Multi-camera skeleton-based activity recognition

Assisted Living and Health Monitoring

Silvia Vecchietti and Ilario Zamberlan

July 13, 2020

UNIVERSITY OF TRENTO

Contents

1 Introduction

Assisted living and health monitoring are technological services finalized to improve the life quality of the polulation who need a support in daily life, for example old people, people with disabilities or person in rehab after a surgery. These kind of systems are important and they are going to become increasingly demanded in particular because of the life expectancy is gradualy enlarging: it is indispensable to extend the life period in which old people are autonomous. Another important function of these technologies, as said before, is the rehabilitation. For example with these kind of system, for a just discharged hospital patient it is possible to resume motory activity while staying at home.

In these kind of applications information and communication technologies are used, in particular computer vision and artificial intelligence, in order to monitor the daily-life activities and make them safer.

The goal of our project is to realize a computer vision system which, using two or more cameras, is able to define the activities of a person in his home using the 3D person's skeleton. We selected four main activities: sit, stand, walk and eat or drink. Our targets are old people or people in their rehab, so these are the most relevant activities for our purposes. The idea is to obtain, as a output of the system, a list of activities and the respective duration that can be used by the doctor to analize the patient's situation, without the latter have to leave his home.

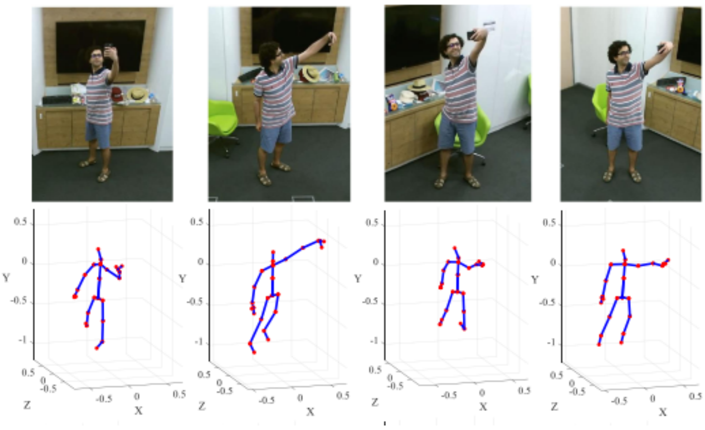
Regarding the realization, we used as a starting point a project [1] in which was implemented a neural network to take over the following actions: stand, walk, run, jump, sit, squat, kick, punch and wave. These starting software didn't give very good results so we tried to improve it by moving from a 2D to a 3D skeleton model and also adding some feature like the angle between joints and body velocity. We also adapt the actions at our application idea. Four actions are considered in this work: walk, stand, have meal and sit.

2 3D model and dataset

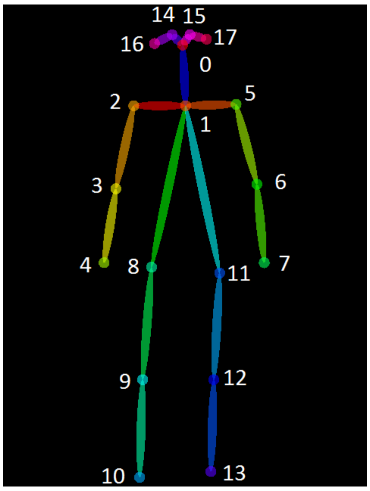
In this section we explain how the software works and we highlight the salient points.

2.1 3D model and feature

The main idea of our project is to switch from a 2D to a 3D model in order to increase the performance of the action recognition. This change allows the system to have more precised data and to be not effected to the body rotation.
Using a 2D skeleton, the same action can be seen changing due to different viewpoints as we can see in the figure[2]:



The depicted action is "doing a selfie" and it is the same in all the four photos. If we observe the skeletons below the four matching skeletons seems really different one from the other because of the different rotation. This is a hard problem for action recognition.
Using a 3D model this problem is easly resolved.
We used openpose skeleton which has 18 joints and for each joint are associated three coordinates: x, y and z:



To train the neural network we extract the following feature:

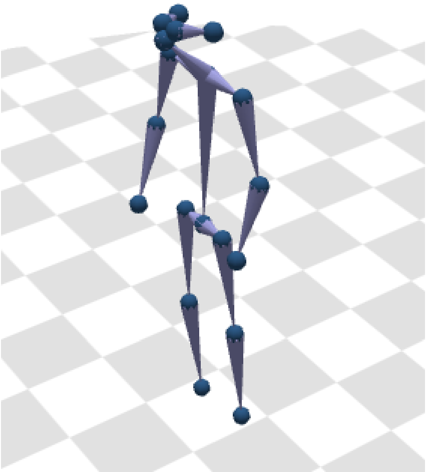
- Joint position
- Joint velocity
- Angle between joints

