# Kinect and Episodic Reasoning for Human Action Recognition *

Ruben Cantarero[1], Maria J. Santofimia[1], David Villa[1], Roberto Requena[1], Maria Campos[1], Francisco Florez-Revuelta[2], Jean-Christophe Nebel[2], Jesus Martinez-del-Rincon[3], and Juan C. Lopez[1]

[1] School of Computing Science, University of Castilla-La Mancha, Spain
[2] Digital Imaging Research Centre, Kingston University, London, UK
[3] The institute of Electronics, Communications and Information Technology (ECIT), Queens University of Belfast, UK
{ruben.cantarero,mariajose.santofimia,
david.villa,juancarlos.lopez}@uclm.es
{Roberto.Requena,Maria.Campos1}@alu.uclm.es
{F.Florez,J.Nebel}@kingston.ac.uk
j.martinez-del-rincon@qub.ac.uk

**Abstract.** This paper presents a method for rational behaviour recognition that combines vision-based pose estimation with knowledge modeling and reasoning. The proposed method consists of two stages. First, RGB-D images are used in the estimation of the body postures. Then, estimated actions are evaluated to verify that they make sense. This method requires rational behaviour to be exhibited. To comply with this requirement, this work proposes a rational RGB-D dataset with two types of sequences, some for training and some for testing. Preliminary results show the addition of knowledge modeling and reasoning leads to a significant increase of recognition accuracy when compared to a system based only on computer vision.

## 1 Introduction

Human action recognition has been a major concern for areas such as computer vision, robotics, machine learning, or ambient intelligence. Traditionally, vision methods for action recognition were mainly based on RGB images [16], sometimes enriched with body sensors [14]. However, with the availability of affordable devices such as Microsoft Kinect or ASUS Xtion Pro, depth information (D) has also came into play. This type of devices facilitates the extraction of 3D points of body joints, from which skeletal models can be constructed. Different methods have been employed to perform action recognition from RGB-D data [12]. These approaches are based on the use of depth maps [19] [20] [25], skeleton joints [23][24] or hybrid methods [22][15][13].

* This work has been partly funded by the Spanish Ministry of Economy and Competitiveness under project REBECCA (TEC2014-58036-C4-1-R) and by the Regional Government of Castilla-La Mancha under project SAND (PEII_2014_046_P).

Despite accuracy improvement in comparison to RGB-based methods [26], more elaborated mechanisms are required to support action recognition. Inspired in how humans tackle this task, different sources of information have to be combined: sensory information (visual, acoustic, smell, etc.), common-sense and context knowledge. Rationality, knowledge, and senses, therefore prove essential for any attempt to replicate human ability for action recognition. This work caters for these three elements in a novel approaches that combines body-pose estimation and common-sense reasoning.

Sensory perception is here limited to vision. RGB-D images recorded from a Microsoft Kinect device are fed to a state-of-art body-pose estimation algorithm. Rationality is achieved by proposing a scenario that has been set up to promote actions aimed to an end. This will allow us to overcome the lack of rationality that most of non-real scenario datasets lack from. Finally, common-sense and context knowledge is here managed by means of an efficient knowledge-base system with reasoning capabilities. This work combines these three elements to propose a five-stage framework for action recognition.

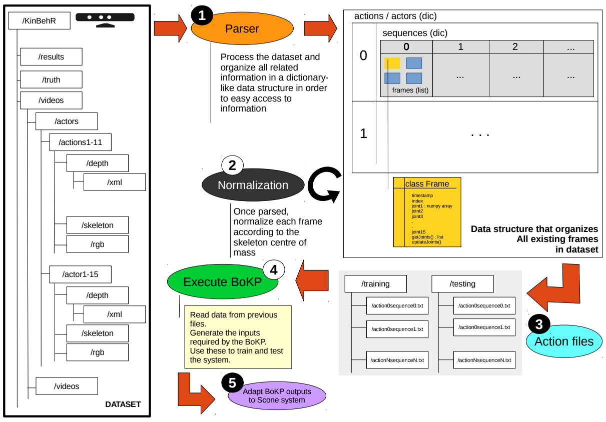
The rest of the paper is organized as follows. First, the methodological approach proposed in this work is detailed in Section 2. Then, Section 3 evaluates the proposed methodology. Finally, Section 4 summarizes the conclusions drawn from this work.

## 2 Methodology

The method for human action recognition proposed here follows a five-stage approach, as depicted in Figure 1. First stage records data from an RGB-D device such as Microsoft Kinect. Then, these data are organized and made available in a public dataset. The 3D-points of body joints will afterwards go through a parsing process first and a normalization process later in order to homogenize different heights, angles, or perspectives. Once normalized, these data, organised as action files, are fed to a body-pose estimation algorithm known as Bag of Key Poses (BoKP) [3] in charge of identifying the action being performed. The BoKP method returns a list of actions ordered by probability, which are different possibilities of the actual action happening in the file. Finally, it is the role of the reasoning system, implemented in Scone [11], to determine whether the actions provided by the vision system need to be corrected based on knowledge premises. The following subsections provide thorough details of the aforementioned stages.

