Domain generalization for activity recognition: Learn from visible, infer with thermal

Yannick Zoetgnande, Jean-Louis Dillenseger

Domain generalization for activity recognition: Learn from visible, infer with thermal

Yannick Zoetgnande and Jean Louis Dillenseger[1]
[1] *Univ Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France*
yannickzoet@live.fr, jean-louis.dillenseger@univ-rennes1.fr

Keywords: Fall detection; Activity recognition; Domain generalization; Thermal images

Abstract: We proposed a solution based on I3D and optical flow to learn common characteristics between thermal and visible videos. For this purpose we proposed a new database to evaluate our solution. The new model comprises an optical flow extractor; a feature extractor based on I3D, a domain classifier, and an activity recognition classifier. We learn invariant characteristics computed from the optical flow. We have simulated several source domains, and we have shown that it is possible to obtain excellent results on a modality that was not used during the training. Such techniques can be used when there is only one source and one target domain.

1 Introduction

Nowadays, in most countries, the population is becoming older and older. Aging is one of the main reasons for frailty (Ambrose et al., 2013). As pointed by (Khan and Hoey, 2017), there are 2.6 falls per person per year in retirement homes. And given the increasing number of seniors and the limited number of retirement homes, some seniors must stay at home. Thus, fall detection will concern domestic context than retirement homes. To prevent falls, it is necessary to survey seniors for a long time. But there are a lot of issues regarding this surveillance, such as non-invasive and privacy.

Vision-based sensors have proven to be very efficient while being passive. Many works were focused on visible images (RGB). But this type of image poses a huge problem of privacy intrusion. Some other vision sensors (depth or thermal sensors) do not allow explicit recognition of people and are therefore relevant for activity monitoring while respecting privacy. Thermal sensors have many other advantages; they work day and night, are easier to analyze than normal images in daily activities classification (Vadivelu et al., 2016); but are generally very expensive. To reduce this cost, some authors proposed to use low-cost thermal cameras, but with very low resolution (80x60 pixels), of type Lepton 2 from FLIR (Zoetgnande et al., 2020; Zoetgnandé et al., 2019).

The problem that arises then is how to use this type of image in the context of deep learning. Indeed, if there is a lot of work in this field, most people have mainly focused on RGB or depth (usually Kinect) sensors. Thus, many comprehensive databases are available for these modalities. On the other hand, only a few data exist for thermal images and even less for low-resolution ones. The main dataset on thermal images are (Vadivelu et al., 2016), (Peñafort-Asturiano et al., 2018) and (Martínez-Villaseñor et al., 2019). In (Vadivelu et al., 2016), the authors provide a dataset captured from FLIR ONE thermal cameras. The resolution of images was 640×480. In (Peñafort-Asturiano et al., 2018), the authors acquired a dataset named FALL-UP by combining 6 infrared sensors, 2 normal cameras, 1 EEG, and 5 wearable sensors. In (Martínez-Villaseñor et al., 2019), the authors provide a dataset named UP-Fall by combining 6 infrared sensors, 2 cameras, 5 IMUs, a gyroscope, ambient light, and 1 EEG. But even they are using thermal cameras, they only provided the thermal images features and no raw data.

In the literature, activity recognition domain generalization is most of the time focused on IXMAS (Weinland et al., 2006). This dataset has been widely used as a cross-view action recognition benchmark (Li et al., 2018; Li et al., 2019). The videos have been collected from five different views. We think that generalization through different views does not demonstrate perfectly the performance of a domain generalization algorithm. In this dataset, different persons perform the same actions in different ways, so it is common that a model might not be able to recognize

actions performed by new subjects not seen during training. Also, we cannot expect a model trained using indoor data to work well outdoors. This is why in this paper, we propose that the source and the target model do not share either the same persons, either the same views and even more neither the same modality (visible spectrum vs. thermal).

