Human Activity Recognition and Prediction

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

1 RESEARCH PROBLEM

Human activity recognition (HAR) has become one of the most active research topics in image processing and pattern recognition (Aggarwal, J. K. and Ryoo, M. S., 2011). Detecting specific activities in a live feed or searching in video archives still relies almost completely on human resources. Detecting multiple activities in real-time video feeds is currently performed by assigning multiple analysts to simultaneously watch the same video stream. Manual analysis of video is labor intensive, fatiguing, and error prone. Solving the problem of recognizing human activities from video can lead to improvements in several applications fields like in surveillance systems, human computer interfaces, sports video analysis, digital shopping assistants, video retrieval, gaming and health-care (Popa et al., n.d.; Niu, W. et al., n.d.; Intille, S. S., 1999; Keller, C. G., 2011).

This area has grown dramatically in the past 10 years, and throughout our research we identified a potentially underexplored sub-area: Action Prediction. What if we could infer the future actions of people from visual input? We propose to expand the current vision-based activity analysis to a level where it is possible to predict the future actions executed by a subject.

We are interested in interactions which can involve a single actor, two humans and/or simple objects. For example try to predict if “a person will cross the street” or “a person will try to steal a hand-bag from another” or where will a tennis-player target the next volley. Using a hierarchical approach we intend to represent high-level human activities that are composed of other simpler activities, which are usually called sub-events which may themselves be decomposable. We expect to develop a system capable of predicting the next action in a sequence initially using offline-learning to bootstrap the system and then with self-improvement/task specialization in mind, using online-learning.

2 OUTLINE OF OBJECTIVES

The summarized objectives are:

- Detecting relevant human behavior in midst of irrelevant additional motion;
- Recognizing the detected actions among several pre-learned actions;
- Given the current recognized action, predicting the next most likely action or behavior that will occur in a near future.

This research intends to create a system that can, in real-time, accurately and robustly predict complex human activities before they occur. The system will evolve and learn new actions over time. We will be using the Kinect sensor to extract 3D skeleton data. The system should be robust to changes in relative distance between the body and the sensor, skeleton orientation, and speed of an action. Some of the existing approaches try to predict the current action in a short term before it has been concluded as seen in (Ryoo, M., 2011) while others try to predict in more long term situations (Li, K. et al., 2012). We foresee several useful applications such as predicting an ongoing stealing activity as early as possible based on live video observations, in sports trying to predict to which side of the goal the player is going to score the penalty or in tennis guessing to which side of the court the player is going to shot the ball, or in health, trying to detect signs of confused or dangerous behavior in patients with diseases that cause the degeneration of the central nervous system.

3 STATE OF THE ART

We separated the state of the art in two sections: the first is related to Human Activity Recognition, while the second focuses on Human Activity Prediction.

Jardim D., Nunes L. and Dias M..
Human Activity Recognition and Prediction.

3.1 Human Activity Recognition

Human activity recognition is a classification problem in which events performed by humans from video data are automatically recognized. Some of the earliest work on extracting useful information through video analysis was performed by O'Rourke and Badler (O'Rourke, J. and N. I. Badler, 1980) in which images were fitted to an explicit constraint model of human motion, with constraints on human joint motion, and constraints based on the imaging process. Also Rashid (Rashid, Rick. 1980) did some work on understanding the motion of 2D points from which he was able to infer 3D positions. Driven by application demands, this field has seen a relevant growth in the past decade. Applied in surveillance systems, human computer interfaces, video retrieval, gaming and quality-of-life devices for the elderly. Initially the main focus was recognizing simple human actions such as walking and running (Dariu M. Gavrila, 1999). Now that that problem is well explored, researchers are moving towards recognition of complex realistic human activities involving multiple persons and objects. In the review written by (Aggarwal, J. K. and Ryoo, M. S. 2011) an approach-based taxonomy was chosen to categorize the activity recognition methodologies which are divided into two categories.

