Identity Preserve Transform: Understand What Activity Classification Models Have Learnt

Jialing Lyu[1], Weichao Qiu[2] and Alan Yuille[2]

[1]University of Science and Technology of China [2]Johns Hopkins University
sirine@mail.ustc.edu.cn, {qiuwch, alan.l.yuille}@gmail.com

Abstract

Activity classification has observed great success recently. The performance on small dataset is almost saturated and people are moving towards larger datasets. What leads to the performance gain on the model and what the model has learnt? In this paper we propose identity preserve transform (IPT) to study this problem. IPT manipulates the nuisance factors (background, viewpoint, etc.) of the data while keeping those factors related to the task (human motion) unchanged. To our surprise, we found popular models are using highly correlated information (background, object) to achieve high classification accuracy, rather than using the essential information (human motion). This can explain why an activity classification model usually fails to generalize to datasets it is not trained on. We implement IPT in two forms, i.e. image-space transform and 3D transform, using synthetic images. The tool will be made open-source to help study model and dataset design.

1. Introduction

We have witnessed the rapid development of video activity classification models thanks to the thriving of Convolution Neural Networks in recent years [14][5][19]. However, a video activity classification model trained on one dataset often failed to generalize to another [10]. Activity classification model can solve the task easily by fitting to correlated factors. For example, the model can associate the activity with typical backgrounds where the activity happen or clothes style, rather than trying to parse the human pose. This poses a challenge for the vision model, which requires the model to successfully extract abstract information, rather than using low-level features.

Human activity, in most cases, can be defined by the human motion instead of **nuisance factors**, such as the human appearance and the environment. Capturing the **essential factors** (human motion) is vital for a robust activity classification model. Psycho-physical experiments show that human can reliably recognize the activity by watching moving dots. We expect a well-performing activity classification model to have similar properties. Researchers created larger datasets[5][10] and carefully selected the combination of nuisance factors [3][9] to study the model robustness. In addition to these efforts, we propose a framework to understand a model.

We propose a method called **Identity Preserve Transform (IPT)** to inspect what an activity classification model has learnt. IPT is a transform which manipulates nuisance factors of an image, while keeping the essential factors the same. This type of transform is inspired by the image generation procedure. It includes two types: image-space transform and 3D transform. Image-space transform is applied on a generated image, while 3D transform affects an image by directly changing the underlying nuisance factors. We use both image-processing techniques and a computer vision model providing prior knowledge (semantic transform) for image-space transform. In order to achieve 3D transform, we implemented a synthetic data generation pipeline,which takes a combination of rendering parameters to render a realistic virtual human and nuisance factors can be directly manipulated.

Identity Preserve Transform unveils interesting properties of state-of-the-art models. Take Temporal Segment Network (TSN) as an example. The model is sensitive to small perturbation created by image-processing. Note that we did not specifically target the model with adversarial attack [6]. More interestingly, the quantitative result shows the model makes decision mainly on the object or background of the video, rather than using the human pose. This can be further validated with visualization technique [20]. This explains why the high performance on the trained dataset does not generalize to other datasets and real-world

scenarios. Our powerful synthetic data pipeline enables us to further analyze the relationship between the model performance and certain factors. The observation for synthetic data can be easily verified using a small real dataset. The IPT operates only on the input data regardless of the model architecture, so it can easily be adopted to study other models.

Our contribution can be summarized as follows: 1. We propose identity preserve transform, a method to inspect an activity model using data probes and independent of model architecture. 2. We analyze a state-of-art model and showed it does not classify video activities according to human motion. 3. We collect a synthetic activity video dataset and develop a diagnostic toolkit to perform IPT. The source code will be available to help others understand and develop activity classification models.

2. Related Work

There are both qualitative and quantitative methods for understanding deep CNNs-based models, the qualitative type mainly focuses on visualizing intermediate layer feature maps and visual saliency, while the quantitative type explains the model through feature importance scores or factor importance scores.

