# VISUAL UNDERSTANDING AND PERSONALIZATION FOR AN OPTIMAL RECOLLECTION EXPERIENCE

ANA GARCÍA DEL MOLINO

Ph.D. ORAL DEFENCE

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

29TH OCT 2019

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

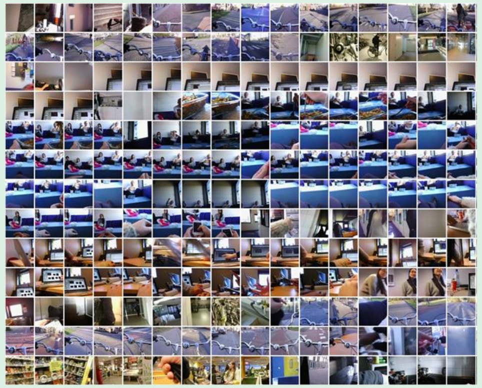
Agency for Science, Technology and Research SINGAPORE

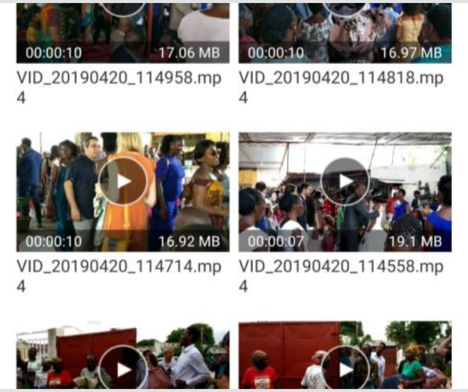
A picture is worth a 1000 words

Lifelogging cameras record every detail

What about a video?

Memory overload problem

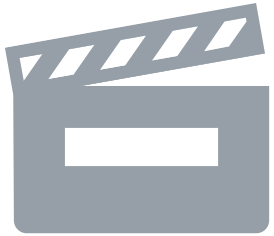





# HOW CAN WE MAKE THOSE MEMORIES ACCESSIBLE?

Take less pictures!

Clean
the gallery periodically



Edit
into a short video

# CHALLENGES

Content needs to be grouped by class (Five Ws)

Good storytelling requires finding relations

Content must be of good visual quality and aesthetic

Result must be adapted to each individual preferences

# OVERVIEW AND THESIS CONTRIBUTIONS

### Segment into event units

- Semantic clusters
- Episodic events

CES

### Select units

- Task-driven
- Story coherence
- Customization

CES*
CRF*
PHD

### Adapt from user feedback

Improve customization through interaction

AVS

* CES for lifelog summarization
* CRF for video summarization

Segment into event units

- Semantic clusters
- Episodic events

CES

Select units

- Task-driven
- Story coherence
- Customization

Adapt from user feedback

Improve customization through interaction

# SEGMENT INTO EVENT UNITS

## State of the Art

- Temporally linked events
  - Use of motion cues [Kitani et al.,Varini et al.]
  - Windowed feature similarity (action change points) [Bettadapura et al., Poleg et al.]
  - Variations in semantic tags (e.g. location) [Furnari et al.]

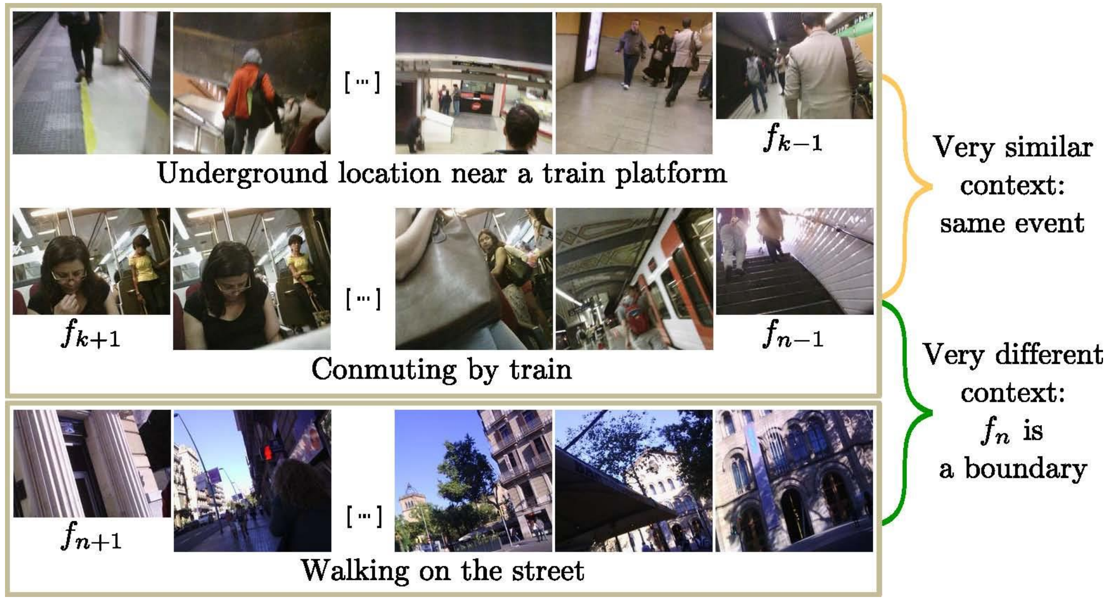
- Grouping by event class
  - Clustering methods by feature [Xu et al.]
  - Semantic consistency [Dimiccoli et al.]

# SEGMENT INTO EVENT UNITS

## Limitations

- Motion cues are not available in Low Time Resolution.
- Heterogeneous events may contain many action change points.
- Event segmentation frequently needs supervision.
- Semantic tags may be costly to annotate.
- Number of events or classes are not known for clustering methods.

# CONTEXTUAL EVENT SEGMENTATION




Episodic event segmentation must be ...

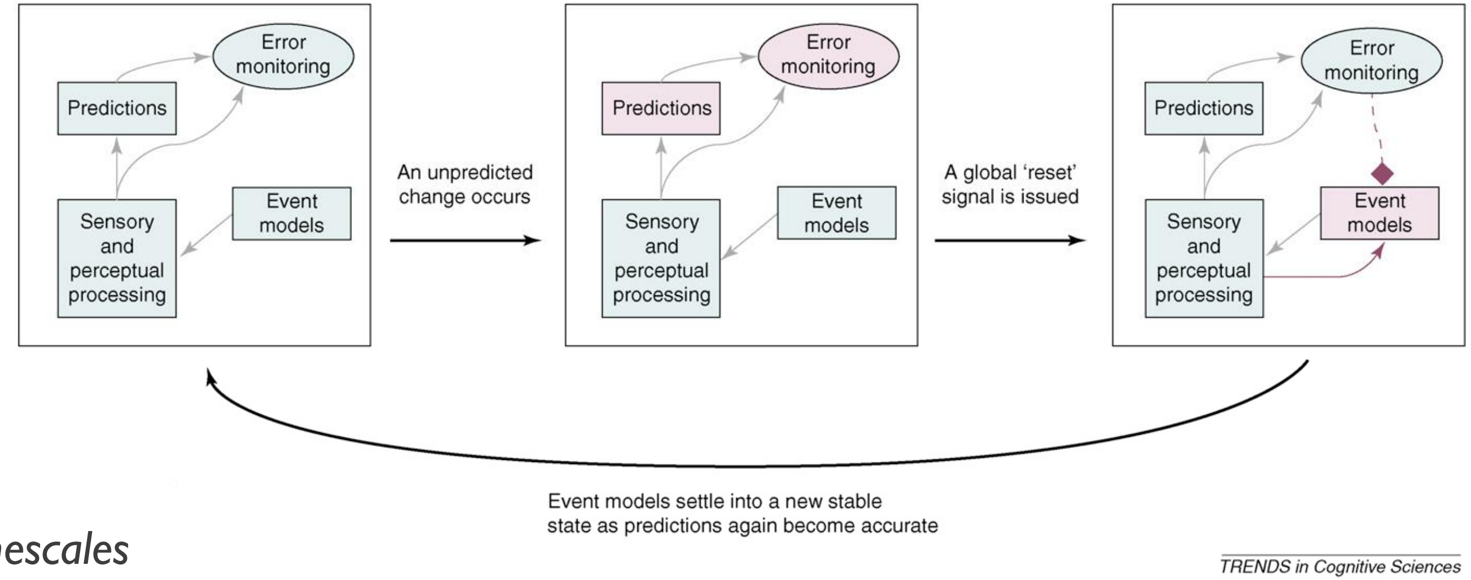
… insensitive to occlusions and short distractions.

… able to detect boundaries between heterogeneous events.

**Garcia del Molino, A.**, Lim, J. H., & Tan, A. H. (2018). Predicting Visual Context for Unsupervised Event Segmentation in Continuous Photo-streams. In *ACM International Conference on Multimedia*.

# EVENT PERCEPTION THEORY

- An *event model* is constructed for each episodic event.
- Depends on *perceptual prediction*
  - Guided by the *event model*
  - Conditioned by *prior knowledge*
- Depends on change (*error monitoring*)
- Happens simultaneously on *multiple timescales*
- Long-term memory links event models by their causal relations.

