.76 ± *5.26*
CNN	**66.4**	60.0	16.6	6.0	2.4	9.4	21.7	6.6	17.7	30.2	4.7	21.96 ± *5.91*
dense trajectories	58.5	**85.2**	**38.0**	12.7	6.2	**28.0**	19.5	11.6	**18.8**	**40.4**	1.8	29.15 ± *6.12*
MFCC	28.1	56.3	4.5	**17.7**	**36.2**	3.8	**35.4**	**15.6**	17.3	26.5	0.0	21.95 ± *13.97*
Face descriptors	47.9	58.1	8.6	12.7	11.4	17.3	9.3	3.2	6.2	22.3	**4.7**	18.35 ± *10.50*
linear score combination	63.9	89.2	32.3	14.0	11.4	18.6	26.0	12.1	18.0	44.3	1.8	30.15 ± *6.72*
+ CRF	76.0	91.2	57.6	19.9	1.0	41.4	43.1	9.6	25.1	44.8	0.0	37.25 ± *9.94*
Leave 1 franchise out												
linear score combination	57.8	83.6	13.0	14.9	9.6	3.8	28.0	5.2	18.2	44.3	0.0	25.32 ± *7.40*
+ CRF	75.4	87.4	31.3	15.8	0.0	12.7	33.4	5.7	23.2	43.7	0.0	29.89 ± *12.11*

- Dense trajectories — best single modality
- MFCC — more important than for TRECVID MED

Localization results (average precision)

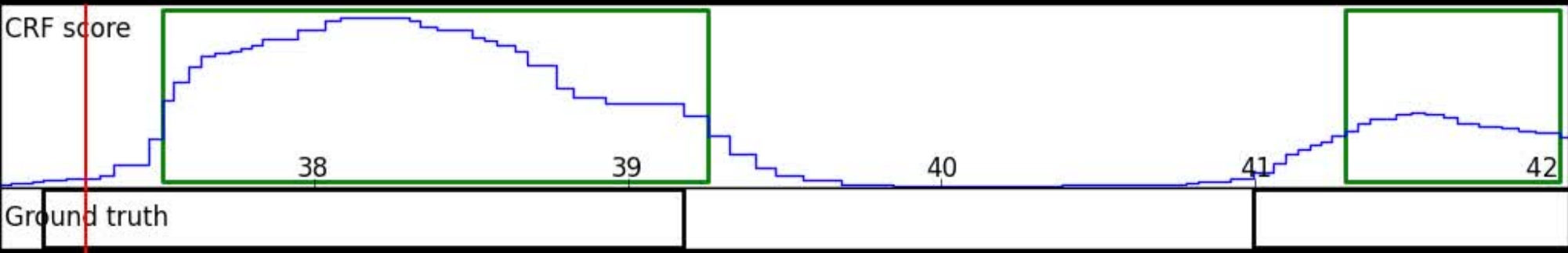
	pursuit	battle	romance	victory good	victory bad	battle preparation	despair good	joy bad	good argue bad	good argue good	bad argue bad	mean accuracy
Leave 4 movies out												
SIFT	32.8	27.7	14.8	11.5	1.2	**24.8**	2.0	3.1	9.3	4.8	0.0	12.00 ± *6.28*
CNN	30.2	21.1	11.2	4.8	0.5	8.6	1.1	2.0	**10.3**	12.7	0.0	9.33 ± *3.80*
Dense trajectories	**39.6**	**40.7**	**18.6**	**15.6**	**8.4**	21.8	3.4	**5.7**	8.1	13.8	0.9	16.05 ± *7.15*
MFCC	34.1	33.5	11.0	5.8	3.1	6.5	**11.5**	1.6	4.7	**21.4**	**22.8**	14.18 ± *8.31*
Face descriptors	20.6	12.9	10.4	2.1	0.1	0.4	4.6	2.9	6.9	9.5	0.1	6.41 ± *4.13*
All features	34.6	38.9	22.6	14.6	4.4	26.7	6.4	4.6	12.2	16.9	0.6	16.59 ± *6.82*
Leave 1 franchise out												
All features	36.8	36.5	28.9	14.3	4.5	1.7	4.2	5.2	6.5	13.5	3.7	14.16 ± *6.84*

