Human Action Recognition Based on Body Segmentation Models

Catherine Huyghe, Nacim Ihaddadene, Thomas Haessle, Chabane Djeraba

Human Action Recognition Based on Body Segmentation Models

Catherine Huyghe*†, Nacim Ihaddadene†, Thomas Haessle‡, Chabane Djeraba*
* *University of Lille, CNRS, Centrale Lille,*
9189 CRIStAL, USR 3380 IRCICA
F-59000 Lille, France
chabane.djeraba@univ-lille.fr
† *JUNIA - ISEN*, F-59000 Lille, France
{catherine.huyghe, nacim.ihaddadene}@junia.com
‡ *CareClever SAS*, Roubaix, France
thaessle@cutii.io

***Abstract*—Human action recognition in videos is an important issue in computer vision. We propose an approach based on the integration of partial or global human body segmentation in the classification process to deal with partial movements and immobility. Experimentation on UCF101 public dataset output competitive recognition accuracy related state of the art.**

***Index Terms*—Actions recognition, smart surveillance, ambient assisted living.**

I. Introduction

Our goal is to make the recognition of actions and postures related to interactions between a user and a robotic system that is of assistance to the autonomy at home, based on computer vision. The majority of behaviors recognized are of the personal environment type. The typical situation is an elderly person living alone. The main objective is to increase safety through intelligent monitoring by detecting dangerous situations (domestic accidents, falls, discomfort, dangerous postures, immobility, etc.). Dangerous situations may happen in everyday activities like during physical exercises, rehabilitation exercises or bad use of medication. Several different types of actions are recognized, including actions containing movements such as falls and others containing little or no movement such as discomfort or people fainting. So here it is a question of posture and movement recognition.

The most popular technology for action and posture recognition is deep learning, it requires large amount of data. Furthermore, several targeted actions and postures such as falls, discomforts and dangerous postures, are rare and few data available. So we need to find out how to construct data for these rare events.

The final solution will be implemented on an assistance robot. So we have the constraints related to the embedded system such as limited resources and energy constraints. We have to build a solution that is fast and resource efficient to be functional on an embedded system.

* We would like to thank the European Regional Development Fund, ERDF and the company CareClever for their support of this project.

The paper addresses the question: How to recognize actions and postures from video streams, considering at the same time, movement, immobility and rare events, for embedded system?

In this paper, firstly we will study existing approaches of deep learning for action recognition from video streams in section II. Then we will present a new approach based on the segmentation of the human part, normally used for the detection of human postures, in section III. Next we will present synthetic data in section IV and our experimentation in section V. To conclude, we will summarize our results and discuss future prospects in sections VI and VII.

II. Action recognition from image sequences

In our study, we are interested at techniques based on the use of deep neural networks. Different architectures for action recognition are proposed in the literature [1]. They are distinguished by :

The input streams: Some methods use only the input image sequence, while others pre-process their image sequence. The most commonly used pre-processing is 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 [2]. Learning is then based on the movement of people without taking into account the human. 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 [3] and those based on the detection of the human skeleton [4]. Methods based on models of the human body allow a better consideration of the human, its position and posture.

The spatio-temporal dimension: This dimension has been treated in different approaches. Firstly, the analysis of images by 2D convolutions to learn it the spatial dimension, is combined with the use of Long Short Time Memory (LSTM) in the upper layers of the architecture to learn the temporal

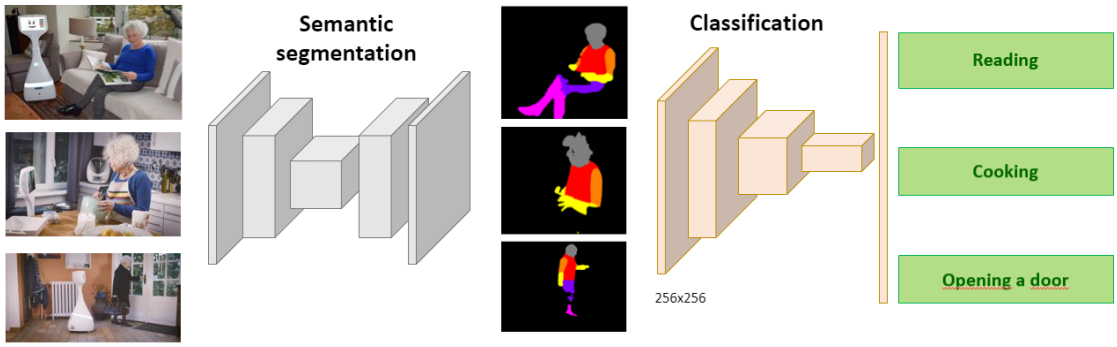

Fig. 1. Our approach to action recognition based on human part segmentation.

dimension. Another approach consists in using 3D convolutions on videos. Thus, the core of the 3D convolutions directly captures the spatio-temporal dimension. The disadvantage is that 3D convolutions require more parameters than 2D convolutions, making them difficult to train and deploy. For all these methods, the focus is on the analysis of the person (whole image analysis) and the absence of immobility treatment.

