Common-Sense Knowledge for a Computer Vision System for Human Action Recognition *

Maria J. Santofimia[1], Jesus Martinez del Rincon[2], and Jean-Christophe Nebel[3]

[1] Computer Architecture and Network Group, School of Computing Science, University of Castilla-La Mancha, Spain
[2] The institute of Electronics, Communications and Information Technology (ECIT), Queens University of Belfast, BT3 9DT, UK
[3] Digital Imaging Research Centre, Kingston University, London, KT1 2EE, UK
MariaJose.Santofimia@uclm.es
j.martinez-del-rinconj.martinez-del-rincon@qub.ac.uk
J.Nebel@kingston.ac.uk

Abstract. This work presents a novel approach for human action recognition based on the combination of computer vision techniques and common-sense knowledge and reasoning capabilities. The emphasis of this work is on how common sense has to be leveraged to a vision-based human action recognition so that nonsensical errors can be amended at the understanding stage. The proposed framework is to be deployed in a realistic environment in which humans behave rationally, that is, motivated by an aim or a reason.

Keywords: Common-Sense Reasoning, Action Recognition, Computer Vision

1 Introduction

Different approaches have been devised to tackle the problem of human action recognition from the computer vision perspective, just to name a few [1][2][3][4]. However, few works combine video-based strategies with anthropological aspects or knowledge about human and social behaviour[5]. The reason for that is also the same reason for autonomous and intelligent systems lack of success, as known: replicating human intelligence is a task that requires an extremely large amount of knowledge. However, it is neither expert nor specific knowledge that needs to be improved in these systems. On the contrary, the focus should be placed at collecting everyday knowledge, also known as common sense. For example, Cyc [6] has been gathering for over 25 years all common-sense knowledge held by humans.

In this sense, this work is intended to prove the hypothesis that computer vision systems for human action recognition could be enhanced with common-sense knowledge overcoming the occurrence of nonsensical recognized actions.

* This research was supported by the Spanish Ministry of Science and Innovation under the project DREAMS (TEC2011-28666-C04-03).

This work places the focus on the role that common-sense knowledge plays on the overall task of human action recognition. Although expert knowledge has been applied to many fields and applications, here we concentrate on common sense, understood as the everyday knowledge that describes how the world works.

Two main difficulties have to be overcome in order to verify the working hypothesis: on the one hand, to date, computer vision systems are not yet capable of recognizing whichever human action performed in video sequence recorded from real scenarios [7]; and, on the other hand, collecting the relevant common-sense knowledge held by humans is far from being a feasible task. These two constraints require reducing the number of considered human actions and assuming a reduced amount of available knowledge for reasoning.

It can be tempting to think that hand-crafted representation of expert knowledge can, at some point, replace the role of common-sense knowledge. In fact, the following quotation, extracted from [6], discusses this issue:

> " *It is often difficult to make a convincing case for having a consensus reality knowledge base, because whenever one cites a particular piece of common sense that would be needed in a situation, it's easy to dismiss it and say "well, we would have put that into our expert system as just one more (premise on a) rule." For instance, in diagnosing a sick twenty-year-old coal miner, the program is told that he has been working in coal mines for 22 years (the typist accidentally hit two 2s instead of just one). Common sense tells us to question the idea of someone working in a coal mine since age -2. Yes, if this sort of error had been foreseen, the expert system could of course question it also. The argument is, however, that we could keep coming up with instance after instance where some additional piece of common sense knowledge would be needed in order to avoid falling into an inhumanly silly mistake.*"

Obviously, a more careful representation of information could take into consideration that the age of a person cannot be a bigger number than the number of years the same person has been working in coal mines. However, the work presented here is more concerned with describing the knowledge that would allow the system to achieve that conclusion on its own. The counterpart is that the amount of information required to do so is huge. For that reason, the approach followed by this work consists in minimizing the common-sense knowledge involved in the recorded scenario by constraining the context in which actors can perform. However, these constrains should not be equated to the approach followed by expert systems.

According to Davidson [8], the reason that motivates an action also rationalizes it. Consequently, if motivations could be heuristically guided and restricted, types of performed actions would also be limited. This approach pursues a twofold aim: first, avoiding rule-based and expert system strategies; and second, minimizing the directions given to actors. The Davidsonian theory of actions and events provides a theoretical justification for the proposed approach. The two pursued aims can be achieved by creating the appropriate atmosphere

that makes actors prone to perform certain actions but allowing them, at the same time, to behave in a rational manner.

The limited number of actions that can be recognized by computer vision systems justifies the need for a set up scenario. There, actors are surrounded by suitable elements that encourage them to perform a predefined and expected set of actions such as punch or kick a punching-ball or read a book. The negligible probability of an actor performing those activities without the presence of interactive objects for fighting or reading makes them neccesary for capturing the actions of interest.