- same overall ranking as for classification


pursuit
romance
victory bad
despair good
good argue bad
bad argue bad
battle
victory good
preparation
joy bad
good argue good
NULL
good argue good
ground truth
SVM output
CRF output

battle


Controversy between heroes and villains


CRF score
38
39
40
41
42
Ground truth

Conclusion

- **Action Movie Franchises** — video analysis benchmark for multiple tasks:
  - shot classification, beat-event localization
- Main advantages:
  - dense shot and beat-event level annotation
  - multiple evaluation protocols, 5 train-test splits
- experiments with state-of-the-art classifiers and descriptors, in multiple modalities

Publication

- D. Potapov, M. Douze, J. Revaud, Z. Harchaoui, C. Schmid "Beat-Event Detection in Action Movie Franchises", arXiv, 2015
- **Action Movie Franchises** dataset online

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

Outline

Conclusion

- Video summarization
  - weakly-supervised approach
  - novel temporal segmentation algorithm
  - MED-Summaries dataset
- Beat-event detection in action movie franchises
  - Action Movie Franchises dataset with 20 movies and 11 beat-categories
  - experimental evaluation for classification and localization tasks

Publications

D. Potapov, M. Douze, Z. Harchaoui, C. Schmid
"Category-specific video summarization", ECCV 2014

D. Potapov, M. Douze, J. Revaud, Z. Harchaoui, C. Schmid
"Beat-Event Detection in Action Movie Franchises", arXiv, 2015

Future work: Using context for video summarization

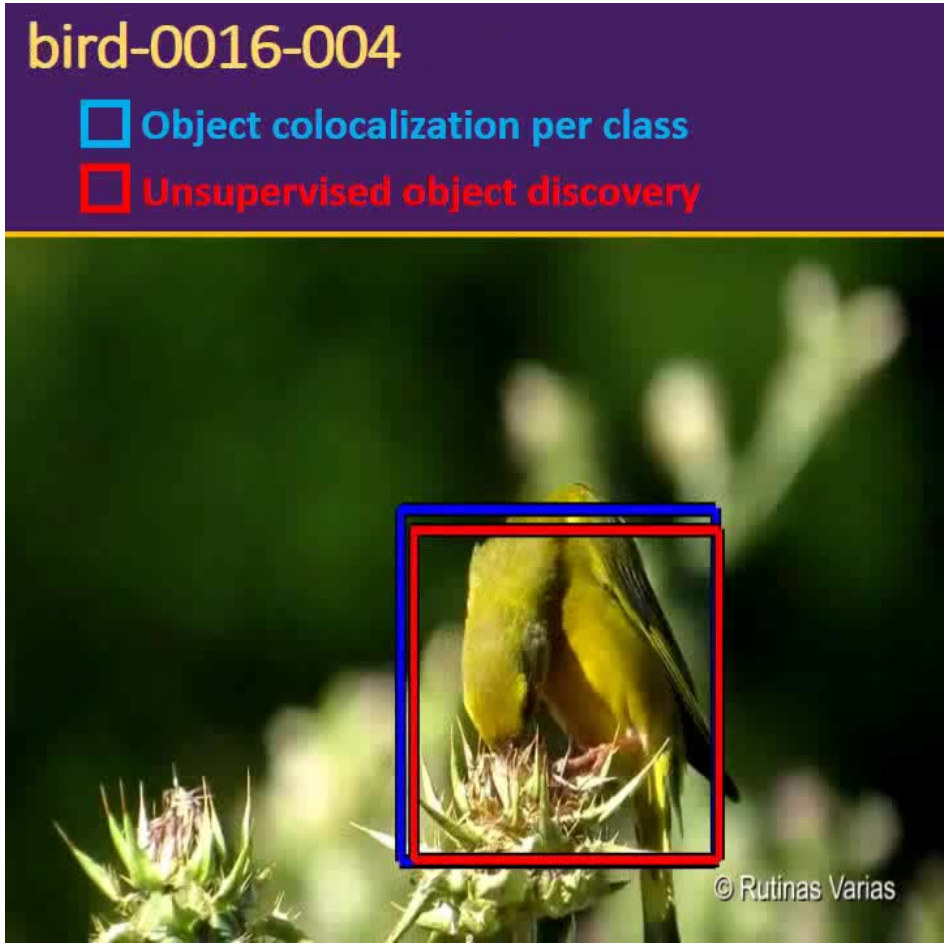
- subject of the video Song et al. [2015]
- summarization goal Kim et al. [2014]
- target audience Khosla et al. [2013]

