Unsupervised Anomaly Detection of the First Person in Gait from an Egocentric Camera

Mana Masuda(✉), Ryo Hachiuma(✉), Ryo Fujii(✉), and Hideo Saito(✉)

Keio University, Tokyo, Japan
{mana.smile,ryo-hachiuma,ryo.fujii0112,hs}@keio.jp

Abstract. Assistive technology is increasingly important as the senior population grows. The purpose of this study is to develop a means of preventing fatal injury by monitoring the movements of the elderly and sounding an alarm if an accident occurs. We present a method of detecting an anomaly in a first-person's gait from an egocentric video. Followed by the conventional anomaly detection methods, we train the model in an unsupervised manner. We use optical flow images to capture ego-motion information in the first person. To verify the effectiveness of our model, we introduced and conducted experiments with a novel first-person video anomaly detection dataset and showed that our model outperformed the baseline method.

Keywords: Assistive technology · Egocentric video · Unsupervised learning · Adversarial training · Optical flow

1 Introduction

As the population ages, remotely monitoring the elderly has become invaluable for providing independent living. As their physical and cognitive skills decrease, the older population faces increased risk for potentially life-threatening accidents while they walk, such as falling down and stumbling. Thus, the ability to monitor their mobility and be alerted of potential dangers (abnormalities) is extremely useful for caregivers in the prevention of injuries and the provision of swift emergency care (Fig. 1).

In the field of computer vision, the problem of detecting anomalous events in human activity has been extensively studied using surveillance cameras [6,14,18]. Unfortunately, this approach suffers from visual occlusions, difficulty handling multiple subjects, and the need to extrapolate spatio-temporal parameters when the full-body cannot be seen. Moreover, they are restricted to fixed areas. Considering that gait is characterized by moving the body from one location to another, daily-life data on gait are difficult to capture without using multiple cameras attached to the environment.

An alternative approach is to use wearable sensors attached to the subject's body. Particularly, anomaly detection systems using inertial measurement unit

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12510, pp. 604–617, 2020.
https://doi.org/10.1007/978-3-030-64559-5_48

Fig. 1. The proposed method predicts the camera wearer's anomaly in gait from a chest-mounted egocentric camera. The images with blue frame shows the normal action (gait), and the images with red frames shows the abnormal action (falling down). (Color figure online)

(IMU) sensors for gait assessment [4] and fall detection [21] have been shown to be highly effective in the detection of anomalous activity. However, accelerometer values from IMU sensors cannot capture the spatial information of the environment [19].

It is reasonable to expect that wearable cameras, such as smart glass or a head-mounted display (HMD), will be readily available in the near future [26]. Based on this, egocentric video analysis has recently attracted increased attention in many applications in assistive technologies, such as personalized object recognition [11], object usage guidance [5], 3D pose estimation [29], and video summarization [7], in an attempt to understand human behavior from a first-person perspective.

In this paper, we aim to detect anomalies in gait from the perspective of the first person (the person wearing the camera) using an egocentric camera images. Note that the aim of our work differs from gait assessment studies that measure the potential risk of falling down. Rather, our work aims to detect any abnormal activity in gait, such as falling down, stumbling, or swaying.

Due to the difficulty of collecting a sufficient number of videos of anomalous activity, an anomaly detection model must be trained in an unsupervised manner. That is to say, the model should be trained to capture the distribution of the data (normal activity) and detect anomalous data as out-of-distribution during testing. Modern image-based anomaly detection systems [3,13,31] employ an autoencoder-based model to learn the manifold for the normal class at the time of training and calculate the difference between the input image and the reconstructed image to calculate the pixel-level anomaly score during testing.

This method, however, cannot be used directly in our task for several reasons. First, because the location in which the subject walks is always changing, it is difficult to model the distribution of normal data from a raw RGB image when it is being compared to the anomaly activity detected in the surveillance camera. Second, the anomaly score is usually computed using the difference between the Euclidean space in the reconstructed and original images; however, the anomaly is reflected in the image globally (if the person falls down, the camera will also

fall down), not locally (pixel-level), and a pixel-wise anomaly calculation might detect novel objects in the image as anomalies.

