Enhancing temporal segmentation by nonlocal self-similarity

MARIELLA DIMICCOLI[1], HERWIG WENDT[2]

[1] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2] CNRS, IRIT, University of Toulouse, Toulouse, France

ICIP 2019, Taipei

Institut de Robòtica i Informàtica Industrial

CSIC

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

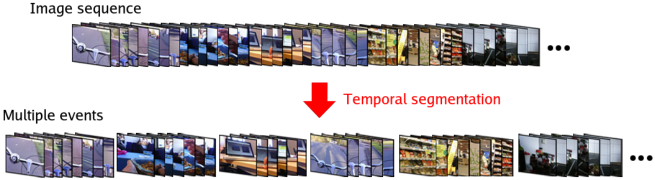
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Institut de Recherche en Informatique de Toulouse

Université Fédérale Toulouse Midi-Pyrénées

Motivation and Goal

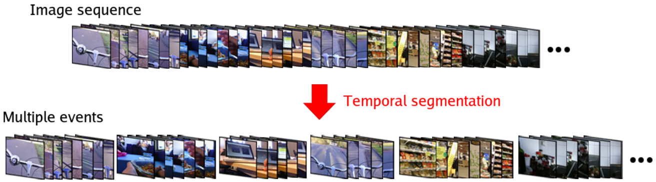
- Temporal segmentation of untrimmed image sequences
  - unconstrained videos (youtube)
  - remotely sensed data (land cover)
  - ...



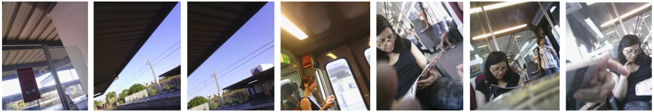
- Egocentric photo stream event segmentation
  - very low frame rate (2fpm)

Motivation and Goal

- Temporal segmentation of untrimmed image sequences
  - unconstrained videos (youtube)
  - remotely sensed data (land cover)
  - ...



- Egocentric photo stream event segmentation
  - very low frame rate (2fpm)



Temporal Segmentation of videos and photostreams

1. Feature extraction
2. Actual segmentation

- **Videos**
  - semantic features
  - motion features
- **Egocentric photostreams**
  → no motion information
  → abrupt appearance changes even in adjacent frames
  - semantic features
  - learnt event representations (NN, LSTM) state-of-the-art [Dias19,Molino18]

Temporal Segmentation of videos and photostreams

1. Feature extraction
2. Actual segmentation

- **Videos**
  - semantic features
  - motion features
- **Egocentric photostreams**
  - → no motion information
  - → abrupt appearance changes even in adjacent frames
  - semantic features
  - learnt event representations (NN, LSTM) state-of-the-art [Dias19,Molino18]

Temporal Segmentation of videos and photostreams

1. Feature extraction
2. Actual segmentation

- **Videos**
  - semantic features
  - motion features
- **Egocentric photostreams**
  - → no motion information
  - → abrupt appearance changes even in adjacent frames
  - semantic features
  - learnt event representations (NN, LSTM) state-of-the-art [Dias19,Molino18]

Temporal Segmentation of videos and photostreams

1. Feature extraction
2. Actual segmentation

- **Videos**
  - semantic features
  - motion features
- **Egocentric photostreams**

  → no motion information

  → abrupt appearance changes even in adjacent frames
  - semantic features
  - learnt event representations (NN, LSTM) state-of-the-art [Dias19,Molino18]

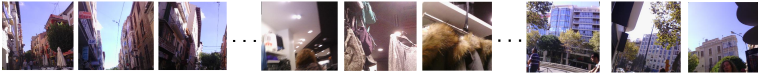
Temporal Segmentation of videos and photostreams

1. Feature extraction → **nonlocal temporal self-similarity**
2. Actual segmentation

- **Videos**
  - semantic features
  - motion features
- **Egocentric photostreams**
  - → no motion information
  - → abrupt appearance changes even in adjacent frames
  - semantic features
  - learnt event representations (NN, LSTM) state-of-the-art [Dias19,Molino18]

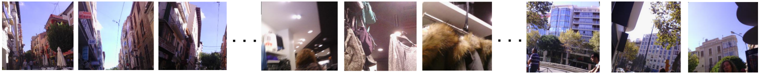
Model assumptions and intuitions

- photostream $\sim$ stationary random process
  - small temporal segment $\rightarrow$ similar segments in same sequence



Model assumptions and intuitions

- photostream $\sim$ stationary random process
  - small temporal segment $\rightarrow$ similar segments in same sequence



- intuitively true for semantic representations
  - e.g. contextual features, objects, concepts, ...

