Predicting Human Activities in Sequences of Actions in RGB-D Videos

David Jardim[1,2,3,4], Luís Nunes[2,3,4], Miguel Dias[2,3,4]
Microsoft Language Development Center, Lisbon, Portugal
Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
Instituto de Telecomunicações, Lisbon, Portugal
ISTAR-IUL, Lisbon, Portugal

ABSTRACT

In our daily activities we perform prediction or anticipation when interacting with other humans or with objects. Prediction of human activity made by computers has several potential applications: surveillance systems, human computer interfaces, sports video analysis, human-robot-collaboration, games and health-care. We propose a system capable of recognizing and predicting human actions using supervised classifiers trained with automatically labeled data evaluated in our human activity RGB-D dataset (recorded with a Kinect sensor) and using only the position of the main skeleton joints to extract features. Using conditional random fields (CRFs) to model the sequential nature of actions in a sequence has been used before, but where other approaches try to predict an outcome or anticipate ahead in time (seconds), we try to predict what will be the next action of a subject. Our results show an activity prediction accuracy of 89.9% using an automatically labeled dataset.

Keywords: human motion analysis, recognition, segmentation, clustering, labeling, Kinect, prediction, anticipation

1. INTRODUCTION

The ability to recognize what a human is currently doing is useful in several applications fields like, surveillance systems, human computer interfaces, sports video analysis, digital shopping assistants, video retrieval, gaming and health-care [13, 3, 12, 15, 6]. Human activity recognition (HAR) has become one of the most active research topics in image processing and pattern recognition [**Error! Reference source not found.**] and has grown dramatically in the past years and recently has evolved into anticipation or forecasting of future human actions. This paper addresses the problem of recognizing and predicting high-level human activities using supervised classifiers trained with automatically labeled data. Given the current recognized action in a sequence of actions, our approach proves that it is possible to predict the next most likely action or behavior that will occur.

1.1 Action Recognition

The initial approaches used computer vision (CV) techniques to extract meaningful features from 2D video data. Motion capture data (MOCAP) has also been used in this field [19] where Zhou et al. were able to achieve competitive detection performances (77\%) for human actions in a completely unsupervised fashion. Using MOCAP data has several advantages mainly the accuracy of the extracted features but the cost of the sensor and the required setup to obtain the data is often prohibitive. With cost in mind Microsoft released a sensor called Kinect, which captures RGB-D data and is also capable of providing joint level information. A previous study using Kinect [8] consider the problem of extracting a descriptive labeling of the sequence of sub-activities being performed by a human, and more importantly, of their interactions with the objects in the form of associated affordances. Their method obtained an accuracy of 79.4\% for affordance, 63.4\% for sub-activity and 75.0\% for high-level activity labeling. There are some approaches which combine motion information and object properties [16, 18]. In [16] the authors abstract the problem in two stages. First, by recognizing general motions such as moving, not moving or tool used. Second, by reasoning about more specific activities (Reach, Take, etc.) given the current context, i.e. using the identified motions and the objects of interest as input information. They've obtained an accuracy classification of 92\%. More directly related to our research, [12] developed a system called Kintense which is a real-time system for detecting aggressive actions from streaming 3D skeleton joint coordinates obtained from Kinect sensors. In two multi-person households it achieves up to 90\% accuracy in action detection.

1.2 Action Prediction

Anticipation or forecasting future human actions has been the focus of few recent works. In [17] the authors tried to construct an intelligent system which will perform early recognition from live video streams in real-time, introducing two new human activity prediction approaches which are able to cope with videos from unfinished activities. In [7] the authors address the task of inferring the future actions of people while modeling the effect of the physical environment on the choice of human actions with prior knowledge of goals. Li et al. [11] propose a framework for long-duration, complex activity, prediction by discovering the causal relationships between constituent actions and the predictable characteristics of activities. This approach uses the observed action units as context to predict the next possible action unit, or predict the intension and effect of the whole activity. The efficiency of their method was tested on the complex activity of playing a tennis game and predicting who will win the game (0.65 of certainty with 60% of observed game). Recently [9] developed a framework where their goal is to enable robots to predict the future activities as well as the details of how a human is going to perform them in short-term (e.g., 1-10 seconds). With an anticipatory temporal conditional random field (ATCRF), they start modeling the past with a standard CRF but augmented with the trajectories and with nodes/edges representing the object affordances, sub-activities, and trajectories in the future. Their algorithm obtains an activity anticipation accuracy of 84.1%, 74.4% and 62.2% for 1, 3 and 10 seconds.