This paper will then try to answer the following question: is it possible to learn from classical modality datasets (e.g., RGB) and infer thermal data. The main idea is then to train some networks on data collected and labeled for a fairly complete study of activity recognition from RGB images (Tran et al., 2018b) and verify the accuracy of the inference on a much smaller database of low-resolution thermal images. To our best knowledge, it is the first time domain generalization is performed from visible to thermal. The source code is open[1].

Our contributions are threefold:

- This paper fills a gap in the literature on fall detection and activity recognition with thermal cameras. Indeed, most of the previous works learn and infer on the same database. In this paper, we propose to go further.
- We proposed to use different modalities. We proposed a new model for domain generalization from visible videos to thermal videos.
- We proposed a new toy activity recognition/fall detection dataset with thermal images. Even if this new dataset is relatively small compared to other state-of-the-art datasets, it allows us to prove the efficiency of our method.

2 Related works

2.1 Activity recognition network architectures

Like in many other fields, methods based on deep learning have got outstanding results in activity classification. Most of these methods have been designed for visible or depth videos classification. In (Carreira and Zisserman, 2017), the authors propose I3D a two-stream architecture based on the image classification inception architecture (Szegedy et al., 2016).

Most of the time, it is not easy to find the right architecture. A neural architecture search can tackle such a problem. In (Piergiovanni et al., 2019), the authors propose a hybrid architecture based on inflated Temporal Gaussian Mixture (iTGM) based on ResNet (He et al., 2016). The iTGM is based on the 1D Temporal Gaussian Mixture (TGM) proposed in (Piergiovanni and Ryoo, 2019). Some parameters of their ResNet are fixed while others evolve; thus, the search space is limited. In (Ryoo et al., 2019) the authors propose a similar approach by evolving a multi-stream neural network. Their network is composed of a convolution block by alternating 2D and (2+1)D residual modules.

While I3D-based approaches are accurate but computationally costly, in (Tran et al., 2018a), Tran et al. propose to separate spatial and temporal components in a new spatiotemporal convolutional block named R(2+1)D based on ResNet architecture. Like in (Piergiovanni et al., 2019), the authors use a 2D convolution for spatial dimension and a 1D convolution for temporal dimension. They show that this architecture, based on Factorized Spatio-temporal convolutional Networks (Wang et al., 2017), is easier to optimize than 3D convolutions. Technically they replace ν_i 3D convolutional filters of size $(\nu_{i-1} \times t \times d \times d)$ with μ_i (2+1)D blocks composed of μ_i 2D convolutional filters of size $(\nu_{i-1} \times 1 \times d \times d)$ and ν_i 1D convolutional filters of size $(\mu_{i-1} \times 1 \times d \times d)$. They choose $\mu_i = \left\lfloor \frac{td^2\nu_{i-1}\nu}{d^2\nu_{i-1}+t\nu_i} \right\rfloor$. They show that their ResNet-based approach beats many state-of-the-art methods with lower computational complexity. As a conclusion of this review of deep learning activity recognition methods, we believe that I3D and R(2+1)D architectures are the most suited for our problem.

2.2 Domain generalization

2.2.1 Visible domain and related

One of the problems we faced for thermal-based activity recognition is the scarcity of datasets. Original domain generalization is used to adapt a model trained with a dataset into another dataset. Domain generalization is linked to domain adaptation and few-shot learning. But while in these techniques, the model sees the target data during training, in domain generalization, the model does not see the target dataset during training.

The main reason why we want to apply domain generalization to our problem is that data acquisition takes time and annotation is time-consuming. Training a model using RGB video and inferring on Thermal videos will save us time. Indeed, there is a huge number of datasets for RGB images, but thermal datasets are sparser.

[1]The code is at https://github.com/2021submissions/ICPRAM_submission

A jth domain is defined as $\{(d_i, x_i, y_i)\}_{i=0}^{n} \sim (\mathcal{D}_j, \mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are the set of data inputs and labels. Generally we have m sources domains. The main idea of domain generalization is to learn to classify a data from unseen domain $d_{us} \notin D_m$ with D_m the set of available domain during training.

