Towards Human Activity Reasoning with Computational Logic and Deep Learning

Ioannis Prapas
NCSR "Demokritos", Greece
National Technical University of Athens, Greece
iprapas@protonmail.com

Georgios Paliouras
NCSR "Demokritos", Greece
paliourg@iit.demokritos.gr

Alexander Artikis
University of Piraeus, Greece
NCSR "Demokritos", Greece
a.artikis@iit.demokritos.gr

Nicolas Baskiotis
Université Pierre et Marie Curie, France
nicolas.baskiotis@lip6.fr

ABSTRACT

We approach the problem of human action recognition in videos by distinguishing between simple and complex actions. To recognize simple actions, we take advantage of the latest advances with 3D convolutional networks, which are able to offer a generic video snippet descriptor. For the complex ones, which involve interaction between more than one individual, we use the recognized simple human actions of the previous step to generate Event Calculus theories. This way, we aim to achieve a high-level human action understanding, combining the opaque effectiveness of deep learning and the transparent reasoning of computational logic. Our experimental results on a benchmark activity recognition dataset encourage further research towards this direction.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Neural networks**; **Inductive logic learning**; *Supervised learning by classification*; Online learning settings;

KEYWORDS

Activity Recognition, Activity Reasoning, C3D features, Event Calculus, Inductive Logic Programming

ACM Reference Format:
Ioannis Prapas, Georgios Paliouras, Alexander Artikis, and Nicolas Baskiotis. 2022. Towards Human Activity Reasoning with Computational Logic and Deep Learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/

1 INTRODUCTION

Paraphrasing [4], human activities are the most basic human-centered interactions with meaning. Human activity recognition is considered to be the epitome of all computer vision tasks not only due to its inherent complexity as a task, but also because of its numerous prospective applications and the implications that they will have in our daily life. A lot of modern work has focused in finding good generic representations of activities or developing end-to-end black box architectures that achieve a high accuracy in recognizing human activities. Recent advances with deep convolutional neural networks as well as the creation of big annotated datasets has led to improved results, in some cases achieving near perfect performance.

However, a key disadvantage of these types of approaches, compared to human cognition, is model interpretability. While a human can analyse the properties of a complex activity and decompose it to the individual actions that constitute it, these models can only reply with relatively high accuracy if a given activity is happening or not. In this work, we would like to add some model transparency to the activity recognition setup.

To this end, we see the problem of human action recognition as consisting of two parts: 1) simple action recognition, whereby the task is to detect 'simple activities' on individual video frames, such as walking, running, being active or inactive, and 2) a complex action (consisting of simple actions) recognition, where we aim to recognize 'complex activities' between two or more tracked entities, such as meeting and moving together. For the first part, we want to take advantage of the latest advances with 3D convolutional networks [5, 13], which are able to offer a generic video snippet descriptor. For the second part, we use the recognized simple human actions to generate Event Calculus [7] theories for more complex ones, which involve interaction between two or more individuals.

The contribution of this work lies mainly in bridging the gap between the opaque (black box nature) effectiveness of deep learning to make sense out of raw data and the transparent reasoning of Event Calculus, which also allows us to embed human knowledge in the problem-solving process. Because observations capture only certain aspects of the real world, but logic often represents generic knowledge, we would like to encourage research to move away from end-to-end black box architectures and towards hybrids with computational logic.

The rest of the paper is structured as follows: Section 2 presents current trends in activity recognition. Section 3 presents the necessary background to continue to Section 4, in which we describe our experimental setup. In Section 5 we present our experimental results. Finally, in Section 6, we summarize our work and give hints for future relevant research plans.

2 RELATED WORK

Activity recognition is a highly researched field, thus offering a rich bibliography with a great variety of proposed solutions. Central role to almost every proposed solution plays the pursuit for a good

Conference'17, July 2017, Washington, DC, USA
2022.