Methods based on feature importance scores alter individual features (pixels, super-pixels, word-vectors, etc) through removal or perturbation [13] for each input to the model to approximate the importance of each feature for model's prediction. They have been proved susceptible to human confirmation biases [8]. Whereas Aubry *et al.* [2] analyzed CNN feature responses corresponding to different scene factors (object color, object style, lighting, etc.) by controlling them via rendering using a large database of 3D CAD models. In our approach, we abstract influential elements from the generation of human activity videos instead of selecting scene factors.

Recently, researchers started to use synthetic data (data generated through computer graphics) to understand vision models. This is mainly due to the high cost and difficulty to collect and annotate large numbers of controlled real data. Synthetic data has been used to study the sensitivity to rendering parameters, such as viewpoint [17][1], material property [18]. The controllability of the synthetic data also enables studying the invariance and equivariance property of a model [2].

3. Method

3.1. Identity Preserve Transform

Our design of Identity Preserve Transform is inspired by the data generation process. There are various factors influencing what the data looks like. Theoretically, all factors can be controlled during the data generation process.

Given a computer vision task, the entire set of factors can be split into 2 subsets. One is those factors directly related to the task which represent the most essential information for the task, such as the human motion for human activity classification in our case, or object shape and texture for object detection. We denote this subset as p. The other subset comprises of nuisance factors not directly related to the task, and a model should be insensitive to their change or even to their absence, such as viewpoint, human appearance for human activity classification and background color for object detection. We denote the latter subset as θ.

Therefore, the data generation process can be modelled by Eq. 1, where G denotes generation function, I denotes the data generated.

$$I = G(p, \theta) \tag{1}$$

If we manipulate p with τ_p and θ with τ_θ during data generation process, effects on the generated data I is equivalent to a transform T after the generation, as in Eq. 2.

$$T(I) = G(\tau_p(p), \tau_\theta(\theta)) \tag{2}$$

A well performing model capable of learning the essential information for a specific class (denoted by p) should be invariant or insensitive to changes of the nuisance factors (denoted by θ). It means that as long as p keeps constant, no matter how τ_θ changes θ, a good model should yield same results, as in Eq. 3. Here we denote the model as f.

$$f(I) = f(T_\theta(I)) = f(G(p, \tau_\theta(\theta))) \tag{3}$$

Identity Preserve Transform (IPT) refers to transformations that preserves p. It can be implemented in the image space or in the 3D space. Hence, IPTs can be categorized into two types: one is image-space transforms, which operate on the generated data I, the other is 3D transforms, which operate on θ during data generation process.

The identity factor p for video activity classification is human motion. It is the most essential information a model should learn from training data. Compared with human motion, background, human appearance, viewpoint and lighting, etc. are all nuisance factors for classifying video human activities, they belong to θ. If a human activity classification model can develop a mechanism to model and infer human motion, it should be insensitive to the change in all these nuisance factors.

3.2. Image-space Transform

Image-space transform is IPT operating on generated data to simulate the change of nuisance factors during data generation process, as in Eq. 4 Image-space means it edits image extracted from a video.

It can be realized by applying image processing techniques commonly used for data augmentation, such as

