Supplementary Material for FAR: Fourier Aerial Video Recognition

Divya Kothandaraman[1][0000−0002−6276−4968], Tianrui Guan[2,3][0000−0002−6892−9778], Xijun Wang[3], Shuowen Hu[3], Ming Lin[3][0000−0003−3736−6949], and Dinesh Manocha[3][0000−0001−7047−9801]

[1] University of Maryland College Park, United States
[2] Army Research Laboratory, United States
dkr@umd.edu
https://gamma.umd.edu/far

1 Datasets

We describe the UAV datasets used for evaluating FAR.

UAV Human RGB [7]: UAV Human is the largest UAV-based human behavior understanding dataset. Split 1 contains 15172 and 5556 images for training and testing respectively. This challenging dataset covers human actions captured under varying illumination, time of day (daytime, nighttime), different subjects and backgrounds, weathers, occlusions, etc, across 155 diverse human actions. UAV Human RGB is collected by drones with an Azure Kinect DK camera. The videos are of resolution 1920×1080. The dataset is available at https://sutdcv.github.io/uav-human-web/.

UAV Human Night Camera [7]: UAV Human Night Camera contains videos similar to UAV Human RGB captured using a night-vision camera. The night vision camera captures videos in color mode in the daytime, and grey-scale mode in the nighttime. The resolution of the videos is 640×480. The dataset is available at https://sutdcv.github.io/uav-human-web/.

Drone Action [8]: Drone Action is an outdoor drone video dataset captured using a free flying drone. It has 240 HD RGB videos with 66919 frames, across 13 human actions. The dataset is available at https://asankagp.github.io/droneaction/.

NEC Drone [3]: NEC Drone dataset is an indoor UAV video dataset with 16 human actions captured by a DJI Phantom 4.0 pro v2 drone, performed by human subjects in an unconstrained manner. The dataset contains 2079 labeled videos at a resolution of 1920×1080. It has 10 single person actions such as walk, run, jump, etc, and 6 two person actions such as shake hands, push a person, etc. The dataset is available at https://www.nec-labs.com/ mas/NEC-Drone/.

2 Implementation Details

In the interest of reproducibility, we will make all code and pretrained models publicly available upon acceptance of the paper. We also attach the codes used in our experiments with the supplementary zip folder submitted for review.

Backbone network architecture: We benchmark our models using two state-of-the-art video recognition backbone architectures (i) I3D [2] (CVPR 2017) (ii) X3D-M [4] (CVPR 2020). I3D is a 3D inflated CNN, based on 2D CNN inflation, and enables the learning of spatial-temporal features. X3D is also a 3D inflated CNN, and progressively expands a 2D CNN along multiple network axes such as space, time, width and depth. For both X3D and I3D, we extract mid-level features after the second layer.

Training details: Our models were trained using NVIDIA GeForce 1080 Ti GPUs, and NVIDIA RTX A5000 GPUs. Initial learning rates were {0.01, and 0.001} across datasets. We use cosine annealing and poly annealing for learning rate decay in X3D and I3D respectively, We use the Stochastic Gradient Descent (SGD) optimizer with weight decay of 0.0005 and momentum of 0.9, and cosine/ poly annealing for learning rate decay. The final softmax predictions of all our models were constrained using multi-class cross entropy loss.

3 Fourier Disentanglement

Videos depicting human action have four types of entities: moving salient regions (typically corresponding to moving object), static salient regions (typically corresponding to static object), moving non-salient regions (typically corresponding to dynamic background), and static non-salient regions (typically corresponding to static background). Robust action recognition systems should learn features that heavily amplify moving objects, followed by static objects (that provide contextual cues and are relevant to the prediction). This should be followed by background entities. According to our formulation, dynamic salient regions are amplified the most. This is because the Fourier mask highlights dynamic regions, and the features learnt by the network have a higher amplitude at the salient regions. Static non-salient regions are at the other end of the spectrum because the Fourier mask suppresses these regions, as well as the features learnt by the network have a lower amplitude at the non-salient regions. Static-salient and dynamic salient regions lie at the middle of the spectrum. The final equation for Fourier disentanglement uses the $l2$ operation in the computation of M_{FO} and linear application of f. This implies that static salient regions have a higher amplitude than the dynamic non-salient regions. Thus, the ordering of amplitudes that is formed as: dynamic-salient > static-salient > dynamic-non-salient > static-non-salient, in concordance with the relevance for decision making for action recognition. Thus, static as well as dynamic background regions have lower amplitudes than static and dynamic regions of the object executing action.

In addition, the video may contain noise (light noise or otherwise) and camera movement. In regions of the video where there is noise, the amplitude of the feature map depicting saliency will be low. Hence, noise gets suppressed. Any movement of non-salient pixels due to camera motion gets suppressed since they are a part of dynamic non-salient regions. Moreover, camera motion is generally uniform across the spatial dimensions of the video (covering salient as well as non-salient regions). Thus, it doesn't impact the decision making ability of the aerial video recognition system.