They are many ways to perform domain generalization: domain invariant features, hierarchical models, data augmentation and optimization algorithms.

We will be more focused about domain invariant features. The main idea is to learn a representation Φ such that $P(\Phi(x^d))$ is the same whatever the domain.

In (Albuquerque et al., 2019), the authors also propose to make training distributions indistinguishable. They define a domain as $\langle \mathcal{D}, f \rangle$ where $\mathcal{D}$ is the probability distribution over $\mathcal{X}$ and f the deterministic labeling function defined as $f : \mathcal{X} \to \mathcal{Y}$. They also define $h \in \mathcal{H}$ with $\mathcal{H}$ a set of candidate hypothesis with $f : \mathcal{X} \to \mathcal{Y}$.

A risk R or source error can associated to h on domain $\langle \mathcal{D}, f \rangle$ as follows (Ben-David et al., 2010):

$$R[h] = \mathbb{E}_{x \sim \mathcal{D}} \mathcal{L}[h(x), f(x)] \tag{1}$$

with $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to R_+$ representing how different $h(x)$ and $f(x)$.

2.2.2 From visible to thermal

In the literature, there are some works regarding visible-to-transfer learning. In (Akkaya et al., 2021), the authors proposed a domain adaptation model with three phases pre-training, warm-up, and pseudo-labeling. Their proposed model is composed of three modules: Source CNN, Target CNN, Classifier, and Discriminator.

In (Liu et al., 2018), the author proposes a work that is close to ours. They extract features using iDTs (Wang and Schmid, 2013) + LLC + PCA method. Then they used kernel manifold alignment to create a common latent space of thermal and visible features. Finally, aligned-to-generalized encoders are used to classify activities. Their method is more related to transfer learning given that the model is trained with visible and thermal video, while we want to only train the model with visible video and infer with thermal video.

2.3 Activity monitoring dataset

Regarding data availability, there is a lot of data for human activity recognition (Carreira et al., 2018; Carreira et al., 2019). Unfortunately, most of these datasets concern visible images. In the specific case of monitoring human activities as a goal of fall detection, the literature is also prolific regarding visible image (Charfi et al., 2012) and recently about Kinect images (Adhikari et al., 2017).

In (Tran et al., 2018b), the authors propose a fall detection dataset with 8 falls types, 50 participants (20 women and 30 men). To collect the dataset, they use 7 overlapped Kinect and visible cameras and 2 WAX3 wireless accelerometers. We used this dataset as the source and collected a new dataset as the target.

There are a few dataset using thermal images. In (Martínez-Villaseñor et al., 2019), the authors propose UP-Fall a fall detection dataset composed of 255 videos. They used 17 participants (8 men and 9 women). To acquire the dataset, they use 6 infrared sensors, 2 visible cameras, 5 IMUs with an accelerometer, a gyroscope, an ambient light, and 1 EEG.

In (Peñafort-Asturiano et al., 2018), the authors propose a new fall detection dataset composed of 255 videos. This dataset is acquired using 6 infrared sensors, 2 cameras, 1 EEG, 5 wearable inertial sensors.

3 Materials and methods

3.1 Dataset

Given that the dataset available in the state of the art was not satisfactory, we decided to collect a new dataset for thermal activity recognition. This is one of the core contributions of our paper. To acquire the images, we use Lepton 2.5 FLIR cameras (Fig 1). In the Table 1, there are the specifications of lepton 2.5. These cameras are characterized by low resolution, noise, sudden brightness change (due to clutter refreshment), and halo effect. The main advantage is that these cameras are low-cost, thus can be used for industrial applications or for anyone wanting to install them in their homes.