Single-layered approaches (Bobick, A.F. and Wilson, A.D. 1997; Yamato, J. et al., 1992; Starner, T. and Pentland, A., 1995) typically represent and recognize human activities directly based on sequences of images and are suited for the recognition of gestures and actions with sequential characteristics. Hierarchical approaches represent high-level human activities that are composed of other simpler activities (Aggarwal, J. K. and Ryoo, M. S. 2011). Since we are interested in recognizing action sequences we will focus on the hierarchical approaches and interactions between humans and objects. Hierarchical approaches can be seen as statistical, syntactic and description-based (Damen, D. Hogg, D., 2009; Gupta, A., 2009; Intille, S. S. and Bobick, A. F., 1999; Pinhanez, C.S. and Bobick, A.F., 1998; Yu, E. and Aggarwal, J.K., 2006).

3.1.1 Statistical Approaches

This approach uses multiple layers of statistical state-based models (usually two) such as Hidden Markov Models (HMMs) and dynamic Bayesian networks (DBNs) to recognize activities with sequential structures. At the lower-layer, atomic actions are recognized from sequences of feature vectors which are converted to a sequence of atomic actions. Then, the upper-layer treats this sequence of atomic actions as observations generated by the upper-layer models. For each model, a probability of the model generating a sequence of observations is calculated to measure the likelihood between the activity and the input image sequence.

One of the most fundamental forms of the hierarchical statistical approach was presented by (Oliver, N et al., 2002) using layered Hidden Markov Models (HMM). In this approach, the bottom layer HMMs recognize atomic actions of a single person by matching the models with the sequence of feature vectors extracted from videos. The upper layer HMMs represent a high-level activity as a sequence of atomic actions. The authors of (Nguyen, 2005) have also constructed hierarchical HMMs to recognize complex sequential activities.

These approaches are especially suited to recognize sequential activities (Damen, D. and Hogg, D., 2009; Yu, E. and Aggarwal, J.K., 2006). With enough training data, statistical models are able to reliably recognize activities even with noisy inputs. The major limitation of statistical approaches is their inability to recognize activities with complex temporal structures, such as an activity composed of concurrent sub-events (Ivanov, Y.A. and Bobick, A.F., 2000).

3.1.2 Syntactic Approaches

Syntactic approaches model human activities as a string of symbols, where each symbol corresponds to an atomic-level action which has to be recognized first. Human activities are represented as a set of production rules generating a string of atomic actions, and they are recognized by adopting parsing techniques from the field of programming languages such as context-free-grammars (CFGs) and stochastic context-free grammars (SCFGs).

A hierarchical approach to the recognition of high-level activities using SCFGs was proposed by (Ivanov, Y.A. and Bobick, A.F., 2000) where they divided the framework into two layers: the lower layer used HMMs for the recognition of simple actions, and the higher layer used stochastic parsing techniques for the recognition of high-level activities. The authors in (Moore, D., n.d.) extended the work described by (Ivanov, Y.A. and Bobick, A.F., 2000) using SCFGs for the recognition of activities, focusing on multitasked activities. They were able to recognize human activities happening in a blackjack card game, such as "a dealer dealt a card to a player" with a high accuracy level.

This approach also struggles to recognize concurrent activities. Syntactic approaches model a high-level activity as a string of atomic-level activities that compose them. The temporal ordering of these atomic-level activities has to be strictly sequential. Therefore, they tend to have difficulties when an unknown observation interferes with the system.

3.1.3 Description-based Approaches

This recognition approach explicitly maintains spatio-temporal structures of human activities. It represents a high-level human activity in terms of simpler activities as sub-events, describing their temporal, spatial and logical relationships. The recognition of the activity is performed by searching the sub-events satisfying the relations specified in its representation.

In description-based approaches, a time interval is usually associated with an occurring sub-event to specify necessary temporal relationships among sub-events. Many researchers (Pinhanez, C.S. and Bobick, A.F., 1998; Nevatia et al. 2003; Vu, V. et al., 2004; Ryoo, M.S. and Aggarwal, J.K., 2006) have adopted the temporal predicates specified by (Allen, J. F. and Allen, J. F., 1983). These predicates are: before, meets, overlaps, during, starts, finishes and equals. Researchers (Pinhanez, C.S. and Bobick, A.F., 1998) have created a system that recognizes the top-level activity by checking which sub-events have already occurred and which have not. They were able to recognize cooking activities in a kitchen environment such as “picking up a bowl”. The atomic-level actions were manually labelled from the video in the experiments, and recognition was successful even when one of the atomic actions was not provided.

