Accepted for presentation at EPIC@CVPR2019 workshop

From Third Person to First Person: Dataset and Baselines for Synthesis and Retrieval

Mohamed Elfeki*, Krishna Regmi*, Shervin Ardeshir, Ali Borji
University of Central Florida, Center for Research in Computer Vision (CRCV)
{elfeki, ardeshir}@cs.ucf.edu, kregmi@knights.ucf.edu, aliborji@gmail.com

Abstract

In this effort, we introduce two datasets (synthetic and natural/real) containing simultaneously recorded egocentric (first-person) and exocentric (third-person) videos. We also explore relating the two domains in two aspects. First, we train a conditional GAN model to synthesize (hallucinate) images in the egocentric domain from its exocentric correspondent frame. Second, we explore the possibility of performing a retrieval task across the two views. Given an egocentric query frame (or its momentary optical flow), we retrieve its corresponding exocentric frame (or optical flow) from a gallery set. We show that performing domain adaptation from the synthetic domain to the natural/real domain, is helpful in tasks such as retrieval. The code and dataset are publicly available.[1]

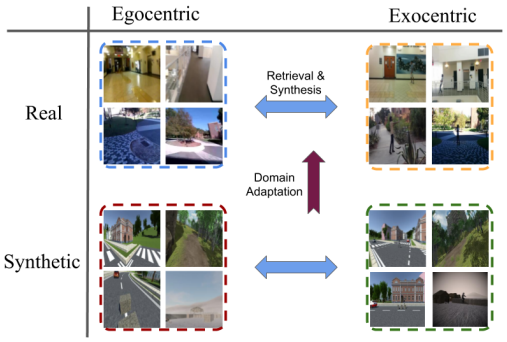

Figure 1: We study the relationship between first person and third person videos, in synthetic and natural domains. Domain adaptation from synthetic to real is helpful when we have limited real data, which is difficult to collect compared to synthetic data.

1. Introduction

First-person (egocentric) and third-person (exocentric) domains, although drastically different, can be related together. In this work, we take a step towards exploring this relationship. Our contributions in this work are three folds:

Dataset: We collect two datasets (synthetic and real/natural), each containing simultaneously recorded egocentric and exocentric video pairs, where the egocentric is captured by body mounted cameras and the exocentric is captured by static cameras, capturing the egocentric camera holders performing diverse actions covering a broad spectrum of motions. We collect a large scale synthetic dataset generated using game engines, and provide frame-level annotation on egocentric and exocentric camera poses, and the actions being performed by the actor. We also collect a smaller scale dataset of simultaneously recorded real egocentric and exocentric videos.

Image Synthesis: Given an exocentric side-view image, we aim to generate an egocentric image hallucinating how the world would look like from a first person perspective. This is a very challenging task, as the images in two domains often do not have a significant overlap in terms of their fields of view. As a result, transforming the appearances across the two views is non-trivial.

Retrieval: Given an exocentric video frame or its momentary optical flow (with respect to the previous frame), we explore retrieving its corresponding egocentric frame (or optical flow). We train a two-stream convolutional neural network seeking view-invariant representations across the two views given a momentary optical flow map (2 channel input). We also train another network for RGB values (3-channel input). We perform domain adaptation across synthetic and real domains, proving that using synthetic data improves the retrieval performance on real data.

In the past, the relationship between egocentric and exocentric information has been explored in tasks such as human identification [1, 3, 5], and action classification[13, 2]. Also, GANs have been used in conditional settings to synthesize images controlled by different parameters, such as labels of digits [8], images [7, 11, 12], textual descriptions [10]. However, synthesis and retrieval problems between these two domains has yet to be fully explored.

* equal contributions
[1] www.github.com/M-Elfeki/ThirdToFirst

		Training Pairs		Validation Pairs		Testing Pairs		Total Number of Pairs	
		# Vid	# Frames	# Vid	# Frames	# Vid	# Frames	# Vid	# Frames
Real	Ego-Side	124	26,764	61	13,412	70	13,788	255	53,964
	Ego-Top	135	28,408	68	12,904	73	14,064	276	55,376
Synth.	Ego-Side	208	119,115	109	6,702	95	6,778	412	132,595
	Ego-Top	208	119,115	109	6,702	95	6,778	412	132,595

Table 1: Dataset Statistics.

2. Dataset

To further examine the relationship between the first and third person views, we collect real and synthetic data of simultaneously recorded ego and exocentric videos. In both data, we isolate the egocentric camera holder in the third person video and thus, collect videos in which there is only a single person recorded by an exocentric video. Videos of different views are temporally aligned providing simultaneous ego/exo pairs. We provide frame level action labels for the videos in each view. Details about both real and synthetic data can be found in Table 1.