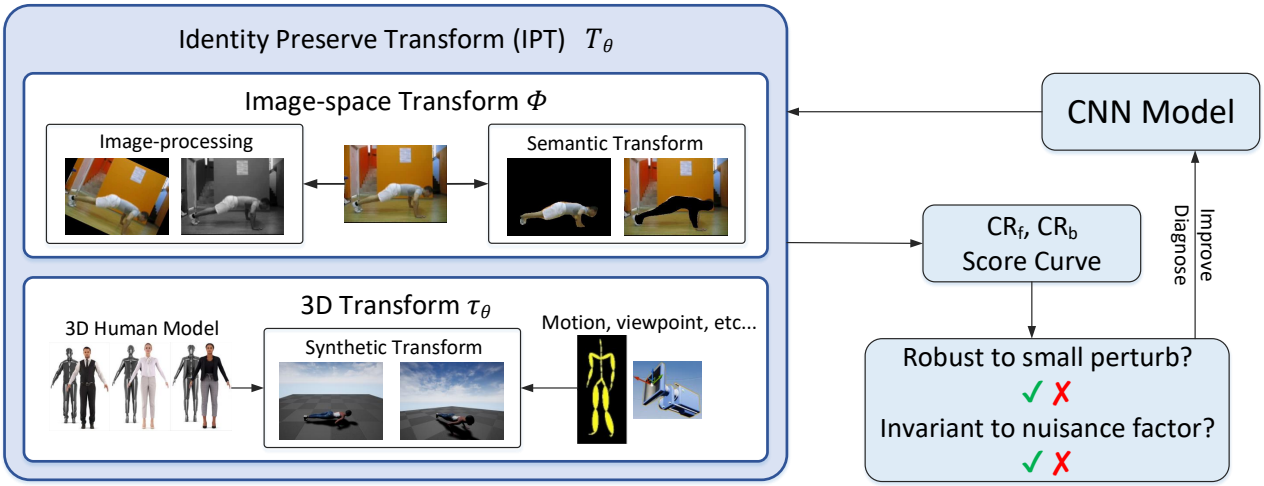

Figure 1: This figure illustrates an overview of Identity Preserve Transform (IPT). IPT consists of two types. It can take an activity classification model as input and be used to analyze properties of the model. The analysis can be used to improve model design.

adding noise, blurring, we call it image processing transform. It can also be realized by a computer vision model providing prior semantic knowledge, such as segmenting objects in the image, super resolution, etc., called semantic transform.

$$T_\theta(I) = \Phi * G(p, \theta) \tag{4}$$

3.2.1 Image Processing Transform

Image processing transform influences how a video look like while keeping background and human motion the same. For example, we can lower the image resolution by applying a blurring filter, or increase the image brightness through histogram equalization. However, lowering the image resolution or increasing brightness would not change people's body motion inside the video.

If a video activity classification model is robust enough, we expect its performance not to be harmed by image processing transform on the test data. Once its classifying accuracy drops due to image processing transforms, we should be alarmed that it overfits to image pattern without really learning human motion.

3.2.2 Semantic Transform

Semantic transform edits image by an additional computer vision model that provides semantic knowledge while preserving human motion information.

We used Mask-RCNN [7] to implement Semantic Transforms in our experiments. As shown in Fig. 3, Mask-RCNN detects and segments regions of people. With this additional model we can obtain a foreground-only video and a background-only video from the original video by superimposing black masks onto background and foreground respectively.

We can test a model on the original video set, foreground-only set and background-only set to get three classifying accuracy Acc_o, Acc_f and Acc_b, then we can compute foreground-only accuracy changing rate and background-only changing rate by Eq. 5.

$$\begin{cases} CR_f = \frac{Acc_o - Acc_f}{Acc_o} \\ CR_b = \frac{Acc_o - Acc_b}{Acc_o} \end{cases} \tag{5}$$

CR_f is expected to be very close to 0, if the model has really learnt human motion and use human motion to classify. An ideal segmentation of the human-centric foreground well preserves all kinds of information on the person in foreground-only videos, while leaving a silhouette of the person in background-only videos. Therefore, if the model is capable of infering human motion but has no reliance on background, CR_b should have a positive value closer to 1 rather than 0.

3.3. 3D Transform

A 3D transform is an IPT that directly manipulates nuisance factors during data generation process, as in Eq. 6. It has access to every factor in the data and can manipulate each factor separately.

$$T_\theta(I) = G(p, \tau_\theta(\theta)) \tag{6}$$

Controlling a certain nuisance factor to change while keeping other factors consistent, we can study how sensitve a model is to this controlled factor and whether the models response follow some kind of rule. For example, by changing the viewpoint gradually in different videos and record the classification scores, we can derive the curve reflecting how the model reacts to viewpoint change. If we repeat the viewpoint control experiment on different human appearances, we can further know how the model is influenced by human appearance.

Conclusions drawn from synthetic data can be verified by real data. Synthetic domain and real domain share lots of human motion information that is essential for recognizing human activity. When it is expensive and difficult to collect a large number of control variable videos in real domain , we can instead enlarge the increasing or decreasing step of the controlled variable in different videos, and emphasize on important values that reflect the main trend.

However, the conditions for achieving 3D-transform in reality are very strict. Previously, researchers use robotic arm in a lab setting to control viewpoint and lighting. Such realization of 3D transform is usually expensive and difficult to set up. Recently, due to the popularity of synthetic data, meaning data generated through computer graphics, researchers started to use synthetic data to control factors of an image, such as viewpoint [17][1], and material property [18].