Therefore, we present an anomaly detection method that uses a 2D Convolution Neural Network (CNN) and the input of optical flow images. Our network is inspired by GANomaly [1]. To more easily capture the distribution of normal data, our model calculates the dense optical flow of successive frames and used this as input instead of raw RGB images. The use of optical flow means our model involves time series data using a 2D CNN, allowing it to work with time-series data at a lower computational cost than if it were to use a 3D CNN. During testing, we calculated the anomaly score in the latent vector space instead of the image space to capture the anomaly of the image globally.

As there is no publicly available dataset that contains first-person videos of normal and anomalous gait, we introduced a novel dataset that consists of two-hour egocentric video sequences of normal gait and five types of anomalous activities: squatting, stumbling, staggering, falling down, and collision. The sequences were captured by cameras on three different individuals at different places. Unlike the conventional anomalous action detection dataset [14], this dataset focuses on the anomalous events that happen, not on the person themselves. The experiments were conducted using this dataset. Our codes and the dataset are available from our repository[1].

The contributions of this paper are summarized as follows:

- We aim to detect the anomalous events in the gait of the first-person (camera wearer) from an egocentric video. To the best of our knowledge, this is the first work tackling this problem. We present a 2D CNN-based anomaly detection network trained in an unsupervised manner with input in the form of optical flow images.
- We introduce a novel first-person video anomaly detection dataset that consists of normal gait and five different anomalous events using an egocentric camera, and the dataset is now publicly available. Unfortunately, we cannot provide the raw RGB images due to privacy issues. We will also make public the reproducible results, training, and evaluation code.
- We experiment with this anomaly detection dataset and show that our model outperforms the baseline method. Moreover, we conduct an ablation study on different hyperparameter settings to verify the effectiveness of our approach.

2 Related Work

2.1 Egocentric Vision

A typical problem in egocentric vision is the recognition of the activities of the first-person (the person wearing the camera). Recent works have primarily focused on action-forecasting [30] and person localization [27,29]. [29] predicts the camera wearers place in a future frame, and [27] predicts the future locations

[1] https://github.com/llien30/ego-ad.

of people in first-person videos. To forecast first-person behavior, most models use pose prediction: [30] used 3D human pose prediction previously used for third-person pose estimation tasks. In contrast, our method does not need to estimate the pose to detect the abnormal behaviors of the first-person, reducing the time to inference.

2.2 Anomaly Detection

Anomaly detection is a well-known task within the field of machine learning, with the major areas of real-world application being fraud detection, biomedical, video surveillance, etc. Anomaly detection is also referred to in previous studies as "novelty-detection" and "out-of-distribution detection". The basic method of anomaly detection is to identify whether or not the data is out of normal data distribution. The traditional anomaly detection methods are distance-based, using the distance between the normal and abnormal data.

Recent studies on anomaly detection have involved the use of deep neural networks, with methods using auto-encoder and variational auto-encoders to train a model to reconstruct normal images and detect abnormal images as samples with high reconstruction errors. Since the adversarial-learning-based method was proposed in [24], many new methods using adversarial-learning have been used, and various methods have been introduced to the generative model. Efficient GAN [31] combines the auto-encoder to the generative model and reduces the inference time. AnoVAEGAN [3] introduces VAEGAN [13] as the generative model. In GANomaly [1], Akcay *et al.* proposed an adversarial network such that the generator comprises encoder–decoder–encoder sub-networks. The objective of our model is not only to minimize the distance between the original and reconstructed normal images, but also to minimize the difference between their latent vector representations. Skip-GANomaly [2] introduced skip connection to reconstruction, and the accuracy of anomaly detection was improved by reconstructing the image more precisely.