2. PROPOSED PIPELINE

A modular framework was built with several task-oriented modules organized in a workflow (Figure 1) as follows:

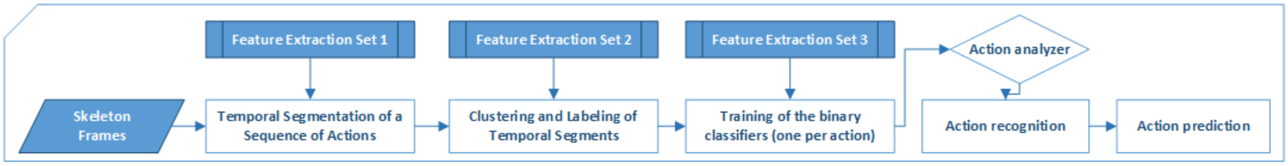
Feature Extraction Set 1
Feature Extraction Set 2
Feature Extraction Set 3
Action analyzer
Skeleton Frames
Temporal Segmentation of a Sequence of Actions
Clustering and Labeling of Temporal Segments
Training of the binary classifiers (one per action)
Action recognition
Action prediction

Figure 1 - Proposed pipeline of our modular framework responsible for segmentation; recognition and prediction of human activity

2.1 Feature Extraction

Kinect is capable of tracking 20 joints of a subject's skeleton. Skeleton frames are generated at the rate of 30 frames per second, and each frame consists of the 3D coordinates of 20 body joints along with their tracking states (tracked, inferred, or not tracked). We perform feature extraction from 4 main joints (*wrist-right, wrist-left, ankle-right and ankle-left*) as shown in Figure 2. Several features were calculated for the selected joints: relative velocity in X, Y, Z; total relative distance traveled in X, Y, Z and angles of the elbows and knee joints. The 3D coordinates are with respect to a frame of reference centered at Kinect. Frames from the camera are converted into feature vectors which are invariant to relative position of the body. We achieved this by re-calculating all the joints positions relative to the hip joint.

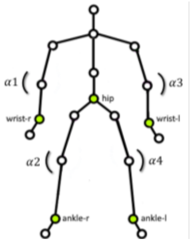

Figure 2 - Visual representation of body relative features used in our system where the green joints were selected for feature extraction along with the angles of the elbows and knees

2.2 Action Classification

In our previous research [4] we proved that given a sequence of contiguous actions it is possible to automatically divide the sequence into what we called temporal segments that correspond to individual actions. With a hierarchical clustering algorithm, we were able to automatically assign a label to an action. This allowed us to create an automatically labeled training set. Our next contribution [5] would be to compare the performance of our action recognition framework trained with data automatically labeled versus data manually labeled. The results proved that, for a dataset of simple combat actions, obtained with a standard Kinect camera with no special acquisition conditions, a temporal segmentation and clustering algorithm can be used to label identical actions performed by different users. Also, we have established that this labeling can be used to train supervised classifiers that will be capable of identifying specific actions in a RGB-D video feed, with a minor loss of precision relative to training with data manually labeled.

2.3 Action Prediction

There are several activities were a human subject perform certain actions as a sequence of actions. With that premise we would like to prove that given the observations of a scene containing a human performing an action *a* for time *t*, it is possible to predict the possible action *a+1* in a sequence of actions. Our framework is capable of recognizing the current action that it is being performed by a subject. Prediction is performed based on that information and the history of previous recognized actions. Instead of using manually labeled data to create our prediction training set, we used our binary classifiers trained with automatically labeled data to perform action recognition and reconstruct the sequences of actions that compose our dataset. We propose two approaches in a parallelism with computational linguistics where a sequence of actions can be seen as a sequence of text and each action seen as a word:

- We train several supervised classifiers: Multilayer Perceptron (MLP) as in [10]; Support Vector Machines (SVM) using pairwise classification [14], Random Forests (RF) [2] with *n*-grams of variable size.
- Conditional Random Fields (CRFs) [1] suited for labeling structured data they model rich contextual relations and are capable of learning and inferring a small and discrete label space such as our sequences of actions.