2.1. Real Data

Containing 531 video pair, each pair is an egocentric and an exocentric (side or top-view) video. Each pair is collected by asking an actor to perform a range of actions (walking, jogging, running, hand waving, hand clapping, boxing, and push-ups), covering various motions and poses. Some examples are shown in rows 1 and 2 in Fig. 3.

2.2. Synthetic Data

Simultaneously recorded videos of different views are lacking in real-life. Collecting such data from the web and in large scale is usually not feasible. To attain a large number of samples, we collect a synthetic data using graphics engines. Several environments and various actors were used in Unity 3D platform, programmed to perform actions such as walking, running, jumping, crouching, ... etc. A virtual egocentric camera was mounted on actor's body, while static virtual top/side view camera was also positioned in the scene. To better simulate real data, we added slight random rotations to virtual cameras. We have a total of four environments with five, seven, ten and ten scenes. Scene refer to a location where the actions are recorded. For each environment, we use two scenes for testing and the rest for validation and training. Rows 4 and 5 of Fig. 3 show some examples of synthetic dataset.

2.3. Datasets Value

We believe that the relationship across views (egocentric and exocentric) and modalities (synthetic and real data) can be explored in many aspects. Given that the dataset contains simultaneously recorded videos, and it contains frame-level annotations in terms of action labels and camera poses, we believe that it could be used for many tasks such as video retrieval and video synthesis, for which we provide some baselines. Also this relationship could be explored in other tasks such as action recognition, camera pose estimation, human pose estimation, 3D reconstruction, etc.

3. Framework

3.1. Image Synthesis

We use Generative Adversarial Networks(GANs) [6] to synthesize realistic-looking images. In addition to adversarial loss, we also apply $L1$ distance, which was shown to increase image sharpness in generation tasks [7, 11, 9]. Similar to [7, 11, 12], the GAN is conditioned on an image of one view to synthesize the other view. Particularly, we use an exocentric view as a conditional input to synthesize the ego view; $I'_{ego} = G(I_{exo})$. The conditioning view is paired with real/synthesized image and both are fed to D, which in turn predicts whether the image pair is real or fake. We use the pre-trained model of [11], and fine-tune it for 15 epochs on our real and synthetic datasets. We resized our model's input to 256×256 for generative tasks.

3.2. Retrieval

Given an egocentric video frame, we aim to retrieve the corresponding exocentric frame from all frames of the entire exocentric video set. Thus, a video frame at time t of ego and exocentric frame constitutes a positive pair. Every other pair of video frames is considered to be a negative pair, i.e., egocentric frame t_1 and exocentric frame t_2; $t_1 \neq t_2$. Such a network extracts view-specific features for each stream and encourage a view-invariant embedding by setting the difference between simultaneously recorded pairs to zero.

We examine the retrieval based on two aspects: appearance features (RGB) and motion features (momentary optical flow between two consecutive frames). Our appearance retrieval uses a two-stream network optimized using contrastive loss, shown in Fig. 2. Similarly, motion retrieval

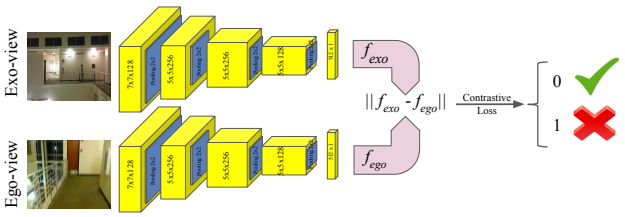

Figure 2: Appearance and Motion Retrieval Networks.

uses the same architecture with the only difference in the input size: 3 channels for RGB, 2 channels for Optical Flow. As a preprocessing step, we apply Gaussian smoothing over time to obtain a more consistent flow maps, reducing the random noise commonly found in optical flow maps.

Adapting Synthetic to Real. First, we train a network on the synthetic training pairs, and test it on the synthetic test split. Then, we perform a similar experiment on the real dataset where we train and test on the real train and test split correspondingly. In both RGB and optical flow, we observe that the retrieval performance on real data is not as favorable as synthetic data. This is because the latter is often less noisy, is in a more controlled environment, and has more training data than the former. Since synthetic and real data are of different modalities, we train a third retrieval network. We initialize the networks with the weights trained on synthetic data then fine-tune its convolutional layers on real data. Adapting synthetic domain to real data in the fine-tuned network results in a significant improvement when tested on the real data.