2.3 Video Anomaly Detection

Video anomaly detection has received considerable attention in computer vision and robotics. Many methods have been proposed for third-person video specifically. [9] proposed a 3D convolutional auto-encoder (Conv-AE) to model regular frames, and [15] proposed a stacked RNN for temporally-coherent sparse coding (TSC-sRNN). In [14], the generator was trained using an optical flow image to reconstruct the next frame image and detect anomalies by looking for differences between a predicted future frame and the actual frame. In [18], both the RGB image and optical flow image were reconstructed to calculate the anomaly score. When detecting anomalies from a first-person perspective, however, it is hard to reconstruct either the current or future RGB frames due to the first-person's motion. [28] localized the potential anomaly participants to detect traffic accidents from first-person videos under the assumption that the anomalous roadway event can be detected by looking for deviations between the predicted and actual

locations of objects. However, the motion of the egocentric camera on a walking person, which is described by 6-Dof, is significantly more complex than that of a dash-cam, which can be determined using the forward velocity and yaw angle. Also, the abnormal event is only limited to the collision between objects (e.g., cars, bikes, pedestrians).

Previous approaches to person-centric anomaly detection and prediction relied on 2D pose estimation [17], such as in [10], where a method was proposed for predicting falls that consisted of a pose-prediction module and a falls classifier. To apply this to our scenario would require the pose of the first-person, which is not as readily available as the third person 2D pose estimation of a static camera. Therefore, we leverage optical flow containing the first-person's ego-motion information. Qiao *et al.* [22] reconstructed optical flow images to detect abnormal actions from third-person video; however, to the best of our knowledge, this is the first work that uses only optical-flow images to detect anomaly actions in the first person using a reconstruction-based approach.

3 Method

If the abnormal actions, such as falling down or struggling, should occur, the image of the first-person's view would drastically change. Therefore, we take into consideration that there should be some difference between the optical flow calculated from the egocentric video for normal behavior and the optical flow for abnormal behavior. For these reasons, we use optical flow to detect first-person incidents from the first-person video.

Also, we do not aim to detect anomalies of the image in the pixel-level but rather to use the entire image to determine the existence of an anomaly. Thus, we compare the feature vectors of the original and reconstructed images instead of comparing the original and reconstructed images themselves as in previous studies. To achieve this, we adopt a sub-network, named the reconstructor, with two encoders and one decoder. This structure is inspired by GANomaly [1]. The details of the structure of the reconstructor are described in detail in Sect. 3.1. We use PWC-Net [25] to generate the optical flow image from an egocentric video.

Problem Definition. Our objective is to train an unsupervised network that detects anomalies using optical flow images. The definition of our problem is as follows: The training dataset is denoted as $\mathcal{I}$ which is composed of m normal egocentric videos while the person is walking. From these videos, we calculate M optical flow images X_i, as in

$$\mathcal{I} = \{X_1, \cdots, X_M\}. \tag{1}$$

The test dataset is denoted as in $\hat{\mathcal{I}}$, which is composed of n normal and abnormal videos. From these videos, we calculated N optical flow images $\hat{X}_i$ and labeled $y_i \in [0, 1]$ for the evaluation, as in

$$\hat{\mathcal{I}} = \{(\hat{X}_1, y_1), \cdots, (\hat{X}_N, y_N)\}. \tag{2}$$

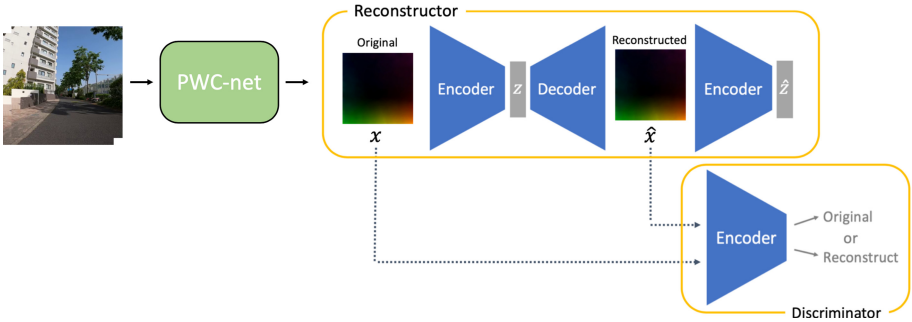

Fig. 2. Network overview. Our model consists of two networks: reconstructor R and discriminator D. The reconstructor consists of two encoders and a decoder.