Comparisons with motion-based methods. Motion-based methods either model spatial and temporal information separately using two-stream 2D CNNs [6] or use motion representation as an auxiliary guiding factor to 3D CNNs. The latter is very expensive [9]. In contrast, we jointly model space and time using a 3D backbone, and then disentangle the moving human actor from the background using FO. Prior work has demonstrated the superiority [5,4] of 3D CNNs over two-stream 2D CNNs. FO imparts a relative improvement of 22.93% over the 3D I3D backbone and can be used with any 3D CNN to achieve state-of-the-art performance.

4 Fourier Attention

Lemma 1. *Given an input matrix A, Fourier attention as well self-attention [10,1] encapsulate long-range relationships for global mixing by computing outer products.*

Proof **Self-attention:** Without loss of generality, let $[a_{ij}]$ denote the elements of a square matrix A (with dimensions N) in $2D$. f, g, h represent 1×1 convolutions for key, query, value computations in self-attention. Hence, key, query and value vectors are $[fa_{ij}]$, $[ga_{ij}]$ and $[ha_{ij}]$ respectively. The first step of self-attention is the computation of sub-attention, which is the matrix multiplication of the transpose of query with key, which is $[ga_{ij}]^T \odot [fa_{ij}]$, which is equal to $\sum_{i=1}^{N} ga_{mi} \times fa_{in}$. The next step is the computation of self-attention, which is the matrix multiplication of the value vector with the transpose of sub-attention, which is equal to $[ha_{ij}] \odot \sum_{k=1}^{N} ga_{lk} \times fa_{kn}$. Hence, the self-attention matrix S_{mn} is:

$$S_{mn} = \sum_{l=1}^{N} ha_{ml} \sum_{k=1}^{N} [ga_{lk} \times fa_{kn}] \tag{1}$$

Fourier-attention: Without loss of generality, let $[a_{ij}]$ denote the elements of a square matrix A (with dimensions N) in $2D$. The Fourier transform is $\sum_{i=1}^{N} \sum_{j=1}^{N} \exp(-2\pi mi/N) \exp(-2\pi nj/N)$. Multiplication of the Fourier transform with its conjugate transpose, and inverse FFT gives us $\sum_{b=1}^{N} \sum_{c=1}^{N} \exp(-2\pi mc/N)$ e $\{\sum_{j=1}^{N} \sum_{i=1}^{N} \exp(-2\pi j(b-c)/N) a_{ij} \times \exp(-2\pi i(c-b)/N) a_{ij}\}$. Finally, weighted multiplication of the above term with $[a_{ij}]$ and a careful rearrangement of the terms involved leads us to the final expression for Fourier attention. Fourier attention

F_{mn} is:

$$F_{mn} = \sum_{b=1}^{N}\sum_{c=1}^{N} \overbrace{\exp(-2\pi mc/N)\exp(-2\pi nb/N)}^{h_{mn}(b,c)} a_{mn} \times \{\sum_{j=1}^{N}\sum_{i=1}^{N} \underbrace{\exp(-2\pi j(b-c)/N)}_{f_{mn}(b,c)} a_{ij} \times \underbrace{\exp(-2\pi i(c-b)/N)}_{g_{mn}(b,c)} a_{ij}\} \quad (2)$$

f, g, h in Equation 1 are 1×1 convolutions, and that the exponential terms span the entire spectrum of frequencies lets us define f, g, h for Fourier attention as shown in Equation 2. Thus, the equation for Fourier attention can be simplified as:

$$F_{mn} = \sum_{b=1}^{N}\sum_{c=1}^{N} h_{mn}(b,c) a_{mn} \times \{\sum_{j=1}^{N}\sum_{i=1}^{N} f_{mn}(b,c) a_{ij} \times g_{mn}(b,c) a_{ij}\} \quad (3)$$

In self-attention, f,g,h are learnable. In contrast, in Fourier attention, f,g,h are pre-defined by the Fourier spectrum. Nonetheless, they exhaustively cover the Fourier spectrum. Moreover, the terms involved and the structure of computations (multiplications followed by summation) in Equations 1 and 3 are similar, both promote global mixing and encapsulate long-range relationships.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021) 3
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 2
3. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1717–1726 (2020) 1
4. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020) 2, 3
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019) 3
6. Lee, M., Lee, S., Son, S., Park, G., Kwak, N.: Motion feature network: Fixed motion filter for action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 387–403 (2018) 3
7. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16266–16275 (2021) 1