2.2 Dataset, training and testing

We used the following dataset for training:

- CMU Panoptic Dataset [3]
- CAD-120 [4]
- Two videos [5] [6]

In the CMU panoptic Dataset the output skeleton is different respect to OpenPose:



To the other Dataset and the videos the skeletons are obtaied using ”Lifting from the deep” [7]. Lifting from the deep generate another kind of skeleton.
Starting from this skeletons it is possible to obtain the OpenPose format applying a simple conversion.

Index	OpenPose	Panoptic	Lifting from the deep
0	Nose	Neck	Medium(Lhip, Rhip)
1	Neck	Nose	Lhip
2	Rsho	Medium(Lhip, Rhip)	Lkne
3	Relb	Lsho	Lank
4	Rwri	Lelb	Rhip
5	Lsho	Lwri	Rkne
6	Lelb	Lhip	Rank
7	Lwri	Lkne	Chest
8	Rhip	Lank	Neck
9	Rkne	Rsho	Nose
10	Rank	Relb	Head
11	Lhip	Rwri	Lsho
12	Lkne	Rhip	Lelb
13	Lank	Rkne	Lwri
14	Leye	Rank	Rsho
15	Reye	Reye	Relb
16	Lear	Leye	Rwri
17	Rear	Rear	
18	-	Lear	

In training phase the code use about seventy percent of the dataset for training and the other thirty for test.
For test part we made a video in which we did all the four actions(walk, stand, have meal and sit). We extracted 3D point from the video uscing ”lifting from the deep”.

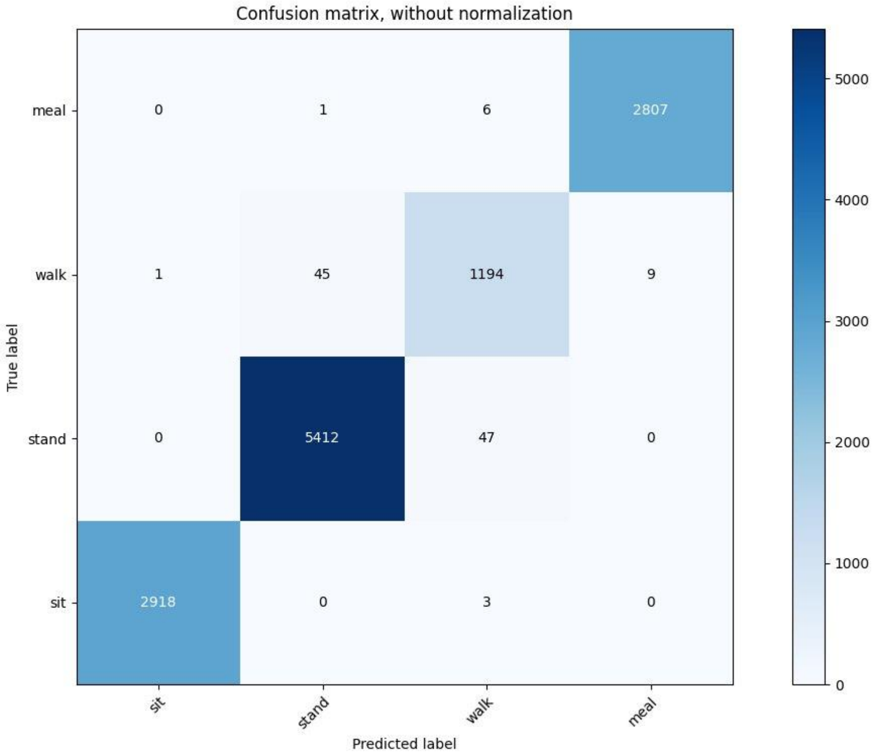
3 Results

First of all we tested our system using all the labeled data: walk, stand, have meal and sit. Number of sample per action:

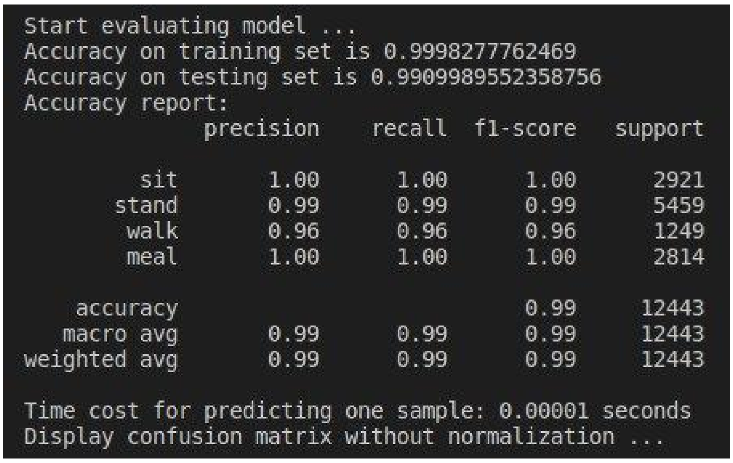
Action	Number of sample
sit	31265
stand	42633
walk	5937
meal	9587

Table 1: Number of each samples present in the dataset

3.1 Confusion matrix



3.2 Accuracy



```
Start evaluating model ...
Accuracy on training set is 0.9998277762469
Accuracy on testing set is 0.9909989552358756
Accuracy report:
              precision    recall  f1-score   support

         sit       1.00      1.00      1.00      2921
       stand       0.99      0.99      0.99      5459
        walk       0.96      0.96      0.96      1249
        meal       1.00      1.00      1.00      2814

    accuracy                           0.99     12443
   macro avg       0.99      0.99      0.99     12443
weighted avg       0.99      0.99      0.99     12443

Time cost for predicting one sample: 0.00001 seconds
Display confusion matrix without normalization ...
```

3.3 Test with *'meal'*

In the table below it's written the number of actions recognized in the video test.

Action	Predicted frame
sit	0
stand	16
walk	472
meal	2138

Table 2: Number of frame with respective action classified in the test video

In order to have a better output we stabilized it. We supposed that the action can change every three seconds, not before.
This is for remove impossble sequence like: "walk, walk, walk, walk, walk, walk, walk, walk, walk, sit, sit, walk, walk, walk, walk, walk, walk, walk, walk, walk". This sequence is not possible in a 30 fps video.
The sequence taken as an example become: "walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk, walk".

Action	Predicted frame
sit	0
stand	0
walk	450
meal	2182

Table 3: Number of predicted frame with average correction

The percentages of predicted actions are:

- Normal output (not stabilized)

Type	percentages
Right	46.43 %
Missclassified	53.56 %

Table 4: percentages of predicted actions

- Stabilized output

Type	percentages
Right	43.65 %
Missclassified	56.34 %

Table 5: percentages of predicted actions - stabilized

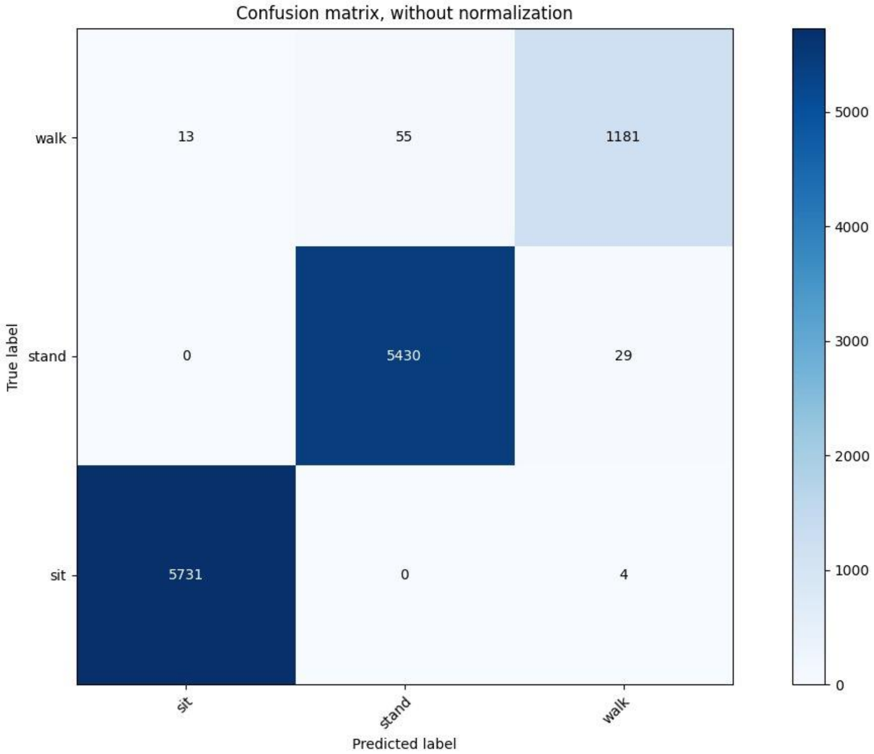
We noticed that in most of cases the actions stand, sit and walk are missclassified as *meal*. So we tried to training the neural network replacing the action meal with the acrion sit.