3. EXPERIMENTS

3.1 Data

We use PRECOG dataset[1] which has 72 RGB-D videos of 12 different subjects performing sequences of combat movements. The data is annotated with action labels within each sequence. The set of actions are: *right-punch; left-punch; elbow-strike; back-fist; right-front-kick*; *left-front-kick; right-side-kick; left-side-kick*. Using combinations of those 8 actions we created 6 distinct sequences (each sequence contains 5 actions). Of the 12 subjects recorded, each subject performed 6 different sequences.

3.2 Action Recognition Results

Given a temporal segment from a sequence of actions and several features extracted from the position of the skeleton joints we perform action recognition per segment. We report the results obtained by a 10-fold cross validation. Using MLP, SVM and RF classifiers trained with automatically labeled data we achieved an average performance of 85,5%, 90,0% and 91,0% respectively in recognizing the occurring action in a temporal segment extracted from a sequence of actions. These results can be examined in more detail in [5].

3.3 Action Prediction Results

In this section we explain our experimental results using our dataset to two different approaches. Several classifiers were trained to compare the results. These classifiers were trained to predict an action based on the previous history of recognized actions. All the experiments were performed using *k*-fold cross validation of 10 folds.

3.3.1 N-Gram Action Prediction

An *n*-gram is an *n*-character slice of a longer string. In our case the string represents a sequence of actions. The *n*-grams are composed by combinations of the actions of the sequence where the last action is the attribute to be used as the class. For example, the sequence "*right-punch, left-punch, side-right-kick, side-left-kick, front-left-kick*" would compose the following *n*-grams:

- tri-grams: "*right-punch, left-punch, side-right-kick*"
- quad-grams: "*right-punch, left-punch, side-right-kick, side-left-kick*"
- penta-grams: "*right-punch, left-punch, side-right-kick, side-left-kick, front-left-kick*"

n-gram	*Ground-Truth*	*SMO*	*RF*	*MLP*

[1] https://github.com/DavidJardim/precog_dataset_16

3	83,3%	77,7%	79,2%	77,7%
4	91,7%	86,8%	88,8%	89,6%
5	100%	95,8%	95,8%	95,8%

Table 1 - Prediction results of future action with different classifiers and different number of actions as input

To perform action prediction with this method, we require knowledge of at least the two previous actions. The second column of **Table 1** shows the probability of accurately predicting the next action using the Bayes' theorem with noiseless data. The following columns show the performance of our approach using several classifiers trained with data automatically labeled. The results are very similar between columns 3-5 and as expected the accuracy improves as we add more actions as input. Compared with column 2 we notice a loss of performance. Since our recognition algorithm at its best has a 91,0% of accuracy some of the actions are mislabeled. For example, we might have a re-constructed sequence as follows: "*right-punch, left-punch, NONE, side-left-kick, front-left-kick*" where the third action of the sequence labeled as NONE would affect negatively the training.

3.3.2 CRF Action Prediction

CRFs model rich contextual relations conditioned on several features as input. It is widely used in Natural Language Process (NLP) tasks like: word breaker, POS tagging, named entity recognized, etc. Our approach is to use CRFs for labeling the next action given the current action performed and the history of actions performed. To create the training data, we gather all the existing sequences of actions and for each sequence, we perform all the possible combinations of *current action – history of actions – next action*, computing a distribution over the possible future actions. Each record of the training corpus represents a sequence of actions (like a matrix) and each row describes an action to be predicted. The first N-1 columns are used as input data to generate the binary features and train the model. The Nth column is the action that the model should predict. As our application recognizes a new action we ask what is the most likely action that will occur next?

Method	*Manually labeled data*	*Automatically labeled data*
CRF	91,7%	89.9%

Table 2 - Prediction results of future action using CRFs

We used manually labeled data and automatically labeled data to train two classifiers. From the results in Table 2 and as expected the classifier trained with data manually labeled performs better, using the data automatically labeled resulted in a decrease of performance of 1,8%. This loss in performance is acceptable if we take into account that with this approach we are able of building a framework capable of recognizing and predicting actions in a fully unsupervised fashion.

4. CONCLUSION

In this paper, we described a framework capable of labeling, recognizing and predicting human actions using several supervised classifiers and CRFs to label structured data such as sequences. We have recorded a dataset of sequences of actions with Kinect since most datasets have only isolated actions. Unlike other approaches that take into account the context of the scene or object affordances in order to obtain more information, our approach relies solely on the features extracted from the movement of the joints of the subject's skeleton. Assuming that patterns of sequences of actions exist in our daily activities, this approach could have several applications. Our results proved that, for a dataset of simple sequences of combat actions, obtained with a standard Kinect camera with no special acquisition conditions, we are capable of recognizing the current action performed and predict the future action.

