ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1632

Impact of Automated Action Labeling in Classification of Human Actions in RGB-D Videos

David Jardim [1234] and **Luís Nunes** [234] and **Miguel Dias** [124]

Abstract.
For many applications it is important to be able to detect what a human is currently doing. This ability is useful for applications such as surveillance, human computer interfaces, games and healthcare. In order to recognize a human action, the typical approach is to use manually labeled data to perform supervised training. This paper aims to compare the performance of several supervised classifiers trained with manually labeled data versus the same classifiers trained with data automatically labeled. In this paper we propose a framework capable of recognizing human actions using supervised classifiers trained with automatically labeled data in RGB-D videos.

1 Introduction

The goal of human activity recognition is to successfully classify an action performed by an individual or a group of people from a video observation. Although significant progress has been made, HAR remains a challenging area with several problems to solve. Manual analysis of video is labour intensive, fatiguing, and error prone. Solving the problem of recognizing human activities from video can lead to improvements in several application fields like surveillance systems, human computer interfaces, sports video analysis, digital shopping assistants, video retrieval, gaming and healthcare [8, 3, 7, 10, 5]. We are interested in recognizing high-level human activities and interactions between humans and objects, ideally our recognition algorithm should be robust to changes in relative distance between the body and the sensor (Kinect), skeleton orientation, and speed of an action. In order to abstract ourselves from computer vision problems the Kinect sensor will be used to extract 3D skeleton data. Usually manually labeled data is used to perform some kind of training of classifiers that will then recognize the human activities. What if this labeling could be achieved automatically? This paper shows that automating the data labeling process for the type of actions studied results in a minor loss in accuracy.

2 Proposed Pipeline

According to [1] human activity can be categorized into four different levels: gestures, actions, interactions and group activities. This paper will focus on the actions and interactions category. We recorded a dataset containing sequences of actions performed by a 12 different subjects. We used Kinect to record the dataset with sequences of combat movements composed of 8 different actions: *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. A total of 72 sequences, 360 actions was recorded. The dataset [5] is available for public usage. A modular framework was built with several task-oriented modules organized in a work-flow (Fig.1).

2.1 Temporal Segmentation and Action Labeling

In our previous work [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 that would latter be automatically labeled by a clustering algorithm.

2.2 Action classification

At this point, using our temporal segmentation approach and an off-the-shelve algorithm to perform action clustering, we were able to automatically assign a label to an action. In order to verify the accuracy of our automatically labeled training set, the original dataset was manually labeled to be used as our ground truth. Kinect is able to track 20 joints of a subject's skeleton. Of those 20 joints, only four were selected to extract features (wrist-right; wrist-left; ankle-right; ankle-left). The 3D coordinates are with respect to a frame of reference centered at Kinect. Frames from Kinect are converted into feature vectors which are invariant to relative position and orientation of the body and will be used to train the classifiers.

3 Experiments

We experimented with the following classifiers: Multilayer Perceptron (MLP) as in [6]; Support Vector Machines (SVM) using pairwise classification [9] and Random Forests (RF) which are a combination of tree predictors [2]. Eight binary supervised classifiers were trained using manually labeled data for recognizing the eight aggressive actions contained in our dataset. Binary classifiers produced the best results in [7] using SVM classifiers.

Comparing Table 1 with Table 2 shows the difference of using a manually labeled training set versus a training set labeled by our automatic labeling pipeline. As expected the usage of automatic labeling has affected the accuracy of the classifiers. This can be explained by the error that our automatic labeling method introduces. Finally,

[1] Microsoft Language Development Center, Lisbon, Portugal, email: t-dajard, midias@microsoft.com
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal, email: luis.nunes@iscte.pt
[3] Instituto de Telecomunicações, Lisbon, Portugal
[4] ISTAR-IUL, Lisbon, Portugal