Pedestrian crossing, people, cars, bikes,
traffic light, building, trees

Temporal nonlocal self-similarity: Definition

quantify similarity between a temporal patch centered at k and a temporal patch centered at j

- time $k = 1, \ldots, K$:
  - $u(k) \in \mathbb{R}^P$ - **image feature vector**
  - **temporal patch** $u(\mathcal{N}_k)$
    $\mathcal{N}_k = \{k - M, \ldots, k - 1, k + 1, \ldots, k + M\}$
- **temporal self-similarity function** of $u(k)$

$$S^{NL}(k,j) = \frac{1}{\mathcal{Z}(k)} \exp\left(-\frac{d(u(\mathcal{N}_k), u(\mathcal{N}_j))}{h}\right)$$

- $d(u(\mathcal{N}_k), u(\mathcal{N}_j)) = \sum_{i=1}^{2M} ||u(\mathcal{N}_k(i)) - u(\mathcal{N}_j(i))||^2$
- h: bandwidth parameter
- $\mathcal{Z}(k)$: normalization s.t. $\sum_j S^{NL}(k,j) = 1$
  $\longrightarrow$ conditional probability of $u(j)$ given $u(\mathcal{N}_k)$

Temporal nonlocal self-similarity: Definition

quantify similarity between a temporal patch centered at k
and a temporal patch centered at j

- time $k = 1, \dots, K$:
  - $u(k) \in \mathbb{R}^P$ - **image feature vector**
  - **temporal patch** $u(\mathcal{N}_k)$
    $\mathcal{N}_k = \{k-M, \dots, k-1, k+1, \dots, k+M\}$
- **temporal self-similarity function** of $u(k)$

$$S^{NL}(k,j) = \frac{1}{\mathcal{Z}(k)} \exp\left(-\frac{d(u(\mathcal{N}_k), u(\mathcal{N}_j))}{h}\right)$$

- $d(u(\mathcal{N}_k), u(\mathcal{N}_j)) = \sum_{i=1}^{2M} ||u(\mathcal{N}_k(i)) - u(\mathcal{N}_j(i))||^2$
- h: bandwidth parameter
- $\mathcal{Z}(k)$: normalization s.t. $\sum_j S^{NL}(k,j) = 1$
  $\longrightarrow$ conditional probability of $u(j)$ given $u(\mathcal{N}_k)$

Nonlocal temporal self-similarity features

- replace features $u(k)$ with new set of N nonlocal features

$$u^{NL}(k) = \{S^{NL}(k,j)\}_{j=k\pm 1,2,\ldots} \in \mathbb{R}^N,$$

($N = K - 1 \rightarrow$ similarity with all other temporal patches)

- similarity of $u^{NL}(k)$ and $u^{NL}(k')$:
  - large if k and k' belong to the same event
  - small if k and k' belong to two different events

$\rightarrow$ suitable for temporal segmentation

Nonlocal temporal self-similarity features

- replace features $u(k)$ with new set of N nonlocal features

$$u^{NL}(k) = \{S^{NL}(k,j)\}_{j=k\pm 1,2,\dots} \in \mathbb{R}^N,$$

$(N = K - 1 \rightarrow$ similarity with all other temporal patches$)$

- similarity of $u^{NL}(k)$ and $u^{NL}(k')$:
  - large if k and k' belong to the same event
  - small if k and k' belong to two different events

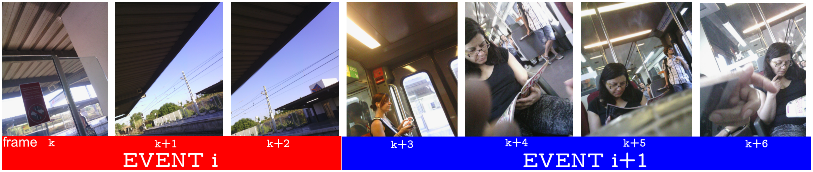
$\rightarrow$ suitable for temporal segmentation

Dataset and performance evaluation

- EDUB-Seg dataset:
  - wearable photo-camera image sequences (2 fpm)
  - subset of ten sequences for five different users
  - ground truth event segmentation

Dataset and performance evaluation

- EDUB-Seg dataset:
  - wearable photo-camera image sequences (2 fpm)
  - subset of ten sequences for five different users
  - ground truth event segmentation
- Event segmentation performance:
  - structured hierarchical clustering algorithm
  - F-measure (tolerance ± 5 frames)
  - number of temporal segments chosen to maximize F-measure