To detect the abnormal events, the number of images in the training dataset is larger than the number in the testing dataset ($N \ll M$).

Using the dataset, our goal is to first model $\mathcal{I}$ to learn its manifold $\mathcal{X}$, then detect the abnormal samples that are not on the learned manifold $\mathcal{X}$ in $\hat{\mathcal{I}}$ as outliers during the inference stage. To detect the abnormal data in the feature vector space, the model learns the distribution of the normal data $\mathcal{X}$ and learns to encode the similar feature vector using two encoders.

Network Structure. The overview of the network is depicted in Fig. 2. Our network consists of two main networks—a reconstructor network and a discriminator network—and a PWC-Net that generates an optical flow image from two successive images. The reconstructor consists of two encoders and one decoder.

The first sub-network is the first encoder in the autoencoder network. The encoder network learns the input optical-flow representation in the latent space. This network input is $x \in \mathbb{R}^{w \times h \times 2}$ and downscales x by compressing it to a latent vector z with the use of 2d convolutional layers followed by batch-norm and LeakyReLU activation [16]. z is a bottleneck features of the reconstructor. The decoder consists of 2D conv-transpose layers followed by a batchnorm layer and a ReLU activation layer, except for the last layer that is only a 2D conv-transpose layer.

Network detail is shown in Fig. 3. The left side shows the detail of the encoder network and the right side shows the detail of the decoder network. The number of channels in the middle layer of the encoder was set to $\{2(input), 8, 16, 32, 64, z(output)\}$.

3.1 Reconstructor Network

We use a reconstructor to learn the manifold of normal samples and detect anomalies. To compare the feature vector of the original image to that of the reconstructed image, we adopt an encoder–decoder–encoder network for the reconstructor, as shown in Fig. 2. As with existing methods [1,2,24,31], we train

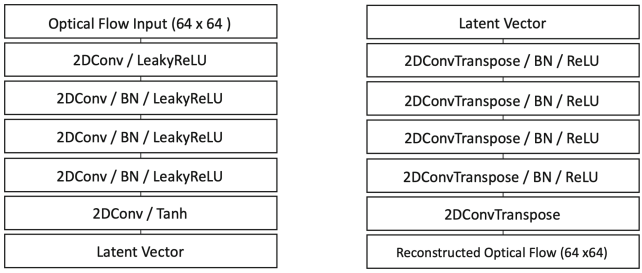

Fig. 3. Network details. The left side shows the details of the encoder network and the right side shows the details of the decoder network.

the reconstructor to perform anomaly detection based on image reconstruction by training it to learn adversarially with the discriminator so that the reconstructor is able to reconstruct images that resemble the original image. To ensure that the feature vectors of the original and reconstructed images are different for the abnormal images, we train the network to make the feature vectors of the original and reconstructed images similar. Because of the structure of the reconstructor, we only need to put the image in the reconstructor once to perform the test.

3.2 Loss Functions

Similar to GANomaly [1], we use three losses—adversarial loss, contextual loss, and encoder loss—to train the reconstructor. We use adversarial loss for the loss function of the discriminator [8]. To explain the loss function, let m be the size of the latent vectors.

Adversarial Loss is the loss calculated from the output of the discriminator. We use "feature matching adversarial loss", which is often used in anomaly detection models [1,24,31]. To explain the loss function, let D_{feat} be a function that outputs just before the last layer of the discriminator D. We use MSELoss to calculate the adversarial loss. The adversarial loss function is defined as

$$L_{adv} = \frac{1}{m}\sum_{i=1}^{m}(D_{feat}(x) - D_{feat}(\hat{x}))^2. \tag{3}$$