We used two cameras set in stereo to take advantage of the depth estimation. We set the stereo baseline to 16 cm for a tradeoff between congestion and the field of the view of the cameras.

We collected 713 videos following the same actions described in (Tran et al., 2018b): *walk*, *hand pick up*, *fall*, *crawl*, *sit then stand up*, *sit then fall*, *lie then sit up* and *lie then fall*. Contrary to many other datasets in literature, our dataset has been acquired in real apartment situations. These rooms were furnished and therefore presented many occlusions to the cameras. Moreover, these data have been acquired for various room temperatures.

Someone might suggest that our dataset is not big

enough. Indeed, compared to other works in the literature we do not have enough data. However, we do not have the same objectives as these works. Indeed we place ourselves in the situation where a person could train a model a big dataset and deploy this model. So knowing that the performance of his model will not decrease even if the inference dataset is different from the training dataset.

Someone might suggest that our database is not large enough and does not contain enough views to do domain generalization. Compared for example to the IXMAX database (which is the standard database for domain generalization in activity detection), we have enough videos. Indeed, authors usually use only the first 5 actions (compared to our 8 actions). Moreover, they usually use 6 people (Alba, Andreas, Daniel, Hedlena, Julien and Nicolas) not taking into account irregular actions.

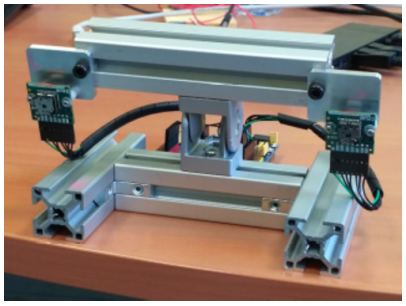


Figure 1: The device we used to capture the images. We registered the two views but used only the left views in this paper. In future work, we will include the second view.

Dimension	8.5 x 11.7 x 5.6 mm
Resolution	80 (h) x 60 (v) pixels
Pixel size	17 μm
Field of view	51° (h) x 37.83° (v)
Thermal sensitivity	<50 mK
Accuracy	± 5°C
Frame rate	9 Hz
Dynamic range	-10 to 140°C
Price	<200$

Table 1: Characteristics of Lepton 2.5

3.2 Domain generalization

One solution could be to train directly using the dataset we have collected for fine-tuning. Unfortunately, we considered that we didn't have enough videos (713 videos) to evaluate. This is why we proposed to learn from visible videos and infer what we have learned from thermal videos.

action id	action name	Number of videos
1	walk	91
2	hand pick up	91
3	lie then sit	81
4	sit then stand up	91
5	crawl	91
6	lie then fall	81
7	sit then fall	91
8	fall	91

Table 2: Number of videos per class for our dataset Baga

3.2.1 Source dataset

For source dataset, we use the dataset CMDFALL provided by (Tran et al., 2018b). We choose this dataset because it involves 50 people and seven viewpoints with 20 activities. We modified the dataset for our purpose. First, they use 20 activities with some activities such as right fall and left fall. We have grouped with the activities that only differ in the sense that they have been performed *left* or *right*. Besides, we did not consider one of the cameras because it was placed on the ceiling.

Moreover, given that their frame rate is 20 fps, we temporally down-sampled the video to 5 fps. As a result, we delete videos where the activities lasted less than 3 seconds. In the end, we got 6235 videos with some imbalance between classes (Table 3).