We collected a 3D animation dataset from Epic Game marketplace as well as the CMU Mocap dataset [4], and created a toolbox for generating synthetic human activity videos. The synthetic data generation pipeline is built on top of Unreal Engine, a popular game engine for creating 3D video game. The toolbox developed from unrealcv [11] can be used to record and render images. After setting human motion with an 3D animation, we could use the toolbox to configure and control each nuisance factor separately, including the environment, lighting, human appearance and viewpoint.

4. Experiment

In this section Identity Preserved Transform is applied to the trained Temporal Segment Networks (TSN) introduced in [14] and trained Inflated 3D ConvNet (I3D) introduced in [5]. TSN is the representative of state-of-the-art video classification models that uses 2D convolutional kernels in neural network. I3D is the pioneer in 3D CNNs, giving rise to many adaptations such as S3D [16], I3D-GCN [15], etc.

We adopted TSN parameters trained on UCF101 [12] provided by OpenMMLab and I3D parameters trained on Kinetics [5] provided by deepmind publicly on Github.

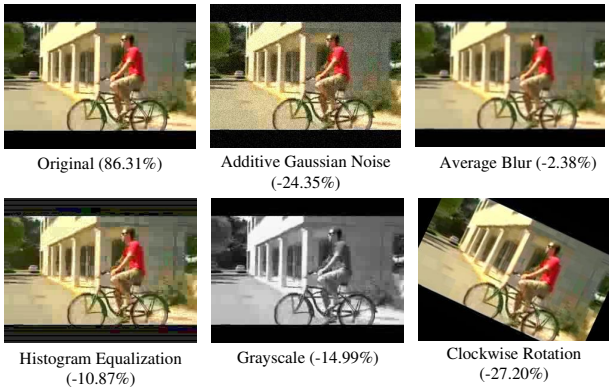

Figure 2: Applying Image-space Transform to videos in UCF101. Though these image processing techniques preserve human motion information in the video, they lead to serious dropping of TSN's classification accuracy.

Both models parameters are pre-trained on ImageNet, input modality to both models is RGB without optical flow.

4.1. Implementing Image-space Transfrom

Image-space transform is an Identity Preserve Transform operating on image to simulate nuisance factor change. It can be achieved in the form of image processing and semantic transform.

4.1.1 Image Processing Transform

Image processing transform includes image processing techniques widely used for data augmentation. We take 5 common image processing techniques in our experiment. They are applied on UCF101 test set to obtain the top-1 accuracy and top-5 accuracy of TSN classification result on each transformed test set.

Results in Tab. 1 shows that the model is susceptible to image processing techniques, even if they preserve complete human motion information. Histogram equalization drops the top-1 accuracy by over 10%, and adding Gaussian noise drops the top-1 accuracy by about 25%. Despite the fact that images transformed by these two techniques still look similar to the original image for human.

The accuracy drop caused by image processing transform conflicts with the expectation for a robust video activity classification model. The reason might be that TSN trained on UCF101 exhibits high reliance on the image pattern exclusive to UCF101.

4.1.2 Semantic Transform

Semantic transform edits generated images and videos with a an additional computer vision model that pro-

Image-space Transform	Top-1	Top-5
Identical Transform	86.31%	97.99%
Average Blurring	83.93%	96.78%
Histogram Equalization	75.44%	93.39%
Grayscale	71.32%	90.72%
Additive Gaussian Noise	61.96%	85.41%
Clockwise Rotation by 25	59.11%	80.52%

Table 1: Top-1 and top-5 classifying accuracy of trained TSN on a UCF101 test set transformed by 5 different image processing techniques. Identical transform means the video remains as its original version.