Contextual Loss. Contextual loss is the loss calculated as the difference between the original image and the reconstructed image and is the loss used to train the autoencoder. We used L_1 distance between the original image and the deconstructed image. The contextual loss function is defined as

$$L_{con} = \frac{1}{m}\sum_{i=1}^{m}(x - \hat{x}). \tag{4}$$

Encoder Loss is the loss calculated as the difference between the original image's feature vector z and the reconstructed image's feature vector $\hat{z}$. This loss is especially important in our model because we employed this difference in feature space when computing the anomaly score. We use MSELoss to calculate the encoder loss. To explain the loss function, let $E_{original}$ be a function of the first encoder network and E_{recon} be a function of the second encoder network. The encoder loss function is defined as

$$L_{enc} = \frac{1}{m}\sum_{i=1}^{m}(E_{original}(x) - E_{recon}(\hat{x}))^2. \tag{5}$$

Our objective function for the reconstructor is defined as

$$L = \lambda_{adv}L_{adv} + \lambda_{con}L_{adv} + \lambda_{enc}L_{enc}, \tag{6}$$

where λ_{adv}, λ_{con}, and λ_{enc} are the weighting hyperparameters that adjust the impact of individual losses on the overall objective function.

3.3 Anomaly Action Detection

We score abnormalities using the difference between z and $\hat{z}$. The anomaly score $\mathcal{A}$ of the testing image x is defined as

$$\mathcal{A}(\mathrm{x}) = \frac{1}{m}\sum_{i=1}^{m}(E_{original}(\mathrm{x}) - E_{recon}(\hat{\mathrm{x}}))^2. \tag{7}$$

The evaluation criteria for the anomaly score $\mathcal{A}$ is to threshold (ϕ) the score, where $\mathcal{A}(\mathrm{x}) > \phi$ indicates an anomaly.

4 Experiments

4.1 Dataset

As there is no publicly available dataset that contains anomaly actions captured from an egocentric video, we created our own. Three people mounted a GoPro camera (60 FPS, 1920 × 1080 resolution) on their chest. For the training data, the video is recorded for about 30 min while each person walked in an outdoor environment, with each environment being different from the others. For the test data, they were asked to perform anomalous action in their gait. Five types of anomaly actions were used: squat down, stumble, stagger, fall down, and collision. Around five minutes of video was captured for each person. We annotated abnormal and normal labels for each frame in the test videos. The RGB images from the dataset with abnormal labels are shown in Fig. 4. To generate the optical flow images, we centered and cropped the RGB image to 1080 × 1080, then resized it to 224 × 224. We used PWC-Net to generate the optical flow images. An example of the annotation of a dataset is shown in Fig. 5. Images surrounded by blue are normal and images surrounded by red are abnormal.

Fig. 4. Examples of abnormal actions. From the left to right: squat down, stumble, stagger, fall down, and collision.

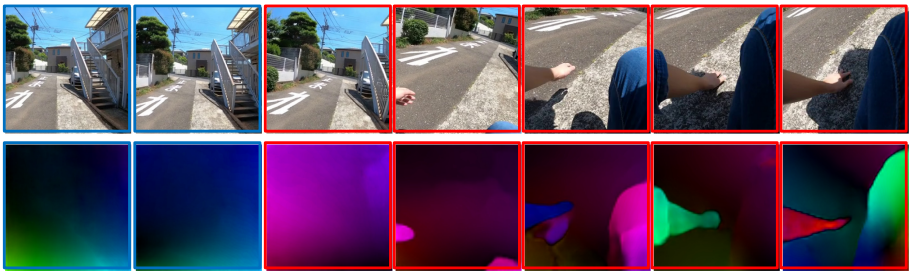

Fig. 5. Comparison of the optical flow and corresponding RGB images. The optical flow is color coded for better visualization. Images surrounded by blue are annotated as normal and images surrounded by red are annotated as abnormal. (Color figure online)

4.2 Baseline Method

In the experiment, we aimed to verify the effectiveness of detecting the anomaly from the optical flow instead of the raw RGB image. Since optical flow is generated from two successive RGB images, it is unreasonable to compare optical flow-based anomaly detection with single RGB-based anomaly detection. Therefore, we concatenated two consecutive RGB images in the channel direction and created a 6-channel image. We used the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) as an evaluation metric.