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

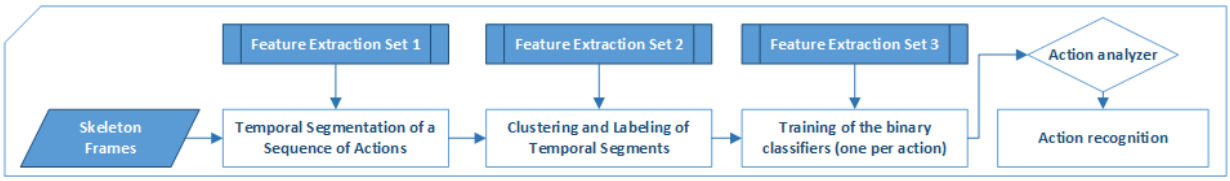

Figure 1. Modular framework for action recognition

Table 1. Classification accuracy (%) of the binary classifiers using manually labeled data and corresponding standard deviation between trials

Action	MLP	SVM	RF
right-punch	94,24 ±0,44%	91,52 ±0,17%	90,08 ±0,37%
left-punch	89,09 ±0,44%	92,50 ±0,26%	92,21 ±0,37%
front-right-kick	88,14 ±0,96%	87,95 ±0,21%	93,20 ±0,53%
front-left-kick	89,96 ±0,79%	90,42 ±0,28%	91,97 ±0,48%
side-right-kick	91,22 ±0,16%	91,92 ±0,07%	94,53 ±0,57%
side-left-kick	83,62 ±0,97%	84,76 ±0,23%	91,74 ±0,51%
backfist	92,55 ±0,32%	92,77 ±0,00%	93,58 ±0,46%
elbow-strike	95,02 ±0,28%	96,66 ±0,00%	96,66 ±0,00%

Table 2. Classification accuracy (%) of the binary classifiers using automatic labeled data and corresponding standard deviation between trials

Action	MLP	SVM	RF
right-punch	83,82 ±0,81%	88,29 ±0,16%	89,40 ±0,48%
left-punch	82,43 ±1,31%	90,20 ±0,00%	90,84 ±0,33%
front-right-kick	81,22 ±0,74%	90,75 ±0,07%	90,00 ±0,49%
front-left-kick	89,99 ±0,76%	87,91 ±0,13%	90,99 ±0,25%
side-right-kick	82,80 ±1,18%	87,88 ±0,07%	89,57 ±0,57%
side-left-kick	84,99 ±0,86%	90,28 ±0,05%	90,56 ±0,68%
backfist	83,09 ±1,44%	87,60 ±0,00%	90,05 ±0,41%
elbow-strike	95,90 ±0,31%	96,83 ±0,00%	96,83 ±0,00%

in Table 3 we calculate the difference in performance for each classifier accuracy using manually labeled data and automatically labeled data.

Table 3. Difference in performance (%) between the two approaches (manual vs automatic) for each binary classifier per action

Action	MLP	SVM	RF
right-punch	-10,42 %	-3,23 %	-0,68 %
left-punch	-6,66 %	-2,30 %	-1,37 %
front-right-kick	-6,92 %	2,80 %	-3,2 %
front-left-kick	0,03 %	-2,51 %	-0,98 %
side-right-kick	-8,42 %	-4,04 %	-4,96 %
side-left-kick	1,37 %	5,52 %	-1,18 %
backfist	-9,46 %	-5,17 %	-3,53 %
elbow-strike	0,88 %	0,17 %	0,17 %
average	-4,95 %	-1,09 %	-1,97 %

4 Conclusion

In summary, our 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 without relying on any human resources, with a minor loss of precision relative to training with human labeled data. Although this research area has grown dramatically in the past years, we identified a potentially under explored sub-area: action prediction. In future work, we would like to expand the current vision-based activity analysis to a level where it is possible, at some points, to predict a future action executed by a subject in the context of a sequence of actions.

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] J K Aggarwal and M S Ryoo, 'Human Activity Analysis: A Review', *ACM Computing Surveys*, **43**(3), 1–43, (2011).
[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, pp. 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*, p. in press. Springer International Publishing, (2016).
[5] 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).
[6] Miroslav Kubat, 'Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7', in *The Knowledge Engineering Review*, volume 13, pp. 409–412. Cambridge Univ Press, (1999).
[7] 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*, pp. 2–10. IEEE Press, (2014).
[8] 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*, pp. 719–722. IEEE Press, (2004).
[9] John C. Platt, 'Fast Training of Support Vector Machines Using Sequential Minimal Optimization', in *Advances in Kernel Methods - Support Vector Learning*, pp. 185 – 208. MIT Press, (1998).
[10] 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, pp. 91–100. Springer, (2012).