Dense saliency-based spatiotemporal feature points for action recognition

Kostas Rapantzikos, Yannis Avrithis and Stefanos Kollias

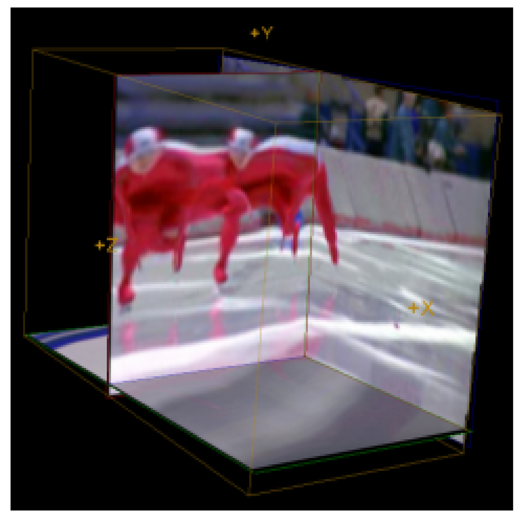
National Technical University of Athens, School of Electrical and Computer Engineering

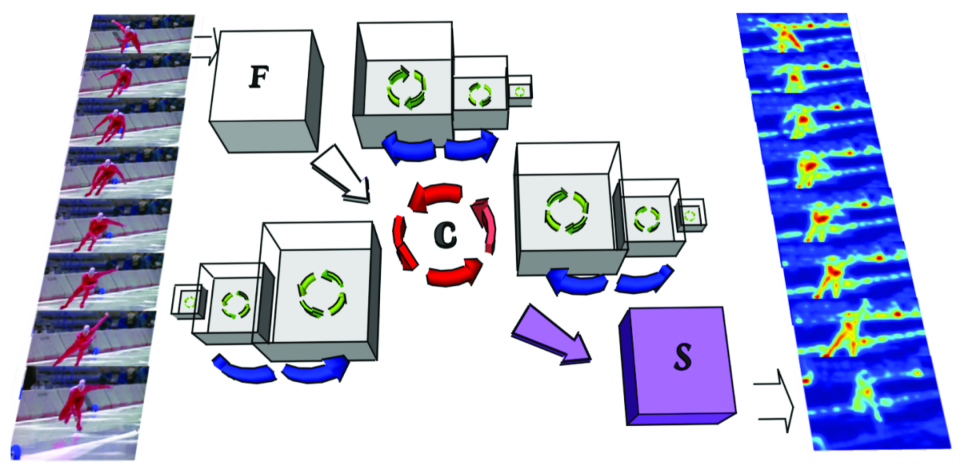
Motivation & Approach

Derive a spatiotemporal point detector that is based on richer information than cornerness (points of velocity change), periodicity or motion activity measures.

- Represent the video as a volume in space-time
- Define a saliency measure for each voxel using spatial proximity, scale and feature conspicuity
- Detect space-time-scale maxima of the saliency distribution

Problem

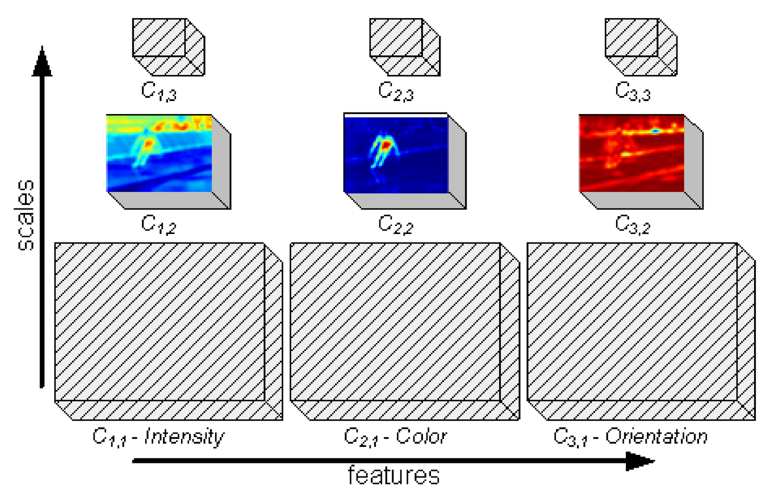




- V: video volume defined on a set of points Q with $q = (x, y, t)$ being a space-time point (voxel).
- $\mathbf{F} = \{F_i\}$: set of feature volumes (intensity, color, motion)
- $\mathbf{C} = \{C_{i,\ell}\}$: set of conspicuity volumes, each one decomposed into scales ℓ
- S: final saliency distribution