4.3 Experimental Setups

We conducted the experiment in three setups to verify the robustness of our method. The setups are as follows:

- **Person-specific:** In this setup, the model is trained with and for a single person. Even though the person during training and test is the same, the place-captured training and test data are different.
- **Person-generic:** In this setup, the model is trained and tested with all three people.

– **Person-out:** In this setup, the model is trained using data from two people and then tested with data from the remaining person. This setup evaluates the model's robustness against an unknown person and place.

4.4 Network Training

The size of the feature vector was set to 64 through the experiments. Following the previous anomaly detection model [1,24,31], our adversarial training is also based on the standard Deep Convolutional GAN (DCGAN) approach [23]. We implement our approach in PyTorch [20] (v1.4.0+cu100, with python 3.8.0) and run it on NVIDIA quadoro GV100 or P6000 processing unit using CUDA 10.0. We optimize the network by using Adam [12] with an initial learning rate $2e^{-3}$ and the momentums $\beta_1 = 0.5$, $\beta_2 = 0.999$. Our model is optimized based on the loss L (defined in Eq. 6) using the weighted values $\lambda_{adv} = 1$, $\lambda_{con} = 50$, and $\lambda_{enc} = 1$. We train the model for under 150 epochs for all cases. Note that the weights of PWC-Net is fixed during the training. The pretrained weight is downloaded from the official repository[2].

5 Results

5.1 Qualitative Evaluation

The graph of the time series of anomaly scores is shown in the following Fig. 6. The upper graph shows the temporal transition of the anomaly score by RGB images and the lower graph shows the transition of anomaly score by optical flow image. The graph and images show that even if the image does not contain any abnormal objects (e.g., hands, feet, etc.) that do not appear in a normal image, the motion of the first person (camera wearer) is captured in the optical flow image and the anomal score is increased. Unlike the optical flow image, the anomaly scores by RGB images do not differ much between abnormal and normal. This shows that optical flow images can identify anomalies more accurately than RGB images.

5.2 Quantitative Evaluation

The results of this experiment are shown in Table 1. The results show that optical flow is better for detecting anomalies than using RGB images. We also found that the accuracy of the RGB images varied greatly between training with each data set and combining a few different data sets. When we used optical flow images, however, we were able to achieve an accuracy of over 0.9 for all conditions. Therefore, it can be said that optical flow images are more robust to domain changes than consecutive RGB images.

The comparison of the AUC for each abnormal behavior is also shown in Table 1. The size of the feature vector was set to 64 for all cases. The results are

[2] https://github.com/NVlabs/PWC-Net.

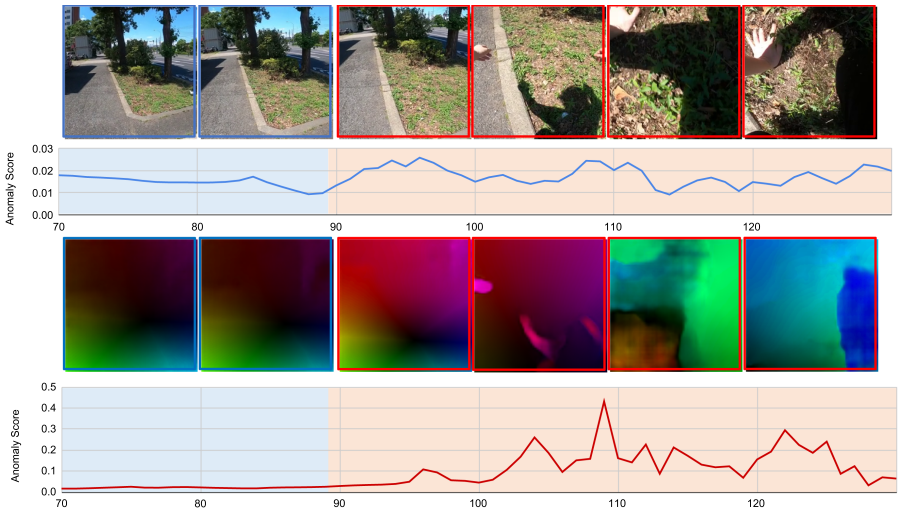

Fig. 6. The temporal transition of the anomaly score by RGB images and optical flow images. The x-axis shows the frame index and the y-axis shows predicted anomaly score. The upper graph shows the anomaly score by RGB images and the lower graph shows the anomaly score by optical flow images. Images surrounded by blue are annotated as normal and images surrounded by red are annotated as abnormal. (Color figure online)