This article is organized as follows. Section 2 describes the semantic model proposed here to enable the integration of common-sense knowledge and computer vision. Section 3 describes the implementation details and some additional aspects of the proposed framework. Finally, Section 4 summarizes the most relevant conclusions drawn from this work.

2 Knowledge Modeling

Human action recognition is a task that should not be tackled in isolation from the context in which that actions are taking place. However, despite the importance of the notion of context, this concept has not yet been universally formalized. On the contrary, the fact that this concept is a relevant issue for different fields of knowledge such as natural language understanding, linguistics, context-awareness, or knowledge representation among others, makes it difficult to provide a common and unique definition of what context is.

Fig. 1. Knowledge modeled using Scone

The concept of context is here understood as the set of facts or propositional knowledge that describes a specific state of the world, in the same way that J. Allen refers to the concept of world in [9]. This concept is represented by a set of descriptions of both the static and dynamic aspects of the world, therefore modeling what is known about the past, present, and future.

Additionally, these propositions need to be semantically enhanced by associating a meaning to each of them. However the meaning of these propositions is unavoidably associated to the context in which they are being considered. In this sense, meaning is expected to be something more elaborate than just mere conventions about what other concepts state their significance to be.

The philosophical theory of *possible worlds* has tackled the problem of associating different meanings or truth value to the same proposition without worrying about inconsistencies or incongruousness. This theory has been successfully translated into a computational model by means of what S. Fahlman has come to call *multiple-contexts* [10].

The need for the multiple-context mechanism, for the purpose that concerns us here, is justified as a way of maintaining parallel action sequences. The multiple-context mechanism is provided as an essential feature of the Scone Knowledge-Base system[11]. As for the possible world theory, the multiple context mechanism allows the representation of different states of affairs, which simultaneously co-occur in the same knowledge-base, without leading to inconsistencies.

The proposed video based system relies on a primary classification of the actions being recognized, as depicted in Fig. 1, according to visual clues.This initial classification is then provided as an input to the knowledge system that reasons about the rationality of the recognized actions. Rather than yielding just one action, the computer vision system provides an ordered list of actions (the optimum number of actions will be discussed later on this paper). The similarity of actions such as wave or check the watch might lead the computer vision system to an erroneous identification. For that reason, actions ranked after the first one are not discarded, but on the contrary, are considered true in parallel contexts. In this sense, the multiple-context mechanism provides a way of holding the propositional knowledge that, *a priory* should have been discarded.

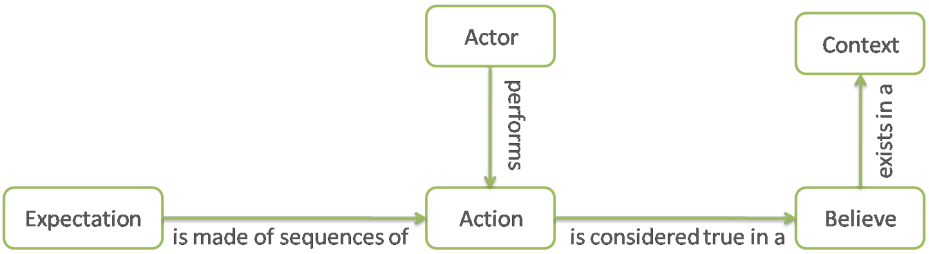

Fig. 2. A semantic model for video-based human action recognition

Since different contexts contain propositional knowledge about different plausible situations, an additional mechanism is therefore required to determine which context is considered true at a specific moment in time. Whenever a new action is recognized by the system, it outputs a ranked list of plausible actions. Each possibility is therefore asserted to a given context. Whenever the system is asked for the sequence of actions that describe the actor behaviour, only one context and its inherent propositions, should be active. Context activation is determined by the occurrence of action events that provides the most rational explanation. Explanations are considered here as the propositional knowledge