4. Experiments

4.1. Synthesis

The qualitative results on the real and synthetic datasets are shown in Fig. 3. The generated frames show that the network is successful at transforming the semantic information across the views. The generated images show blurriness for real dataset which is primarily because egocentric domain experiences motion in the frame rather than on the actor. The last two columns show some failure cases. The first failure case for real dataset shows the network is not able to learn the direction the person is facing so it is not able to generate the railings on right side of the person. The failure case for synthetic images show that the network is not able to hallucinate the textures in the scene.

We conduct the quantitative evaluation of synthesized images using the following metrics: Inception Score [14], Structural-Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Sharpness difference. Refer [4] for details about these metrics. Higher the better for all the metrics.

The inception scores are shown in Table 2. The higher inception scores for the real dataset is expected as the network was pretrained on natural images (Places dataset). SSIM, PSNR and Sharpness Difference scores are reported in Table 3. All of the scores are higher for the Synthetic dataset compared to the real dataset. This is mainly due to the fact that the synthetic dataset has a controlled environment with less motion blur compared to egocentric frames in real dataset.

Images	**Inception Score**		
	all classes	Top-1 class	Top-5 classes
Real Synthesized	3.8280	2.0315	3.4186
Real Ground-Truth	6.3787	2.6652	5.2608
Synthetic Synthesized	3.4320	2.1045	3.5042
Synthetic Ground-Truth	4.5353	2.3815	4.3695

Table 2: Inception Scores for data and model distributions on Real and Synthetic Datasets.

Dataset	**SSIM**	**PSNR**	**Sharp Diff**
Real	0.4822	18.1694	19.8142
Synthetic	0.5153	20.8976	20.5758

Table 3: SSIM, PSNR and Sharpness Difference between real data and generated samples for Real and Synthetic Datasets.

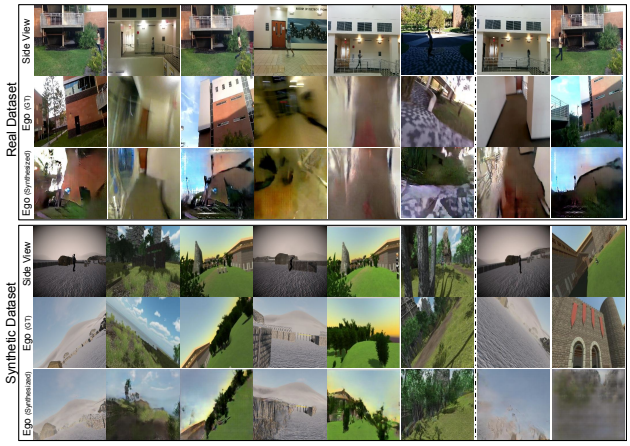

Figure 3: Qualitative Results for synthesis on Real (upper block) and Synthetic Datasets (lower block). In each block, first row shows images in exocentric (side) view, second row shows their corresponding ground truth egocentric images and the third row shows egocentric images generated by our method.

4.2. Retrieval

We evaluate the retrieval performance using the cumulative matching curve (CMC). The area under curve (AUC) is used as a quantitative measure. We evaluate retrieval based on optical flow and RGB images, and report the results in

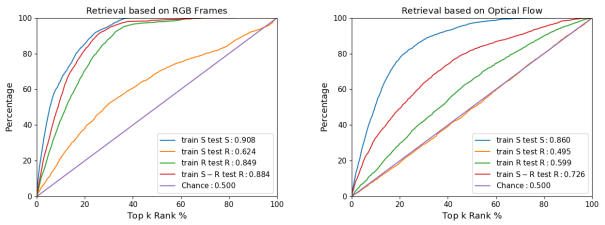

Figure 4: Retrieval performance based on RGB (left) and optical flow (right). S stands for synthetic data and R stands for real data.

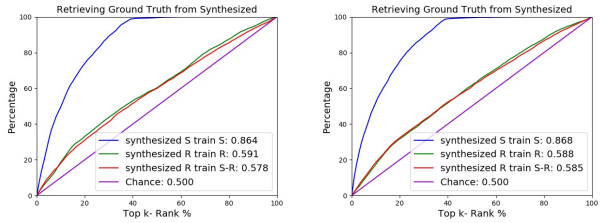

Figure 5: Retrieving the ground-truth egocentric, and exocentric images from the the synthesized images (left and right respectively). S stands for synthetic data and R stands for real data.

Fig. 4.