A description-based approach to analyze plays in American football was designed by (Intille, S. S. and Bobick, A. F., 1999). Using simple temporal predicates (before and around), they have shown that complex human activities can be represented by listing the temporal constraints in a format similar to those of programming languages. This representation was done using three levels of hierarchy: atomic-level, individual-level and team-level activities. More recently (Ryoo, M. S. and Aggarwal, J. K., 2008) proposed a probabilistic extension to their framework that is able to compensate for the failures of its low-level components. Description-based approaches are fragile when their low-level components are noisy. This limitation has been overtaken by (Ryoo, M. S. and Aggarwal, J. K., 2008), where they have used logistic regression to model the probability distribution of an activity, and used it to detect the activity even when some of its sub-events have been misclassified.

Human activities with complex temporal structures can be represented and recognized by description-based approaches which can successfully handle concurrent organized sub-events.

The major drawback of description-based approaches is their inability to compensate for the failures of low-level components (e.g., gesture detection failure). This issue has been addressed in some recent work done by (Gupta, A. et al., 2009) and (Ryoo, M. S. and Aggarwal, J. K., 2008) where they introduce a probabilistic semantic-level recognition to cope with imperfect lower-layers.

3.2 Human Activity Prediction

Human activity prediction (HAP) is a process of inferring ongoing activities from videos (Ryoo, M., 2011). It can be applied in surveillance systems (Ziebart, B., 2009), safety systems (Keller, C. G. et al., 2011), autonomous vehicles and shopping assistances (Popa, M. et al., n.d.).

The problem of predicting unknown variables had a major breakthrough in 1961 with the work published in (Kalman, R. E. and Bucy, R. S., 1961) commonly known as the Kálmán filter. This algorithm works in a two-step process. In the prediction step, the Kálmán filter produces estimates of the current state variables, along with their uncertainties. Once the outcome of the next measurement (including random noise) is observed, these estimates are updated using a weighted average, with more weight being given to estimates with higher certainty. It has been applied in guidance, control of vehicles and time series analysis. The Kálmán filter can also be applied in HAP as we´ve seen in (Pentland, A. and Liu, A., 1999; Ziebart, B. D. et al., 2009).

One of the earliest approaches that we’ve found tried to model and predict human behavior when driving an automobile was (Pentland, A. and Liu, A., 1999). The goal is to recognize human driving behaviors accurately and anticipate the human’s behavior for several seconds into the future. They consider the human as a device with a large number of internal mental states, each with its own particular control behavior and interstate transition probabilities. The states of the model can be hierarchically organized to describe both short-term and longer-term behaviors; for instance, in the case of driving an automobile, the longer-term behaviors might be passing, following, and turning, while shorter-term behaviors would be maintaining lane position and releasing the brake. The authors introduced the concept of multiple dynamic models (MDM) which defends that the most complex model of human behavior is to have several alternative models of the person's dynamics. Then at each instant they make observations of the person's state, decide which model applies, and give a response based on that model. This multiple model approach produces a generalized maximum likelihood estimate of the current and future values of the state variables. With this approach they have accurately categorized human driving actions very soon after the beginning of the action.

Another type of prediction was addressed by (Ziebart, B., 2009) where a robot should predict the future locations of people and plan routes that will avoid disrupting the person's natural behavior due to the robot's proximity, while still efficiently achieving its objectives using a soft-max version of goal-based planning. They represent the sequence of actions that lead to a person's future position using a deterministic Markov decision process (MDP) over a grid representing the environment. People do not move in a perfectly predictable manner, so the robot has to reason probabilistically about their future locations. By maximizing the entropy of the distribution of trajectories, which are subject to the constraint of matching the reward of the person's behavior in expectation, they obtain a distribution over trajectories. One interesting feature is the fact that the feature-based cost function learned using this approach allows accurate generalization to changes in the environment. Although to successfully predict the future trajectory of a person through an environment the authors require a setting where the human behavior is fully observable and not very crowded.

Another work by Ryoo (Ryoo, M., 2011) tries to construct an intelligent system which will perform early recognition from live video streams in real-time. They introduce two new human activity prediction approaches which are able to cope with videos from unfinished activities. Integral bag-of-words is a probabilistic activity prediction approach that constructs integral histograms to represent human activities. Simply putting it, the idea is to measure the similarity between a video and the activity model by comparing their histogram representations. The other approach is called Dynamic bag-of-words which considers the sequential nature of human activities, while maintaining the bag-of-words advantages to handle noisy observation. The motivation is to divide the activity model and the observed sequence into multiple segments to find the structural similarity between them. That is, the bag-of-words paradigm is applied in matching the interval segments, while the segments themselves are sequentially organized based on their recursive activity prediction formulation. They've managed to correctly predict ongoing activities even when the provided videos contain less than the first half of the activity.