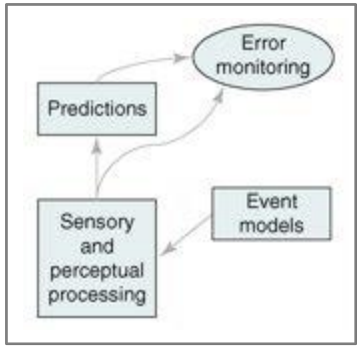


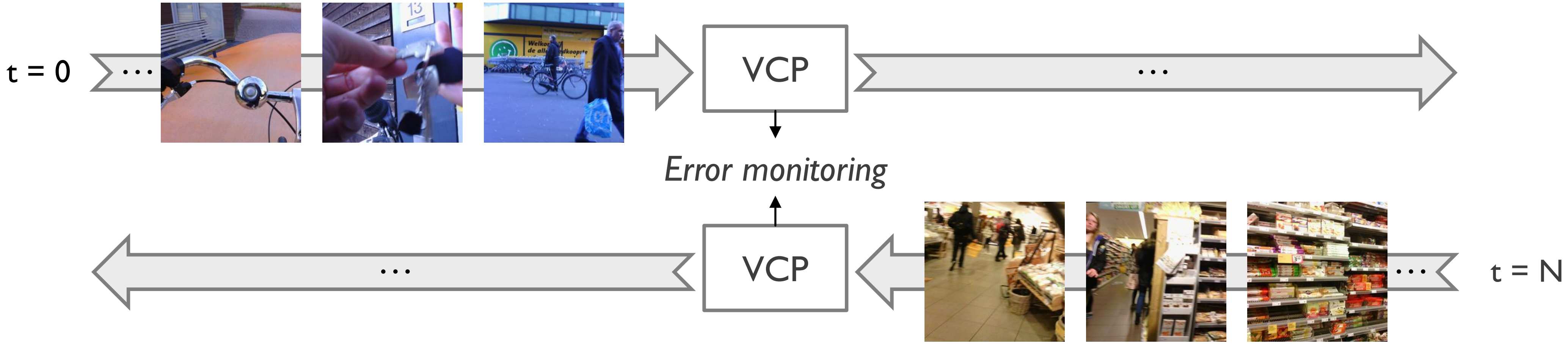


Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72–79.
Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

# CONTEXTUAL EVENT SEGMENTATION AS AN EMULATION OF THE COGNITIVE MODEL

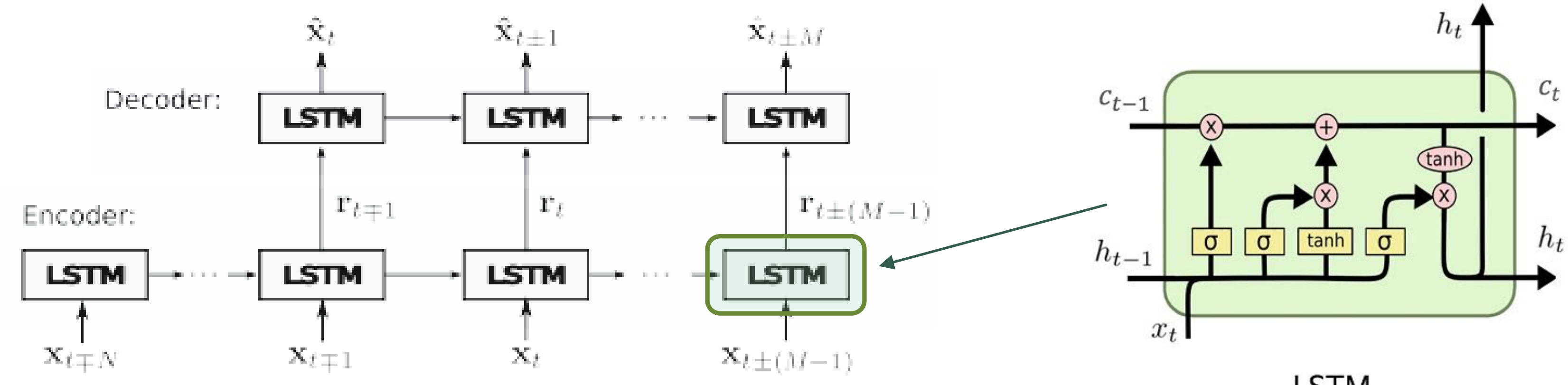
- The Visual Context Predictor builds the *event models* and outputs the *perceptual prediction.*
- *Prior knowledge* is acquired from 13k hours of daily life activities
- *Error monitoring:* imbalance between past and future *perceptual prediction*
- *Timescale granularity*: controlled by the error threshold



# VISUAL CONTEXT PREDICTOR

- The Visual Context Predictor is trained using an autoencoder architecture fed with lifelog image sequences:




LSTM
(Long-Short Term Memory)

$$\begin{pmatrix} \underline{i} \\ \underline{f} \\ \underline{o} \\ \underline{g} \end{pmatrix} = \begin{pmatrix} sigm \\ sigm \\ sigm \\ tanh \end{pmatrix} \left( \begin{bmatrix} \mathbf{x_t} \\ \mathbf{h_{t-1}} \end{bmatrix} \mathbf{W} \right)$$

$$\mathbf{c_t} = \underline{f} \circ \mathbf{c_{t-1}} + \underline{i} \circ \underline{g}$$

$$\mathbf{h_t} = \underline{o} \circ tanh(\mathbf{c_t}),$$

- At test time, the encoder module is used to encode the *event models* from the input image sequences
- The Visual Context Predictor can make predictions from forward and backward sequences.

# BOUNDARY DETECTOR

1. Get future (forward) and past (backward) *perceptual prediction* from the Visual Context Predictor:

$$\mathbf{rf}(t-1) \leftarrow \text{predicted from } [\mathbf{x}_k]_{\forall\ 0 \leq k < t}$$
$$\mathbf{rp}(t+1) \leftarrow \text{predicted from } [\mathbf{x}_k]_{\forall\ \mathbf{len}[\mathbf{x}] \geq k > t}$$

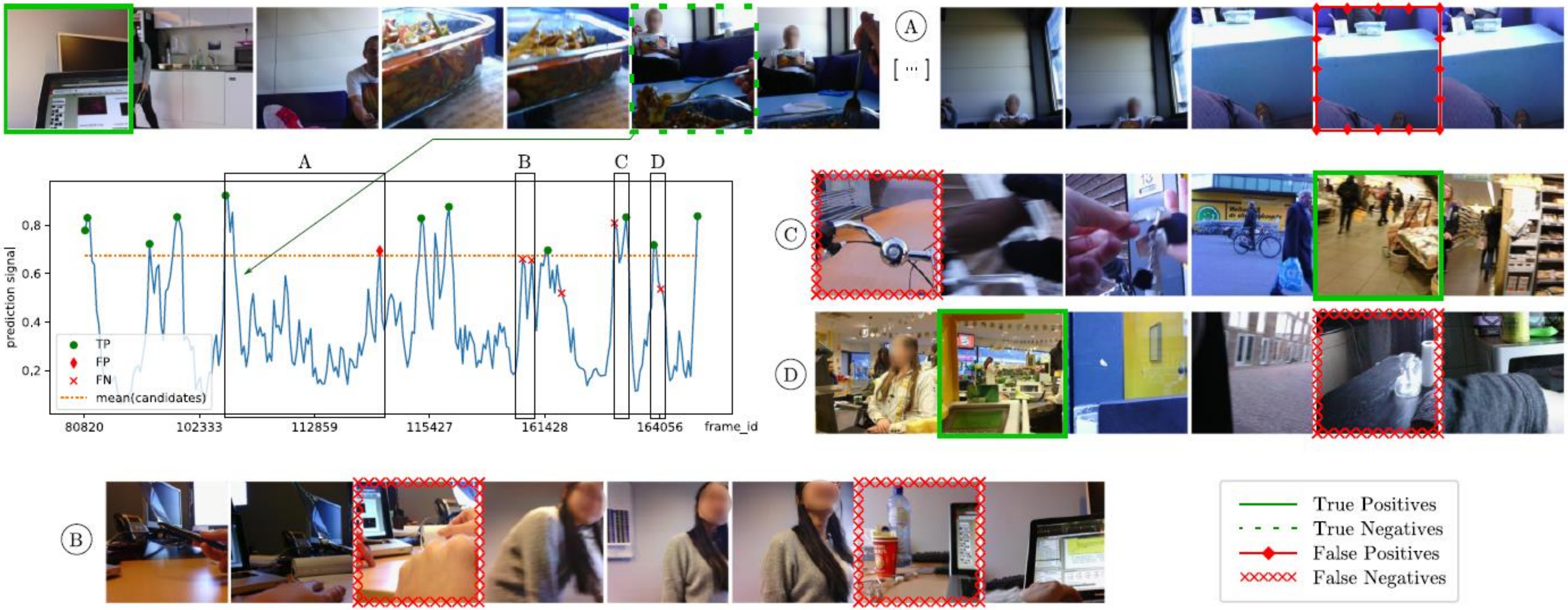
2. Detect boundary candidates analyzing imbalance between Past and Future context (*error monitoring*):

$$pred(t) = cos\_dist(\mathbf{rf}(t-1), \mathbf{rp}(t+1))$$
$$b = \{t \mid (\frac{\delta pred}{\delta t} = 0)\}$$