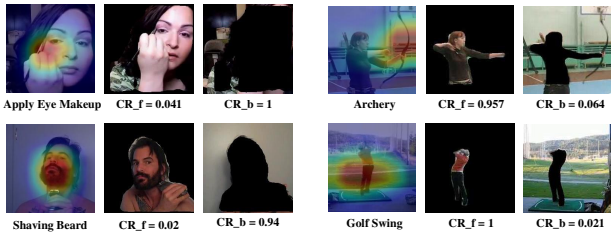

Figure 3: Combined analysis with semantic transform and class activation mapping. Left is the case that CR_f being close to 0 while CR_b being close to 1, the model classify these classes by detecting objects, textures, etc. using human motion information in the foreground. Right is the case that CR_b being close to 0 while CR_f being close to 1, the model overfits to background correlated with the activity classes, these classes can be found easily with semantic transform.

vides prior semantic knowledge. In our experiment we used Mask-RCNN to generate foreground-only videos and background-only videos for UCF101, then tested TSN to get classification accuracy changing rates for each activity class. The segmentation outcomes of Mask-RCNN are reasonably good. For rare cases when Mask-RCNN fails to detect human-centric foreground, leading to a full black foreground image and an unmasked background image, we dropped these image pairs. If full black foreground repeatedly appears in one video, we delete video pairs from the test split to ensure the accuracy of our quantitative results.

For a model capable of doing activity classification according to human motion, CR_f is expected to approximate zero while CR_b is expected to have a value closer to 1. However, the real performance is far from perfect. We plotted the CR_f and CR_b given by TSN over all classes of UCF101 and sorted according to the value of CR_f in Fig. 4. From Fig. 4 it is easy to tell that the model has significant performance drop over most activity classes when only given foreground (high CR_f). Furthermore, there emerges 3 types of results from all UCF101 classes: (1) $1 \approx CR_f \gg CR_b \approx 0$ (2) $0 \approx CR_f \ll CR_b \approx 1$ (3) otherwise.

Emergence of the first type indicates that for these activity classes, the model has overfit to background rather than learnt human motion. Approximation of CR_b to 0 means the model can classify these activities correctly using only background. Because silhouette of the main person remains in background, if the model can recognize human activities according to the silhouette, it should also achieve high accuracy on foreground-only videos, which better preserves human shape and pose. However, the corresponding CR_f is close to 1, which means accuracy sharply drops on foreground-only videos. Therefore, the model actually does classification using objects and textures in the background.

The quantitative results obtained with semantic transform are supported by class activation mapping [20]. We generated the class activation maps (CAM) corresponding to the class that gets highest classification score for each video. CAMs of activity classes belonging to type (1) show that the model focuses on objects or textures in the background instead of the human silhouette. For example, for Golf Swing the model focuses on the white line on the lawn, and for Archery the model focuses on the bow, as shown in Fig. 3.

Simultaneously, the second type of results can be further explained by CAMs. CR_f being close to 0 while CR_b being close to 1 means the model can correctly classify those activity classes with merely foreground, but what specific information is crucial remains unclear. CAMs indicate that the model uses human appearance, objects and textures rather than human motion in the foreground to achieve high accuracy over these classes. For instance, for Apply Eye Makeup, the model focuses the person's face and the eye brush in the foreground, for Shaving Beard it focuses on the person's mouth, as shown in Fig. 3.

The analysis above points out that TSN trained on UCF101 classify video activities by detecting objects or textures correlated with each activity class in training data. It has not developed a mechanism to model and infer human motion.

Semantic transform is complementary to class activation mapping. Our approach can be easily scaled up while CAM requires researchers to check a large number of video data with their eyes. Given a large scale video dataset, we can first employ semantic transform to check if the model under examination has overfit to background over some activity classes. After filtering with semantic transform, reviewing work with CAM will significantly decrease.

4.2. 3D Transform

3D transform is IPT that directly manipulates nuisance factors during data generation process. An ideal 3D transform has access to all factors and can control each one sep-