Also, we have established that data automatically labelled can be used to train our prediction classifiers with a minor loss of precision relative to training with human labeled data. We would like to replicate these results using other existing datasets and explore if action prediction can be used to improve action recognition with objects and scene information.

Acknowledgements: This research was sponsored by a joint scholarship between Fundação para a Ciência e a Tecnologia (FCT) and Microsoft Portugal under the grant SFRH/BDE/52125/2013. The project was hosted by Microsoft Language Development Center (MLDC) in Lisbon, Portugal.

REFERENCES

[1] David M Blei, Andrew Y Ng, Michael I Jordan, H M Wallach, Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. Conditional random fields: An introduction. *Neural Computation*, 18:1–9, 2004.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Stephen S Intille and Aaron F Bobick. A Framework for Recognizing Multi-agent Action from Visual Evidence. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference*, number 489, pages 518–525. AAAI Press, 1999.

[4] David Jardim, Luís Nunes, and Miguel Sales Dias. Automatic human activity segmentation and labeling in rgb-d videos. In *Intelligent Decision Technologies: KES-IDT 2016*, pages SIST 56, p. 383 ff. Springer International Publishing, 2016.

[5] David Jardim, Luís Nunes, and Miguel Sales Dias. Impact of automated action labeling in classification of human actions in rgb-d videos. In *22nd European Conference in Artificial Intelligence: ECAI 2016*, page in press. IOS Press, 2016.

[6] Christoph G. Keller, Thao Dang, Hans Fritz, Armin Joos, Clemens Rabe, and Dariu M. Gavrila. Active pedestrian safety by automatic braking and evasive steering. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1292–1304, 2011.

[7] Kris M. Kitani, Brian D. Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7575 LNCS, pages 201–214, 2012.

[8] Hema Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning Human Activities and Object Affordances from RGB-D Videos. In *The International Journal of Robotics Research*, volume 32, pages 951–970. SAGE Publications, 2013.

[9] Hema S. Koppula and Ashutosh Saxena. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016.

[10] Miroslav Kubat. Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. In *The Knowledge Engineering Review*, volume 13, pages 409–412. Cambridge Univ Press, 1999.

[11] Kang Li, Jie Hu, and Yun Fu. Modeling complex temporal composition of actionlets for activity prediction. *Computer Vision, ECCV 2012*, pages 286–299, 2012.

[12] Shahriar Nirjon, Chris Greenwood, Carlos Torres, Stefanie Zhou, John a. Stankovic, Hee Jung Yoon, Ho Kyeong Ra, Can Basaran, Taejoon Park, and Sang H. Son. Kintense: A robust, accurate, real-time and evolving system for detecting aggressive actions from streaming 3D skeleton data. In *International Conference on Pervasive Computing and Communications*, pages 2–10. IEEE Press, 2014.

[13] W Niu, J Long, D Han, and Y F Wang. Human activity detection and recognition for video surveillance. In *International Conference on Multimedia and Expo*, pages 719–722. IEEE Press, 2004.

[14] John C. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning*, pages 185 – 208. MIT Press, 1998.

[15] Mirela Popa, Alper Kemal Koc, Leon J M Rothkrantz, Caifeng Shan, and Pascal Wiggers. Kinect sensing of shopping related actions. In *Communications in Computer and Information Science*, volume 277 CCIS, pages 91–100. Springer, 2012.

[16] Karinne Ramirez-Amaro, Michael Beetz, and Gordon Cheng. Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artificial Intelligence*, (June 2016), 2015.

[17] MS Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. *Computer Vision (ICCV)*, IEEE Press, 2011.

[18] Mirko Waechter and Tamim Asfour. Hierarchical segmentation of manipulation actions based on object relations and motion characteristics. *Proceedings of the 17th International Conference on Advanced Robotics, ICAR 2015*, 270273:549–556, 2015.

[19] Feng Zhou, F De La Torre, and Jk Hodgins. Hierarchical Aligned Cluster Analysis (HACA) for Temporal Segmentation of Human Motion. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 35, pages 1–40. Citeseer, 2010.