the average of the three data sets except for the person-generic setup. This result shows the accuracy is higher when detecting fast-moving anomalies (stumbling), and lower when detecting slow-moving anomalies (staggering). Even if slow-moving anomalies, however, we were able to achieve an accuracy of over 0.91 on average.

5.3 Comparison of Feature Vector Size

The size of the feature vectors is very important because our method compares the feature vectors of the original and the reconstructed images to predict whether the first person is conducting a normal or abnormal action. Therefore, we gradually increased the vector size from 32 and performed comparative experiments for sizes 32, 64, 128, and 256.

The results of this experiment are in Table 2. The experimental results show that the small vector sizes of 32 and 64 outperformed the large vector sizes of 128 and 256. It can be seen that if the size of the feature vector is large, the feature vector represents the image in detail, whereas if the size of the feature vector is small, the feature vector represents the global information of the optical flow image. As we aim to detect the abnormal events of the first-person, the anomaly is referred to the image globally. Therefore, the small vector size is suitable for our task.

Table 1. AUC results for all setups.

Method	Person-specific					
	Squat down	Stumble	Stagger	Fall down	Collision	Average
Baseline	0.679	0.514	0.850	0.919	0.593	0.711
Ours	**0.981**	**0.915**	**0.831**	**0.969**	**0.951**	**0.929**
Method	Person-generic					
	Squat down	Stumble	Stagger	Fall down	Collision	Average
Baseline	0.639	0.589	0.593	0.813	0.697	0.666
Ours	**0.939**	**0.933**	**0.807**	**0.985**	**0.915**	**0.916**
Method	Person-out					
	Squat down	Stumble	Stagger	Fall down	Collision	Average
Baseline	0.659	0.813	0.689	0.850	0.409	0.684
Ours	**0.945**	**0.990**	**0.922**	**0.998**	**0.892**	**0.949**

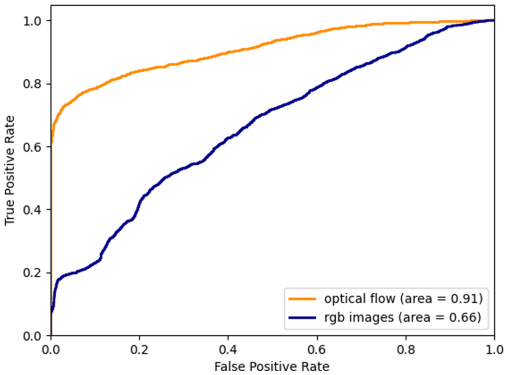

Fig. 7. Comparison of the ROC curve of an optical flow image and consecutive RGB image.

Table 2. Comparison of feature vector size. We compared the feature vectors with different sizes (32, 64, 128, and 256).

Feature vector dim.	Accuracy
32	0.927
64	**0.934**
128	0.908
256	0.877

6 Conclusion

We presented an anomaly detection model that detects first-person abnormalities in gait from an egocentric video. By employing an encoder–decoder–encoder network, we made a model that uses features extracted from an egocentric video and detects anomalies by comparing the features of the original with those of reconstructed images. We also produced a novel first-person video anomaly detection dataset using an egocentric camera. Experiments using our dataset showed that our model outperformed the baseline. The recording of our dataset, however, may have violated other people's privacy, and future work should consider models that more effectively support people in ways that do not violate the privacy of others. Future research should also be done on the possibility of forecasting an individual's abnormal actions using an egocentric camera.