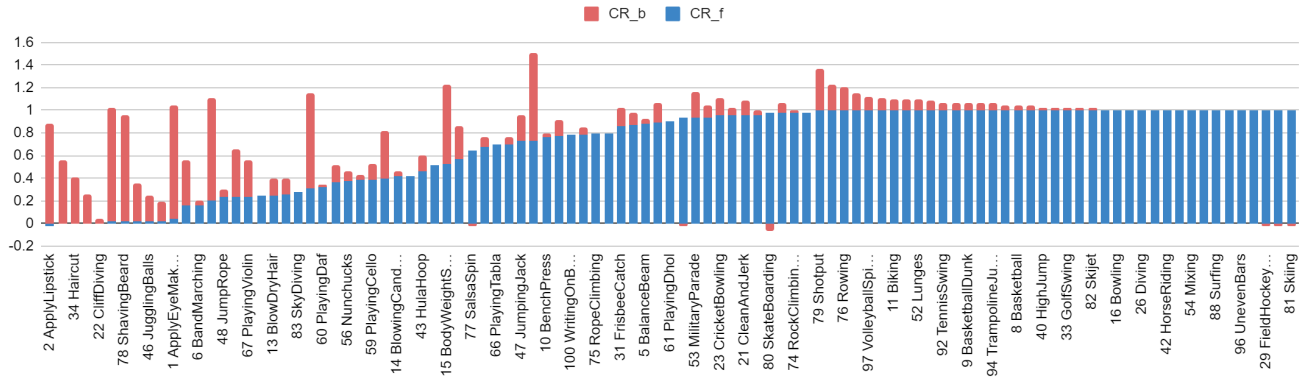

Figure 4: Plot of CR_f and CR_b for all classes in UCF101, sorted according to the value of CR_f. The figure clearly shows that most the model has significant performance drop over most activity classes when only given foreground (high CR_f). Only a few activity classes have low CR_f, which we further inspected with CAM [20] visualization.

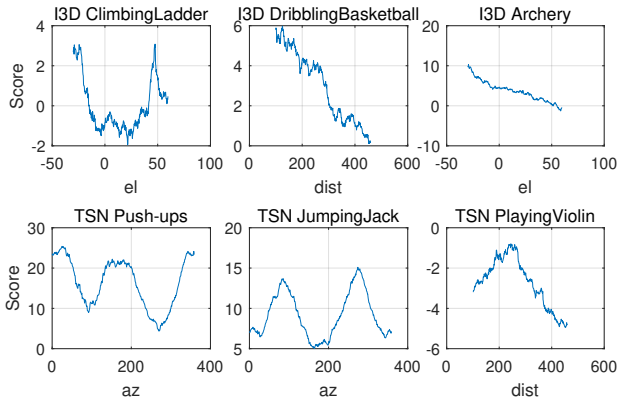

Figure 5: Classification score curves corresponding to different controlled factors obtained by 3D transform. If the model is insensitive to viewpoint, the score curve should be stable about a certain score or fluctuate within a relatively small range. However, many score curves yielded by TSN and I3D display trends other than this.

arately, however, it is difficult and expensive to achieve 3D transform on real data. Thus, we first conduct experiments on synthetic data, then verify conclusions drawn from synthetic data with real data.

4.2.1 Influence of Viewpoint

In our synthetic video, the camera is set towards the person, thus viewpoint is determined by three variables: camera azimuth angle, camera elevation angle (altitude angle) and camera distance. Keeping two variables constant while increasing the value of the third one iteratively, we generated a synthetic data split for each viewpoint. Simultaneously, other factors, including human motion, background environment, and human appearance are consistent through out a data split.

Appearance	Girl_e	Girl_f	Girl_g	Girl_i
TSN	0.19%	0.46%	7.31%	15.09%
I3D	48.61%	72.50%	70.83%	61.39%
Appearance	Alison	Carla	Claudia	Eric
TSN	48.06%	39.91%	16.20%	11.57%
I3D	51.11%	61.20%	23.24%	48.89%

Table 2: Top-1 classification accuracy of TSN and I3D on different synthetic human appearances. We aggregated the azimuth, elevation and distance splits generated on the same human appearance into a larger set, yielding 8 human appearance specific sets in total, then tested with TSN and I3D respectively.

Record a model's classification score corresponding to each viewpoint variable value over a data split, we can obtain a score curve for this variable, as shown in Fig. 5.

If the model is insensitive to viewpoint, the score curve should be stable about a certain score or fluctuate within a relatively small range. However, many score curves yielded by TSN and I3D for different activities are not like this. Instead, there emerged many other trends in these score curves. For example, testing TSN on azimuth split of Jumping Jack videos yields a score curve with two peaks; testing I3D on elevation split of Archery videos yields a score curve monotonically decreasing. Control variable experiment for multiple activity classes indicate that I3D and TSN are sensitive to viewpoint, the sensitivity extent and trend depends on the controlled viewpoint factor and activity class.