Table 1: The basic predicates and domain-independent axioms of the Event Calculus

Predicate	Predicate Meaning
happensAt(E, T)	Event E occurs at time T
initiatedAt(F, T)	At time T, a period of time for which fluent F holds, is initiated
terminatedAt(F, T)	At time T, a period of time for which fluent F holds, is terminated
holdsAt(F, T)	Fluent F holds at time T

Axioms

holdsAt(F, $T+1$) ← initiatedAt(F, T). (1)

holdsAt(F, $T+1$) ← holdsAt(F, T), not terminatedAt(F, T). (2)

and generic representation of actions. A good representation would be invariant to semantically meaningless changes of the input.

For this, holistic representations [3, 14, 15] have been proposed, which try to capture pose, articulated movement, but have been blamed to be too rough, unable to capture fine-grained movements that are a part of actions. Therefore, a big part of the research community has moved to generic local descriptors[16, 17]. In recent years, though, flavours of convolutional neural networks have managed to take over the field, offering the advantage of learning the good representations directly from the data, rather than relying in expert knowledge and intuition. Thus, the generality of the learned representation is only limited by the generality of the training data. Many times it is left to those models to do the full recognition of complex actions. However, it is often highly important for a model to be interpretable and this is a part where neural networks truly fall short.

To make up for this, we look into producing Event Calculus theories of complex activities. Existing works are using ground truth or stochastically noisy data of simple event streams as input [1, 2, 10, 11]. Our work differs in that it explores the whole pipeline: From raw data to simple actions with convolutional networks and from the produced simple actions to Event Calculus.

3 BACKGROUND

3.1 Convolutional neural networks

Deep convolutional neural networks are currently the state-of-the-art algorithm for most computer vision tasks[8, 12]. Their power lies in their capacity to learn powerful representations directly from data, in contrast to approaches with hand-crafted features which rely on the intuition of field experts on what constitutes a good representation. A deficit of traditional convolutional neural networks when it comes to videos, is their inability to capture time dependencies. 3D convolutional networks manage to incorporate short time dependency with a very simple extension from 2D (spatial), to 3D (spatiotemporal) convolutional filters.

3.1.1 Feature extraction. Trained in enough data, deep neural networks can learn in their first layers generic representations [13] that can be useful for datasets different from the ones, which were used for their original training. Thus, it is often a good choice to use pre-trained models as feature extractors. In this paper, we will use C3D[13], a state-of-the-art 3D convolutional network, trained on SPORTS1M datasets which contains more than 1 million sports videos collected from Youtube.

3.2 Event Calculus

The Event Calculus (EC) [7] is a temporal logic used for reasoning about events and their effects. Its ontology consists of *time points* (integer numbers); *fluents*, i.e. properties that have different values in time; and events, i.e. occurrences in time that may alter fluents' values. The axioms of EC incorporate the *common sense law of inertia*, according to which fluents persist over time, unless they are affected by an event. We use a simplified version of the EC that has been shown to suffice for event recognition [1]. The basic predicates and its domain-independent axioms are presented in Table 1. Axiom (1) states that a fluent F holds at time T if it has been initiated at the previous time point, while Axiom (2) states that F continues to hold unless it is terminated. Definitions for initiatedAt/2 and terminatedAt/2 predicates are given in an application-specific manner by a set of *domain-specific* axioms and guide the event recognition process. In this paper, we will avoid constructing such rules, using OLED [6], a state-of-the-art Inductive Logic Programming system, able to handle noisy data by relaxing the requirement to produce optimal theories covering all the input examples.

4 EXPERIMENTAL SETUP

We aim to achieve a high-level activity recognition by distinguishing between simple and complex action recognition. This separation is presented in Figure 1.

4.1 Simple activity classification

For the simple activity recognition we take as granted the detection of individuals in videos given as bounding boxes. We then extract the C3D features for the tracked objects and train a Support Vector Machine (SVM) to classify feature vectors as simple activity classes. This procedure is illustrated in the *Simple Action Recognition* part of Figure 1.