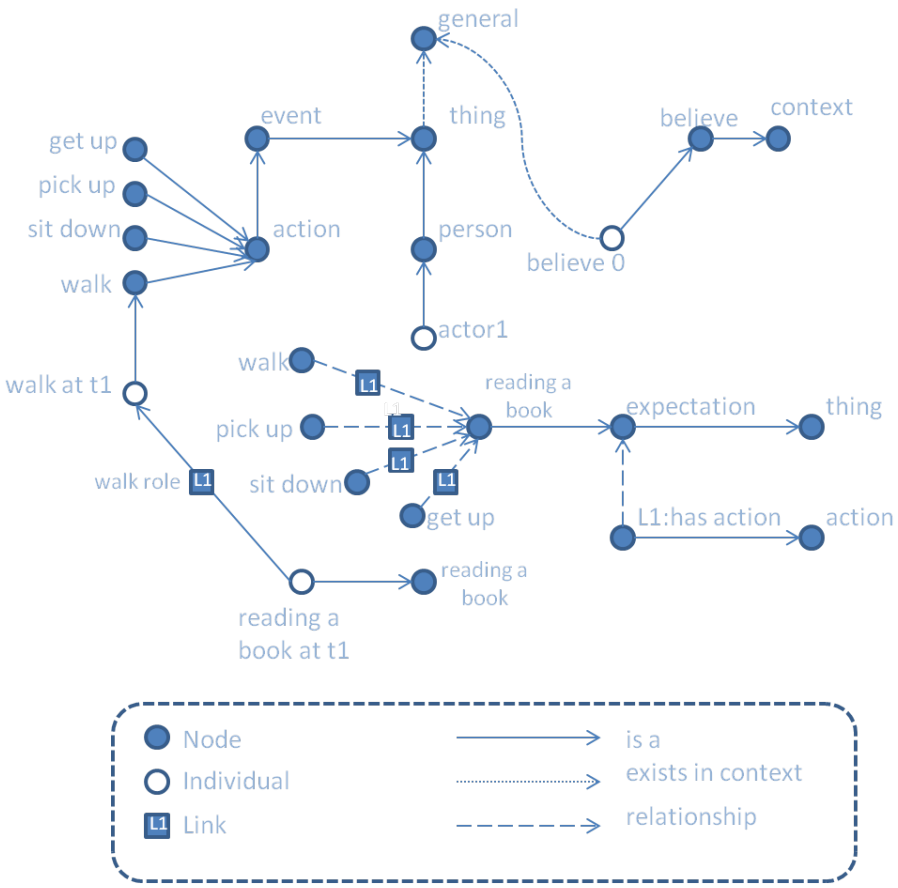

Fig. 3. Knowledge modeled using Scone

contained in a given context. For example, it is more rational or it makes more sense to pick up a book for reading it than for throwing it overhead, although both of them might be plausible.

After having justified how common-sense knowledge should be handled in order to be combined with a computer vision system for human action recognition, a more formal description of the semantic modelis provided. Fig. 2 depicts the semantic model for visual-based human action recognition whereas Fig. 3 presents the implementation of such semantic model using Scone and its multiple-context mechanism.

3 Implementation Details and Validation

The main contribution of this work consists in leveraging common-sense capabilities to a computer vision system. Fig. 1 depicts the different modules involved in the proposed framework.

Although many existing datasets are commonly used to evaluate video based action recognition systems, their artificial nature made them unsuitable for testing the proposed system. First, actions had to be performed for a reason whereas actors are told what to do in the artificial scenarios. Moreover, actions have to be part of a comprehensive story, so that performed actions make sense with regard to the aims or the reasons that motivate actors to behave like they do. Finally, the most compelling reason to produce a new dataset[4] was the exiting dissociation among the IXMAS [12] actions used to trained the system and existing datasets.

Video sequences are provided to a Bag Of Words framework that outputs an ordered list of several actions, printed in a text file, in which each line contains the ranked list of actions. The empirically estimated optimum number of actions is five. Then, the propositional knowledge involved in that action events is modeled using the semantic model proposed in previous section and asserted to the knowledge base according to the algorithm outlined in Fig. 4.

There are some common-sense aspects that enable the system to make corrections without having to consider the output provided by the classifier. For example, given the fact that the set up scenario reproduces a waiting room, actors are expected to enter and leave the waiting room. In this sense, first and last actions are expected to be the *walking towards* action.

The proposed system performance has been assessed in [13] in which a comparison analysis shows how the proposed approach considerably improves results obtained by computer vision systems under realistic scenarios, in which humans are rationally motivated.

4 Conclusions

This paper presents a novel knowledge management approach that enables common-sense capabilities to be incorporated to a computer vision system for human action recognition. In order to do so a semantic model for human action recognition is proposed.