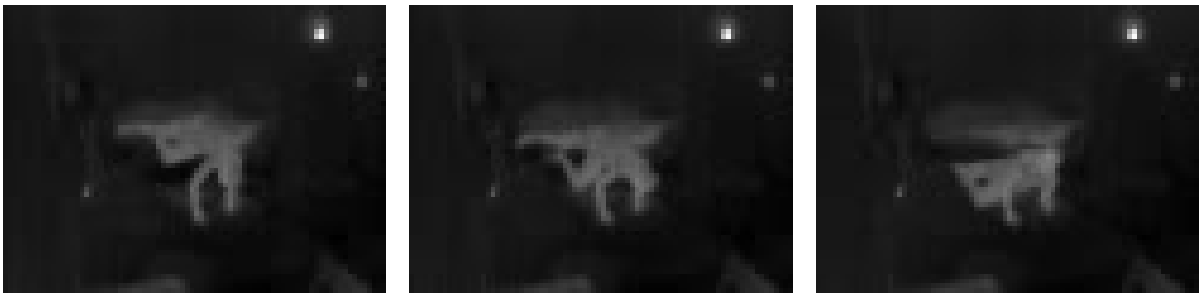
acton id	action name	number of videos
1	walk	1972
2	hand pick up	319
3	lie then sit	311
4	sit then stand up	1017
5	crawl	277
6	lie then fall	834
7	sit then fall	537
8	fall	958

Table 3: Number of videos per class for the original cleaned dataset CMDFALL (Tran et al., 2018b)

3.2.2 Data augmentation

As shown in Table 3, one of the problems we faced is that the cleaned CMDFALL dataset is imbalanced. There are many solutions to learn from imbalanced dataset (Johnson and Khoshgoftaar, 2019): data-level methods, algorithm-level methods, and hybrid methods. We chose to apply the data-level method through over-sampling with additive data augmentations in order to reach 1972 videos per class.

The main problem of oversampling is over-fitting. Thus, we had a random affine transformation and ran-



(a) N° 12 (b) N° 13 (c) N° 14

Figure 2: Baga dataset: frames during a fall from a bed

dom occlusion to any video besides traditional data-augmentation methods. The main idea is to be able to get an infinite number of videos from one video. The additive data augmentation techniques are described as follows:

Random occlusion Randomly generate a rectangle with a random size $r_m \times r_n$ with $r_m \in [0,m]$ and $r_n \in [0,n]$. The rectangle is randomly moved during the video.

Random affine transformation Given a video $v = \{I_0, \cdots, I_{t-1}\}$ with a frame size of $m \times n$. For a given video a transformation matrix is randomly generated and applied to the frames.

3.2.3 From one source to three sources

To perform domain generalization, it is possible to consider one source domain and one target domain (Qiao et al., 2020). In this work, we decide to create artificially three source domains from one source. Given a visible RGB video (**source 0**), we simulated two more sources by computing Sobel edges (**source 1**) and Laplacian edges (**source 2**).

3.2.4 Domain generalization model

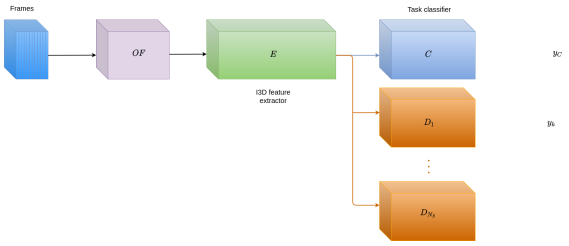


Figure 3: Domain generalization model ($Model_{dm}$)

For domain generalization, we adapted the model proposed by (Albuquerque et al., 2019). Our model, called $Model_dm$ is composed of a pretrained flownet model to extract optical flow, a feature extractor based on I3D model, a task classifier, and domain classifiers.

Our main idea is to make the distributions we created indistinguishable. A domain is defined as $\langle \mathcal{D}, f \rangle$ where $\mathcal{D}$ is the probability distribution over $\mathcal{X}$ and f the deterministic labeling function defined as $f : \mathcal{X} \to \mathcal{Y}$. They also define $h \in \mathcal{H}$ with $\mathcal{H}$ a set of candidate hypothesis with $f : \mathcal{X} \to \mathcal{Y}$.