In (Kitani, K. M. et al., n.d.) 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. They've focused on the problem of trajectory-based human activity analysis exploring the interplay between features of the environment and pedestrian trajectories. To integrate the aspects of prior knowledge into modeling human activity, they've leveraged recent progress in semantic scene labeling and inverse optimal control. This kind of labeling provides a way to recognize physical scene features such as pavement, grass, tree, building and cars, playing a critical role in advancing the representational power of human activity models. Inverse optimal control is also called Inverse Reinforcement Learning which expands the horizon of vision-based human activity analysis by integrating the impact of the environment and goals on future actions. The authors propose a Hidden variable Markov Decision Process (HMDP) model which incorporates uncertainty (e.g., probabilistic physical scene features) and noisy observations (e.g., imperfect tracker) into the activity model to express the dynamics of the decision-making process. Since the proposed method encapsulates activities in terms of physical scene features and not physical location, it is also able to generalize to novel scenes transferring knowledge. They are able to forecast possible destinations of the pedestrians through a path, but this evaluation is limited to the physical features of the environments.

More recently (Koppula, H. S., 2013) consider the problem of detecting past activities as well as anticipating which activity will happen in the future and how. They start by modelling the rich spatio-temporal relations between human poses and objects using a conditional random field (CRF). The key idea is to sample a few segmentations that are close to the ground-truth segmentation using the CRF model instantiated with a subset of features, and then explore the space of segmentation by making merge and split moves to create new segmentations. Done by approximating the graph with only additive features, which lends to efficient dynamic programming. With that they can reason about the possible graph structures for both past and future activities. From their experiments with over 120 activity videos (*making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food*), they achieved an accuracy of 70.3% for sub-activity labeling and 83.1% for high-level activities respectively for detection. Furthermore, they obtained an accuracy of 49.6% for anticipating sub-activities in future time-frames.

In the research of (Koppula, H. and Saxena, A., 2013) the 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). This is achieved by modelling three aspects of the activities. First, they model the activities through a hierarchical structure in time where an activity is composed of a sequence of sub-activities. Second, model their interdependencies with objects and their affordances. Third, it is necessary to anticipate the motion trajectory of the objects and humans, which will tell how the activity can be performed. For anticipation, they present an anticipatory temporal conditional random field (ATCRF), where 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.

They've used a dataset containing 120 RGB-D videos of daily human activities, such as microwaving food, taking medicine, etc. The results show that anticipation improves the detection of past activities: 85.0% with vs 82.3% without. Their algorithm obtains an activity anticipation accuracy (defined as whether one of the top three predictions actually happened) of (75.4%, 69.2%, 58.1%) for predicting (1.3, 10) seconds into the future.

The authors in (Hoai, M. and De la Torre, F., 2013) propose Max-Margin Early Event Detectors (MMED), a novel formulation for training event detectors that recognize partial events, enabling early detection. MMED is based on structured output SVM but extends it to accommodate the nature of sequential data. The key idea behind MMED is that given a training time series that contains a complete event, they simulate the sequential arrival of training data and use partial events as positive training examples. Experiments on datasets of varying complexity, from synthetic data and sign language to facial expression and human actions, showed that their method often made faster detections while maintaining comparable or even better accuracy.

Prediction is also important in the field of human-robot collaboration where (Hawkins, K. and Vo, N., 2013) created a system whose goal is to predict in a probabilistic manner when the human will perform different subtasks that may require robot assistance in a human-robot collaboration. The robot must determine the state of the collaborative task being performed and it must infer both what to do and when to do it. The representation is a graphical model where the start and end of each subtask is explicitly represented as a probabilistic variable conditioned upon prior intervals. This formulation allows the inclusion of uncertain perceptual detections as evidence to drive the predictions. Next, given a cost function that describes the penalty for different wait times, a planning algorithm was developed which selects robot-actions that minimize the expected cost based upon the distribution over predicted human-action timings.