Energy formulation

- Initial intensity & color conspicuity obtained by color opponent theory
- Orientation conspicuity computed using spatiotemporal steerable filters tuned to respond to moving stimuli
  - Conspicuity volumes interact in order to produce a single saliency measure for each voxel
  - Each voxel interacts with its space-time neighborhood at the same scale, at neighboring scales and at the rest of the conspicuity volumes



- Competition is realized by minimizing energy E:

$$E(\mathbf{C}) = \lambda_d \cdot E_d(\mathbf{C}) + \lambda_s \cdot E_s(\mathbf{C})$$

- composed of *data term* E_d:

$$E_d(\mathbf{C}) = \sum_i \sum_l \sum_q (C_{i,\ell}(q) - C^0_{i,\ell}(q))^2,$$

- and *smoothness term* E_s:

$$E_s(\mathbf{C}) = E_1(\mathbf{C}) + E_2(\mathbf{C}) + E_3(\mathbf{C}).$$

Constraints

- **intra-feature** E_1: defines the interaction among neighboring voxels of the same feature at the same scale and enhances voxels that are incoherent with their neighborhood

$$E_1(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{|N_q|} \sum_{r \in N_q} C_{i,\ell}(r) \right)^2$$

- **inter-feature** E_2: defines the interaction among different features so that voxels being conspicuous across all feature volumes become salient

$$E_2(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{M-1} \sum_{j \neq i} C_{j,\ell}(q) \right)^2$$

- **inter-scale** E_3: defines the interaction across different scales. If a voxel retains a high value along all scales, then it should become more salient.

$$E_3(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{L-1} \sum_{n \neq l} C_{i,n}(q) \right)^2$$

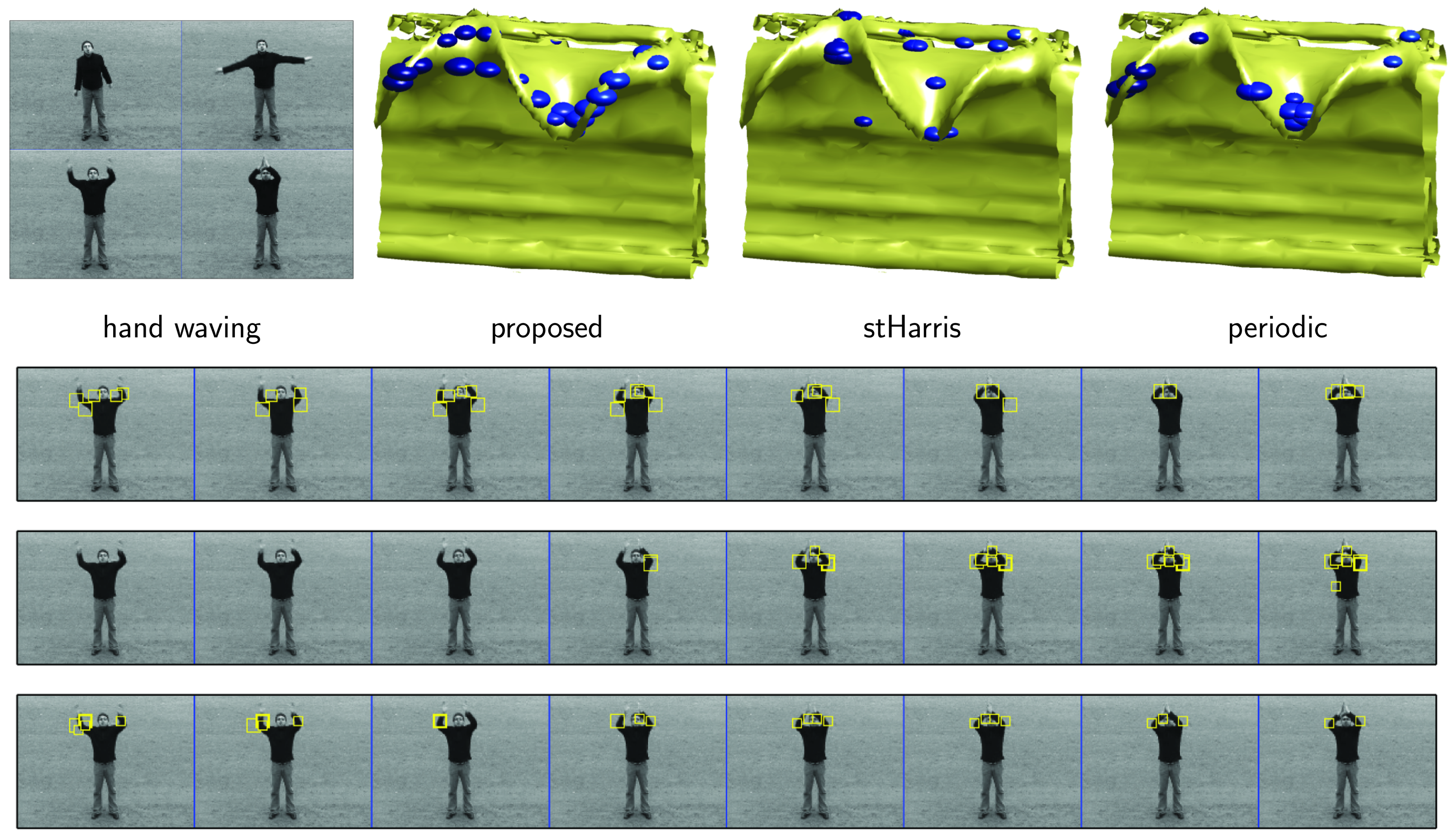
Spatiotemporal saliency and feature points