A source error h is defined on domain $\langle \mathcal{D}, f \rangle$ as follows:

$$R[h] = \mathbb{E}_{x \sim \mathcal{D}} \mathcal{L}[h(x), f(x)] \tag{2}$$

Thus the $\mathcal{H}$-divergence is defined as follows:

$$d_{\mathcal{H}}[\mathcal{D}_S, \mathcal{D}_T] = 2sup_{\eta \in \mathcal{H}} \times \left| Pr_{x \sim \mathcal{D}_S}[\eta(x) = 1] - Pr_{x \sim \mathcal{D}_T}[\eta(x) = 1] \right| \tag{3}$$

While in domain adaptation, there is only one source and one target, in domain generalization, there are many sources and one target domain. Using a one-vs-all classification allows reducing the number of combinations and the complexity of the model.

Given a video input *vid*, our domain generalization model is composed of 4 modules (Fig 3):

1. A optical flow module $OF(vid)$ that computes the optical flow online. We used flownet2. The parameters θ_{opt} of this module are frozen.
2. A feature extraction E (with parameters θ_E) module must be able to extract features that must be invariant in terms of action classification and dissimilar in terms of domain classification.
3. A task classifier C (with parameters θ_C) that is trained to classify $E(OF(vid))$ in terms of activities.
4. Domain classifiers D_{N_S} that is trained to classify $E(OF(vid))$ in terms of domains.

The loss function of the model is defined as follows:

$$\min_{\theta_E,\theta_C}\max_{\theta_1,\cdots,\theta_{N_S}} \mathcal{L}_C\left(C(E(x;\theta_E);\theta_C),y_C\right) - \sum_{k=1}^{N_S}\mathcal{L}_k\left(D_k(E(x;\theta_E);\theta_k),y_k\right) \quad (4)$$

where y_C is the activity label $\in \{0,1,2,3,4,5,6,7\}$ of x, and y_k is the domain label that is to say $y_k = 1$ if $x \sim \mathcal{D}_S^k$,0 otherwise. $\mathcal{L}_C$ is the error between the predicted activity and the ground truth activity while $\mathcal{L}_k$ is the error between the predicted domain and the ground truth domain.

4 Experiments

4.1 Settings

4.1.1 Dataset

As aforementioned, we used two types of datasets. A visible dataset was provided in the literature that we cleaned and a thermal dataset we collected Table 2. The frame rate of the two datasets is set to 5fps and the resolution is 80x60. Each activity lasts at least 3 seconds and for training (and testing) 15 frames are randomly selected. In Fig 2, some example of Baga dataset are displayed. These frames represent a person falling from bed.

Regarding the source dataset, videos have been originally recorded from 50 subjects. We use the first 40 persons for training and the remaining 10 subjects from testing. The source and the target datasets can be downloaded through this link tinyurl.com/eupz32ns

4.1.2 Training settings

We use a pretrained I3D backbone on the Kinetics dataset. The domain classifier is composed of two fully connected layers, followed by a ReLU and a dropout. There is also a last fully connected layer, followed by a softmax layer. The task classifier is composed of an average pooling followed by a dropout, a 3D convolutional layer, and a softmax layer. The model is trained using stochastic gradient descent for 20 epochs, and the initial learning rates are 0.01 for both the task and the domain classifier.

We evaluated three models:

1. *Baseline*: This model is composed of the traditional training. Here, we followed the procedure presented in (Carreira and Zisserman, 2017). We first extracted the optical flow using TVL1 and then the model is trained normally without the domain generalization. That is to say this model is composed only of E and C of Fig 3. We used traditional data augmentation techniques: flipping, rotation and random cropping. In this model we used optical flow of visible videos as source and optical flow of thermal videos as target. The optical flow data can be download through this link shorturl.at/jnsFM.

2. $Model_1$: This model is composed of the module OF, E and C. In this model we used optical flow of visible videos as source and optical flow of thermal videos as target. OF is the optical flow model which is a pretrained version of flownet2. E is the features extractor and C the classifier.