References

1. Akçay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: semi-supervised anomaly detection via adversarial training. In: ACCV, pp. 622–637 (2018)
2. Akçay, S., Atapour-Abarghouei, A., Breckon, T.P.: Skip-GANomaly: skip connected and adversarially trained encoder-decoder anomaly detection. In: IJCNN, pp. 1–8 (2019)
3. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 161–169. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_16
4. Brodie, M., Lord, S., Coppens, M., Annegarn, J., Delbaere, K.: Eight-week remote monitoring using a freely worn device reveals unstable gait patterns in older fallers. IEEE Trans. Bio-Medical Eng. **62**, 2588–2594 (2015)
5. Damen, D., Leelasawassuk, T., Haines, O., Calway, A., Mayol-Cuevas, W.: You-do, i-learn: discovering task relevant objects and their modes of interaction from multi-user egocentric video. In: BMVC (2014)
6. Doshi, K., Yilmaz, Y.: Continual learning for anomaly detection in surveillance videos. In: CVPR Workshops (2020)
7. Furnari, A., Farinella, G.M., Battiato, S.: Recognizing personal contexts from egocentric images. In: ICCV Workshop, pp. 393–401 (2015)
8. Goodfellow, I., et al.: Generative adversarial nets. In: NeurIPS, pp. 2672–2680 (2014)
9. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR, pp. 733–742 (2016)
10. Hua, M., Nan, Y., Lian, S.: Falls prediction based on body keypoints and seq2seq architecture. In: ICCV Workshop, pp. 1251–1259 (2019)
11. Kacorri, H., Kitani, K.M., Bigham, J.P., Asakawa, C.: People with visual impairment training personal object recognizers: feasibility and challenges. In: CHI Conference on Human Factors in Computing Systems, pp. 5839–5849 (2017)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

13. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: International Conference on Machine Learning, pp. 1558–1566 (2016)
14. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection-a new baseline. In: CVPR, pp. 6536–6545 (2018)
15. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: ICCV, pp. 341–349 (2017)
16. Maas, A., Hannun, A., Ng, A.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013)
17. Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: CVPR, pp. 11988–11996 (2019)
18. Nguyen, T.N., Meunier, J.: Anomaly detection in video sequence with appearance-motion correspondence. In: ICCV, pp. 1273–1283 (2019)
19. Nouredanesh, M., Li, A.W., Godfrey, A., Hoey, J., Tung, J.: Chasing feet in the wild: a proposed egocentric motion-aware gait assessment tool. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11134, pp. 176–192. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11024-6_12
20. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
21. Phillips, L., et al.: Using embedded sensors in independent living to predict gait changes and falls. West. J. Nurs. Res. **39**, 78–94 (2017)
22. Qiao, M., Wang, T., Li, J., Li, C., Lin, Z., Snoussi, H.: Abnormal event detection based on deep autoencoder fusing optical flow. In: Chinese Control Conference, pp. 11098–11103 (2017)
23. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR (2016)
24. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: IPMI, pp. 146–157 (2017)
25. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: CVPR, pp. 8934–8943 (2018)
26. Tadesse, G.A., Cavallaro, A.: Visual features for ego-centric activity recognition: a survey. In: ACM Workshop on Wearable Systems and Applications, pp. 48–53 (2018)
27. Yagi, T., Mangalam, K., Yonetani, R., Sato, Y.: Future person localization in first-person videos. In: CVPR, pp. 7593–7602 (2018)
28. Yao, Y., Xu, M., Wang, Y., Crandall, D.J., Atkins, E.M.: Unsupervised traffic accident detection in first-person videos. In: IROS, pp. 273–280 (2019)
29. Yuan, Y., Kitani, K.: 3D ego-pose estimation via imitation learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 763–778. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_45
30. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time PD control. In: ICCV, pp. 10082–10092 (2019)
31. Zenati, H., Foo, C.S., Lecouat, B., Manek, G., Chandrasekhar, V.R.: Efficient GAN-based anomaly detection. arXiv preprint arXiv:1802.06222 (2018)