The final saliency distribution is obtained by minimizing the following energy

$$\frac{\partial E(\mathbf{C})}{\partial C_{k,m}(s)} = \lambda_d \cdot \frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} + \lambda_d \cdot \frac{\partial E_s(\mathbf{C})}{\partial C_{k,m}(s)}$$

The output is a set of modified conspicuity multi-scale volumes $\hat{\mathbf{C}} = \{\hat{C}_{i,\ell}\}$ and saliency is computed as $\mathbf{S} = \{S_\ell\} = \frac{1}{M} \cdot \sum_{i=1}^{M} \hat{C}_{i,\ell}$

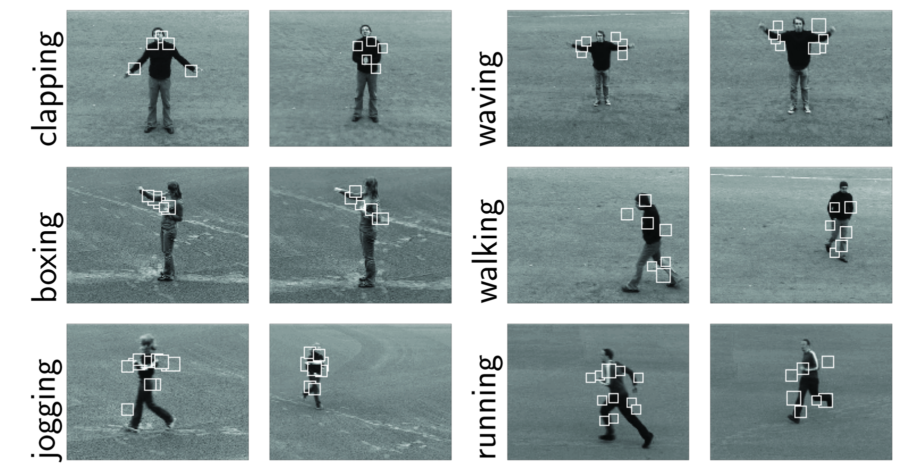
- Feature points are extracted as the local maxima of the response
- Detected points are located at regions that exhibit high compactness (*proximity*), are consistent across scales (*scale*) and pop-out from their surroundings (*conspicuity*)