III. Our Approach

Our objective is to develop a model that can be deployed on an ambient assistance robot capable of recognizing different types of actions (daily activities, physical rehabilitation exercises, falls and immobility).

We propose in our approach to combine human part segmentation with ConvLSTM cells in the classification architecture (Fig. 1).

Methods for human part segmentation (total or partial) have been developed in recent years [5], without reaching the semantic level of actions. The human part segmentation consists in classifying each pixel of an image into several categories: pixels that correspond to the background and pixels that correspond to a given part of the human body (legs, arms, head, torso, etc.). Each part of the human body then constitutes a distinct class. The human part segmentation gives a better view of the location and position of each human body part present in the scene, whether it is moving or immobile.

We propose in our approach to integrate this human part segmentation in a first phase. The obtained images are then introduced into an action classifier composed of one convLSTM cell. The convLSTM cells [6] are a variant of the LSTM cells. They contain convolutions inside the LSTM cell itself. Thus, the convolution is performed inside the LSTM cell. The convLSTM cells allow to enter spatial and temporal characteristics at the same time.

This approach allows to process spatio-temporal data with the convLSTM cell and to process static actions or partial movements with the human part segmentation.

IV. Synthetic data for action recognition

We have a large number of different actions and postures to recognize, but there is little or no data to recognize rare phenomena such as falls, discomfort or dangerous postures.

Much work explores the use of complementary synthetic data for the estimation of human posture and recognition of actions [7].

Our approach is to create a new set of synthetic data allowing the recognition of rare phenomena such as falls. We are focused on modeling it. To create this action, we first modeled a man and a woman in 3D and human part segmented them into 12 distinct parts visible in Fig. 2. We unbalanced our subject to be able to make it fall. These imbalances resulted in 4 types of falls: forward falls, backward falls, rightward falls and leftward falls. To increase the diversity of this new data, the different falls were recorded from 12 points of view placed all around the 3D model.

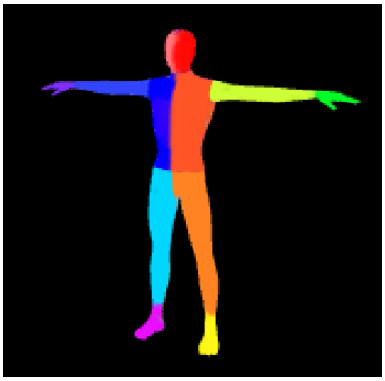

Fig. 2. Segmented 3D modeling of a man for synthetic data creation.

V. Experimentation

We first experimented our approach on the UCF101 dataset [8]. It is a dataset for action recognition composed of 13,320 videos from YouTube divided into 101 classes. It covers actions related to sports, human interactions, playing a musical instrument, personal care and daily actions such as cooking and related to the well-being of the person. These videos contain camera movements and different points of view, cluttered backgrounds and variations of objects appearance and lighting conditions.

The first step of our experiment was to pre-process each video frame of the UCF101 dataset by human part segmentation. To perform this segmentation, we relied on existing depth learning models for human part segmentation and in particular on "Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation" (CDCL) [5]. This pre-processing gives us output frames of segmented videos where 7 semantic regions of the human part (head, torso, arms, forearms, legs, lower legs, hands) appear Fig. 3. At the same time, the frames are resized in 256px by 256px.

Fig. 3. Example of segmentation of the human body. Each part of the body is represented by a color, and the background in black. Each part of the body constitutes a distinct class.

The training of our classification model composed of a single convLSTM is then done with these segmented frames. To limit the training time, the training of our classification model was carried out on 5 subsets composed of 5 classes from the UCF101 dataset. 80% of the videos of each subset are used for training and the remaining 20% for the evaluation of the model. As the videos do not all have the same duration, we have a different number of frames for each video and the action can be visible only on a limited number of frames or all of it. To do this, we take a sub-sequence of a fixed number of frames, initialized here at 25, and we randomly determine the starting position of this sequence of frames.

To understand the relevance of the semantic segmentation of the human body, we carried out two training sessions for each subset. One with the segmented frames and one with the original frames (non-segmented). The results are given in Tab. I.

We then wanted to test our virtual data by mixing it with 4 other classes from the UCF101 dataset and performing the learning with our approach. Since the human part segmentation of the 3D models is not identical to the human part segmentation obtained by our approach, we first segmented our virtual data with the same model as the data from the UCF101 dataset.

We then compared the results obtained with those obtained on similar classes from the UCF01 dataset. The results are shown in the Tab. II.