3. $Model_2$: This model is composed of the module OF, E, C and $D_i's$ but during training we used traditional data augmentation techniques and the data augmentation techniques we proposed in this paper.

4.2 Results

We reported the results of two models compared to the baseline method. In terms of accuracy and F1 score (Table 4), it is easily noticeable that the *baseline* is not able to categorize videos. Indeed this simple model cannot generalize to a new domain even with the same type of input (*i.e.* optical flow). $Model_1$ is similar to the baseline method except that the optical flow is calculated differently. Even if we are using the same input (optical flow) as for $Model_2$, the model cannot generalize to an unseen domain without the discriminator modules. Indeed we got 60.45% of accuracy and 59.68% of F1-score, which is better than the *baseline* from a big margin. These results are not good compared to those of $Model_2$. Indeed, the model we proposed output 72.93% of accuracy and 71.56% of F1-score. So Table 4 shows that using domain generalization, we can learn activities from a domain to another domain through inferring.

The figures 4 and 5 support this idea. In these figures, we computed the confusion matrices for $Model_1$ and $Model_2$. While $Model_1$ is able to classify some activities such as *fall*, *pick* and *fall* but fails to classify *lie_sit*.

Both models ($Model_1$, $Model_2$) struggle to classify well the action *lie_sit*. $Model_1$ tends to confound this activity with fall. In order to improve the results, we could include the second view of our device Fig 1.

Model	Accuracy/F1 score %
Baseline	44.50/42.56
Model1	60.45/59.68
Model2	72.93/71.56

Table 4: Accuracy/F1 score of the evaluated methods

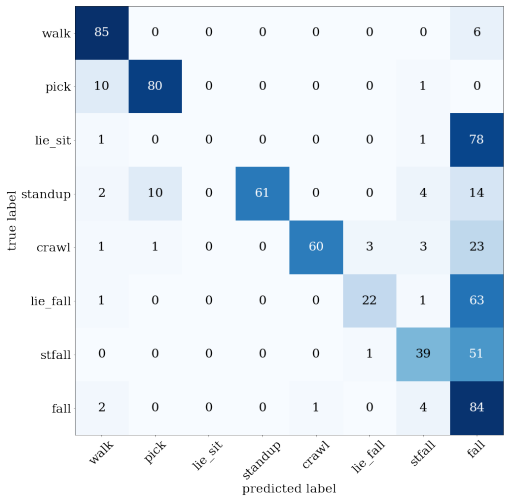


Figure 4: Results on the test dataset with $Model_1$

5 Conclusion

In this paper, we presented a new dataset for fall detection and activity recognition. The dataset has been acquired in real apartments contrary to many datasets in the literature.

Then, we proposed a domain generalization model for transfer learning. Rather than learning directly from raw video, we proposed a model based on optical flow. We showed that using flownet to extract optical flow from simulated three sources (raw video, Sobel videos, and Laplacian videos) we were able to obtain better results compared to the baseline method and $Model_1$.

We wished to acquire more data and compared the domain generalization method vs. domain adaptation and fine-tuning.

It has also been proven in the literature that multi-view frameworks can improve results. Adding more views should bring more information to the model and increase the performance.

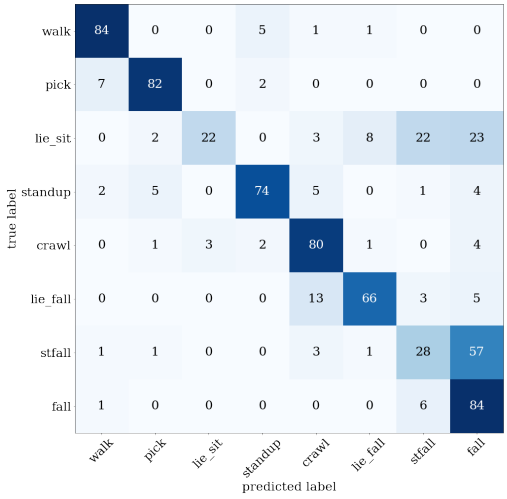


Figure 5: Results on the test dataset with $Model_2$

6 ACKNOWLEDGMENTS

This work was part of the PRuDENCE project (ANR-16-CE19-0015) which has been supported by the French National Research Agency (ANR).