3. Adjust timescale grain:

$$b = \{b_k \mid pred(b_k) \leq average(pred(b))\}$$

# USE CASE EXAMPLE



# EXPERIMENTAL PROTOCOL

## Datasets

- Training:
  - LTR: NTCIR, CLEF, R3
  - HTR: CSumm
- Testing:
  - LTR: EDUB-Seg, EDUB-SegDesc
  - HTR: FPInteraction, HujiEgoSet

## Comparison to the state of the art

- Precision, Recall and F-measure of detected event boundaries
- Benchmark:
  - windowed GIST dist.
  - AC-Color
  - SR-Clustering
  - KTS

## Ablation study

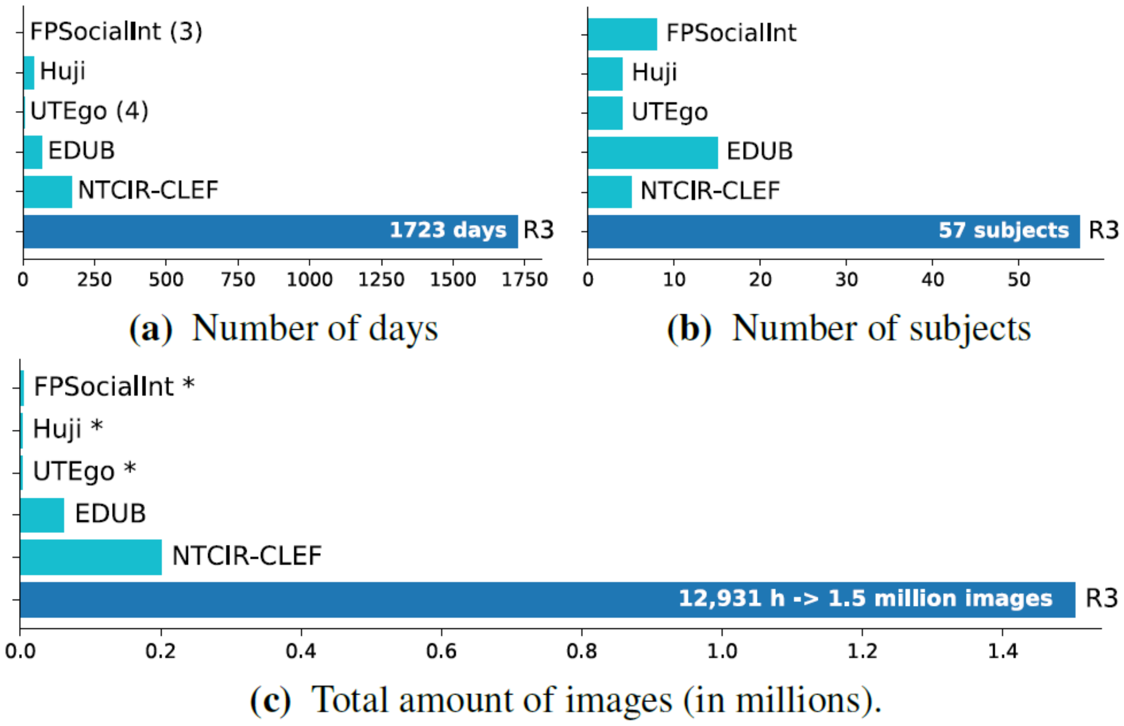
- Predicting the next frame vs predicting the event model
- Use of PCA or mean aggregation instead of VCP
- Use of supervision for candidate pruning


Dimiccoli, M., Bolaños, M., Talavera, E., Aghaei, M., Nikolov, S. G., & Radeva, P. (2017). Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *Computer Vision and Image Understanding*, *155*

Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. *IEEE Conference on Computer Vision and Pattern Recognition*.

Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014). Category-specific video summarization. In *European conference on computer vision*. Springer,

# DATASETS




R3*:

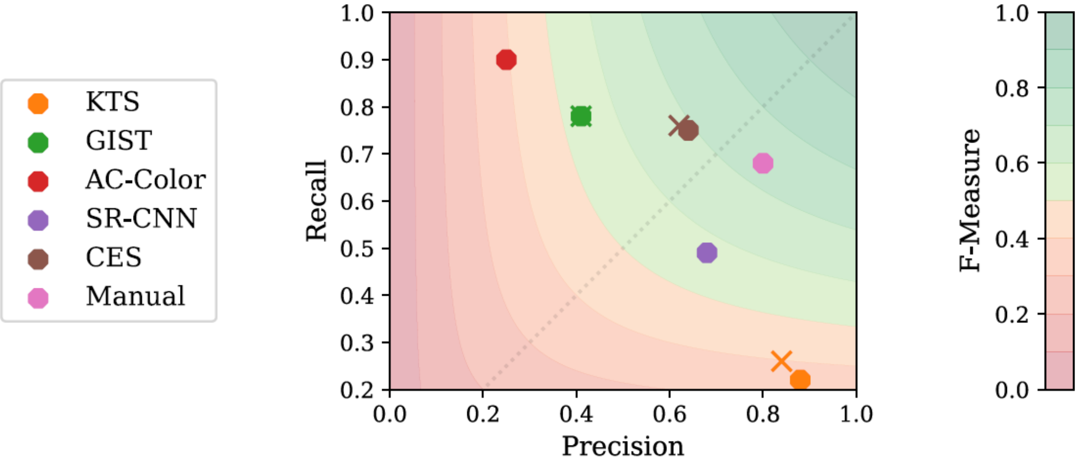
- Recorded with Narrative Clip
- Daily activities of 57 subjects
- Two pictures per minute during 8h daily
- 1.500.890 images
- Wide range of occupations and lifestyles

Ref. Table 3.1

*Visual features publicly available at http://dx.doi.org/10.17632/ktps5my69g.1

# RESULTS FOR LOW TIME RESOLUTION

- CES outperforms all the baselines for LTR videos.
- CES can detect 10% more true boundaries than the average person but will also find a relative 80% more incorrect events.




- Ablation study:
  - Using the imbalance between VCP features outperforms predicting the next video frame (*error*).
  - The VCP feature is more informative than other kinds of temporal aggregations (*mean*, *PCA*)
  - Supervised learning (*w/ SVM*) does not improve the prediction substantially.

| | averaged F1 | averaged Prec. | averaged Rec. |
|---|---|---|---|
| CES-error | 0.42 | 0.45 | 0.49 |
| CES-mean | 0.52 | 0.56 | 0.56 |
| CES-PCA | 0.66 | 0.67 | 0.69 |
| **CES (with VCP)** | 0.69 | 0.66 | **0.77** |
| k-means w/ SVM | 0.67 | 0.70 | 0.67 |
| CES w/ SVM | 0.71 | 0.75 | 0.71 |

# RESULTS FOR HIGH TIME RESOLUTION

- CES outperforms the baselines for long videos (*FP Social Int*), and is competitive for shorter videos (*Huji Ego*).
- For both datasets, the best results are obtained with CES and video frames downsampled at 12 frames per min.
- Lower frame rates are preferred to train the VCP. High frequencies will cause VCP to learn trivial representations.

| Sampling: | HTR: 2 sec. | | | | | | HTR: 5 sec. | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset: | Huji | | | FP Social Int | | | Huji | | | FP Social Int | | |
| method | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R |
| KTS [102] | 0.31 | 0.45 | 0.27 | 0.08 | 0.07 | 0.11 | 0.34 | 0.88 | 0.22 | 0.09 | 0.20 | 0.06 |
| GIST dist [7] | **0.32** | 0.72 | 0.24 | 0.14 | 0.26 | 0.10 | 0.31 | 0.71 | 0.23 | 0.13 | 0.24 | 0.09 |
| CES30 | 0.28 | 0.27 | 0.35 | 0.12 | 0.09 | 0.18 | 0.29 | 0.36 | 0.28 | 0.11 | 0.12 | 0.10 |
| CES30-win | 0.31 | 0.29 | 0.41 | 0.18 | 0.13 | 0.30 | **0.35** | 0.42 | 0.34 | 0.19 | 0.19 | 0.20 |
| CES10-win | 0.30 | 0.29 | 0.40 | **0.18** | 0.13 | 0.31 | 0.32 | 0.42 | 0.31 | **0.23** | 0.22 | 0.25 |
| CES{2, 5}-win | 0.28 | 0.23 | 0.45 | 0.15 | 0.09 | 0.33 | 0.34 | 0.39 | 0.35 | 0.20 | 0.18 | 0.24 |

# CONTEXTUAL EVENT SEGMENTATION

- ✓ Is based on human perceptual reasoning
- ✓ Models the photo-stream sequences and detects changes in the visual context
- ✓ Is insensitive to occlusions and short distractions
- ✓ Detects boundaries between heterogeneous events
- ✓ Leverages unsupervised learning

## Segment into temporal units

- Semantic clusters
- Episodic events

## Select units

- Task-driven
- Story coherence
- Customization

CES*

CRF*

PHD

## Adapt from user feedback

Improve customization through interaction

* CES for lifelog summarization

* CRF for video summarization

# SELECT UNITS

## State of the Art

- Story Coherence
  - Diversity from visual features [Lu et al. , Zhao et al., Varini et al., Shargi et al.]
  - Representativeness [Wang et al., Gygli et al., Xu et al., Ho et al.]
- Interestingness
  - Global [Lee et al., Gygli et al., Yao et al.]
  - Personalized [Ng et al., Varini et al.]
- Task-driven [Okamoto et al.]