4.1.1 C3D features. The C3D network takes as input 16 RGB video frames of size 112x112. From every video we take crops of the bounding boxes that contain individuals and re-size them to 112x112. We feed the C3D network with batches of 16 consecutive frames of size 112x112 pixels (short clips) and use as a feature vector (1x4096) the output of the first fully connected layer (fc6) of the network. Each of these feature vectors finally represents 16 consecutive frames, in which an individual appears in the video. For training our model, we only keep the features in which a constant annotation is given. Given the generated feature vectors and the ground truth annotation, we train a One-Vs-Rest Linear SVM, as normally suggested with the C3D features. To evaluate our method, we perform the Leave-One-Video-Out Cross Validation method and count the total True Positive, False Positive, False Negative for each of the classes to finally compute the precision, recall and f1-score micro-metrics. The results are presented in Table 2.

4.2 Event Calculus theories for complex activities

We define a complex action as consisting of many simple actions performed at a previous time by one or more individuals. We aim to generate Event Calculus theories to recognize complex events, by combining the imperfect simple events that have been recognized

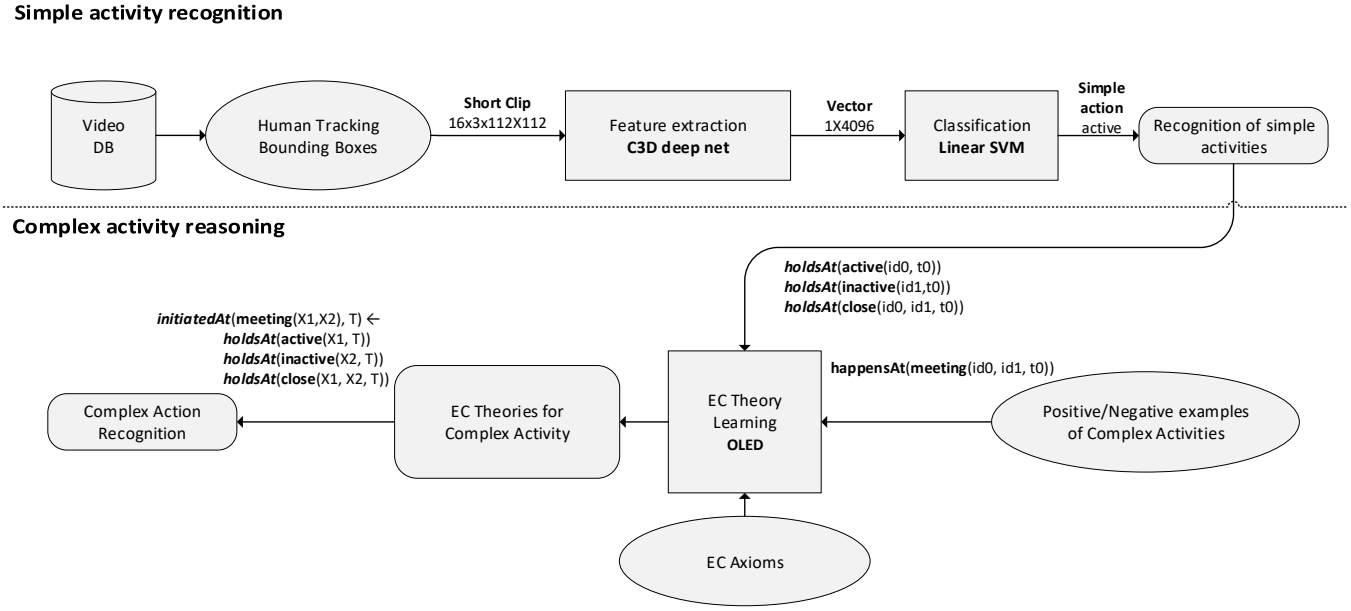

Figure 1: Proposal architecture

by the process presented in the previous subsection. To this end, we use a system for Online Learning of Event Definitions (OLED) [6]. OLED is a general Inductive Logic Programming system that is able to handle noise, by relaxing the requirement to cover all the training examples that it comes across. As depicted in the *Complex Event Reasoning* part of Figure 1, OLED will take as input the domain-independent Event Calculus axioms, the generated simple actions from the previous step, as well as positive and negative examples describing when a complex action is taking place. Its output will be domain-dependent Event Calculus theories, i.e. initiatedAt and terminatedAt rules for the complex actions meant to be recognized.

5 EXPERIMENTAL RESULTS

We illustrate our approach using the CAVIAR project dataset[1], which covers our need for annotation both for simple and complex actions. The CAVIAR dataset consists of videos where actors perform some two types of actions. The first type corresponds to simple actions and consists of knowledge about the activities of a person at a certain video frame/time point, such as *walking*, or being *inactive*. The second action type corresponds to complex events and consists of activities that involve more than one person, e.g. two people *meeting each other*, or *moving together*. The goal is to recognize complex events as combinations of simple events automatically recognized from the raw input videos.

5.1 Simple activity recognition results