### 2.1 The dataset

Rational behavior is implicit in human's daily life activities, since a person performs an action for a reason or aimed to an end [8]. Lab-recorded and synthetic datasets that involve actors performing isolated and unrelated actions [22] [23] are therefore not representative of real scenarios and not valid for the purpose of this work. Rationality is what entitles the reasoning system to understand an ongoing activity which also brings opportunities to correct the computer vision



Process the dataset and organize all related information in a dictionary-like data structure in order to easy access to information

Once parsed, normalize each frame according to the skeleton centre of mass

Read data from previous files.
Generate the inputs required by the BoKP.
Use these to train and test the system.

Data structure that organizes All existing frames in dataset

**Fig. 1.** System-stage overview

system estimation. Other more realistic datasets simply do not provide skeleton joint positions [26] [18].

It is the lack of a benchmarking dataset, with the appropriate configuration, what has motivated the recording of a new one. The KinBehR [2] dataset has been recorded using the second generation of Kinect for Windows and its SDK[4].

The recorded information can be divided into five categories: RGB video with a resolution 1920 x 1080 pixels per frame; 512 x 424 resolution depth video; 512 x 424 infrared video; background substracted videos that yield detected users; 560 x 420 video where the 25 joints of each identified and tracked user are represented. Additionally, one XML file is generated per frame. This XML file contains all the relevant information collected by Kinect and the Microsoft capturing algorithm: the position and orientation of the 25 joints of all tracked users, user state, user hand states, etc.

The resulting dataset has been split in two sets of sequences: training and testing. The generation of this dataset involved 18 actors, performing 13 types of actions: watch clock, crossing arms, scratching head, sit down, get up, turn around, walk, wave, punch, kick, point, take something, throw over head, throw down. These 13 actions have been selected to match those of the public dataset IXMAS [21]. In [2] all the information about this dataset is available.

For each actor, two types of sequences were recorded. First, the actor was told what to do to ensure that all the considered actions were performed, and thus recorded. Second, one more sequence was recorded for every actor, in which the actor was not told what to do. In this second type, actors were given the sole direction of behaving normally. Finally, all sequences were manually labelled, indicating the start and the end of actions occurrences, since automatic segmentation of videos is not considered in this study.

[4] https://dev.windows.com/en-us/kinect

In order to ensure rational behavior in the sequences where actors were freely behaving, the recording scenario was equiped with appropriate elements such as a punching ball, a small ball, a book, and a chair. With this scenario configuration, we were ensuring that performing an action such as kicking, was motivated by the fun obtained from interacting with a punching ball.

### 2.2 Bag of Key Poses

Among the different approaches for human action recognition, this work resorts to a machine learning technique, known as Bag of Key Poses (BoKP), described in [7] [4] and available in [6]. This method relies on a training phase during which salient features are learned. For this phase, training sequences were used where actors were being told what to do. Originally, the BoKP system worked with silhouette-based pose representation, where only the contour points of the silhouettes were used as features. An extension of that work led to consider 3D real-world coordinates for the points comprising the 25 joints of tracked users. Consequently, the number of points to be processed is dramatically reduced [5] while preserving the recognition accuracy by only considering the most representative points of the users in the scene, i.e. their joint points.

However, to ensure the user-characteristic independence, a normalization process has to be carried out. The normalization method employed in this work is described in [1]. During the training stage, skeleton sequences performing an action are provided to the BoKP algorithm. For the skeleton of each frame, the algorithm computes the most similar key pose. Then, at the end of the process, the frame sequence is represented by a sequence of key poses, labeled with the given action. During the testing stage, a similar process is carried out in order to compute the most representative key pose for the skeletons of the sequence. The learned action most similar to that sequence is the recognized one.

The algorithm proposed in [6] has been modified returning a ranking of five recognised actions according to their distance to the trained sequences, instead of returning the most similar one. The ranking might reveal if a test sequence is clearly matched with a particular action or it is similar to several actions. The reasoning system, as explained below, uses that information provided by the five most probable actions at every frame in order to make corrections.

### 2.3 Common-sense reasoning

Finally, the last system stage is intended to select one action, out of the five obtained from the previous stage, in order to complete the understanding of the ongoing activities.

The work in [9] demonstrates the assessment improvement obtained from combining common-sense reasoning with computer vision when trying to recognize sequence of actions that were performed for a reason [10]. In this sense, the representation of this type of sequences, as well as its support for reasoning, is well addressed by the Scone Knowledge-Base [11].