REFERENCES

Adhikari, K., Bouchachia, H., and Nait-Charif, H. (2017). Activity recognition for indoor fall detection using convolutional neural network. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 81–84. IEEE.

Akkaya, I. B., Altinel, F., and Halici, U. (2021). Self-training guided adversarial domain adaptation for thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4322–4331.

Albuquerque, I., Monteiro, J., Falk, T. H., and Mitliagkas, I. (2019). Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*.

Ambrose, A. F., Paul, G., and Hausdorff, J. M. (2013). Risk factors for falls among older adults: a review of the literature. *Maturitas*, 75(1):51–61.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A. (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Carreira, J., Noland, E., Hillier, C., and Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset.

In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Charfi, I., Miteran, J., Dubois, J., Atri, M., and Tourki, R. (2012). Definition and performance evaluation of a robust svm based fall detection solution. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, pages 218–224. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Khan, S. S. and Hoey, J. (2017). Review of fall detection techniques: A data availability perspective. *Medical engineering & physics*, 39:12–22.

Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.-Z., and Hospedales, T. M. (2019). Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. (2018). Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Liu, Y., Lu, Z., Li, J., Yao, C., and Deng, Y. (2018). Transferable feature representation for visible-to-infrared cross-dataset human action recognition. *Complexity*, 2018.

Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., and Peñafort-Asturiano, C. (2019). Up-fall detection dataset: A multimodal approach. *Sensors*, 19(9):1988.

Peñafort-Asturiano, C. J., Santiago, N., Núñez-Martínez, J. P., Ponce, H., and Martínez-Villaseñor, L. (2018). Challenges in data acquisition systems: Lessons learned from fall detection to nanosensors. In *2018 Nanotechnology for Instrumentation and Measurement (NANOfIM)*, pages 1–8. IEEE.

Piergiovanni, A., Angelova, A., Toshev, A., and Ryoo, M. S. (2019). Evolving space-time neural architectures for videos. In *Proceedings of the IEEE international conference on computer vision*, pages 1793–1802.

Piergiovanni, A. and Ryoo, M. (2019). Temporal gaussian mixture layer for videos. In *International Conference on Machine Learning*, pages 5152–5161.

Qiao, F., Zhao, L., and Peng, X. (2020). Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565.

Ryoo, M. S., Piergiovanni, A., Tan, M., and Angelova, A. (2019). Assemblenet: Searching for multi-stream neural connectivity in video architectures. *arXiv preprint arXiv:1905.13209*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018a). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Tran, T.-H., Le, T.-L., Pham, D.-T., Hoang, V.-N., Khong, V.-M., Tran, Q.-T., Nguyen, T.-S., and Pham, C. (2018b). A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1947–1952. IEEE.

Vadivelu, S., Ganesan, S., Murthy, O. R., and Dhall, A. (2016). Thermal imaging based elderly fall detection. In *Asian Conference on Computer Vision*, pages 541–553. Springer.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558.

Wang, Y., Long, M., Wang, J., and Yu, P. S. (2017). Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2-3):249–257.

Zoetgnande, Y. W. K., Cormier, G., Fougeres, A.-J., and Dillenseger, J.-L. (2020). Sub-pixel matching method for low-resolution thermal stereo images. *Infrared Physics & Technology*, 105:103161.

Zoetgnandé, Y. W. K., Fougères, A.-J., Cormier, G., and Dillenseger, J.-L. (2019). Robust low resolution thermal stereo camera calibration. In *Eleventh International Conference on Machine Vision (ICMV 2018)*, volume 11041, page 110411D. International Society for Optics and Photonics.