Seen in Tables 2 and 3, the running instances are never being classified correctly, but this seems to be a result of the very low number of running examples that are available in the dataset. The over-represented class walking is generally classified correctly, but many examples of the other classes get confused with it. This is again a problem commonly presented when dealing with imbalanced classes, but in our case this is not the sole problem. Intuitively, the inactive/walking classes should be easily distinguishable, because in the inactive examples very little motion occurs, while a lot of motion occurs in the walking ones. Qualitatively examining the inactive examples that get confused with walking, we observe that it is a matter of low resolution, relative to the size of input that is expected by the C3D network. Enlarging small-sized crops, augments the noise present in the videos, which is then mistakenly translated to motion in the generated feature vectors. So this lousy performance is mainly a result of bad compatibility of the variable tracked object size in the CAVIAR dataset and the static input that the C3D network expects.

Other work [9] at the CAVIAR dataset achieves near perfect results. However, it is not directly comparable, as they assume that the geometry of the space where the videos have been shot is known.

Table 2: Micro-averaged precision, recall and f1-scores for every class of simple actions.

Class	*Precision*	*Recall*	*F_1-score*	*Sample size*
Inactive	0.80	0.55	0.65	479
Active	0.47	0.39	0.42	203
Walking	0.86	0.97	0.91	1547
Running	0.00	0.00	-	24

Table 3: Confusion matrix for the simple activity classification.

		Predictions			
		Inactive	**Active**	**Walking**	**Running**
Ground Truth	**Inactive**	**263**	72	144	0
	Active	45	**80**	78	0
	Walking	19	26	**1499**	2
	Running	0	0	24	0

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

Table 4: Confusion matrix for the complex activity classification.

Activity	Data	TP	FP	FN	Precision	Recall	f_1
Meeting	Ground Truth[6][2]	2750	226	844	0.92	0.77	**0.84**
	Ours	2787	1159	807	0.71	0.78	0.74
Moving	Ground Truth[6][2]	4700	3314	1583	0.59	0.75	**0.66**
	Ours	3967	5573	2315	0.41	0.63	0.5

This allows them to use simple features, such as the velocity and the shape of the tracking bounding boxes. For the proof-of-concept of our method, we concur that our achieved results suffice for moving forward to complex action recognition.

5.2 Complex activity recognition results

As expected and shown in Table 4, the manual annotation data (ground truth), always produce better performing theories than ours. For the meeting activity we see that the two results are highly comparable, confirming the findings of [1], that the recognition of this activity is resistant to noise. Moving activity recognition shows a substantially lower performance, a result of almost double FPs. This is explained by the type of noise in our data. Simple activities inactive and running are a part of the terminatedAt rules for the moving activity and as they are so much being confused in the simple activity recognition part, the activity is falsely persisting in time.