Future work: More modalities for event recognition

- summarization based on parallel channels of descriptors Truong and Venkatesh [2007]
- advanced fusion methods Yu et al. [2014]
- **object-centered and human-centered cues** Prest et al. [2013]

Future work: Speed and scalability

- fast approaches ignore spatial and temporal grouping
- localization approaches are slow
- intermediate solution could be interesting Cho et al. [2015]

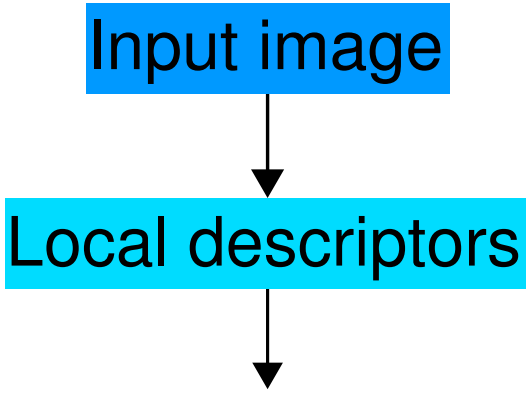


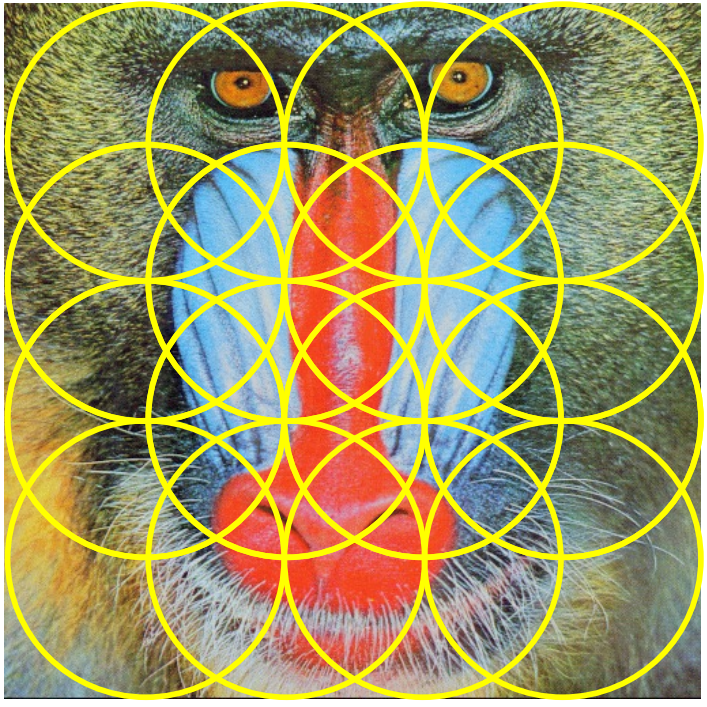
Thank you for your attention!