# SELECT UNITS

## Limitations

- Rarely task or user-driven
- Interestingness predicted globally
- Personalized methods rely on the similarity to a given query, not balancing with the global interestingness, diversity or representativeness.

# CONTEXTUAL EVENT SEGMENTATION FOR TASK-DRIVEN LIFELOG SUMMARIZATION

### GOOD QUALITY IMAGES

Ranking according to color diversity and blurriness

### DIVERSE AND UNIQUE CONTENT

Event clusters defined by contextual event segmentation

### RELEVANT TO QUERY (TASK-DRIVEN)

Relevance score based on a learned linear model

### MAX. INFORMATION IN MIN. LENGTH

Iterative key-frame selection from relevant events

# EXPERIMENTAL PROTOCOL

## Benchmarking

- ImageCLEF 2017 LifeLog Task
- Precision, Recall and F-Measure
- Summaries of different lengths

## Tasks

- Working from home
- Shopping
- Driving
- Lunch at the office
- …

## Ablation study

- Different levels of human intervention
- Different summary lengths
- Use of K-means segmentation against CES

# BENCHMARKING RESULTS

- The proposed method is only outperformed by methods involving human intervention
- CES segmentation outperforms clustering with temporally-constrained k-means

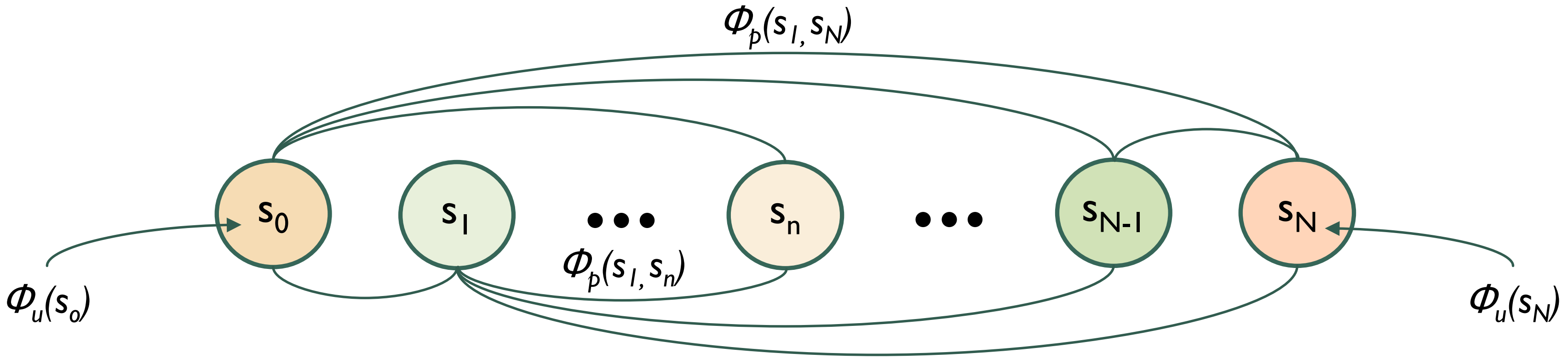
| Summary Size: | $X = 10$ | | | $X = 50$ | | | |
|---|---|---|---|---|---|---|---|
| method | F1 | P | R | F1 | P | R | notes |
| Org_A [147] | 0.19 | N.A. | | N.A. | | | Automatic (NLP) |
| Org_SA [147] | 0.32 | N.A. | | N.A. | | | Keywords |
| UPB [30] | 0.13 | N.A. | | N.A. | | | WordNet filter |
| I2R_KM [24] | **0.50** | 0.70 | 0.43 | **0.51** | 0.53 | 0.58 | Human intervention |
| CRF_KM | 0.30 | 0.41 | 0.28 | 0.37 | 0.34 | 0.49 | Relevance learned from |
| CRF_CES | **0.37** | 0.53 | 0.33 | **0.39** | 0.34 | 0.54 | WordNet propagation |

Nguyen, D., Tien, D., Piras, L., Riegler, M., Boato, G., Zhou, L., & Gurrin, C. (2017). Overview of ImageCLEF Lifelog 2017: lifelog retrieval and summarization.

# CONTEXTUAL EVENT SEGMENTATION AND CONDITIONAL RANDOM FIELDS FOR TASK-DRIVEN LIFELOG SUMMARIZATION

- ✓ Generates informative summaries
- ✓ More accurate event segmentation than other clustering methods
- ✓ Minimal user intervention

# CONDITIONAL RANDOM FIELDS FOR CONSUMER VIDEO SUMMARIZATION




The unary potential enforces that the selected segments are of good visual quality

$$E_\theta(\mathbf{s}) = \lambda \sum_i \underbrace{\phi_u(s_i)}_{\text{unary}} + \sum_{ij} \underbrace{\phi_p(s_i, s_j)}_{\text{pairwise}},$$

The pairwise potential enforces that the selected segments are diverse and representative of the whole video

$$\phi_u(s_i) = \begin{cases} L & \text{if } s_i = 0 \\ Q_i & \text{if } s_i = 1 \end{cases}$$

$$\phi_p(s_i, s_j) = e^{-d(\Psi_i, \Psi_j)} \begin{cases} L\alpha & \text{if } s_i = s_j = 0 \\ -L\beta & \text{if } s_i = s_j = 1 \\ \gamma & \text{if } s_i \neq s_j \end{cases}$$

# EXPERIMENTAL PROTOCOL

## Datasets

- UTEgo
- CSumm

## Comparison to the state of the art

- User survey
- Benchmark:
  - Uniform sampling
  - Manual annotations
  - VMMR
  - Lee et al. (2012)

## User survey

- Informativeness
- Visual quality

Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. *IEEE Conference on Computer Vision and Pattern Recognition*.
Li, Y., & Merialdo, B. (2010). Multi-video summarization based on video-mmr. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE.

# COMPARISON TO THE STATE OF THE ART

- Datasets:
  - CSumm: 10 videos of ~30 min each
  - UTEgo: 4 videos of ~6 h each, split into 7 videos to be at most 3h long
- Amount of videos for which the method on the left is ranked better than the method on top by most users (based on an on-line survey):

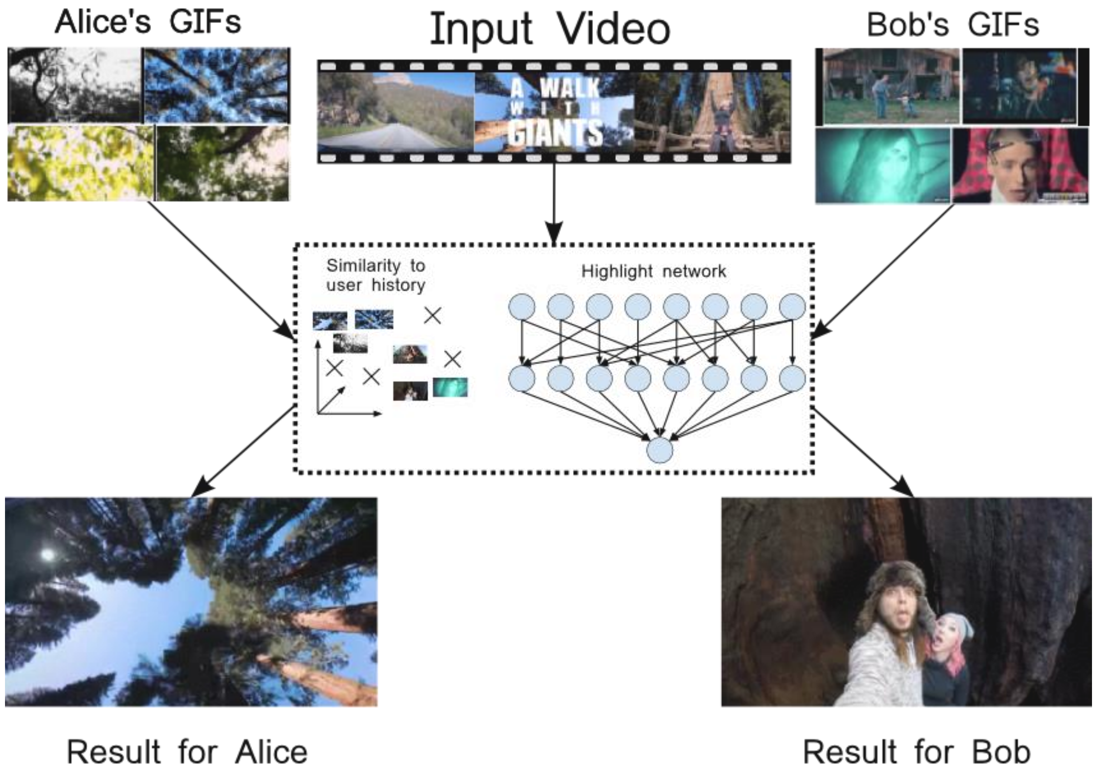
| | CSumm | | | | UTEgo | | | |
|---|---|---|---|---|---|---|---|---|
| | Unif. | Manual | VMMR | CRF | Unif. | CVPR | VMMR | CRF |
| Uniform | - | 3 | 3 | **4** | - | 2 | 2 | 4 |
| Manual/CVPR | 3 | - | 6 | 5 | 5 | - | 5 | 5 |
| VMMR | 3 | 1 | - | 3 | 5 | 2 | - | 3 |
| CRF | **4** | 2 | **5** | - | 3 | 1 | **4** | - |

- Conditional Random Fields are suitable for video summarization. Shorter videos have easier convergence.