4.2.2 Influence of Human Appearance

The model's response to controlled viewpoint variable can also vary when human appearance changes. We generated

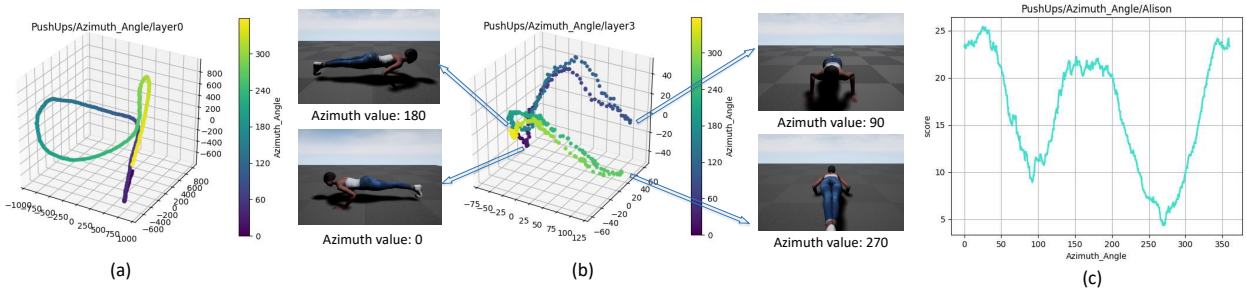

Figure 6: TSN's response to the azimuth split over Push-ups. TSN is most likely to recognize Push-ups when viewpoint is set towards the flank of person with invariance to left-right flipping, and it can easily fail to recognize Push-ups when viewpoint is set right towards the person's face or feet.(a) PCA embeddings derived from features output by max-pooling layer after the third Inception Module in the Spatial ConvNet of TSN.(b) PCA embedding derived from features output by Segmental Consensus layer. (c) Curve of classification scores output by TSN over the whole azimuth split.

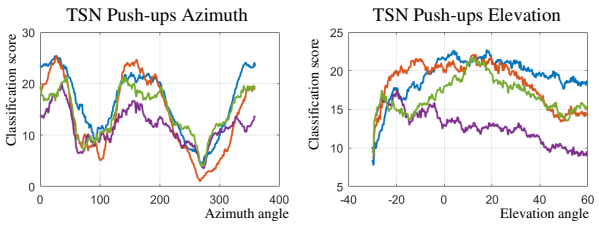

Figure 7: Influence of human appearance on the controlled factor score curve. Each color denotes a synthetic human appearance. (a) Though human appearance differs, the TSN classification scores are always high when azimuth is around 0 and 180, while being low when azimuth is at 90 or 270. (b) The score curves obtained from TSN by controlling elevation angle in Push-ups videos look dissimilar on different human appearances.

the controlled viewpoint variable data splits on different human appearances, then compared the score curves given by the same model over the same activity class.

Outcomes show that, in some cases the score curve trend on one human appearance no longer appears on another, for example, the score curves obtained from TSN by controlling elevation angle in Push-ups videos look dissimilar on different human appearances, as shown in Fig. 7(b).

However, there also exists cases when the score curves look similar. Take azimuth angle controlled in Push-up videos as example, though human appearance differs, the TSN classification scores are always high when azimuth is around 0 and 180, while being low when azimuth is at 90 or 270, as shown in Fig. 7(a).

Classifying accuracy for different human appearances also differs. We computed the top-1 and top-5 classifying accuracy of TSN and I3D on push-ups videos corresponding to eight different human appearances, as reported in Tab. 2. Both models show preference for some specific human appearances.

4.2.3 Manifold Visualization Using PCA

Similar viewpoint score curves corresponding to many synthetic human appearances could probably reflected the models classification pattern. Visualizing intermediate layers' features of the model during a forward pass can help us further understand the outcomes of score curves.