Number of sample per action:

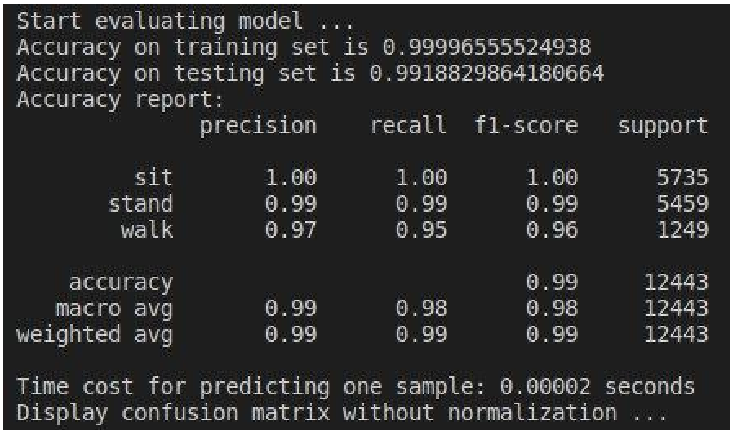
Action	Number of sample
sit	40852
stand	42633
walk	5937

Table 6: Number of each samples present in the dataset

3.4 Confusion matrix (without meal)



3.5 Accuracy (without meal)



```
Start evaluating model ...
Accuracy on training set is 0.99996555524938
Accuracy on testing set is 0.9918829864180664
Accuracy report:
              precision    recall  f1-score   support

         sit       1.00      1.00      1.00      5735
       stand       0.99      0.99      0.99      5459
        walk       0.97      0.95      0.96      1249

    accuracy                           0.99     12443
   macro avg       0.99      0.98      0.98     12443
weighted avg       0.99      0.99      0.99     12443

Time cost for predicting one sample: 0.00002 seconds
Display confusion matrix without normalization ...
```

3.6 Test without *'meal'*

3.6.1 Predicted frame

In the table below it's written the number of actions recognized in the video test.

Action	Predicted frame
sit	2146
stand	7
walk	474

Table 7: Number of frame with respective action classified in the test video

In order to have a better output we stabilized it also in this case.

Action	Predicted frame
sit	2092
stand	0
walk	540

Table 8: Number of predicted frame with average correction

The percentages of predicted actions are:

- Normal output (not stabilized)

Type	percentages
Right	64.16 %
Missclassified	35.84 %

Table 9: percentages of predicted actions

- Stabilized output

Type	percentages
Right	64.17 %
Missclassified	36.83 %

Table 10: percentages of predicted actions - stabilized

As can be seen from the results performance does not improve removing "meal" action.
The code generates an output video in which it is displayed the predicted label at each frame.

4 Conclusions

The system used do not give excelent results.

The main problem is the dataset used. This dataset in fact is not consistent. Except for panoptic which have 3D skeleton's information for the other part of the dataset we start from a 2D image and we obtain a 3D skeleton: this means that the data are not always precised.

We have found difficulties to obtain the data we needed. In our dataset the action "walk" is less rapresented respect to the other actions. Also this is a problem.

Finally we don't take advantage of the fact that inputs frames are correlated over time. We used a simple classifier but it is better used a rnn (for example lstm) which could increase the frame correlation capabilities in succession. How to improve: Improve and expand the dataset and use a rnn.

References

[1] felixchenfy, “Realtime-action-recognition.” https://github.com/felixchenfy/Realtime-Action-Recognition/.

[2] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive neural networks for high performance skeleton-based human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[3] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, “Panoptic studio: A massively multiview system for social interaction capture,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[4] H. S. Koppula and A. Saxena, “Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation,” *Robotics: Science and Systems (RSS)*, 2013.

[5] C. Minusio, “Camminata in salita sul tapis roulant.” `https://www.youtube.com/watch?v=Jn52_g7CKx0`, April 2019.

[6] B. e in forma in 30 minuti, “Allenamento 30 minuti tapis roulant.” `https://www.youtube.com/watch?v=rXjx5Jplfz8`, Dicember 2019.

[7] DenisTome, “Lifting-from-the-deep-release.” https://github.com/DenisTome/Lifting-from-the-Deep-release.