# CONDITIONAL RANDOM FIELDS FOR CONSUMER VIDEO SUMMARIZATION

- ✓ Each segment of the video is defined by a CRF node
- ✓ The optimal summary maximizes the energy cost of the CRF
- ✓ The CRF unaries enforce a summary of good visual quality
- ✓ The CRF pairwise parameters enforce a diverse and informative summary

# PERSONALIZED HIGHLIGHT DETECTION




Not all users are interested in the same content.

Highlight detectors must ...

… take the user into account.

… use minimal user input.

**Garcia del Molino, A.**, & Gygli, M. (2018). PHD-GIFs: Personalized Highlight Detection for Automatic GIF Creation. In *ACM International Conference on Multimedia.*

# PAIRWISE RANKING FOR PERSONALIZED PREDICTIONS

- Personalized Highlight Detection takes two inputs:
  - A video *V* to analyze, formed by segments $\{s_k\}$
  - A user history *G*, formed by the previous GIFs that user generated, i.e. $\{g_i\}$

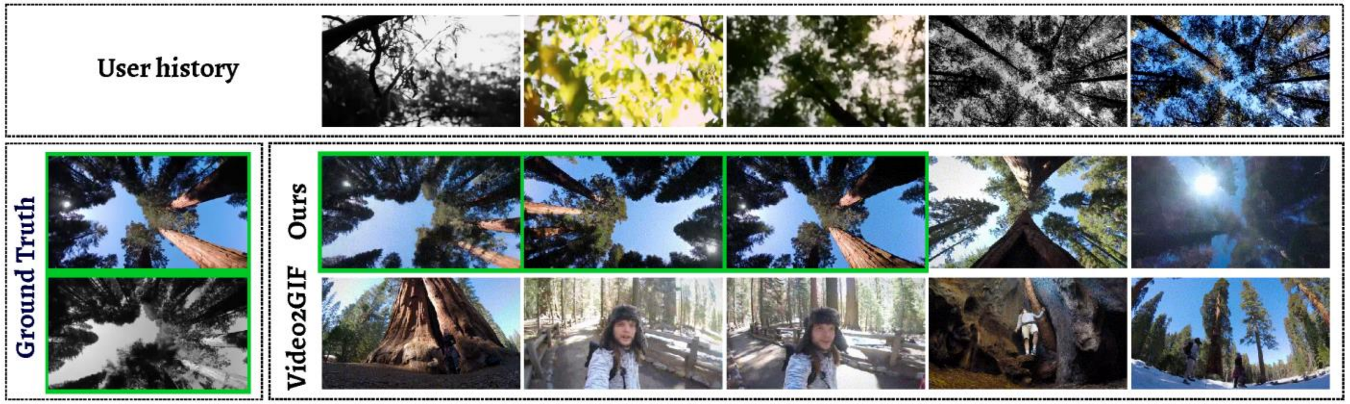
- Two ranking models are combined to predict personalized highlights:

1. Deep ranking on the aggregated history *p* = mean(*G*) : $h_{FNN}(s,\mathscr{G}) = FNN\left(\begin{bmatrix}\mathbf{s}\\ \mathbf{p}\end{bmatrix}\right)$
2. Ranked SVM on the distances *d* between *s* and *G* : $h_{SVM}(s,\mathscr{G}) = \mathbf{w}^{\mathrm{T}}\mathbf{d} + b$

$$h(s,\mathscr{G}) = h_{FNN}(s,\mathscr{G}) + \omega * h_{SVM}(s,\mathscr{G})$$

# USE CASE EXAMPLES

Accurate personalized prediction





Misleading user history

# EXPERIMENTAL PROTOCOL

## Dataset

- Personalized Highlights Dataset

## Comparison to the state of the art

- mAP, MSE and Recall@5
- Generic:
  - Video2Gif
  - SVM ranking
- Personalized:
  - VMMR
  - Residual

## Ablation study

- PHD w/o SVM-D
- SVM-D w/o Deep model
- Impact of the user history size

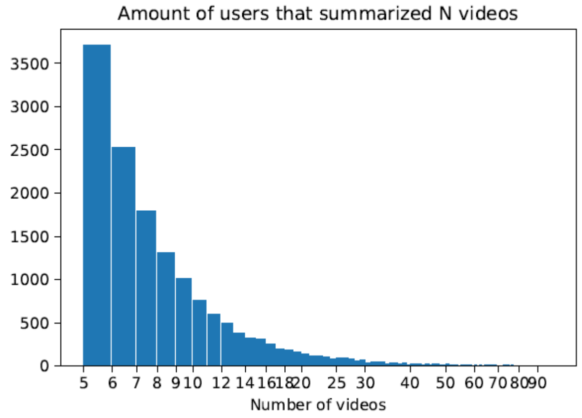

Gygli, M., Song, Y., & Cao, L. (2016). Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Y., & Merialdo, B. (2010). Multi-video summarization based on video-mmr. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE.

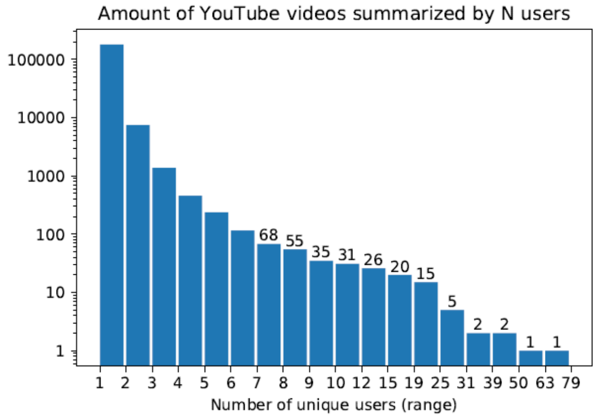
# DATASET

PHD²

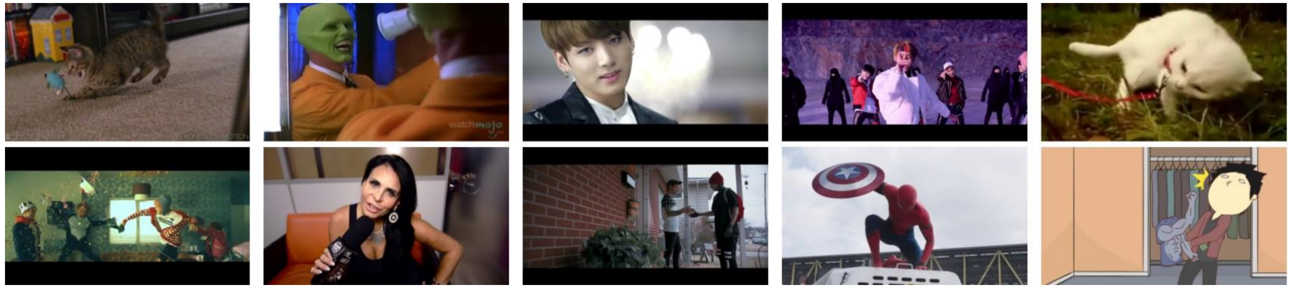
- Labels on what each specific user considers a highlight
- Most users summarized videos from three or less categories
- Close to 14.000 users from gifs.com
- A minimum of 5 videos per user
- More than 222,000 annotated highlights



**(a)** Number of videos per user.



**(b)** User queries per video.



Dataset publicly available at https://github.com/GarciaDelMolino/personalized-highlights-dataset

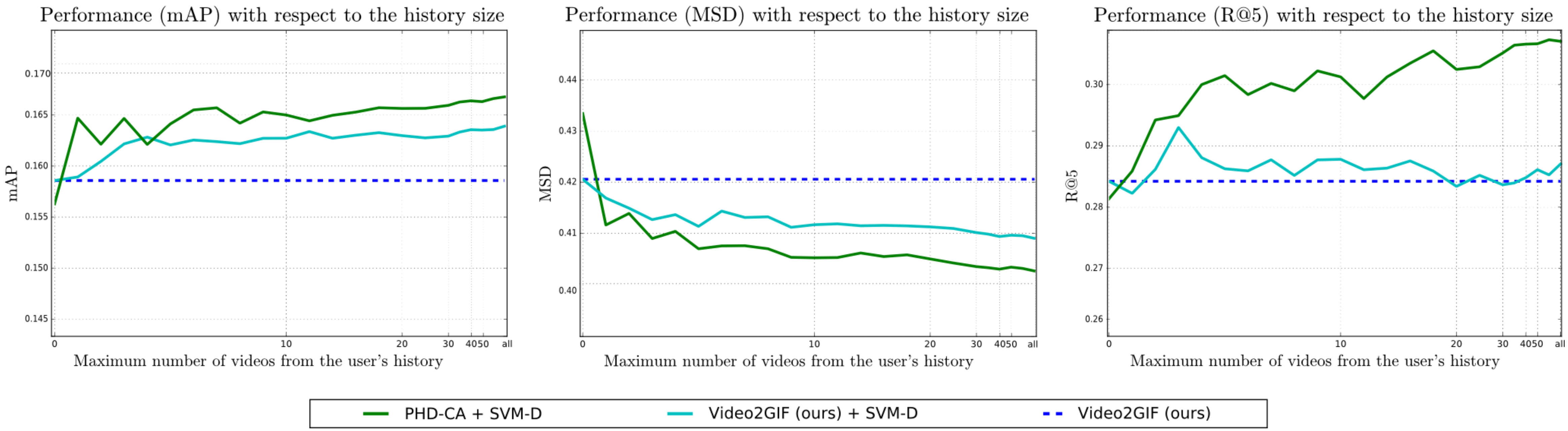
# COMPARISON TO THE STATE OF THE ART

- Tested for 1.000 users
- Models using only generic highlight information (*Video2GIF (ours)*) or only the similarity to previous GIFs (*SVM-D*) perform similar.
- Combining both kinds of information is beneficial.
  - PHD (CA + SVM-D) offers a relative improvement over generic highlight detection of 5.2% in mAP, 4.3% in mMSD and 8% in Recall@5.

