Im2Flow: Motion Hallucination from Static Images for Action Recognition

TEXAS
The University of Texas at Austin

Ruohan Gao Bo Xiong Kristen Grauman

The University of Texas at Austin

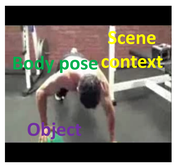
Project page: http://vision.cs.utexas.edu/projects/im2flow

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Static-image Action Recognition

Static-image action recognition exploits various high-level appearance cues such as human body pose/ scene context and objects in the image:

[Thurau & Hlavác, CVPR 2008; Delaitre *et al.*, NIPS 2011; Sener *et al.*, ECCV 2012; Gkioxari *et al.*, CVPR 2015]



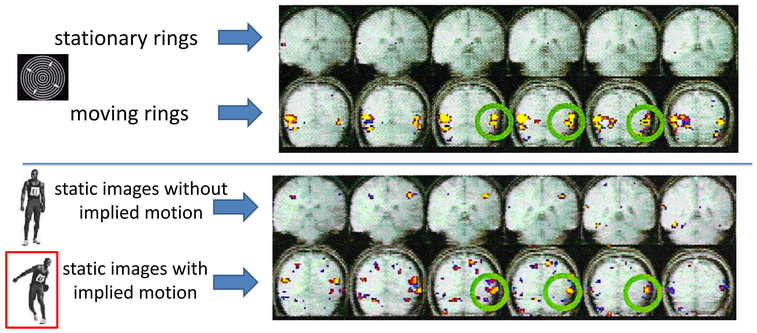
Video-level action recognition methods exploits both appearance and motion:

[Wang & Schmid CVPR 2013; Simonyan & Zisserman, NIPS 2014; Wang *et al.*, ECCV 2016; Girdhar *et al.*, CVPR 2017; Tran *et al.*, CVPR 2018; Feichtenhofer *et al.*, CVPR 2018]

But is motion really absent in static images?

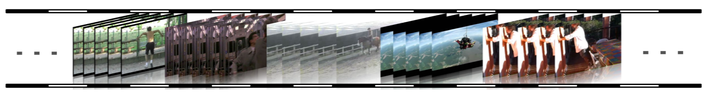
Implied Motion Perception in the Brain

Activation in human's medial temporal / medial superior temporal (MT/MST) cortex by static images with implied motion [Kourtzi & Kanwisher, 2000]:

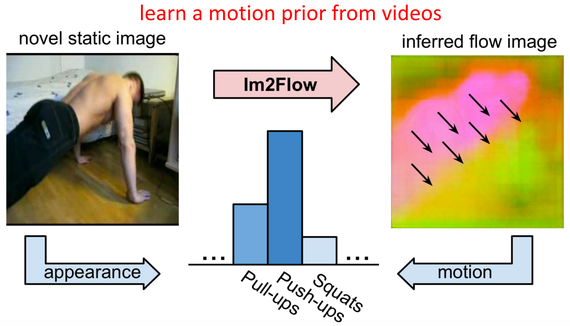


Our Idea

We propose to learn a motion prior from unlabeled videos, and hallucinates motion implied by a single snapshot to help static-image action recognition.



learn a motion prior from videos

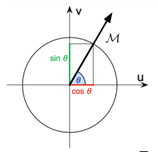


We formulate motion prediction as a novel image-to-image translation framework, and use predicted motion image to aid action recognition.

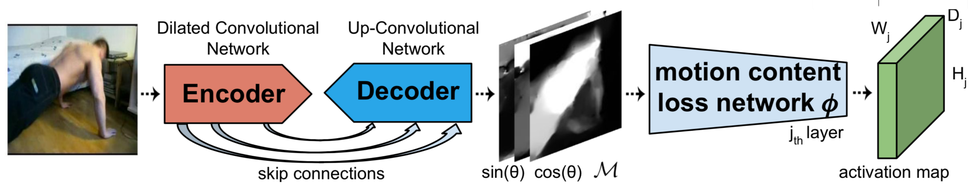
Approach

Motion Encoding: Detangle flow direction and strength, and encodes as a single 3-channel flow image.

$$\mathcal{F}_1 = \sin(\theta) = \frac{v}{\mathcal{M}}; \quad \mathcal{F}_2 = \cos(\theta) = \frac{u}{\mathcal{M}}; \quad \mathcal{F}_3 = \mathcal{M}$$



Im2Flow Network Architecture: An encoder-decoder network, which takes a static image as input and outputs the predicted flow image.

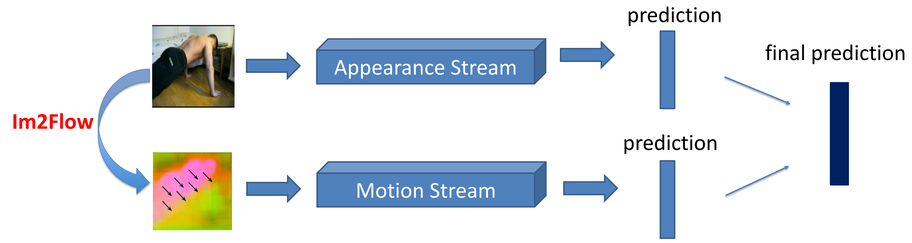


Total Loss: $L = L_{pixel} + \lambda L^{\phi,j}_{content}$

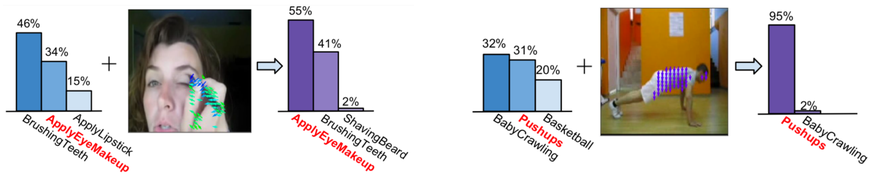
Pixel Error Loss: $L_{pixel} = \mathbb{E}_{p,q\in\{x_i,y_i\}_{i=1}^N}[||y_i - G(x_i)||_2]$