Conventional image classification pipeline

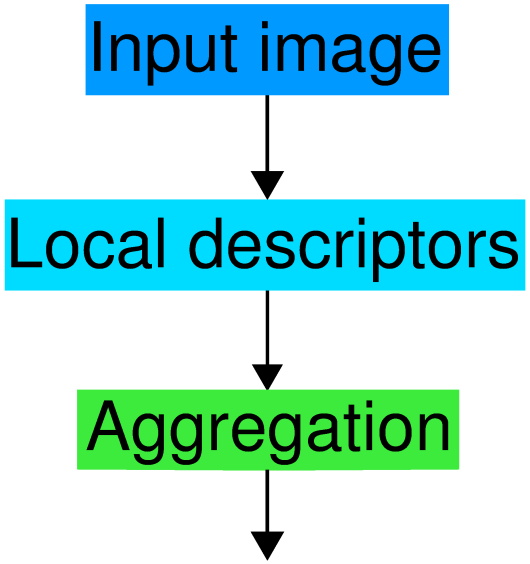
Input image

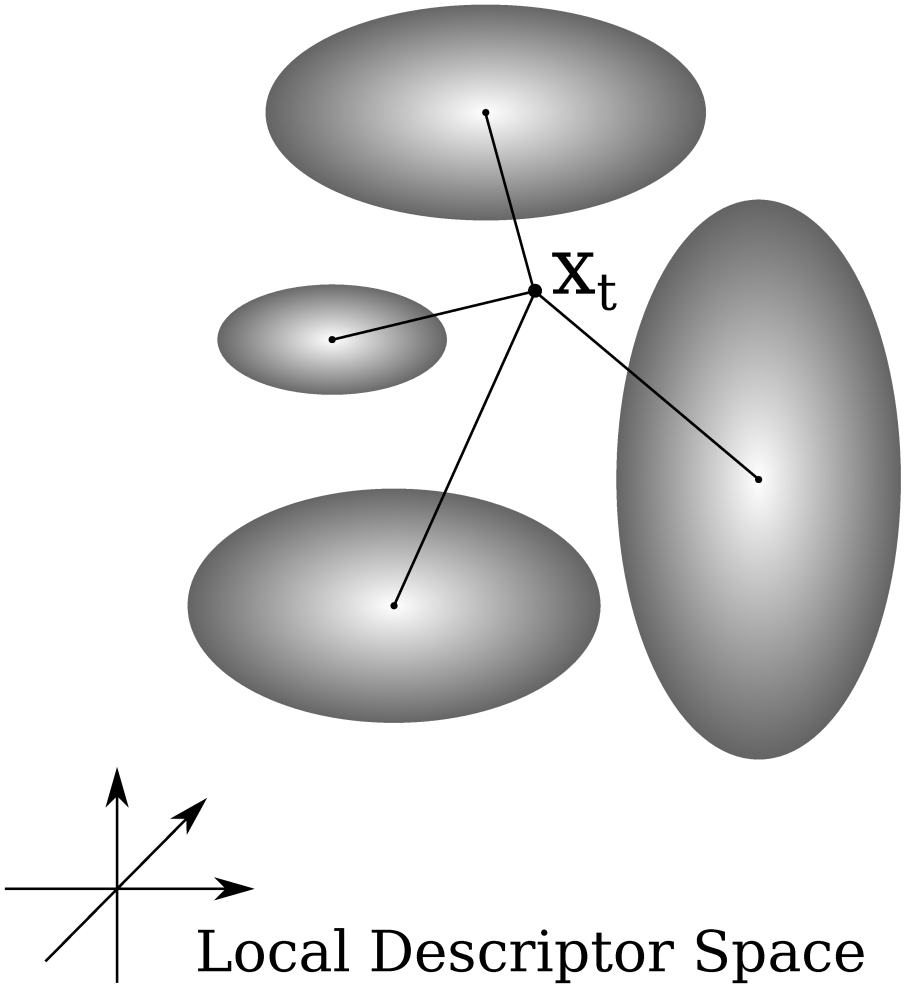
Conventional image classification pipeline



