



HAL
open science

A dataset of human actions for abnormal event recognition in assistive robotic environments

Catherine Huyghe, Nacim Ihaddadene

► To cite this version:

Catherine Huyghe, Nacim Ihaddadene. A dataset of human actions for abnormal event recognition in assistive robotic environments. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun 2023, Vancouver, Canada. hal-04130886

HAL Id: hal-04130886

<https://hal.science/hal-04130886v1>

Submitted on 16 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A dataset of human actions for abnormal event recognition in assistive robotic environments

Catherine Huyghe
JUNIA

catherine.huyghe@junia.com

Nacim Ihaddadene
JUNIA

nacim.ihaddadene@junia.com

Abstract

Human action recognition and segmentation are important tasks in the context of assistive robotics and active assisted living. Multiple datasets of human actions are available and cover a large set of general or daily activities, but they less cover the case of abnormal events. When some specific actions are available, they are not adapted for the context of assistive robotics. In this paper we will describe a new dataset that covers some specific events which are useful in the context of active assisted living. We will also describe an approach for human action recognition based on human body semantic segmentation, to focus on the human body and deal with partial or slow movements and immobility.

1. Introduction

Understanding human actions is still an important task in artificial intelligence. Computer vision-based approaches are the most widely used techniques and lead to different sub-tasks such as: human detection, body part segmentation, pose detection, human action recognition (HAR) and human action segmentation (HAS).

In recent years, the development of robotics allowed the use of robots in active assisted living (AAL). They are used at different levels to allow vulnerable people, due to age, illness, or disability, to gain a certain level of autonomy. They could be used, for example, to alert caregivers or family members in the case of dangerous or abnormal situations.

Mobile assistive robots embed one or multiple camera sensors for localisation, obstacle avoidance and user interaction. These sensors are increasingly used for user behaviour understanding. However, some constraints of mobile assistive robotics make this task more complicated and show the limitations of optical flow based methods, which are currently the most widespread approaches in the state of the art.

There is a large number of available datasets to train and

test action recognition models. They cover many everyday actions, but fewer rare and abnormal events and situations (Falls, fainting, or abnormal and dangerous postures). There are only few cases that could be exploitable for AAL and daily activity monitoring.

2. Related works

Human action understanding in images and video sequences is an important task in computer vision and still a challenging problem. We focus in this work on human action recognition, a classification problem that associates an action to a data entry, and each data entry corresponds to a unique action. This task could be performed from different types of inputs and modalities: Visual data (RGB images and sequences, RGBD or depth images, infrared, 3D point clouds, ...), or non visual data [25] (Accelerometers, audio, and WIFI) provided by on-body devices or from ambient sensors. We distinguish human action recognition from human action detection and segmentation, which consist on searching for an action in a temporal stream or splitting temporal data (Visual or not) into small sequences that are associated semantically to the same action.

2.1. Human action recognition from videos

This section gives an overview of existing approaches for HAR from video sequences. In our context, we are interested by RGB input which is present and widely used in assistive robotics. Independently from action segmentation, action recognition process generally follows two main steps: pre-processing and classification.

Action classification methods have evolved considerably in recent years [1, 5, 22]. We focus in our study on the techniques based on the use of deep neural networks. Different architectures for action recognition are proposed in the literature [5]. They are distinguished by :

The input data: While some methods directly use the input image sequence, others pre-process the input sequence to extract mid-level information. The most used pre-processing operation is the estimation of optical flow, which

is used to locate and characterize motion. It is enriched with a second stream containing the original images allowing it to keep information on the position and general posture of the person. The processing of the two separate streams is then merged [22]. In recent years, methods based on models of the human body have been developed, such as methods based on the detection and localization of people [15] and those based on the detection of the human skeleton [6]. Methods based on models of the human body allow a better consideration of the human, its position and posture.

The spatio-temporal dimension: Different approaches were proposed to deal with the spatial and temporal aspects of human actions. They are based on LSTM [8,26], 3D convolutions [13], or uses optical flow estimation [22] [10] [5]. Recently, transformer based approaches for video classification were proposed [1]. and outperformed convolutional and RNN based approaches. They require large datasets, but active developments tend to adapt them to deal with small-size datasets.

2.2. HAR Datasets

There is a large number of uncontrolled datasets for action recognition. The most commonly used in the literature are HMDB [17], UCF-101 [23], ActivityNet [4], charades [21], HowTO100M [20] and recently Kinetics [14]. Most of the videos in these uncontrolled action recognition datasets are from videos collected from Internet video platforms.

The datasets differ from each other in the number of videos and action classes, the number of subjects in the videos, the background of the videos, the appearance and variations of the actions, the camera movement, the quality of the video, etc. Despite these differences, unconstrained datasets for action recognition remain limited in size and diversity. Many datasets used in the literature deal with similar actions. For example, many datasets deal with actions related to sports, interaction between multiple people, or interaction of a person with an object. However, there is little or no data to recognize rare phenomena such as falls or dangerous postures. Some outdoor datasets are dedicated to the detection of abnormal events [19] and anomalies [11, 24] in public areas. For indoor scenarios, the datasets focus on fall detection [2, 12, 16].