This work was motivated by one of the most intrinsic features of humans: their actions are rationally motivated. The incorporation of common-sense knowledge and reasoning capabilities can reduce errors introduced by computer vision systems that, to human observers, simply do not make any sense. Simple errors such as the fact that to throw away an object you need to hold that object prevents the system from identifying an action as throwing away an object if previously an object have not been picked up. Recall that this approach consists

[4] WaRo11 dataset available at www.arco.esi.uclm.es/~mariaj.santofimia/puff/videos.zip

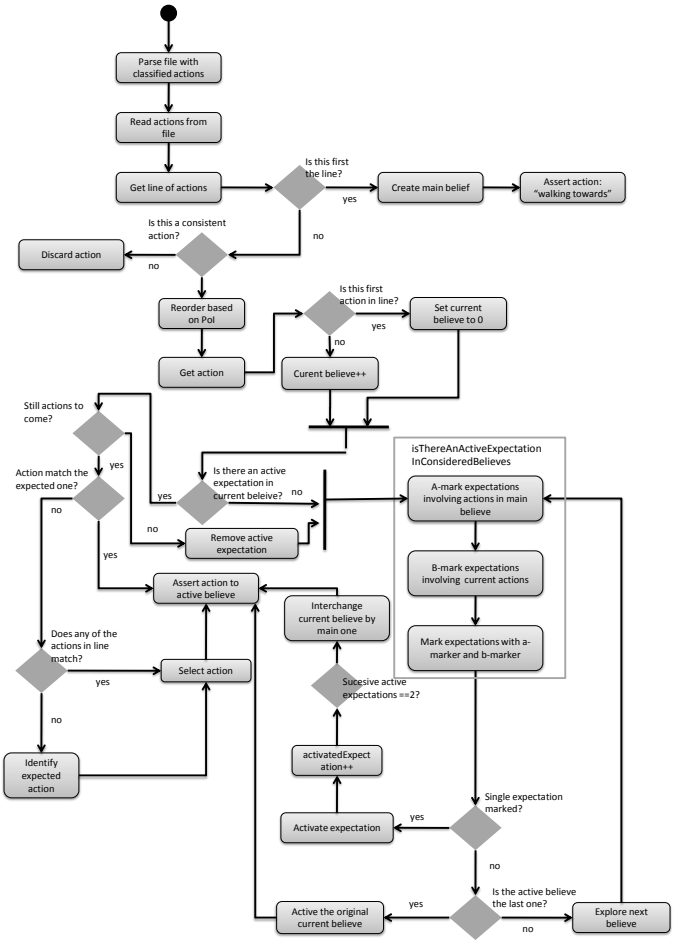

Fig. 4. Activity diagram for the proposed reasoning system

in providing the knowledge base system with the required information to achieve that conclusion on its own, rather than providing it with a basic rule of the type "*if you do not have an object you cannot throw it away*".

Finally, it can be concluded that human action recognition performed under realistic scenarios can greatly benefit from incorporating common-sense capabilities to correct computer vision system mistakes committed due to similarities in movements defining different actions, such as waving and scratching your head, and variability between humans performing the same action.

References

1. Vezzani, R., Baltieri, D., Cucchiara, R.: Hmm based action recognition with projection histogram features. In: Proceedings of the 20th International conference on Recognizing patterns in signals, speech, images, and videos. ICPR'10, Berlin, Heidelberg, Springer-Verlag (2010) 286–293
2. Martinez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S.A.: Recognizing human actions using silhouette-based hmm. In: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS '09, Washington, DC, USA, IEEE Computer Society (2009) 43–48
3. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, IEEE Computer Society (2008)
4. Zhang, J., Gong, S.: Action categorization with modified hidden conditional random field. Pattern Recogn. **43**(1) (January 2010) 197–203
5. Baker, C.L., Saxe, R., Tenenbaum, J.B.: Action understanding as inverse planning. Cognition (2009)
6. Lenat, D.B., Guha, R.V.: Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1989)
7. Nebel, J.C., Lewandowski, M., Thévenon, J., Martínez, F., Velastin, S.: Are current monocular computer vision systems for human action recognition suitable for visual surveillance applications? In: Proceedings of the 7th international conference on Advances in visual computing - Volume Part II. ISVC'11, Berlin, Heidelberg, Springer-Verlag (2011) 290–299
8. Davidson, D.: Actions, reasons, and causes. The Journal of Philosophy **60**(23) (1963) 685–700
9. Allen, J.F.: Towards a general theory of action and time. Artif. Intell. **23** (July 1984) 123–154
10. Fahlman, S.E.: Marker-Passing Inference in the Scone Knowledge-Base System. In: First International Conference on Knowledge Science, Engineering and Management (KSEM'06), Springer-Verlag (Lecture Notes in AI) (2006)
11. Fahlman, S.E.: The Scone knowledge-base project (2010) Available online at: `http://www.cs.cmu.edu/~sef/scone/`. Retrieved on February 28th, 2010.
12. Weinland, D., Boyer, E., Ronfard, R.: Action Recognition from Arbitrary Views using 3D Exemplars. In: International Conference on Computer Vision, Rio de Janeiro, Brazil, IEEE (2007) 1–7
13. del Rincon, J.M., Santofimia, M.J., Nebel, J.C.: Common-sense reasoning for human action recognition. Under Review Process. (2012)