Retrieval based on Optical Flow. The cumulative matching curves for retrieval based on optical flow is shown in Fig. 4 (right). It can be observed that the network trained on synthetic and tested on real (orange) perform as chance level. The effect of adapting the synthetic network to the real data (red curve) is significant. As it can be observed the red curve (trained on synthetic, tuned on real data) outperforms the baselines on real data (green and orange curves). Please note that the blue curve is evaluated on the synthetic data and thus, is not comparable to other curves.

Retrieval based on RGB. The retrieval results based on RGB values are shown in Fig. 4 left. Similar to optical flow based retrieval, the phenomena of synthetic data being helpful in retrieving real data is observed. However, the improvement margin is less significant. This is due to the higher accuracy of the network trained on real data (green).

4.3. Retrieving Synthesized Images

Given an exocentric image I_{exo}, the synthesis network outputs a synthesized image I'_{ego}, and the corresponding ground-truth egocentric frame is called I_{ego}. In this experiment, we explore if the synthesis preserves higher level information that can be useful in retrieval. In order to answer this, we use the RGB retrieval network to extract egocentric features from the synthesized and ground truth egocentric images. In other words, we extract $f_{ego}(I'_{ego})$ and $f_{ego}(I_{ego})$ (where f_{ego} and f_{exo} are shown in Fig. 2.). We store all the features extracted from all synthesized egocentric images in F'_{ego}, the features from the ground-truth egocentric images in F_{ego}, and the features extracted from the exocentric images in F_{exo}. For each synthesized egocentric image in F'_{ego}, we retrieve its corresponding ground truth exocentric feature from F_{exo}. The retrieval results are shown in Fig. 5 (left). We also retrieve its corresponding ground truth egocentric feature from F_{ego}. The results are shown in Fig. 5 (right). In both figures, the blue curve is the retrieval performance on the synthesized synthetic data, and the red and green curves show the retrieval on the synthesized real data using the different networks explained in the retrieval section.

Retrieval Network \View	**Ego.**	**Exo.**	**Ego.+Exo.**
train Synthetic OF	37.71%	21.17%	27.33%
train Synthetic RGB	29.05%	27.29%	28.71%
trained Real OF	33.49%	28.18%	30.82%
trained Synthetic - Real OF	32.31%	32.97%	30.72%
trained Real RGB	42.58%	20.28%	24.16%
trained Synthetic - Real RGB	42.58%	20.43%	23.34%

Table 4: View Invariance-test based on Actions: In the synthetic dataset the chance level is 20% as there are 5 action classes. In the real dataset the chance level is 12% as there are 8 classes.

4.4. View-invariance Test

Here, we test the view-invariance of the retrieval network. We feed the RGB frames and optical flows to the retrieval networks and extract their features from their last fully connected layers (512 dimensions). We train two separate SVM classifiers on the features extracted from each view of the retrieval network: one SVM on egocentric features and action labels, and another on exocentric actions and labels. We then evaluate the performance of each of the SVMs (reported in Table 4 Egocentric view and exocentric view columns). A third SVM is then trained on pool of features from both views, corresponding to each action, independent of the fact that it is coming from the egocentric or exocentric stream. We then evaluate the performance of the third SVM on the first two. The classification performance of the SVM trained on both views does preserve the accuracy, and sometimes even outperforms the separately trained SVMs.

5. Discussion and Conclusion

In this work, we introduce new synthetic and real datasets of simultaneously recorded egocentric and exocentric videos. We also provide some baselines for performing tasks such as retrieval and synthesis from third person to first person. We also observed in our retrieval task, the synthetic data can be leveraged to address the lack of real data.

References

[1] S. Ardeshir and A. Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *European Conference on Computer Vision*, pages 253–268. Springer, 2016.

[2] S. Ardeshir and A. Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 2018.

[3] S. Ardeshir and A. Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 285–300, 2018.

[4] M. Elfeki, K. Regmi, S. Ardeshir, and A. Borji. From third person to first person: Dataset and baselines for synthesis and retrieval, 2018.

[5] C. Fan, J. Lee, M. Xu, K. K. Singh, Y. J. Lee, D. J. Crandall, and M. S. Ryoo. Identifying first-person camera wearers in third-person videos. *arXiv preprint arXiv:1704.06340*, 2017.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[8] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[9] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

[11] K. Regmi and A. Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018.

[12] K. Regmi and A. Borji. Cross-view image synthesis using geometry-guided conditional gans. *CoRR*, abs/1808.05469, 2018.

[13] B. Soran, A. Farhadi, and L. G. Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In D. Cremers, I. D. Reid, H. Saito, and M. Yang, editors, *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V*, volume 9007 of *Lecture Notes in Computer Science*, pages 178–193. Springer, 2014.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826, 2016.