This work employs the knowledge and semantic model presented in [17]. Similarly to that work, rather than considering just one action, the five most probable ones are being considered at the same time. Based on this premise, the reasoning system needs to be capable of simultaneously tracking both the action considered as the most probable and the remaining four. In order to articulate this need, the employed architecture is thoroughly described in [17].

This architecture resorts to the notion of **Action** to represent each of the actions considered in the dataset. The notion of **Belief** encompasses a sequence of actions or episode. Each belief can only hold one action out of the five considered possible. **Expectations** group a set of actions. For example, the expectation used for *reading a book* considers the following sequence of actions: picking up, sitting down, and standing up. The appearance of any of these actions might suggest that an ongoing activity such as reading a book might be taking place. Finally, the **Estimation** concept is used to refer to the explanation standing from the recorded sequence.

These concepts along with the *knowledge about how the world works* and the considered context scenario are modeled and represented in Scone Knowledge-Base.

## 3 Validation and preliminary results

To validate our approach, the previously described KinhBer dataset is used. Figure 2 depicts the different modules involved and their interconnections.

At this stage, the CVS only implements the BoKP algorithm. The extraction of low-level information refers to information that can be potentially extracted from recognizing objects from video sequences in future extensions. The idea is to enhance the AIRS with low-level information regarding objects location in the scene, with regards to the actor. At this stage, only a reduced version of the DSK, WK, and BK is being considered for this prototype. Our recognition process could greatly benefit from such information, since for example, if a sitting down action is being considered but no chair is near the actor, this choice can therefore be rationally discarded.

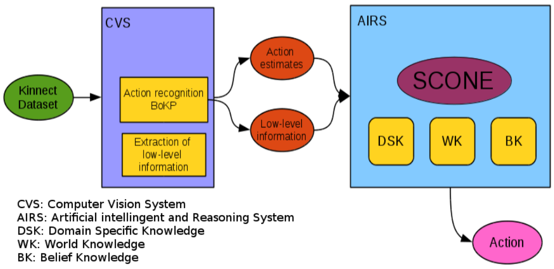

Kinnect Dataset
CVS
Action recognition BoKP
Extraction of low-level information
Action estimates
Low-level information
AIRS
SCONE
DSK
WK
BK
Action
CVS: Computer Vision System
AIRS: Artificial intellingent and Reasoning System
DSK: Domain Specific Knowledge
WK: World Knowledge
BK: Belief Knowledge

**Fig. 2.** System overview

Table 1 summarizes the recognition rate of the actions performed in each sequence obtained by the implemented prototype. A CVS-only system obtains an average accuracy of 31,9% when the first and most likely action of the BoKP algorithm is used as the estimated action to compute the accuracy. The AIRS obtains an accuracy of 41,07% after having considered the five most probable actions in the reasoning to select the most likely action as the final estimation. Here, accuracy is measured in terms of the numbers of actions correctly estimated.

**Table 1.** Preliminary results obtained by the prototype implementation

| | CVS-only | AIRS |
|---|---|---|
| Actor 1 | 33.33% | 33.33% |
| Actor 2 | 66.66% | 66.66% |
| Actor 3 | 42.85% | 52.38% |
| Actor 4 | 13.33% | 23.33% |
| Actor 5 | 16.66% | 66.66% |
| Actor 6 | 11.11% | 11.11% |
| Actor 7 | 36.84% | 42.10% |
| Actor 8 | 52.94% | 58.82% |
| Actor 9 | 18.75% | 18.75% |
| Actor 10 | 30.76% | 30.76% |
| Actor 11 | 25.00% | 50.00% |
| Actor 12 | 33.33% | 43.33% |
| Actor 13 | 42.85% | 50.00% |
| Actor 14 | 22.22% | 27.77% |
| Average | 31.90% | 41.07% |

This first prototype only considers a reduced version of DSK. When the considered knowledge is relevant for any of the sequences, like in Actor 5 or 11, an accuracy improvement is observed. However, actors were not given behavioral instructions and DSK was neither tailored for the recorded sequences. Results therefore show that, when the encoded knowledge responds to recorded behaviour, the recognition rate is improved. Additional efforts need to be dedicated to model DSK, WK, and BK.

## 4 Conclusions

This paper presents a novel methodology for human action recognition that combines RGB-D data, captured from a Microsoft Kinect device, and a commonsense reasoning system. In order to validate the proposed methodology a dataset recording rational behaviour has been constructed and released. A prototype of the system shows encouraging results since, with a reduced version of the required knowledge, the system has improved the average accuracy obtained by the computer vision system.