| | **Model** | **mAP ↑** | **nMSD ↓** | **R@5 ↑** | **Notes** |
|---|---|---|---|---|---|
| Non-personal | Random | 12.97% | 50.60% | 21.38% | |
| | Video2GIF [48] | 15.69% | 42.59% | 27.28% | Trained on [48] |
| | Highlight SVM | 14.47% | 45.55% | 26.13% | |
| | Video2GIF (ours) | **15.86%** | **42.06%** | **28.42%** | |
| Personal | Max Similarity | 15.49% | 44.22% | 26.44% | unsupervised |
| | V-MMR | 14.86% | 43.72% | **28.22%** | unsupervised |
| | Residual | 14.89% | 47.07% | 26.05% | |
| | SVM-D | **15.64%** | **43.49%** | 28.01% | |
| | PHD (CA + SVM-D) | **16.68%** | **40.26%** | **30.71%** | |

# ABLATION STUDY

- PHD outperforms the state of the art of highlight detection with as little as one history element per user:



# PERSONALIZED HIGHLIGHT DETECTOR

- ✓ Is a global ranking model
- ✓ Conditions on the user previous browsing experience
- ✓ No human intervention
- ✓ Is personalized via the inputs
- ✓ New information from the user can trivially be included
- ✓ Proves to be more precise than the state of the art even with just one person-specific example

Segment into temporal units

- Semantic clusters
- Episodic events

Select units

- Task-driven
- Story coherence
- Customization

Adapt from user feedback

Improve customization through interaction

AVS

# ADAPT FROM USER FEEDBACK

## State of the Art

- Personalization via
  - Query [Han et al., Ng et al., Shargi et al., Yang et al.]
  - User profiling from metadata [Varini et al., Jaimes et al.]
  - User profiling from historical data [Peng et al., Yoshitaka et al.]
  - Attention signals [Aizawa et al., Varini et al., Xu et al.]

## Limitations

- The generated summary is not tunable.

# ACTIVE VIDEO SUMMARIZATION


Video editing should be seamless.

Automatic video summarization must ...

… generate diverse and representative videos.

… leverage on user profiles.

… allow for further modification.

**García del Molino, A.** , Boix, X., Lim, J. H., & Tan, A. H. (2017). Active video summarization: Customized summaries via on-line interaction with the user. In *Thirty-First AAAI Conference on Artificial Intelligence*.

# USER INTERACTION GUIDED BY PROBABILISTIC INFERENCE

- AVS asks the user specific questions about segments of the video:
  1. Would you want this segment to be in the final summary?
  2. Would you want to include similar segments?
- The user can also give feedback about the segments in the summary

- AVS can be divided into two independent inference problems:
  I. Infer the customized summary:

$$\mathbf{s}^{\star}_{\theta_t} = \arg\max_{\mathbf{s}} E_{\theta_t}(\mathbf{s})$$
$$E_{\theta}(\mathbf{s}) = \lambda \sum_i \phi_u(s_i) + \sum_{ij} \phi_p(s_i, s_j)$$

  II. Infer the next segment to show:

$$k^{\star} = \arg\max_k S_k$$
$$S_k = E_{\theta_{t+1}}\left[R\left(\mathbf{s}^{\star}_{\theta_{t+1}}, \mathbf{s}^{\star}_{\theta_t}\right) \mid k\text{-th candidate}\right]$$

# UPDATE OF THE CRF PARAMETERS

$$E_\theta(\mathbf{s}) = \lambda \sum_i \underbrace{\phi_u(s_i)}_{\text{unary}} + \sum_{ij} \underbrace{\phi_p(s_i, s_j)}_{\text{pairwise}},$$

$$\phi_u(s_i) = \begin{cases} L & \text{if } s_i = 0 \\ Q_i & \text{if } s_i = 1 \end{cases}$$

Controls visual quality and relevance

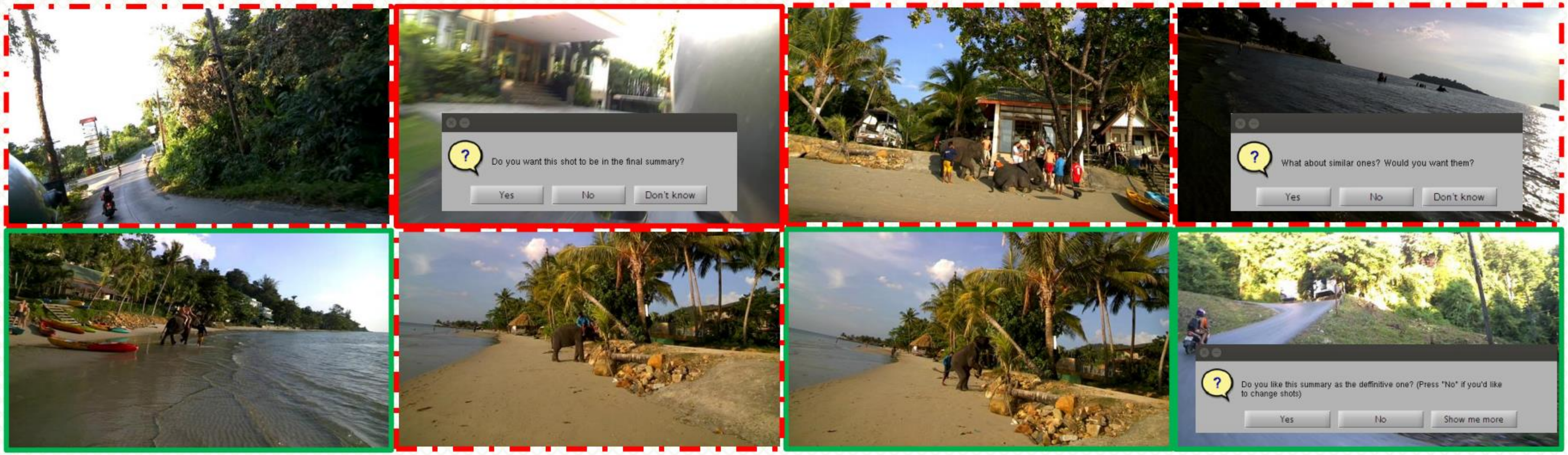
$$\phi_p(s_i, s_j) = e^{-d(\Psi_i, \Psi_j)} \begin{cases} L\alpha_{ij} & \text{if } s_i = s_j = 0 \\ -L\beta_{ij} & \text{if } s_i = s_j = 1 \\ \gamma_{ij} & \text{if } s_i \neq s_j \end{cases}$$

Controls diversity and representativeness

| | Q2 = Yes | Q2 = No |
|---|---|---|
| **Q1 = Yes** ▷ $Q_{k,t+1} = \Delta Q_{k,t}$ | $\{\gamma_{kj,t+1}\}_{\forall j} = \{-K\gamma_{kj,t}\}_{\forall j}$<br>$\{\beta_{kj,t+1}\}_{\forall j} = \{-K\beta_{kj,t}\}_{\forall j}$ | $\{\gamma_{kj,t+1}\}_{\forall j} = \{K\gamma_{kj,t}\}_{\forall j}$ |
| **Q1 = No** ▷ $Q_{k,t+1} = \Delta^{-1} Q_{k,t}$ | $\{\gamma_{kj,t+1}\}_{\forall j} = \{K\gamma_{kj,t}\}_{\forall j}$ | $\{\gamma_{kj,t+1}\}_{\forall j,t+1} = \{-K\gamma_{kj,t}\}_{\forall j}$ |

# USE CASE EXAMPLE

Q1N; Q2Y Q1N; Q2N Q1N; Q2Y Q1N; Q2Y




Q1Y; Q2- Q1N; Q2Y Q1Y; Q2- Q1Y; Q2-

# EXPERIMENTAL PROTOCOL

## Datasets

- UTEgo
- CSumm

## Comparison to the state of the art

- User study:
  - Discovery Task
  - Search Task
- Benchmark:
  - Uniform sampling
  - Manual annotations
  - VMMR
  - Lee et al. (2012)

## Ablation study

- Inferred questions vs random questions
- Impact of the number of questions asked

Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. *IEEE Conference on Computer Vision and Pattern Recognition*.
Li, Y., & Merialdo, B. (2010). Multi-video summarization based on video-mmr. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE.

# COMPARISON TO THE STATE OF THE ART

- Discovery task: the users create a summary from a video they have never seen before
- Evaluation:
  - Subjective preference against other summaries (top)
  - Subjective preference against random selection of questions (center)
  - Time to generate the summary (bottom).
- In 41% of the videos, AVS is considered the best over all tested methods, including summaries manually generated.
- The time to generate a video summary is reduced by four when using AVS against manual editing.