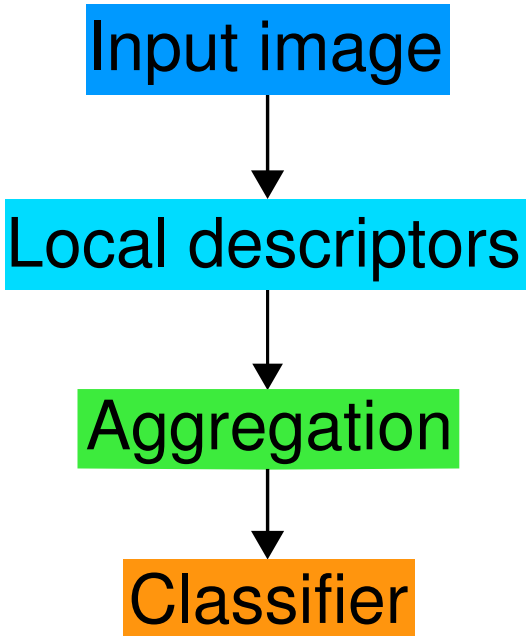


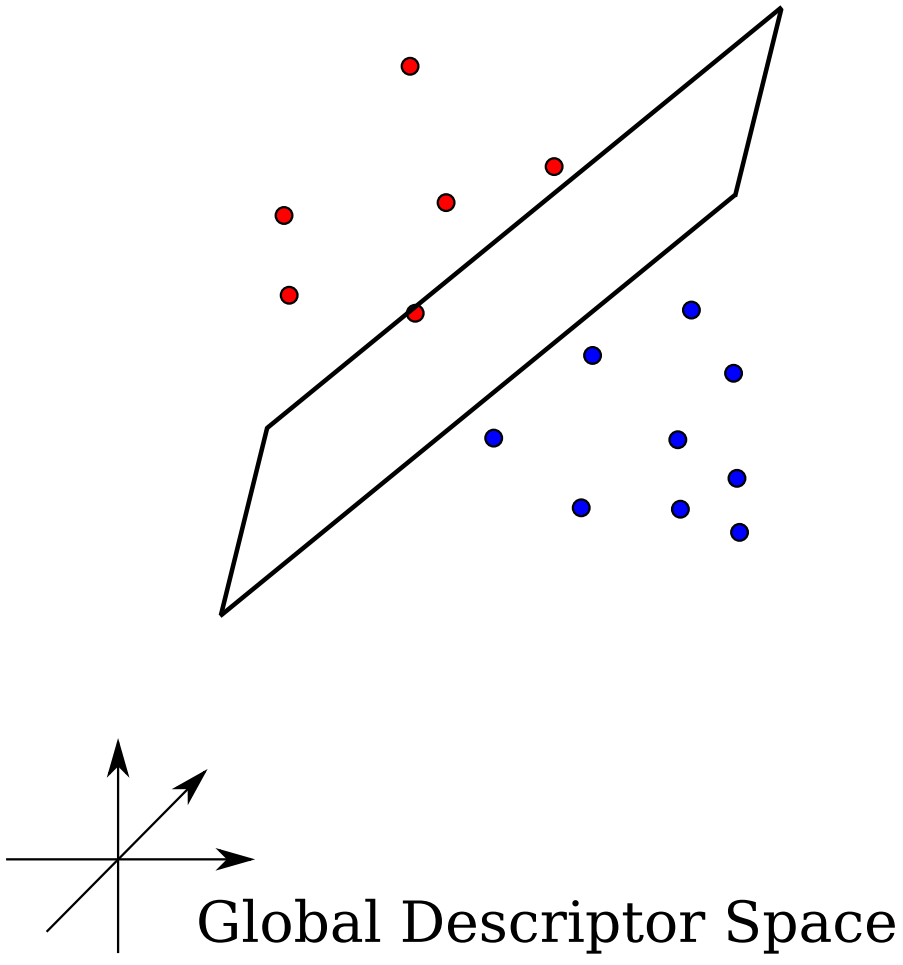
Conventional image classification pipeline





Conventional image classification pipeline





Outline