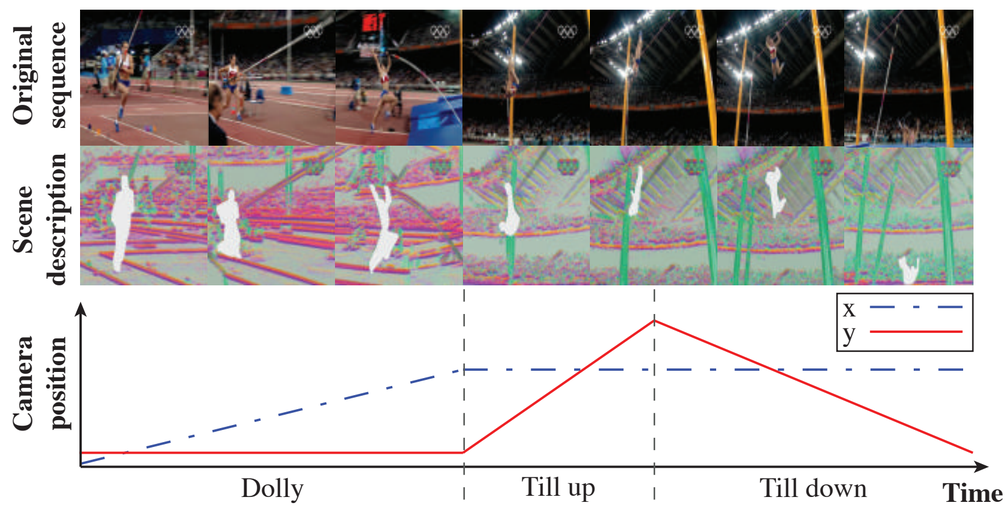
Camera Motion and Surrounding Scene Appearance as Context for Action Recognition

Fabian Caba Heilbron[12], Ali Thabet[1], Juan Carlos Niebles[2], and Bernard Ghanem[1]
[1] King Abdullah University of Science and Technology (KAUST); [2] Universidad del Norte (Colombia)

SUMMARY

This work introduces a framework for recognizing human actions by incorporating a new set of visual cues that represent the context of the action:

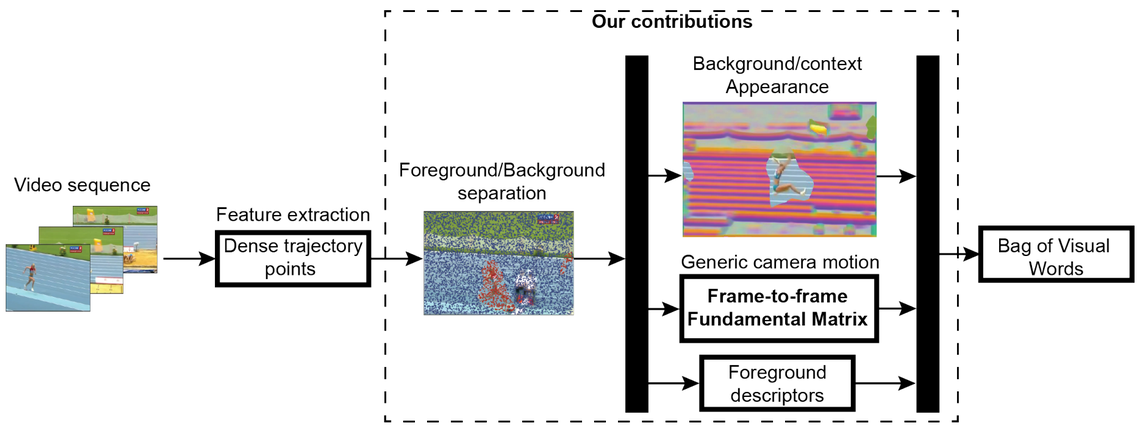


CONTRIBUTIONS

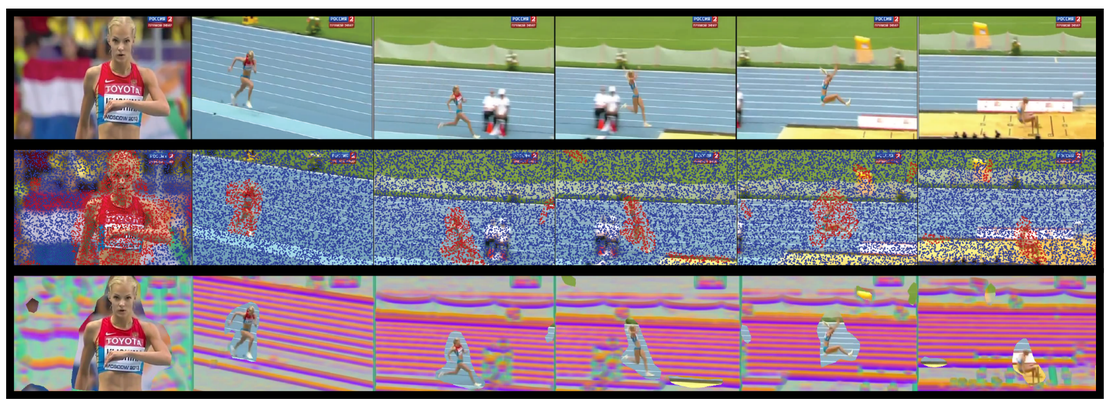
- Weak foreground-background segmentation approach.
- Study of the global camera motion as a cue for action recognition.
- Incorporating appearance from static background.

METHODOLOGY

This work follows the conventional action recognition pipeline. Given a set of labeled videos, a set of features is extracted from each video, represented using visual descriptors, and combined into a single video descriptor used to train a multi-class classifier for action recognition.



- **Foreground-background separation**: Assuming that a background trajectory produces a small frame-to-frame displacement, we associate a trajectory with the background if the overall displacement is more than three pixels.



- **Global camera motion**: We argue and show that the relationship between an estimated camera motion and underlying action can be a useful cue for discriminating certain action classes. As illustrated in the figure below, there is a correlation between how the camera moves and the actor.

- **Background-context appearance**: Beyond local motion and appearance properties, the surrounding in which an action is performed is a critical component to recognize actions. As Figure below illustrates, the background appearance plays an important role to discriminate the action Drumming in the sense that the drummer needs a drum set to perform the action.

- **Implementation details**: We follow two different Bag Of Feature implementations as described in the Table below.

Model	Codebook	Code	Normalization	Classifier
Hard enoding	K-means	VQ	L2	Non-linear SVM
Fisher vectors	GMM	Fisher kernel	L2+PW+IN	Linear SVM

EXPERIMENTAL RESULTS

- **Datasets:** We use state-of-the-art human action datasets and their corresponding protocols.

- **Impact of contextual features**: We note that using Fisher vectors consistently boost the performance of our contextual features. Also, our experiments provide evidence that action recognition performance can be improved when static background appearance and global camera motion is incorporated with foreground features.

- **Comparison with the state-of-the-art**: We set side by side our method with recent methods that address the same application using similar representations, i.e. methods that use dense trajectory points to represent video sequences [2,3,4] in the Table below.

Approach	HMDB51	Hollywood2	Olympic	UCF50
Jiang et al. [3]	40.7%	59.5%	80.6%	N.A.
Jain et al. [4]	52.1%	62.5%	83.2%	N.A.
Wang et al. Non-HD [2]	55.9%	63.0%	90.2%	90.5%
Wang et al. HD [2]	57.2%	64.3%	91.1%	91.2%
Ours: baseline	56.5%	**62.4%**	90.4%	90.9%
Ours: Foreground + SIFT	**59.2%**	63.5%	**91.6%**	**93.3%**
Ours: Foreground + SIFT + CamMotion	**57.9%**	**64.1%**	**92.5%**	**93.8%**

DISCUSSIONS

- **Contextual features:** When combined with foreground trajectories, we show that these features, can improve state-of-the-art recognition on challenging action datasets.

- **Project page:** http://www.cabaf.net/actioncue

References:

[1] Fabian Caba Heilbron, Ali Thabet, Juan Carlos Niebles, Bernard Ghanem. Camera Motion and Surrounding Scene Appearance as Context for Action Recognition. ACCV, Singapore 2014.

[2] Wang, H., Schmid, C. Action recognition with improved trajectories. ICCV, Sydney 2013.

[3] Jiang, Y.G., Dai, Q., Xue, X., Liu, W., Ngo, C.W. Trajectory-based modeling of human actions with motion reference points. ECCV, 2012.

[4] Jain, M., J ́egou, H., Bouthemy, P.: Better exploiting motion for better action recognition.

Acknowledgements: Research reported in this publication was supported by competitive research funding from KAUST. F.C.H. was also supported by COLCIENCIAS. J.C.N. is supported by a Microsoft Research Faculty Fellowship.