Motion Content Loss: $L^{\phi,j}_{content} = \frac{1}{D_j \times H_j \times W_j}\mathbb{E}_{p,q\in\{x_i,y_i\}_{i=1}^N}[||\phi_j(y_i) - \phi_j(G(x_i))||_2]$

Static-image Action Recognition: We adopt the popular and effective two-stream CNN [Simonyan & Zisserman, NIPS 2014] to inject our Im2Flow predictions into action recognition.



How does inferred motion help recognition?



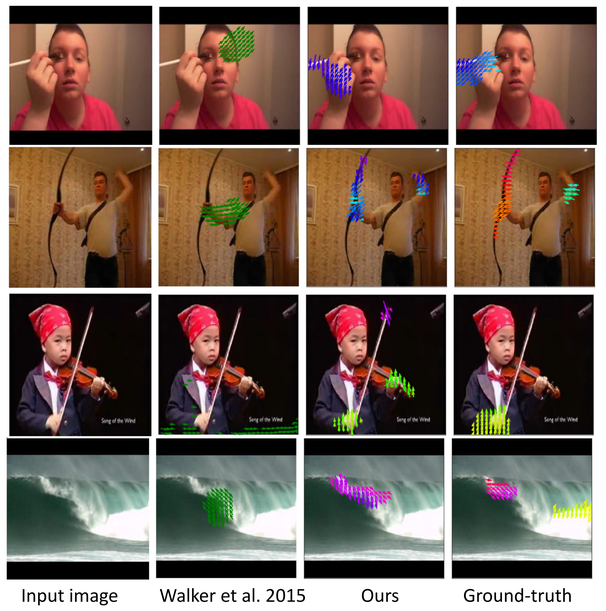
While a classifier solely based on appearance can be confused by actions appearing in similar contexts, the inferred motion provides cues about the fine-grained differences to aid recognition.

Motion Potential: the strength of movement and activity that is poised to happen in a static image.

Motion potential offers a high-level view of a scene's activity, identifying images that are most suggestive of coming events.

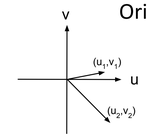
Results

Flow prediction example qualitative results:



Flow prediction quantitative results:

Metrics: End-Point-Error (EPE), Direction Similarity (DS) Orientation Similarity (OS)



$$DS = \frac{u_1u_2 + v_1v_2}{\sqrt{u_1^2+v_1^2}\sqrt{u_2^2+v_2^2}}, \; OS = \frac{|u_1u_2 + v_1v_2|}{\sqrt{u_1^2+v_1^2}\sqrt{u_2^2+v_2^2}}.$$

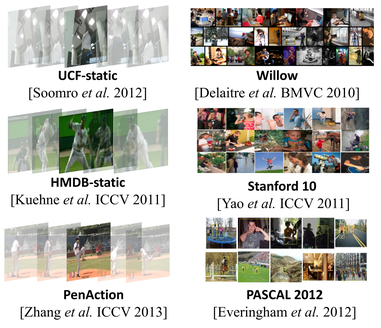
Evaluate prediction results over
- all pixels in the whole image
- masks on canny edges
- masks on foreground (FG) regions

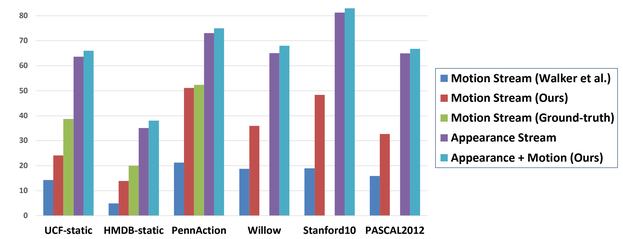
UCF-101	EPE ↓	EPE-Canny	EPE-FG	DS ↑	DS-Canny	DS-FG	OS ↑	OS-Canny	OS-FG
Pintea *et al.* 2014	2.401	2.699	3.233	-0.001	-0.002	-0.005	0.513	0.544	0.555
Walker *et al.* 2015	2.391	2.696	3.139	0.003	0.001	0.014	0.661	0.673	0.662
Nearest Neighbor	3.123	3.234	3.998	-0.002	-0.001	-0.023	0.652	0.651	0.659
Ours	**2.210**	**2.533**	**2.936**	**0.143**	**0.135**	**0.137**	**0.699**	**0.692**	**0.696**

End-Point-Error (lower better) | Direction Similarity (higher better) | Orientation Similarity (higher better)

Results on HMDB-51 and Weizmann datasets are similar. Across all metrics, our method predicts more accurate optical flow.

Static-image Action Recognition





Inferring motion from Im2flow improves the recognition accuracy for static-image action recognition.