As future work, the system will be extended to consider information about the objects as well as additional domain specific knowledge. This additional information will support, for example, corrections based on if no object is nearby the actor, a *picking up* does not make sense.

## References

1. Alexandros Andre Chaaraoui, Jose Ramon Padilla-Lopez, and Francisco Florez-Revuelta. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2013.
2. Ruben Cantarero, Maria J. Santofimia, Jean-Christophe Nebel, Francisco Florez Revuelta, Jesus Martinez del Rincon, and Juan C. Lopez. KinBehR: Kinect for Behaviour Recognition. `arco.esi.uclm.es/public/prj/KinbehR/KinbehrDataset_v2.zip`.
3. Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. An efficient approach for multi-view human action recognition based on bag-of-key-poses. In *Proceedings of the Third International Conference on Human Behavior Understanding*, HBU'12, pages 29–40, Berlin, Heidelberg, 2012. Springer-Verlag.
4. Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799 – 1807, 2013.
5. Alexandros Andre Chaaraoui and Francisco Florez-Revuelta. Adaptive human action recognition with an evolving bag of key poses. *IEEE Trans. on Auton. Ment. Dev.*, 6(2):139–152, June 2014.
6. Alexandros Andre Chaaraoui, Francisco Flórez-Revuelta, and Pau Climent-Pérez. Bag of key poses. `https://github.com/DAIGroup/BagOfKeyPoses`, 2015.
7. Pau Climent-Pérez, Alexandros André Chaaraoui, José Ramón Padilla-López, and Francisco Flórez-Revuelta. Optimal joint selection for skeletal data from RGB-D devices using a genetic algorithm. In *Advances in Computational Intelligence - 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27 - November 4, 2012. Revised Selected Papers, Part II*, pages 163–174, 2012.
8. Donald Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.
9. Jesús Martínez del Rincón, Maria J. Santofimia, and Jean-Christophe Nebel. Common-sense reasoning for human action recognition. *Pattern Recognition Letters*, 34(15):1849–1860, 2013.
10. Chris Edwards. Decoding the language of human movement. *Commun. ACM*, 57(12):12–14, November 2014.
11. Scott E. Fahlman. The Scone knowledge-base project, 2016. Available online at: `http://www.cs.cmu.edu/~sef/scone/`. Retrieved on January 20th, 2016.
12. Adnan Farooq and Chee Sun Won. A survey of human action recognition approaches that use an rgb-d sensor. *IEIE Transactions on Smart Processing and Computing*, 4(4):281–290, 2015.
13. Mehmet Gonen and Ethem Alpaydin. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011.
14. O.D. Lara and M.A. Labrador. A survey on human activity recognition using wearable sensors. *Communications Surveys Tutorials, IEEE*, 15(3):1192–1209, Third 2013.

15. Jinna Lei, Xiaofeng Ren, and Dieter Fox. Fine-grained kitchen activity recognition using rgb-d. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 208–211, New York, NY, USA, 2012. ACM.
16. Hemali S. Mojidra and Viral H. Borisagar. A literature survey on human activity recognition via hidden markov model. *IJCA Proceedings on International Conference on Recent Trends in Information Technology and Computer Science 2012*, ICRTITCS(6):1–5, February 2013.
17. Maria J Santofimia, Jesus Martinez-del Rincon, and Jean-Christophe Nebel. Episodic reasoning for vision-based human action recognition. *The Scientific World Journal*, 2014, 2014.
18. Atsushi Shimada, Kazuaki Kondo, Daisuke Deguchi, Géraldine Morin, and Helman Stern. Kitchen scene context based gesture recognition: A contest in ICPR2012. In *Advances in Depth Image Analysis and Applications - International Workshop, WDIA 2012, Tsukuba, Japan, November 11, 2012, Revised Selected and Invited Papers*, pages 168–185, 2012.
19. Antonio. Vieira, Erickson. Nascimento, Gabriel. Oliveira, Zicheng Liu, and Mario Campos. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In Luis Alvarez, Marta Mejail, Luis Gomez, and Julio Jacobo, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7441 of *Lecture Notes in Computer Science*, pages 252–259. Springer Berlin Heidelberg, 2012.
20. Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7573 of *Lecture Notes in Computer Science*, pages 872–885. Springer Berlin Heidelberg, 2012.
21. Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conference on Computer Vision*, 2010.
22. Ying Wu. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 1290–1297, Washington, DC, USA, 2012. IEEE Computer Society.
23. Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, pages 20–27. IEEE, 2012.
24. Xiaodong Yang and YingLi Tian. Effective 3d action recognition using eigenjoints. *J. Vis. Comun. Image Represent.*, 25(1):2–11, January 2014.
25. Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1057–1060, New York, NY, USA, 2012. ACM.
26. Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012.