2.3. Limitations

It’s crucial to notice that the limitations we’re outlining in this section are related to HAR applications in the context of AAL and assistive robotics. Most approaches consider human action recognition as a video classification problem and focus on global analysis of image sequences instead of targeting human subjects. The emphasis is on analysing the image as a whole, and considering movement rather than focusing on the human body. Also, methods

based on generic feature extractors are interesting for video classification, but less for action recognition. For the case of optical flow based methods, it is more complicated when it comes to applications in AAL. These methods are useless in case of immobility and have bad results for partial movements. Finally, this analysis is not adapted for assistive mobile robotics where robots could move and rotate. These variations and transformations generate a supplementary optical flow in addition of the moving (or static) subjects.

We consider that it is necessary to detect the human in the pre-processing stage and to extract a richer representation that is more consistent with human behaviour understanding.

3. JARD action recognition dataset

In this work, we propose a new dataset ¹ to train, test or validate human action recognition models. The dataset was recorded in an experimental apartment with different rooms. The user panel is composed of 20 persons for the first version of the dataset. Some actions represent daily normal activities and situations, such as sitting or walking. Another subset of actions present dangerous situations and actions, such as falling down and lying on the ground. The dataset contains also some normal actions with a big similarity to the abnormal one to distinguish them. Some actions presents similar postures, but the context make them different. For example, lying on the ground is different from lying on bed.

| Action | Motion | Abnormal |
|--------------------------|--------|----------|
| Falling down | Yes | Yes |
| Heart attack | Yes | Yes |
| Bending down | Yes | No |
| Sitting down | Yes | No |
| Sitting | No | No |
| Lying on bed or similar | No | No |
| Lying on the floor | No | Yes |
| Standing Up | Yes | No |
| Getting up off the floor | Yes | Yes |
| Walking | Yes | No |

Table 1. JARD V1 dataset list of actions.

The camera sensors were mounted on a mobile robot at the heights 30cm and 1m50. This allows a various set of configurations of cameras positions, backgrounds. It allows also to move and rotate cameras during the capture of the actions. When some videos present static backgrounds, others present dynamic backgrounds due to the movements of the robot (Translation, rotation or both).

¹Dataset URL available soon






| | Dangerous situations | Normal situations but similar to dangerous ones | |
|--------------------|--|--|---|
| Dynamic Situations |  <p>Falling</p> |  <p>Bending down</p> |  <p>Sitting down</p> |
| | Static Situations |  <p>Lying on ground</p> |  <p>Lying on bed or similar</p> |

Figure 1. Examples of events from JARD Dataset.

The first version of the dataset contains 9 classes detailed in Table 1. Those classes were chosen to be adapted to the context of AAL. Each action contains about 200 videos. Videos have different durations depending on the nature of action, This goes from 2 seconds to more than 20 for some videos.

4. HAR based on semantic segmentation

In this section, we will present an approach for human action recognition in the context of assistive robotics and AAL. The main objective is to increase safety through intelligent monitoring by detecting abnormal and rare situations (domestic accidents, falls, discomfort, dangerous postures, immobility, etc.).

To deal with optical flow limitations, we replaced the pre-processing step by a human body semantic segmentation [7, 18]. The approach is interesting from the point of view of multi-task networks that are embedded in assistive robots. The semantic segmentation of the scene could be used as input for other models that perform complementary applications. This allows to develop a global understanding model instead of separating and focusing on different tasks.

For the classification step, we used a video vision transformer. The results of attention-based models in the do-

main of natural language processing inspired a lot of approaches in computer vision. Recently, vision transformers [9] appeared to be an efficient alternative to convolutional neural networks (CNNs). The idea is to integrate them in conjunction with convolutional layers or to apply directly a pure transformer on sequences of image patches. For video classification, vision transformer based architectures for video sequence processing were proposed. Video Vision Transformer (ViViT) [1] was proposed as a pure-transformer based models for general video classification and understanding. However, these architectures seem to require more data and stronger regularisation when applied to RGB data directly.

Figure 2 shows some examples of semantic segmentation of images from JCARD dataset. The first and the second columns correspond to static situations (Lying on bed, Lying on ground) where optical flow do not generate significant data. The third and the last columns correspond to bending and falling down events.

In this step, an architecture based on transformers is used to classify the pre-processing output. Multi-headed self-attention layers are used to classify the sequence of tokens obtained by embedding. Multiple embedding strategies are available, such as Uniform Frame Sampling and Tubelet Embedding.

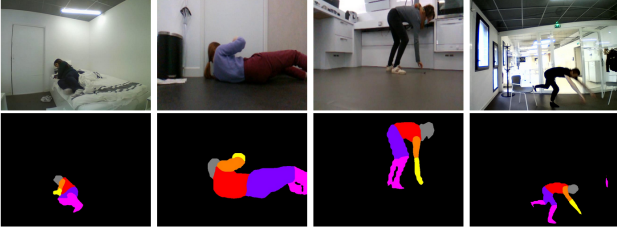


Figure 2. Examples of human body semantic segmentation on videos with mobile camera.