6 DISCUSSION

We presented a novel approach, consisting of two stages of learning for activity reasoning in videos. In the first stage, we extract deep learned features from a state-of-the-art 3D convolutional neural network, with which we train an SVM to classify simple actions. This process is far from perfect, as transformations that add noise are needed to prepare the input data to the pretrained neural network. However, using the recognized simple actions as input to an ILP system capable of learning Event Calculus theories, we demonstrate the robustness of OLED to learn from imperfect data. Computational logic learning approaches can in fact work on par with deep learning ones to output transparent models, but their better fusion in a unified learning process remains a topic that needs further research.

For future work, we would like to experiment with unsupervised approaches that instead of specific classified actions, would generate low level symbols that could potentially be used to learn Event Calculus theories for complex events. Like that, we will be able to try our approach in bigger datasets where annotation for simple actions is not present, as is the general case. Deep learning relies on big datasets and we find this to be the next step towards a smooth fusion with symbolic learning. A unified learning approach can then be the next goal.

ACKNOWLEDGMENTS

This work is partially funded by the H2020 project Track and Know (780754).

We extend our sincere gratitude to Dr. Nikolaos Katzouris for open sourcing the code of OLED[3] and especially for his aid and positive attitude regarding the use of it in our experiments.

[2]Results differ from original paper because rather than using all the different annotations for video 'Fight_OneManDown.mpg', we only use one.

[3]https://github.com/nkatzz/OLED

REFERENCES

[1] Elias Alevizos, Anastasios Skarlatidis, Alexander Artikis, and Georgios Paliouras. 2017. Probabilistic Complex Event Recognition: A Survey. *ACM Comput. Surv.* 50, 5 (2017), 71:1–71:31. https://doi.org/10.1145/3117809

[2] Alexander Artikis, Opher Etzion, Zohar Feldman, and Fabiana Fournier. 2012. Event processing under uncertainty. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*. ACM, 32–43.

[3] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. 2005. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Vol. 2. IEEE, 1395–1402.

[4] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. 2016. Going Deeper into Action Recognition: A Survey. *arXiv preprint arXiv:1605.04988* (2016).

[5] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2013), 221–231.

[6] Nikos Katzouris, Alexander Artikis, and Georgios Paliouras. 2016. Online learning of event definitions. *arXiv preprint arXiv:1608.00100* (2016).

[7] Robert Kowalski and Marek Sergot. 1986. A logic-based calculus of events. *New generation computing* 4, 1 (1986), 67–95.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[9] Pedro Canotilho Ribeiro, José Santos-Victor, and P Lisboa. 2005. Human activity recognition from video: modeling, feature selection and classification architecture. In *Proceedings of International Workshop on Human Activity Recognition and Modelling*. Citeseer, 61–78.

[10] Anastasios Skarlatidis, Alexander Artikis, Jason Filipou, and Georgios Paliouras. 2015. A probabilistic logic programming event calculus. *TPLP* 15, 2 (2015), 213–245. https://doi.org/10.1017/S1471068413000690

[11] Anastasios Skarlatidis, Georgios Paliouras, Alexander Artikis, and George A. Vouros. 2015. Probabilistic Event Calculus for Event Recognition. *ACM Trans. Comput. Log.* 16, 2 (2015), 11:1–11:37. https://doi.org/10.1145/2699916

[12] Suraj Srinivas, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas SS Kruthiventi, and R Venkatesh Babu. 2016. A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. *arXiv preprint arXiv:1601.06615* (2016).

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.

[14] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 588–595.

[15] Chunyu Wang, Yizhou Wang, and Alan L Yuille. 2013. An approach to pose-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 915–922.

[16] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision* 103, 1 (2013), 60–79.

[17] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*. 3551–3558.