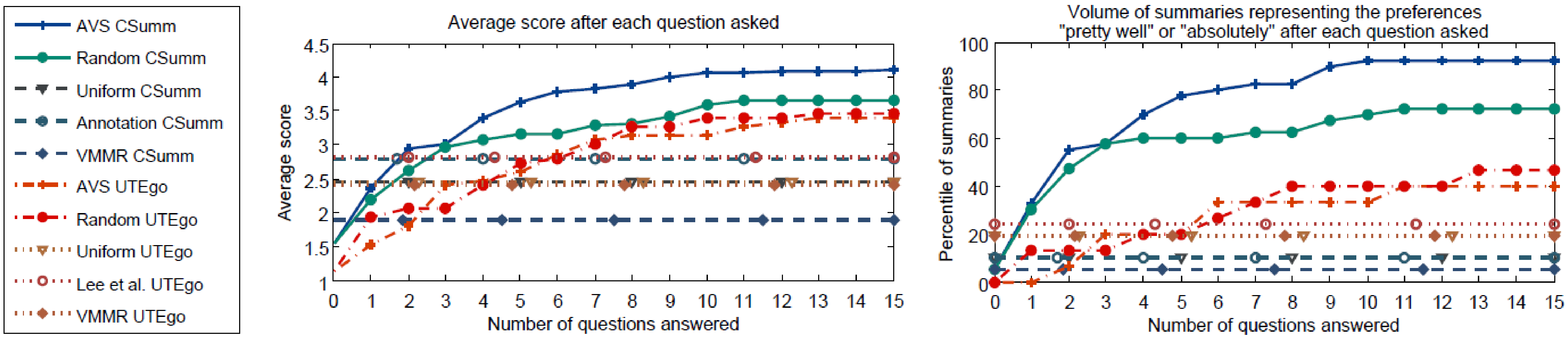
| | CSumm | | | | UTEgo | | | |
|---|---|---|---|---|---|---|---|---|
| | Unif. | Annot. | VMMR | AVS | Unif. | CVPR | VMMR | AVS |
| Unif. | - | 28% | 44% | 25% | - | 29% | 41% | 24% |
| An./CV. | 66% | - | 78% | **50%** | 59% | - | 71% | 41% |
| VMMR | 47% | 19% | - | 19% | 47% | 24% | - | 24% |
| AVS | **59%** | 34% | **66%** | - | **71%** | **53%** | **76%** | - |

| | Much worse | Worse | Similar | Better | Much better |
|---|---|---|---|---|---|
| CSumm: | 5.4% | 16.2% | 18.9% | **43.2%** | 16.2% |
| UTEgo: | 6.7% | 13.3% | 26.7% | **40%** | 13.3% |

| AVS | Manual |
|---|---|
| $5.89 \pm 3.85$ min. | $21.66 \pm 6.59$ min. |

# ABLATION STUDY

- Search task: the user needs to create a summary containing four specific segments.
- Evaluation: "Does the summary include the required segments?", with responses "Not at all" (1), "Not much' (2), "So-so' (3), "Pretty much' (4) and "Absolutely' (5)



# ACTIVE VIDEO SUMMARIZATION

- ✓ Is an interactive approach to gather the user's preferences while creating the summary
- ✓ Uses Conditional Random Fields for summary inference
- ✓ Reduces the user interaction by optimizing the expected reward using the previous feedback
- ✓ Strikes a balance between usability and quality of the summary

# SUMMARY OF CONTRIBUTIONS

| | |
|---|---|
| Contextual Event Segmentation | Segment |
| CES for lifelog summarization<br>CRF for video summarization<br>Personalized Highlight Detector | Select |
| Active Video Summarization | Adapt |
| R3<br>CSumm<br>PHD$^2$ | Novel Datasets |

# OPPORTUNITIES FOR FUTURE WORK

- Homogenization of the ground truth for highlight detection
- Emphasis on aesthetics and enjoyable moments
- Exploitation of the stored user-data
- Use of other multimodal cues

# LIST OF PUBLICATIONS ( I)


**A. García del Molino**, J.-H. Lim, and A.-H. Tan, "Predicting visual context for unsupervised event segmentation in continuous photo-streams," in Proceedings of the 26th ACM International Conference on Multimedia, MM '18, pp. 10–17, ACM, 2018.

**A. García del Molino** and M. Gygli, "PHD-GIFs: Personalized highlight detection for automatic GIF creation," in Proceedings of the 26th ACM International Conference on Multimedia, MM '18, pp. 600–608, ACM, 2018.

**A. García del Molino**, X. Boix, J.-H. Lim, and A.-H. Tan, "Active Video Summarization: Customized summaries via on-line interaction with the user.," in AAAI Conference on Artificial Intelligence, pp. 4046–4052, 2017.

**A. García del Molino**, C. Tan, J.-H. Lim and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," in IEEE Transactions on Human-Machine Systems, vol. 47 (1), pp. 65–76, IEEE, 2017.

## LIST OF PUBLICATIONS (II)

**A. García del Molino**, M. Bappaditya, J. Lin, et al., "VC-I2R at ImageCLEF2017: Ensemble of deep learned features for lifelog video summarization," in CLEF working notes, CEUR, 2017.

J. Lin, **A. García del Molino**, Q. Xu, *et al.*, "VC-I2R at the NTCIR-13 lifelog semantic access task," in Proceedings of NTCIR-13, 2017.

**A. García del Molino**, "First Person View video summarization subject to the user needs," in Proceedings of the 24th ACM International Conference on Multimedia, MM '16, (New York, NY, USA), pp. 1440–1444, ACM, 2016. (Doctoral Symposium)

**A. García del Molino**, Q. Xu, and J.-H. Lim, "Describing lifelogs with convolutional neural networks: A comparative study," in Proceedings of the 1st Workshop on Lifelogging Tools and Applications, pp. 39–44, ACM, 2016.

**A. García del Molino**, B. Mandal, L. Li, and L. J. Hwee, "Organizing and retrieving episodic memories from first person view," in International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6, IEEE, 2015.

THANKS!

# Q&A

# CONTEXTUAL EVENT SEGMENTATION: PERFORMANCE OF THE AUTO-ENCODER

**Table A.2** Performance of the auto-encoder's prediction at test time (mean *mse* amplified $\cdot 10^2$, with $N = 1$, $M = T - 1$ and $T = \mathbf{len}[\mathbf{x}]$) for different training configurations of VCP (on R3 dataset).

| trained with N / M : | | 10 / 10 | | 1 / 40 | | 1/100 | 1/1 | 10/1 |
|---|---|---|---|---|---|---|---|---|
| # neurons : | 256 | 512 | 1024 | 512 | 1024 | 1024 | mean* | |
| mse future pred.: | 1.058 | 1.030 | **1.024** | 1.03 | 1.029 | 1.028 | 1.58 | 1.054 |
| mse past pred.: | 1.059 | 1.029 | **1.024** | 1.03 | 1.029 | 1.028 | | |

*The predicted feature corresponds to the average of the previous $N$ frames, *i.e.* $\hat{\mathbf{x}}(t) = \sum_{n=1}^{N} \mathbf{x}(t-n)/N$.

# CAPABILITIES OF CES: FURTHER EXAMPLES

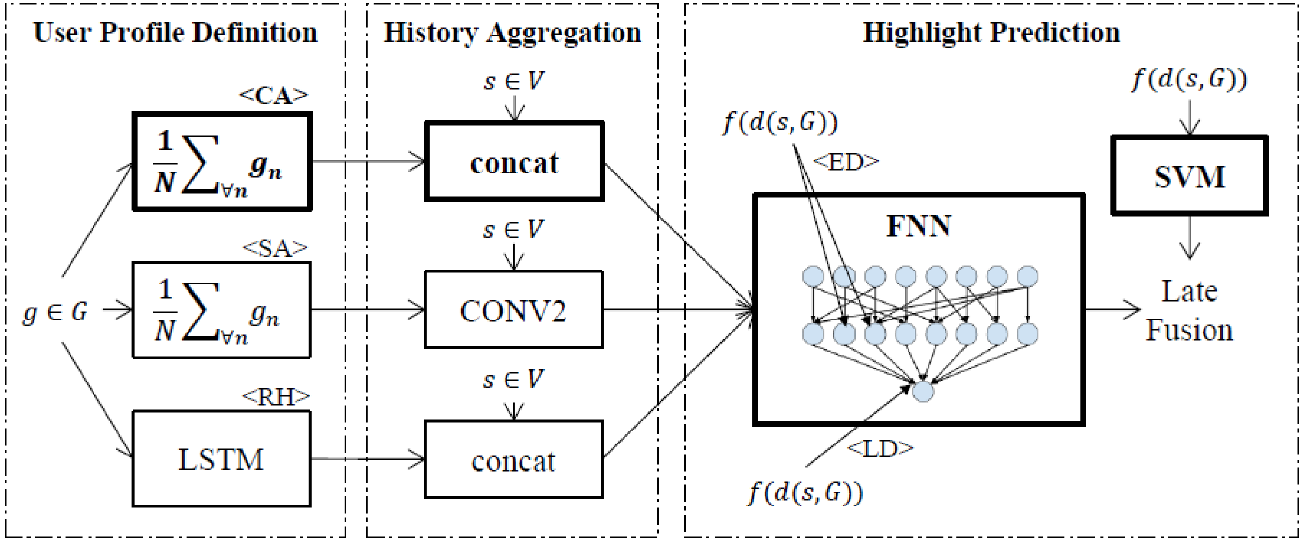

(a) True Positives: CES can model public transportation events, as well as street walking.