VI. Discussion

In our experiment (results in Tab. I) some action classes taken into account contain partial movements where only part of the body moves. This is the case of the actions like Archery, PlayingFlute, PlayingCello, BlowDryHair, PlayingViolin, ShavingBeard, PlayingGuitar, PlayingTabla, ApplyLipstick and PlayingDhol. In these posture type actions, the learning is more about the person's posture than about the movement itself because there is little or no movement. In Tab. I, we can observe that we obtain better results with segmented frames than with non-segmented frames (average of 46.04% for non-segmented frames against an average of 87.66% for segmented frames). The human part segmentation allows us to have a better vision of each human part location and position in the scene while disregarding the rest. Only the person is visible and those who are moving or immobile. Thus Its segmentation makes it possible to take into account immobility and partial movements.

We then compared our approach with the results obtained in the literature in Tab. III.

Our approach gives results close to the models in the literature and human part segmentation plays an important role in our approach because it is the only entry in our classification architecture. Our model therefore only learns about segmented frames. For some action classes, the quality of human part segmentation is variable, especially when the subject is very close to the camera, the areas are poorly defined, or very far from the camera, people are poorly distinguished.

It is difficult to obtain concrete results using existing trained human part segmentation models. They are trained in a certain way. For example, the human body is close to the camera, with good quality images and the whole human body visible. In our UCF101 experiment dataset [8]), we have different adaptation environments. However, the approach gives us interesting and encouraging results for a first experiment because, as this approach is based on the human part segmentation, it allows us to take into account immobility and partial movements for the future implementation of posture recognition.

The results obtained with our synthetic data show that it is a good solution for learning actions when little or no data exists. Nevertheless, the higher success rate with synthetic datasets (92.31% and 93.97% with synthetic data versus 90.43% and 91.47% with only real data (in Tab. II)) is due to the lack of diversity in our synthetic data. It would be interesting to test the models learned with our synthetic data with real fall data to verify that the modelled falls correspond to real falls.

VII. Conclusion

Our objective is to develop a deployable model on an assistance robot capable of recognizing different types of actions and postures such as daily activities, physical rehabilitation exercises, falls and dangerous postures or immobility. Some of these actions are rare phenomena where little real data is available, and others are actions where there is little or no movement.

To take rare phenomena into account we proposed synthetic data and showed that it integrates well with the real data.

To take actions based on posture into account where there is little movement, we proposed a segmentation-based approach and obtained a result of 87.66% on the UCF101 dataset. Our approach can be improved in the future by improving the segmentation of the input data. As segmentation allows a better view of people, their location and posture, this approach is

TABLE I
RESULT OF OUR APPROACH WITH 5 UCF101 CLASSES

UCF-101 classes	With segmentation	Without segmentation
Archery, FloorGymnatics, JumpRope, PlayingFlute, PushUps	88.72	72.93
BlowDryHair, BodyWeightSquats, Bowling, JumpingJack, PlayingCello	91.47	25.62
GlofSwing, PlayingViolin, ShavingBeard, Taichi, Yoyo	82.11	77.24
Knitting, PlayingGuitar, PlayingTabla, ThrowDiscus, WallPushups	85.60	25.60
ApplyLipstick, CleanAndJerk, JugglingBalls, PlayingDhol, WalkingWithDog	93.97	28.83

TABLE II
COMPARISON BETWEEN 5 UCF101 CLASSES AND 4 UCF101 CLASSES WITH 1 SYNTHETIC DATA CLASSES

UCF101 Classes and Synthetic data	With segmentation
ApplyLipstick, JugglingBalls, WalkingWithDog, PlayingDhol, CleanAndJerk	90.43
ApplyLipstick, CleanAndJerk, PlayingDhol, WalkingWithDog, **Falls**	92.31
BlowDryHair, BodyWeightSquats, Bowling, JumpingJack, PlayingCello	91.47
BlowDryHair, BodyWeightSquats, JumpingJack, PlayingCello, **Falls**	93.97

TABLE III
COMPARISON WITH THE STATE OF THE ART

Architecture	Input	UCF101
LRCN [9]	RGB + Optical Flow	82.92
C3D [10]	RGB	85.2
Two-Stream [2]	RGB + Optical Flow	88.0
3D-Fused [11]	RGB + Optical Flow	92.5
Two Stream I3D [1]	RGB + Optical Flow	93.4
Our approach	Human part segmentation	87.66

encouraging and will allow us to take immobility and partial movements into account for future implementation of posture recognition.

Further development should include the improvement of the human part segmentation and action classification to improve the results. Another future task is to improve the synthetic data to make it more robust compared to the real data and to enrich it with other actions.

REFERENCES

[1] 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)*, 2017, pp. 4724–4733.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 1, 06 2014.

[3] O. Köpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified cnn architecture for real-time spatiotemporal action localization," 2019.

[4] 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, pp. 597–600.

[5] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M.-T. Sun, "Cross-domain complementary learning with synthetic data for multi-person part segmentation," 07 2019.

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 802–810.

[7] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017.

[8] K. Soomro, A. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 12 2012.

[9] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

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

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," 2016.