Figure 3 explains this tokenisation process. Smaller tubelet dimensions imply an increasing number of tokens, which increases the computation time. When the temporal information from different frames is fused by the transformer in the case of uniform frame sampling, this information is maintained in the case of tubelet embedding due to volumetric nature of patches.

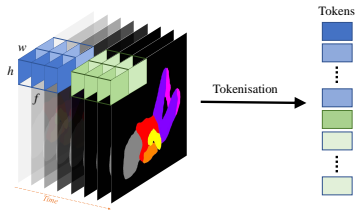


Figure 3. Tubelet embedding of segmentation image sequence.

5. Experimental results

In our experimentations, we aimed to demonstrate the interest of body semantic segmentation in the understanding of similar or static actions, and dealing with optical flow limits. The initial experiments on the JCARD dataset were based on the state-of-the-art methods. The first one was based on the optical flow pre-processing of the entire dataset sequences. A lot of sequences where the robot moves, or the subject was static generated empty or uniform images. For the second experiment we used the BlazePose skeleton extraction model described in [3]. It is a lightweight convolutional neural network architecture for human pose estimation that is tailored for real-time inference on mobile devices. The problem with this model is that it works when the whole body is visible, moving and in standard positions. Otherwise, the model extracts bad skeleton and induces errors in the classification process.

Before experimenting our approach, we tested a video vision transformer on the RGB data. Since the training process requires a large amount of videos, the dataset was augmented by sub-sampling, cropping and mirroring the origi-

nal videos. We also sub-sampled the videos into sequences to adapt the size of input data (Two cases were evaluated : 8 frames and 16 frames). For Tubelet embedding, patches have 8x8 pixel size with a window of 10 frames. We divided JCARD dataset into two sets (Train 66%, and Test 34%). The accuracy of the first obtained model on test data was 77.82%.

| Input data | Top-1 Acc Sub-sampling 8 | Top-1 Acc Sub-sampling 16 |
|------------------|-----------------------------|------------------------------|
| RGB | 70.97 | 77.82 |
| SEG | 58.06 | 72.18 |
| RGB \oplus SEG | 73.79 | 81.03 |

Table 2. Classification results on test data.

Finally, we trained the classification model on semantically segmented images using two configurations. The first one uses only segmented sequences as input, when the second fuses the two streams : RGB and segmentation results. Since the body segmentation removes the contextual information, the classification results of the first configuration were lower than the RGB ones. The model based on the fusion of RGB and segmentation sequences gives the best results since it keeps the contextual information. This model presented an accuracy of 81.03%. Even this accuracy is far from other datasets benchmarks, it demonstrates interest of semantic segmentation in the analysis of static and low/partial motion actions and motivates the development of more efficient models.

6. Conclusion

The main objective of this paper was to share a new dataset that present some constraints of mobile assistive robotics, such as moving cameras and low sensor positions. The dataset presents some actions related to AAL and some abnormal situations that may occur (fall and immobility events). We discussed the limitations of existing datasets and approaches in those particular contexts. The proposed dataset could be used alone or as a supplementary material for larger datasets. We also presented an approach that uses semantic segmentation of human body to avoid optical flow limitations. This method focuses on human body and gives more valuable results in the case of static, partial or low motion actions. In a future work, we will replace the body segmentation image by the latent features to optimise the performances. We will also augment the data with synthetic samples due to the small number of abnormal videos. Finally, we will continue hosting, maintaining, and distributing the dataset and enrich it with new samples of events and situations related to AAL.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. 1, 2, 3
- [2] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall data set. <http://www.iro.umontreal.ca/01> 2011. 2
- [3] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. BlazePose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020. 4
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 2
- [5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 1, 2
- [6] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2017. 2
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 3
- [8] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 3
- [10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2
- [11] Huiwen Guo, Xinyu Wu, Nannan Li, Ruiqing Fu, Guoyuan Liang, and Wei Feng. Anomaly detection and localization in crowded scenes using short-term trajectories. In *2013 IEEE International Conference on Robotics and Biomimetics (RO-BIO)*, pages 245–249, 2013. 2
- [12] M Muñoz J Eraso, E Muñoz and J Pinto. Dataset caucafall, v4, 2022. 2
- [13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 2
- [14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2
- [15] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019. 2
- [16] Michal Kepski. UR fall detection dataset. <http://fenix.univ.rzeszow.pl/mkepski/ds/uf.html> (last access: 11/11/2022). 2
- [17] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2
- [18] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Trans. on Circuits and Systems for Video Technology*, 31(3):1066–1078, 2020. 3
- [19] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 2
- [20] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2
- [21] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 2
- [22] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 1, 06 2014. 1, 2
- [23] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 12 2012. 2
- [24] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [25] Zehua Sun, Qihong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1
- [26] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017. 2