(b) True Negatives: CES remembers previously seen context, and is able to match future and past.


(c) False Positives will raise if the different sight positions span longer than CES' memory span.


(d) False Negatives: two events taking place in the same location could be understood as a single one.

# PERSONALIZED HIGHLIGHT DETECTOR: OTHER ARCHITECTURES




| Model | mAP | nMSD | R@5 |
|---|---|---|---|
| PHD-SA | 15.73% | 42.80% | **28.65%** |
| PHD-RH | 15.74% | 42.75% | 27.45% |
| PHD-CA | **16.58%** | **41.01%** | 28.18% |
| PHD-CA-ED (1st layer) | 16.14% | 41.26% | 29.20% |
| PHD-CA-LD (last layer) | 16.20% | 41.07% | 29.78% |
| Video2GIF (ours) + SVM-D | 16.39% | 40.90% | 28.70% |
| **PHD-CA + SVM-D** | **16.68%** | **40.26%** | **30.71%** |

# PERSONALIZED HIGHLIGHT DETECTOR: IMPACT OF LATE FUSION

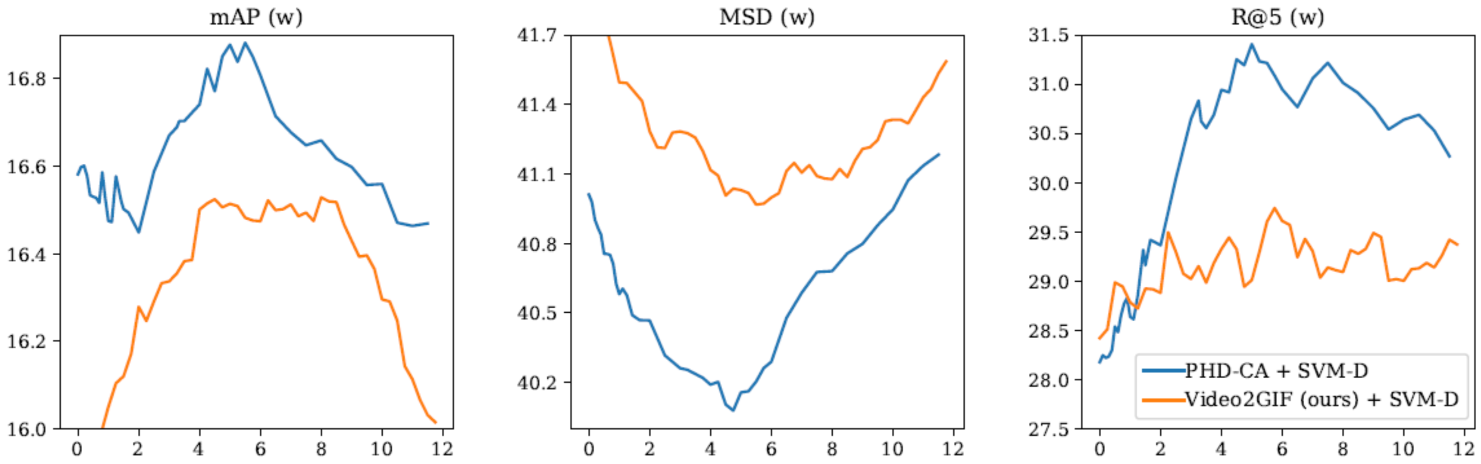



**Figure A.2** Impact of the late fusion weight. Performance of **PHD-CA + SVM-D** and **Video2GIF (ours) + SVM-D** as a function of the late fusion weight. PHD is consistently better than adding the SVM-D model to the baseline.

# VIDEO SUMMARY WITH CRF: USER STUDY

## Video 9

Rate one summary as worst (0) and another one as best (3). You may rate two summaries as equally good/bad (same score).

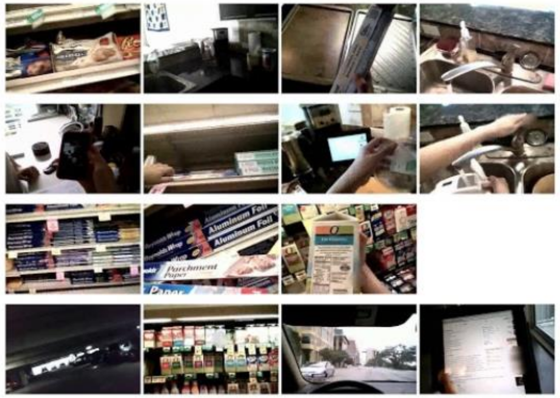
1 - 14hD6a5.jpg
2 - 6CE35sN.jpg
3 - bum6J04.jpg
4 - eex1VTQ.jpg

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 - 14hD6a5 | ○ | ○ | ○ | ○ |
| 2 - 6CE35sN | ○ | ○ | ○ | ○ |
| 3 - bum6J04 | ○ | ○ | ○ | ○ |
| 4 - eex1VTQ | ○ | ○ | ○ | ○ |

## Video 15

Rate one summary as worst (0) and another one as best (3). You may rate two summaries as equally good/bad (same score).

1 - Hn27Vw1.jpg
2 - pTwiu6E.jpg
3 - t0hnMa3.jpg
4 - y4cQiEJ.jpg



| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 - Hn27Vw1 | ○ | ○ | ○ | ○ |
| 2 - pTwiu6E | ○ | ○ | ○ | ○ |
| 3 - t0hnMa3 | ○ | ○ | ○ | ○ |
| 4 - y4cQiEJ | ○ | ○ | ○ | ○ |

# ACTIVE VIDEO SUMMARIZATION: A DEMO

# ACTIVE VIDEO SUMMARIZATION: SEARCH TASK

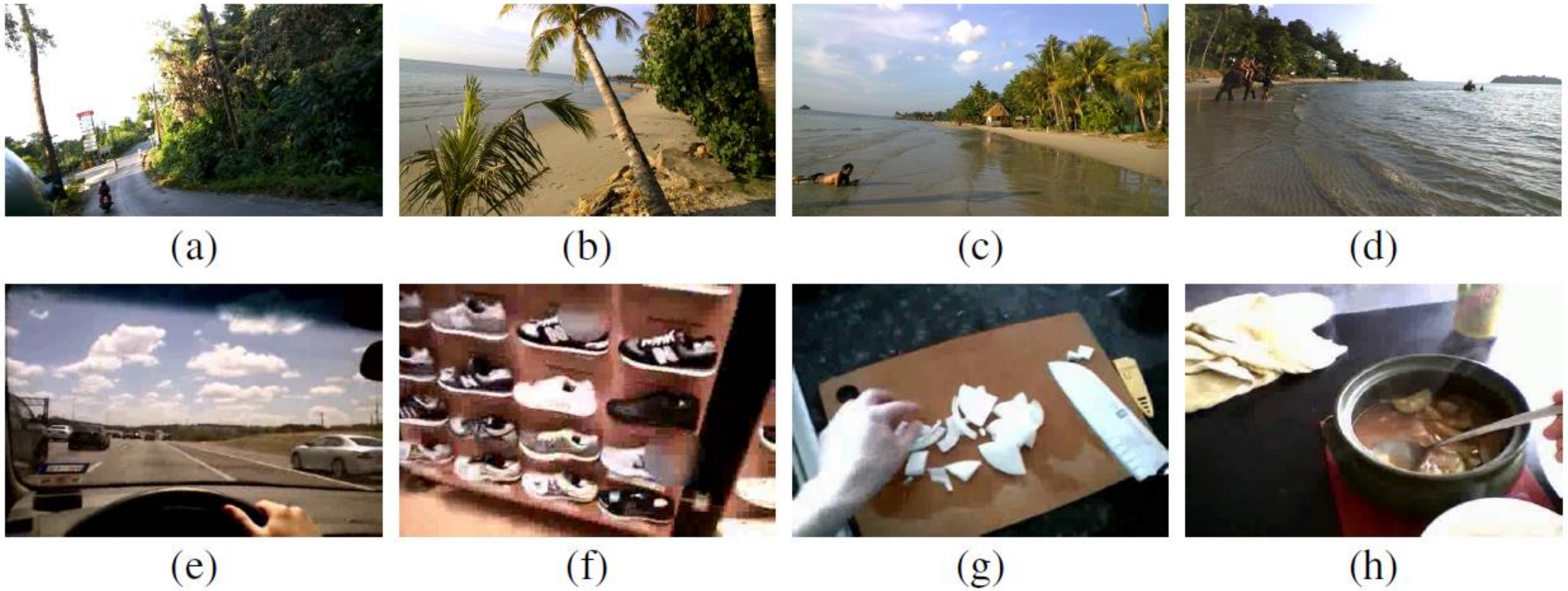


Figure 7.1 Items to be found in Scenario 2 for two example videos. CSumm: (a) Gas station by the road. (b) Beach viewed from the road. (c) Man lying at the shore. (d) Elephants in the water. UTEgo: (e) Driving in highway. (f) Shoe shopping. (g) Chopping vegetables. (h) Serving food.