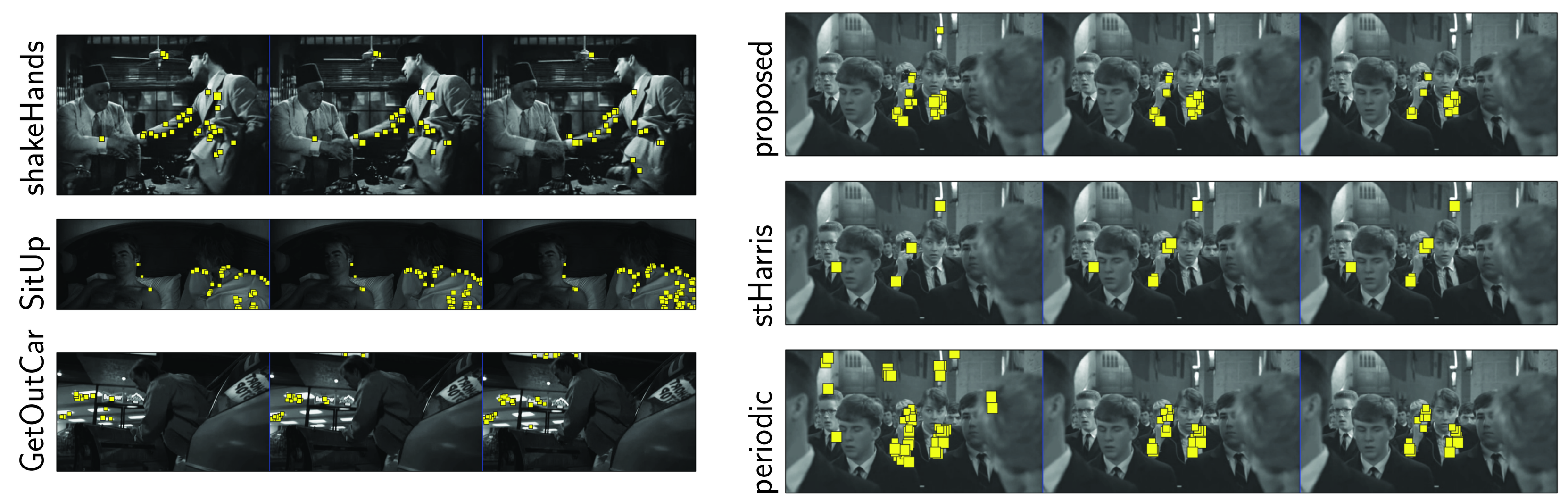


Experiments

Datasets

- **KTH**: six types of human actions (handclapping, handwaving, boxing, walking, jogging and running)
  `http://www.nada.kth.se/cvap/actions/`
- **HOHA**: video samples with human actions from 32 movies (AnswerPhone,GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp and StandUp)
  `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`





Results

Method	Accuracy	Classifier
Schuldt *et al.* (reported in Wong *et al.*)	26.95%	NNC
Schuldt *et al.* (implemented by us)	50.33%	NNC
Oikonomopoulos *et al.* (reported in Wong *et al.*)	64.79%	NNC
Wong *et al.*	80.06%	NNC
Dollár *et al.* (implemented by us)	79.83%	NNC
Dollár *et al.*	81.20%	NNC
Ours	**88.30%**	NNC
Ke *et al.*	80.90%	SVM
Schuldt *et al.*	71.70%	SVM
Niebles *et al.*	81.50%	pLSA
Willems *et al.*	84.36%	SVM
Jiang *et al.*	84.40%	LPBOOST
Laptev *et al.*	91.80%	mc-SVM

ours

	boxing	handclapping	handwaving	jogging	running	walking
boxing	.71	.24	.04	.00	.00	.01
handclapping	.04	.94	.02	.00	.00	.00
handwaving	.02	.03	.95	.00	.00	.00
jogging	.00	.00	.00	.90	.10	.00
running	.00	.00	.00	.10	.90	.00
walking	.00	.00	.00	.10	.00	.90

ours

	SitUp	StandUp	GetOutCar	AnswerPhone	Kiss	SitDown	HugPerson	HandShake
SitUp	.27	.09	.08	.13	.13	.16	.01	.14
StandUp	.03	.39	.03	.10	.09	.17	.07	.11
GetOutCar	.04	.14	.27	.09	.28	.12	.01	.06
AnswerPhone	.03	.14	.04	.32	.16	.17	.03	.11
Kiss	.08	.13	.04	.06	.35	.12	.13	.09
SitDown	.09	.07	.05	.11	.12	.37	.04	.15
HugPerson	.04	.12	.03	.01	.13	.14	.40	.13
HandShake	.13	.17	.04	.07	.15	.09	.02	.32

Conclusions

- A more descriptive spatiotemporal feature point detector
- Compares well to state-of-the-art detectors
- Future work on computational efficiency and more advanced descriptors