Feature space data augmentation for viewpoint-robust action recognition in videos

Carla Geara, Aleksandr Setkov, Astrid Orcesi, Bertrand Luvison

FEATURE SPACE DATA AUGMENTATION FOR VIEWPOINT-ROBUST ACTION RECOGNITION IN VIDEOS

Carla Geara, Aleksandr Setkov, Astrid Orcesi, Bertrand Luvison

Université Paris-Saclay, CEA, List, 91120, Palaiseau, France

ABSTRACT

The ongoing research on human action recognition models is achieving very promising results, and the existing models reach very high performances. However, they still suffer from one major challenge: their performance decreases on viewpoints not seen in the training step of the model. In this paper, we introduce a new approach based on virtual viewpoint augmentation in the feature space to increase the robustness of the action recognition models to different camera viewpoints. This approach was evaluated on two action recognition datasets: DAHLIA and Toyota SmartHome. Our model shows promising results, with a significant performance increase on both datasets for viewpoints not seen during the training step.

***Index Terms*—** Action Recognition, Viewpoint Robustness, Data Augmentation

1. INTRODUCTION

Visual recognition methods (object detection, segmentation of objects and instances, etc.) have grown rapidly in recent years, achieving a very good performance in a lot of different applications. However, they suffer from a great challenge which is the decrease in performance when it comes to viewpoint changes. Indeed, the same scene, object, or action have different appearances when captured from two different viewpoints. This is especially problematic for action recognition because the datasets are often small and do not cover many different viewpoints while the semantic level of analysis is large.

One possible solution to this problem is to use a very rich dataset such as Kinetics [1], which would contain a large number of videos, taken from different viewpoints. However, such datasets do not always exists for all contexts and all action ontologies. For ambiant living analysis, for example, available datasets, such as Toyota SmartHome [2], DAHLIA [3], and NTU RGB+D [4], are made from a few cameras and are not large enough to provide a broad variability in viewpoints for training. Another approach consists in applying data augmentation to increase the model's robustness to viewpoint changes. However, the known data augmentation methods are not efficient against strong camera viewpoint changes. Therefore, we decided to focus on designing approaches that are robust to changes in camera positions.

In this paper, we introduce a new approach based on virtual viewpoint data augmentation in feature space that deals with the view-invariance problem. We call virtual viewpoint, a virtual camera pointing toward the scene from another point of view. We show that a significant performance increase is achieved by our method when evaluated on two multi-view human action recognition datasets, DAHLIA and Toyota SmartHome. Finally, we present an extensive ablation study to show the impact of each parameter on our method.

2. RELATED WORKS

Human Action Recognition: Video action recognition is the task of understanding human actions and behavior in videos, by associating each timestamp with an activity label. It can be accomplished using multiple modalities: RGB videos, the skeleton of the person performing the action in the video, the depth map, or even the optical flow. Most video action recognition models can be categorised into 5 main categories: 2D-CNNs [5, 6], 3D-CNNs [7, 1, 8, 9], RNNs [10, 11], GNN/GCN [12] and Transformers [13, 14]. Our approach focuses on 3D-CNN models, because of their simplicity and efficiency.

View-invariant action recognition methods: Some of the view-invariant approaches rely on the skeleton of the person performing the action [15, 16], where the 3D human pose is represented in a global and common space in order to create an invariant representation. The main drawback of these methods is that 3D pose estimation in videos is a challenging problem, and it becomes even harder when there are multiple people in the scene.

Another approach described in [16] refrains from using the human skeleton, and focuses instead on learning a latent 3D representation of the scene in order to perform a view-invariant action recognition. The method learns 3D video fea-

This work was funded by the project FULGUR and made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council. The research program named FULGUR benefits from a French Research Agency aid (reference ANR-19- STPH-0003). This program is part of the perspective of the Paris 2024 Olympic and Paralympic Games.