Features for event segmentation

- Local state of the art features:
  - Concept vectors: CNN-based indicator vectors for concepts detected in images
  - NNF: simple feed-forward NN autoencoder
  - NNFB: forward-backward NN autoencoder
    temporal depths $n = 1, 2, 3, 4$
  - LSTM: LSTM autoencoder

- Nonlocal self-similarity features
  - temporal patch size ± 2 frames
  - 6 main principal components used

Features for event segmentation

- Local state of the art features:
  - Concept vectors: CNN-based indicator vectors for concepts detected in images
  - NNF: simple feed-forward NN autoencoder
  - NNFB: forward-backward NN autoencoder
    temporal depths $n = 1, 2, 3, 4$
  - LSTM: LSTM autoencoder

- Nonlocal self-similarity features
  - temporal patch size ± 2 frames
  - 6 main principal components used

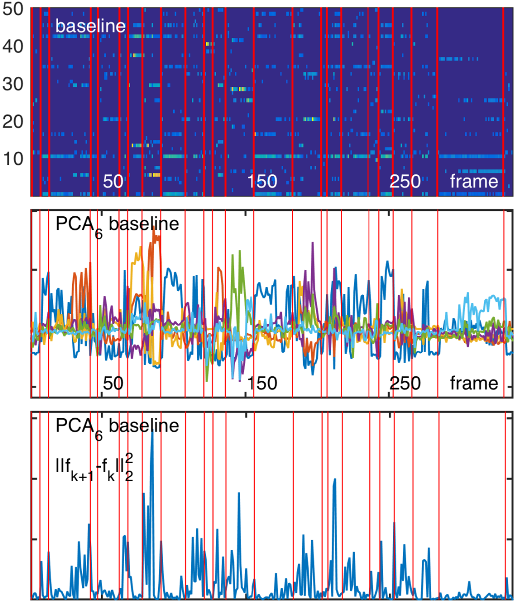
Segmentation performance for a single user

concept vectors

features

PCA comp. 1-6

L2 norm of difference
between neighbors



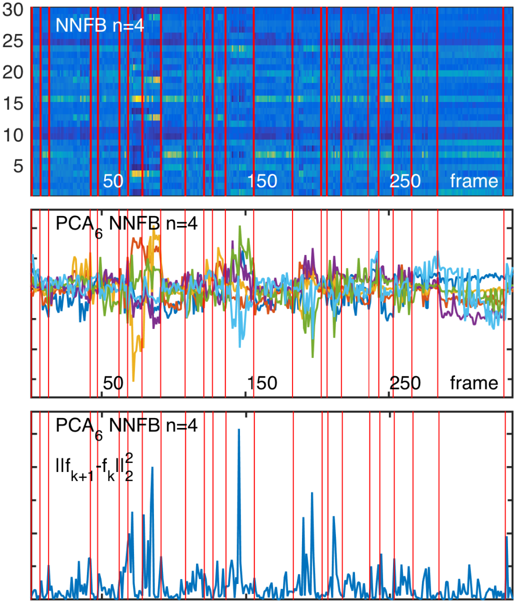
Segmentation performance for a single user

NNFB n=4

features

PCA comp. 1-6

L2 norm of difference
between neighbors



Segmentation performance for a single user

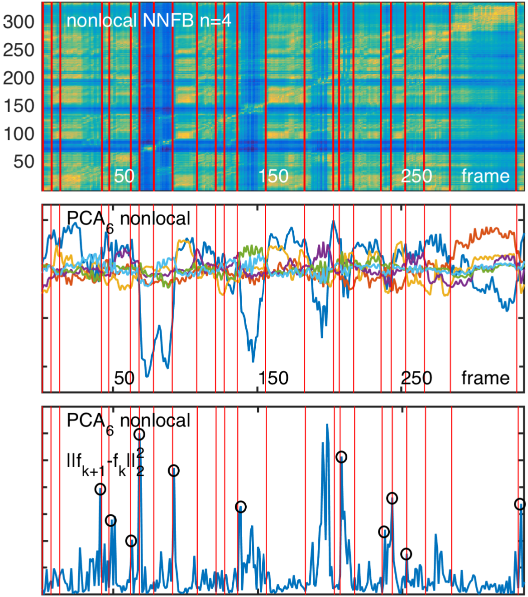
NNFB n=4 **nonlocal**

features

PCA comp. 1-6

L2 norm of difference between neighbors

→ peaks align with true event boundaries



Average segmentation performance

F-measure

	concept vectors	NNF $n=1$	NNFB $n=1$	NNFB $n=2$	NNFB $n=3$	NNFB $n=4$	LSTM $n=1$
L	0.46	0.50	0.54	0.51	**0.56**	0.49	0.53
NL	**0.58**	**0.52**	**0.59**	**0.54**	0.52	**0.54**	**0.56**
Diff.	+0.12	+0.03	+0.05	+0.04	−0.05	+0.04	+0.03

