Invariant Representation of Motion for Gesture Recognition in Daily Life Scenarios

Matteo Saveriano and Dongheui Lee
Fakultät für Elektrotechnik und Informationstechnik,
Technische Universität München

ABSTRACT

Human gesture recognition is of importance for smooth and efficient human robot interaction. If the robot understands what the human is doing, then it can select and adapt its behaviours coherently with human's needs, guaranteeing a safe and useful cooperation.

We aim at making the robot capable of recognizing actions performed by different people in slightly different manners and from different view points, as shown in Fig. 1. To this end, we propose a new compact representation of a 3D motion, which is invariant to rotations, translations and linear scaling factors. This representation consists of two scalar quantities for each rigid body:

$$\gamma(t) = \frac{\|\dot{\mathbf{r}}(t) \times \ddot{\mathbf{r}}(t)\|}{\|\dot{\mathbf{r}}(t)\|^2}, \qquad \xi(t) = \frac{\|\boldsymbol{\omega}(t) \times \dot{\boldsymbol{\omega}}(t)\|}{\|\dot{\boldsymbol{\omega}}(t)\|^2} \tag{1}$$

where $\mathbf{r}(t)$ is the position and $\boldsymbol{\omega}(t)$ the angular velocity of a rigid body, and the symbols $\dot{\mathbf{a}}$ and $\ddot{\mathbf{a}}$ in (1) represent the first and second order time derivatives of $\mathbf{a}$.

The proposed representation, calculated for each part of the human body, becomes the input of a classification algorithm used to recognize human gestures. In particular, Hidden Markov Models based approach (*Invariant HMM*) and Dynamic Time Warping based approach (*Invariant DTW*) are modified by weighting the importance of each body part [1]. The performance of our approach are tested on the *MSR-Action3D* dataset used in [2]. This public dataset consists of 20 actions captured by a depth camera at 15 fps. Each action is performed 3 times by 10 different subjects. The results in Fig. 2 show that the proposed approach outperforms the invariant representation proposed in [2], the so-called *Fourier Temporal Pyramid (FTP) features*.

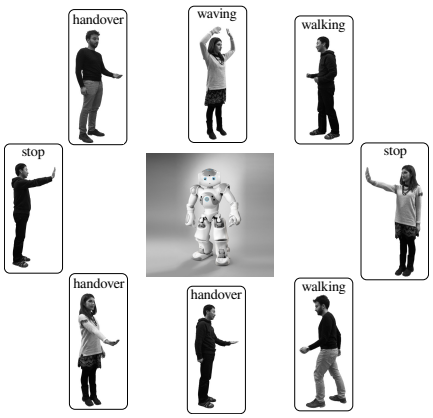

Fig. 1. Gesture recognition in a daily-life scenario.

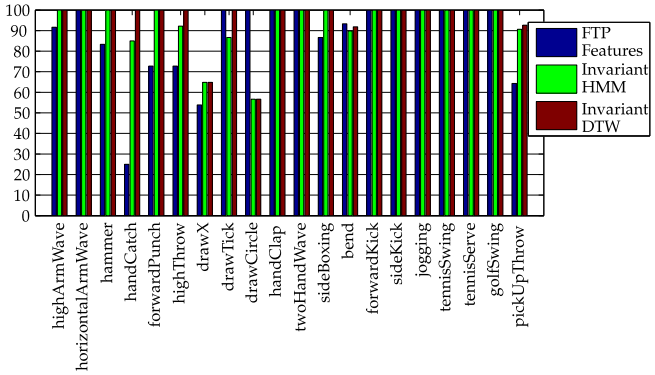

Fig. 2. The comparison between the recognition rates for MSR-Action3D dataset.

The average recognition rate with FTP features is 88.2%, with Invariant HMM is 93.2% and with Invariant DTW is 95.3%. Moreover, using our approach, the recognition rate for each action is always higher (except for the *drawCircle* gesture), even with very similar gestures such as *bend* and *pickUp&Throw*. This results are probably due to the fact that FTP features only use position information together with a technique to select the involved features. On the other hand, our approach combines velocity and acceleration information with a features selection technique. In many cases, this combination makes the gestures more distinctive and easier to recognise.

ACKNOWLEDGEMENTS

This work was partially supported by the DFG excellence initiative research cluster "Cognition for Technical System CoTeSys" and by the European Community within the FP7 ICT-287513 SAPHARI project.

REFERENCES

[1] M. Saveriano and D. Lee, "Invariant representation for user independent motion recognition," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 2013, (accepted).

[2] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.