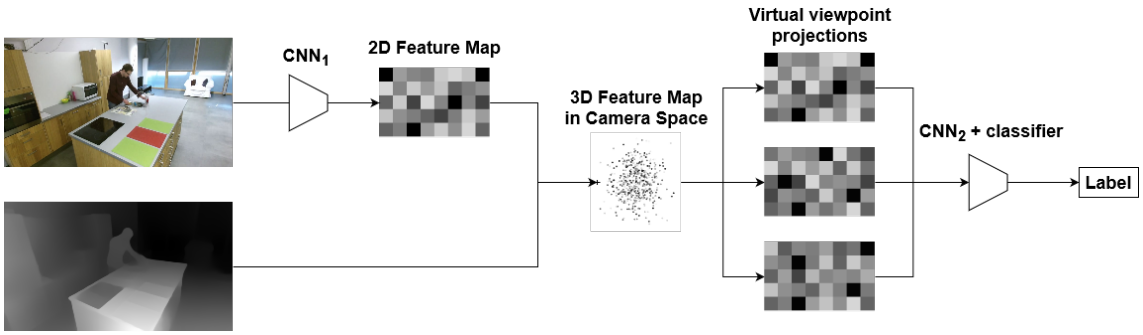

Fig. 1. View Robust Action Recognition Approach. CNN_1 takes the RGB video as input and generates 2D feature maps. These 2D feature maps are then transformed to point clouds using the input's depth map. These point clouds are then rotated and projected back to 2D to generate virtual viewpoints which are used for classification, along with the original 2D feature maps.

tures together with the camera extrinsic parameters which are then used to project the 3D features from the camera to the world space, thus creating a global invariant 3D representation of the video. The model then learns multiple virtual cameras parameters to generate 2D multi-view projections used for classification. However, learning the extrinsic parameters is not a simple task, especially if only few fixed views with little variability are available in the training dataset. In addition to that, the learned virtual cameras are fixed, and common between all the videos. In the case of a dataset with few fixed viewpoints, these fixed projections will not provide enough variability for the view-invariance.

Unlike the aforementioned method, in this paper we address the view-point problem differently. Instead of learning fixed 3D to 2D projections, we apply random transformations in 3D feature space during training, simulating different viewpoints for each input and increasing variability.

3. OUR METHOD

In order to increase an action recognition model's robustness to viewpoint changes, we introduce a new data augmentation protocol, insprired from [16], to simulate new viewpoints during training. Our method is illustrated in Figure 1. From an RGB image, we extract intermediate features using a first CNN (CNN_1) and then bring them from the pixel space to the 3D camera space thanks to the corresponding depth map given in input. To simulate different viewpoints, we apply rotation operations to the 3D feature map and then project them back in the 2D space. The generated 2D projections are fed to the remaining part of the model (CNN_2) and the classifier, along with the original 2D feature maps, to perform the action recognition task. The applied rotations are generated randomly for every batch, which provides a better viewpoint variability.

To move from pixel to camera space, we apply the following transformation:

$$\begin{bmatrix} X & Y & Z \end{bmatrix}^T = Z\, K^{-1} \begin{bmatrix} x & y & 1 \end{bmatrix}^T,$$

where (x, y) are the pixel coordinates, (X, Y, Z) the 3D camera coordinates, K is the intrinsic matrix, and Z is the depth value. Therefore, to perform this operation, we need the intrinsic parameters of the cameras, and the depth maps of the videos. While the former are assumed to be provided by the datasets, the latter are computed using the MiDaS [17] depth estimation model. In fact, generating depth maps of ordinary scenes is now a common task which is significantly easier than predicting extrinsic camera parameters.

Despite our approach being inspired by [16], it differs from this work in several main aspects.

- While Piergiovanni and Ryoo [16] use two input videos, our approach takes only one. This allows us to avoid learning the extrinsic parameters, reducing the model's complexity.
- Unlike the reference approach, in which the transformations are learnt and are fixed for all the videos, we generate random transformations for each batch.
- Instead of learning the 3D feature map as in [16], we apply classical transformations using the generated depth maps.
- Piergiovanni and Ryoo [16] use 3D volumes to represent 3D feature maps. Our formulation allows to store them as point clouds instead, which is more efficient regarding memory consumption.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

Our model is evaluated on two multi-view action recognition datasets: DAHLIA [3] and Toyota SmartHome [2].

The DAily Home LIfe Activity dataset [3], or DAHLIA, consists of 51 different subjects, performing 7 different home activities in the kitchen. These actions are captured from three

different viewpoints (see figure 2). In our experiments, we train on one viewpoint, and then evaluate the model on the three viewpoints in the dataset.

The Toyota Smart Home dataset (TSH) [2] consists of 16115 trimmed videos of 18 different subjects performing various home activities in an apartment. These videos are captured from 7 Kinect v1 cameras, positioned to capture the dining room, the kitchen and the living room. The subjects are filmed performing 31 actions in total. These actions are divided into 19 activities. Some of them contain sub-activities, which makes 31 activities in total. In our case, we decided to focus only on the 19 main activities. In our experiments, we train our model on videos captured from cameras 1,3,5 and 7. The videos taken from cameras 2, 4 and 6 are used to evaluate the performance of our model on viewpoints not seen during training. This camera split ensures that each scene is covered in both the training and evalution set (see figure 3).