- NL features

→ average improvement of up to 12%

→ better temporal segmentations also for each user individually

Average segmentation performance

F-measure

	concept vectors	NNF $n = 1$	NNFB $n = 1$	NNFB $n = 2$	NNFB $n = 3$	NNFB $n = 4$	LSTM $n = 1$
L	0.46	0.50	0.54	0.51	**0.56**	0.49	0.53
NL	**0.58**	**0.52**	**0.59**	**0.54**	0.52	**0.54**	**0.56**
Diff.	$+0.12$	$+0.03$	$+0.05$	$+0.04$	-0.05	$+0.04$	$+0.03$

- NL features

$\rightarrow$ average improvement of up to 12%

$\rightarrow$ better temporal segmentations also for each user individually

Final remarks

- Method to enhance temporal segmentation by nonlocal self-similarity:
  - improves feature representations
  - based on the nonlocal similarity between temporal patches
- Validated on unconstrained image sequence:
  - EDUB-Seg dataset
  - nonlocal representations $\longrightarrow$ consistent performance improvements
- How to next use nonlocal self-similarity within a neural network based learning framework

Final remarks

- Method to enhance temporal segmentation by nonlocal self-similarity:
  - improves feature representations
  - based on the nonlocal similarity between temporal patches
- Validated on unconstrained image sequence:
  - EDUB-Seg dataset
  - nonlocal representations $\longrightarrow$ consistent performance improvements
- How to next use nonlocal self-similarity within a neural network based learning framework

Final remarks

- Method to enhance temporal segmentation by nonlocal self-similarity:
  - improves feature representations
  - based on the nonlocal similarity between temporal patches
- Validated on unconstrained image sequence:
  - EDUB-Seg dataset
  - nonlocal representations $\longrightarrow$ consistent performance improvements
- How to next use nonlocal self-similarity within a neural network based learning framework

https://www.iri.upc.edu/people/mdimiccoli/

https://github.com/mdimiccoli/Nonlocal-self-similarity-1D

Bibliography

- [Dias19] C. Dias et al., *Learning event representations by encoding the temporal context*, Proc. ECCV Workshops, 2019.
- [Molino18] A. Garcia del Molino et al., *Predicting visual context for unsupervised event segmentation in continuous photo-streams*, ACM Multimedia Conference, 2018.
- [Tracey12] B. H. Tracey et al., *Nonlocal means de- noising of ECG signals*, IEEE T. Biomedical Engineering, 2012.
- [Efros99] A. A. Efros et al., *Texture synthesis by non-parametric sampling*, in Proc. IEEE ICCV, 1999.
- [Buades05] A. Buades et al., *A non-local algorithm for image denoising*, in Proc. IEEE CVPR, 2005.
- [Dimiccoli09] M. Dimiccoli et al., *Hierarchical region-based representation for segmentation and filtering with depth in single images*, in Proc. IEEE ICIP, 2009.
- [Dimiccoli17] M. Dimiccoli et al., *Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation*, Computer Vision and Image Understanding, 2017.

Related work: nonlocal self-similarity

local (neighbor frames) $\rightarrow$ **nonlocal** temporal context

- Nonlocal means for image denoising [Buades05]
  $\rightarrow$ image contains many similar patches (upon transformation)

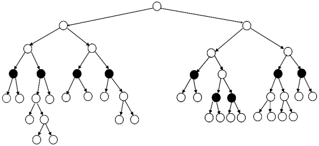
- Spatial nonlocal self-similarity for image segmentation [Dimiccoli09]
  $\rightarrow$ model each pixel by a conditional probability density
  $\rightarrow$ hierarchical segmentation

little explored for time series [Tracey12]

$\longrightarrow$ **use to improve event representation for temporal segmentation**

Structured hierarchical clustering algorithm [Dias19]

- Hierarchical partition
  - finest level $\rightarrow$ initial frames
  - root node $\rightarrow$ entire image sequence



- Tree construction:
  - ascending:
  - join temporally neighboring nodes with smallest distance
  - frame union modeled as average
  - distance = Euclidean norm