In order to better visualize the manifold which impacts the model performance, we extracted features from two layers inside the Spatial ConvNet of TSN, then applied dimension reduction using Principal Component Analysis to get low-dimensional features for visualization. The lower layer is max-pooling layer after the third Inception Module [14], the higher layer is segmental consensus layer. As Fig. 6 shows, 3-dimensional PCA embeddings corresponding to the azimuth angle from 0 to 360 forms a converging manifold, the manifold is twisted through layers between two extraction locations. At Fig. 6(b), embeddings of 0 and 180 azimuth angle have been pulled very close, they represent viewpoint set towards the flank of person, while embeddings of 90 and 270 azimuths have transferred to the other side at 3D feature space, they represent viewpoint set straightly towards the person's face or feet.

The visualization indicates high level feature has a certain degree of invariance to viewpoint change. The feature visualization and score curve together proves that TSN has developed a classification pattern for viewpoint in Push-ups videos. After being trained on UCF101 it has correlated push-ups motion with the human pose looking from a viewpoint at the flank, with invariance to left-right flipping.

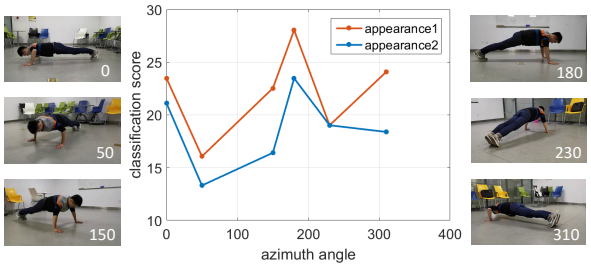

Figure 8: Classification score curves for real human Push-ups videos follow the same trend emerged from synthetic human Push-ups videos.

4.2.4 Classification Pattern Supported by Real Data

The classification pattern derived from synthetic data also take effects when the model does classification on real data. Since it is difficult to collect a large number of push-ups videos with different azimuth angles while keeping other factors constant in the real domain, we first recorded push-ups videos from 6 distinct viewpoint at the same height from the same distance, then found out the azimuth angle each real video corresponds to, they are 0, 50, 150, 180, 230, 310 respectively. As expected, the score curve yielded by TSN for these real videos exhibit a trend similar to those over synthetic data. We conducted azimuth controlled experiment on two real human appearances, both came out to be verified by the classification pattern found with synthetic data.

5. Conclusion

With identity preserve transform we proved that TSN, a state-of-the-art video activity classification model pre-trained on ImageNet and trained on UCF101 is sensitive to factors including the background, related objects, viewpoint and human appearances. Its classification accuracy can be influenced by changes in each of these video factors, therefore TSN probably has memorized the video factor sets for every activity class in UCF101, instead of developing a human motion reasoning mechanism, which is independent of the nuisance factors. Note that our approaches are not only suitable for TSN or I3D. Identity preserve transform is independent of model architecture and thus can be implemented to inspect any trained activity classification model's capabilities, as well as to examine the design of video datasets.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00342. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. The authors would like to thank Xinyue Wei, Yi Zhang and Professor Zheng-Jun Zha for their helpful advices.

References

[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *CVPR*, 2019. 2, 4

[2] Mathieu Aubry and Bryan C Russell. Understanding deep features with computer-generated imagery. In *ICCV*, 2015. 2

[3] Ali Borji, Saeed Izadi, and Laurent Itti. ilab-20m: A large-scale controlled object dataset to investigate deep learning. In *CVPR*, 2016. 1

[4] Carnegie Mellon Graphics Lab. Carnegie Mellon University Motion Capture Database, 2016. 4

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 4

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[8] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*, 2017. 2

[9] Yann LeCun, Fu Jie Huang, Leon Bottou, et al. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR*, 2004. 1

[10] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 2019. 1

[11] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang. Unrealcv: Virtual worlds for computer vision. *ACM Multimedia Open Source Software Competition*, 2017. 4

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 2

[14] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment

networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 4, 7

[15] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 4

[16] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 4

[17] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *CVPR*, 2019. 2, 4

[18] Yi Zhang, Weichao Qiu, Qi Chen, Xiaolin Hu, and Alan Yuille. Unrealstereo: Controlling hazardous factors to analyze stereo vision. In *3DV*. IEEE, 2018. 2, 4

[19] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 1

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 5, 6