Fig. 2. Cameras configuration for DAHLIA

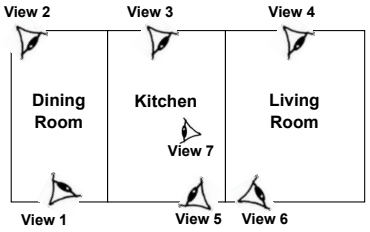

Fig. 3. Cameras configuration for TSH.

4.2. Implementation details

To simplify the virtual viewpoint projections step, we decided to restrict the transformations to rotations of the point cloud before projecting into pixel space. These rotations are applied around the x and y axis only, since rotations around the z axis (depth) seemed unnatural. Moreover, in order to limit the distortions created by the random rotations, we decided to limit the rotation angles generated inside the range [-max_angle, +max_angle]. We also decided to evaluate the model with one virtual viewpoint 2D projection, as it requires less computational and memory resources. For the action recognition model, we use X3D [9] as it is very efficient and performant while having a much lower complexity than other models.

Additionally, we applied data augmentation transformations to the input videos: random horizontal flip, random color jittering, and random crops in random positions. We train the model for 50 epochs using a batch size of 5 for DAHLIA and 8 for TSH. We start by training the model in a warm up phase for the first 3 epochs, where the learning rate increases linearly from 10^{-6} to 10^{-5} for DAHLIA and to 10^{-3} for TSH. We then decrease the learning rate exponentially for the remaining 47 epochs. The metric used for evaluation is the mean per-class precision.

In the following, "Baseline" results refer to the supervised training without our virtual viewpoint data augmentation.

4.3. Quantitative results

Results on DAHLIA: Table 1 shows the quantitative results obtained when training the model on each of the three views, and evaluating on the different views in the dataset. During training, we perform the data augmentation step on the output of the third block of X3D, and we perform one virtual viewpoint projection with a rotation angle range of [-60°,60°]. The choice of parameters used is explained in section 4.4.

Training View		Testing View		
		View 1	View 2	View 3
View 1	Baseline	0.71	0.33	0.29
	[-60°,60°]	0.72	0.47	0.46
	Increase	+1 p.p.	+14 p.p.	+17 p.p.
View 2	Baseline	0.36	0.72	0.62
	[-60°,60°]	0.45	0.73	0.67
	Increase	+9 p.p.	+1 p.p.	+5 p.p.
View 3	Baseline	0.38	0.62	0.75
	[-60°,60°]	0.54	0.62	0.77
	Increase	+16 p.p.	+0 p.p.	+2 p.p.

Table 1. Mean per-class precision on DAHLIA with rotation of one projection with rotation range of [-60°,60°] at 50 epochs. Increase is measured in percentage points (p.p.)

We observe the biggest performance increase when we train on the first view and test on the second and third, or when we test on the first viewpoint after training the model on the second or third view. This is due to the fact that the first view is furthest from the other two (see figure 2). In addition to that, when we train on the second view and test on the third and vice versa, we see no important improvement. This can be explained by the fact that the two viewpoints are already too close to each other, and a rotation range of [-60°,60°] might be too important in this case.

Results on TSH: For TSH, we decided to crop the videos around the person performing the action (more details about our choice of video preprocessing can be found in section 4.4). To perform the person-centered crops, we used Faster-RCNN [18] to detect the person's bounding box in each frame, which is used to specify the cropping window

of the video, with a margin of 10% around the Faster-RCNN bounding box.

	Training Views 1,3,5,7	
	Testing Views 1,3,5,7	Testing Views 2,4,6
Baseline	0.89	0.50
[-60°,60°]	**0.91**	**0.62**

Table 2. Mean per-class precision on person-centered cropped TSH videos.

As we can see in table 2, adding the virtual viewpoint data augmentation step increased the performance of the model by 12 p.p. on the viewpoints not seen in training.

4.4. Ablation Studies

Effect of rotation angle range: As mentioned in section 4.2, the angle of the randomly generated rotations is constrained in the range of [-max_angle, +max_angle]. We ran multiple experiments, varying the range and studying its effect on the model performance. We trained on the first view and tested on all the views in DAHLIA. We fixed the position of the virtual viewpoint data augmentation step after the third block in the X3D pipeline, using only one virtual viewpoint projection. The table 3 summarizes the results with 3 different ranges.