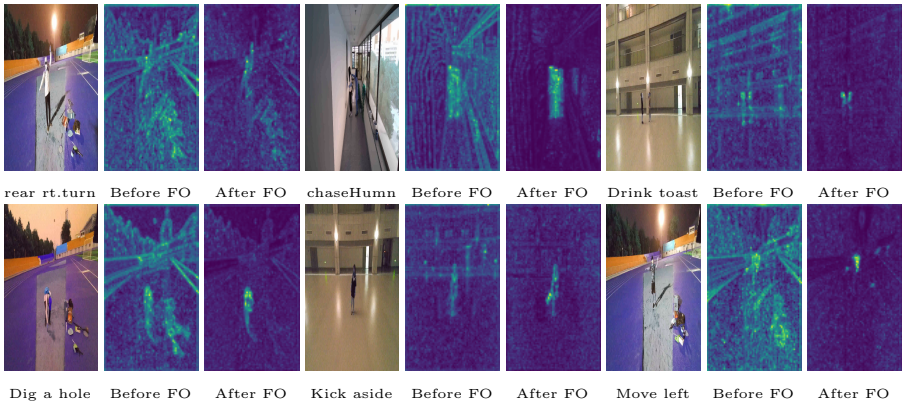
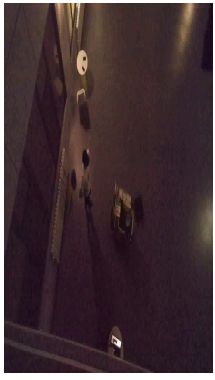
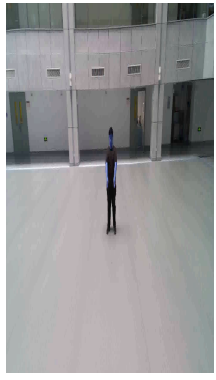


Fig. 1: **Qualitative results on UAV Human RGB.** We show the effect of our Fourier Object Disentanglement (FO) method. In each sample, the images, in order, correspond to a frame from the video, feature representation before disentanglement and the feature representation after disentanglement respectively. Notice the effectiveness of FO in scenes with light noise, dim light, dynamic camera and dynamic background. Regions of the scene corresponding to moving human actor (or salient dynamic) are amplified most (solid yellow). Static background is completely suppressed (solid purple). Static salient regions are slightly amplified, and dynamic backgrounds are suppressed to a great extent. We show videos depicting various complexities along with the predictions in the video file attached with the supplementary.

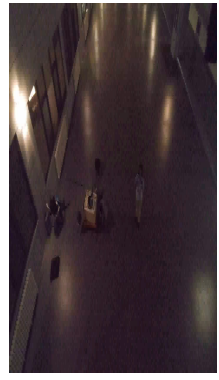
8. Perera, A.G., Law, Y.W., Chahl, J.: Drone-action: An outdoor recorded drone video dataset for action recognition. Drones **3**(4), 82 (2019) 1
9. Piergiovanni, A., Ryoo, M.S.: Representation flow for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9945–9953 (2019) 3
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017) 3



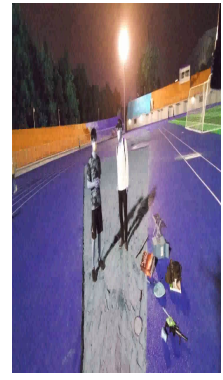
(a) Predicted: Drop something
GT: Put hands on hips



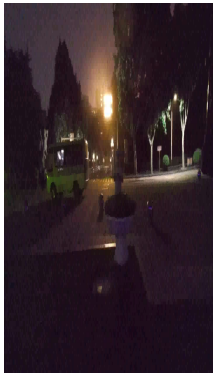
(b) Predicted: Open the bottle
GT: Decelerate



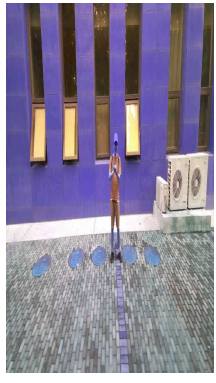
(c) Predicted: Punch with fists
GT: Cheer



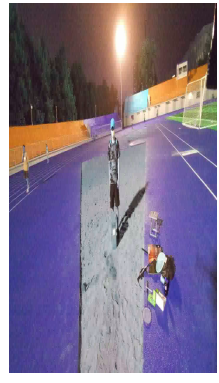
(d) Predicted: Pushing someone
GT: Rob something from someone



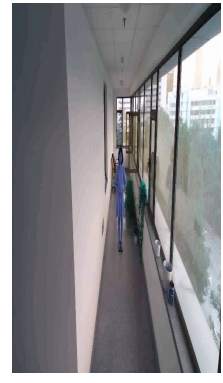
(e) Predicted: Smoke
GT: Apply cream to hands



(f) Predicted: Play with cell phones
GT: Applaud



(g) Predicted: Blow nose
GT: Throw litter



(h) Predicted: Applaud
GT: Cross palms together

Fig. 2: **Failure cases on UAV Human RGB.** We show frames from UAV Human RGB videos where FAR predicts the wrong class. In many cases, we observe that the predicted class has pixel level interactions similar to the ground truth. For instance, in case (d), both, predicted class and GT are two-person actions, and entail one person harming the other. Similarly, in video (h), both actions involve interaction between the two hands of a person. In video (a), both actions correspond to a human standing straight with hands at hip level. It would be interesting to explore learning distinguishable feature representations for the 155 classes as a part of future work.