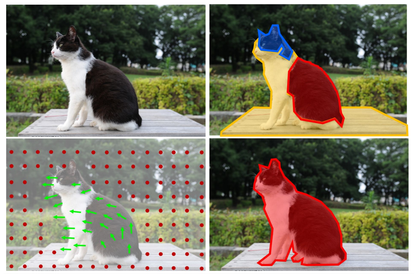
Disentangling Motion, Foreground and Background Features in Videos

Xunyu Lin, Víctor Campos, Xavier Giró-i-Nieto, Jordi Torres, Cristian Canton Ferrer

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

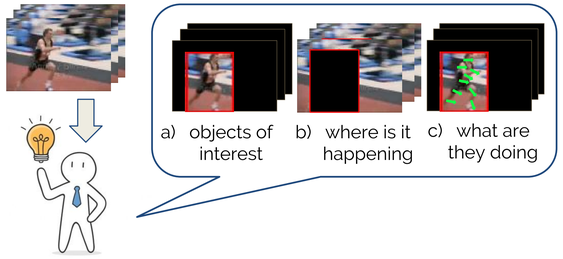
Motivation

Infants tend to group foreground objects by observing motion cues [1].



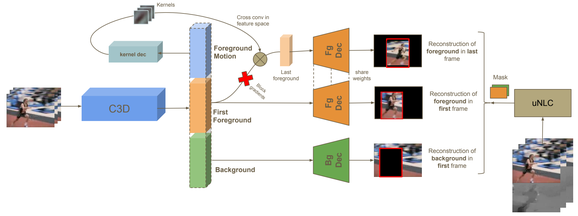
[1] D. Pathak, R. Girshick, P. Dollar, T. Darrell, and B. Hariharan. **Learning features by watching objects move.** In CVPR, 2017

Hypothesis: humans summarize videos by decomposing foreground, background and foreground motion.



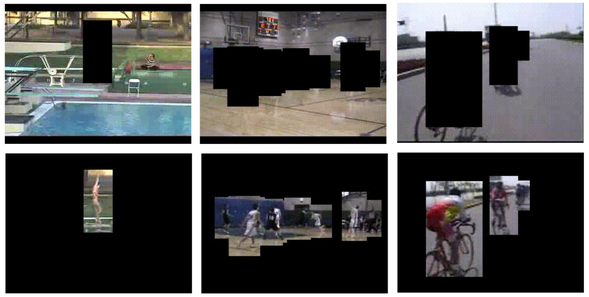
MFB-Net

MFB-Net is proposed to disentangle foreground, background and foreground motion features in videos.



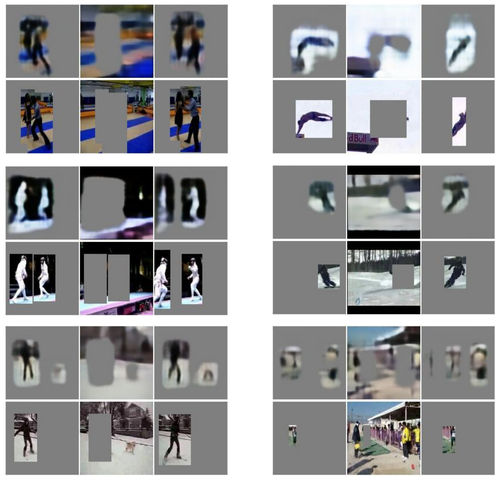
Dataset

UCF-24 with action localization annotations

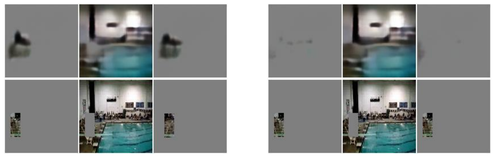


Reconstruction

Reconstruction results on test set. (first row: prediction, second row: ground truth)

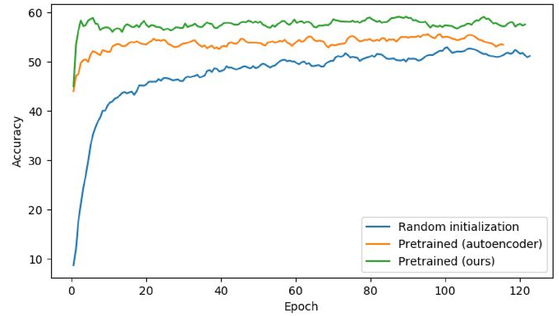


Reconstruction when motion is removed. (left: original input, right: motion removed)



Action Recognition

Action recognition accuracy on validation set with different initialization schemes.



Action recognition accuracy on test set with different initialization schemes.

Method	Accuracy
Random initialization	52.2%
Pretrained (autoencoder)	56.8%
Pretrained (ours)	**62.5%**

Model and source code:
https://github.com/imatge-upc/unsupervised-2017-cvprw

Acknowledgements

Generalitat de Catalunya
Obra Social Fundació "la Caixa"
Gobierno de España
Ministerio de Economía y Competitividad
NVIDIA