	View 1 (training view)	View 2	View 3
Baseline	0.71	0.33	0.29
[-20°,20°]	**0.73**	0.42	0.43
[-40°,40°]	0.72	0.44	0.46
[-60°,60°]	0.72	**0.47**	**0.46**

Table 3. Mean per-class precision on DAHLIA with different rotation angle ranges.

As can be seen in table 3, model accuracy is improved on the views not seen during training (views 2 and 3) for DAHLIA even with a range of [-20°,20°], while maintaining a good performance on the training viewpoint. The improvement becomes larger as we increase the range of the rotation angle up to [-60°,60°].

Effect of the data augmentation step position in the X3D pipeline: In this experiment, we show the results obtained when varying the position of the data augmentation step in the X3D pipeline. We train on the first view of DAHLIA, with one virtual viewpoint projection, and a rotation range of [-60°, 60°]. We run 3 experiments, placing the virtual viewpoint data augmentation step after the second, third and fourth block of the X3D model.

As we can see in table 4, the highest increase in performance is when the data augmentation step is applied after the third X3D block. One assumption is that after the fourth block, the spatial resolution of the feature maps is too small, and therefore the impact of the data augmentations is not as high as for the third block. As for the second block, the features are of high resolution, and they mostly contain information about the gradients of the scene, rather than semantic information. Therefore, applying the virtual viewpoint data augmentation step might create more distortions at this level, than after the third block.

	View 1 (training view)	View 2	View 3
Baseline	0.71	0.33	0.29
Block 2	**0.76**	**0.47**	0.42
Block 3	0.72	**0.47**	**0.46**
Block 4	0.72	0.41	0.29

Table 4. Mean per-class precision on DAHLIA with different X3D block positions.

Effect of video crops on TSH: We tested the effect of the rotation range on TSH with the projection applied after the third X3D block. The results are shown in table 5.

	Training Views 1,3,5,7	
	Testing Views 1,3,5,7	Testing Views 2,4,6
Baseline	**0.90**	**0.47**
[-20°,20°]	0.89	0.45
[-40°,40°]	0.89	**0.47**
[-60°,60°]	0.89	0.43

Table 5. Mean per-class precision on TSH with different rotation angle ranges.

The model seems to have an unstable performance on TSH. One explanation could be that the resolution on the person may be too small in comparison to the whole video. The model is not able to focus on the person's features. Therefore, we repeated the experiments after cropping the videos around the person performing the action. This approach significantly improved the model performance, as seen in table 2.

5. CONCLUSION

We presented a new simple approach that increases the robustness of human action recognition models to camera viewpoint changes, by inserting a virtual viewpoint data augmentation step in the action recognition model pipeline, ensuring a vast variability in viewpoints. Our approach, evaluated on two challenging multi-view human action datasets, shows a performance increase on viewpoints not seen during training.

For future perspective, it would be interesting to extend our approach to other datasets and action recognition models but also to address other tasks than classification ones where the ground truth management is more subtle.

6. REFERENCES

[1] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[2] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca, "Toyota smarthome: Real-world activities of daily living," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 833–842.

[3] Geoffrey Vaquette, Astrid Orcesi, Laurent Lucat, and Catherine Achard, "The daily home life activity dataset: A high semantic activity dataset for online recognition," in *Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, 2017, pp. 497–504.

[4] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684—2701, October 2020.

[5] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[6] 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 *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 20–36, Springer International Publishing.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, dec 2015, pp. 4489–4497, IEEE Computer Society.

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[9] Christoph Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 200–210.

[10] Lin Sun, Kui Jia, Kevin Chen, Dit Yan Yeung, Bertram E. Shi, and Silvio Savarese, "Lattice long short-term memory for human action recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2166–2175.

[11] Wenbin Du, Yali Wang, and Yu Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3745–3754.

[12] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6804–6815.

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3192–3201.

[15] Lingling Gao, Yanli Ji, Gedamu Alemu Kumie, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen, "View-invariant human action recognition via view transformation network," *IEEE Transactions on Multimedia*, vol. 24, pp. 4493–4503, 2021.

[16] AJ Piergiovanni and Michael S. Ryoo, "Recognizing actions in videos from unseen viewpoints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4122–4130.

[17] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.