Depending on the confidence of the model several results were obtained. Though a high confidence detector can occasionally produce little or no wait time, it can also suffer from severe failures. A low confidence detector, however, can produce consistently reasonable results.

The work done by (Li, K. et al., 2012) might be the most related to what we are trying to achieve with our research. Authors 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 key contribution of this work is the idea that causality of action units can be encoded as a Probabilistic Suffix Tree (PST) with variable temporal scale, while the predictability can be characterized by a Predictive Accumulative Function (PAF) learned from information entropy changes along every stage of activity progress. The efficiency of their method was tested on the complex activity of playing a tennis game and predicting who will win the game (65% of certainty with 60% of observed game).

From what we´ve seen prediction of actions can be much improved specially in a mid/long-term prediction in complex activities. We hope that with the use of Kinect 2 and our array of specialized classifiers each of them connected to purpose data filter combined with contextual information such as the scene we might be able to obtain better results.

4 METHODOLOGY

In our setting, our algorithm will be observing a scene containing a human (or two interacting humans) for a certain amount of time, and our goal is to detect current activities and anticipate future activities.

The processing sequence can be outlined as:

- Obtain RGB-D data;
- Extract and calculate useful features such as joint position in 3D, angle of the joints and speed;
- Partition sequences into small actions (or movements). These may be full body or part-of-body actions. In some cases more than one part-of-body action can be active simultaneously.
- Detect actions using an action recognition module (includes filtering data, running it through an array of parallel classification models and merge classifications producing one a set of matches with each of the known actions); Eventually, this module may create new classifiers for movements that do not match any of the known actions.
- Predict future actions based on the probabilities of the previously observed action sequence.

Initially our tests will be performed with our own recorded dataset (part of which was already acquired) and later applied on other datasets to verify the performance of our framework. In order to partition the sequences in sub-activities we have to be able to detect the occurring actions. One approach would be to label the actions manually and then train a classifier with labelled actions. However we will aim at creating a mechanism that can successfully label activities automatically by finding patterns of movement. This could be possible by analyzing the obtained metrics (3D position, angle and velocity of the joints) and using a clustering algorithm, such as K-means or clustered HMM's. If we successfully label the low level activities we will proceed to train an array of classifiers as illustrated in Figure 1, each of which recognizes a specific action. We will also include the possibility to learn new actions. If the system is presented with an unknown action it should create a new class of actions, thus creating a library of actions.

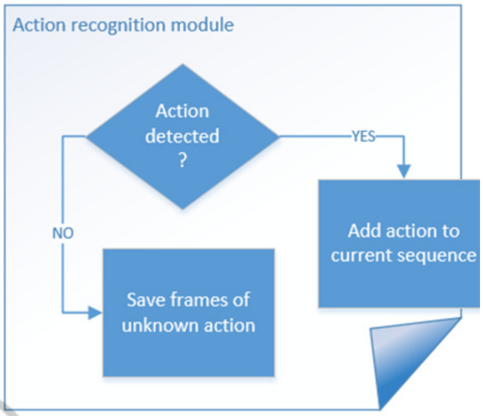
Action recognition module
Action detected ?
YES
NO
Add action to current sequence
Save frames of unknown action

Figure 1: Action recognition module.

As for the action prediction module (Figure 2), given the current detected action and the context of the current sequence of actions it will predict with a probability from 0 to 1 of confidence which action will be performed next. This module is dependent of the action recognition module, after having the actions labeled and discovered our first approach will be to use conditional random fields (CRFs) to recognize patterns and perform a structured prediction. This is often used to labeling or parsing sequential data, which is the case of our data.

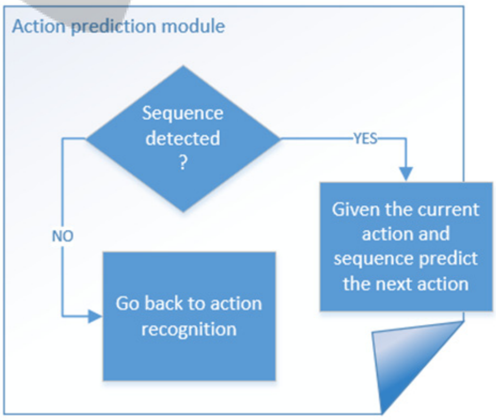
Action prediction module
Sequence detected ?
YES
NO
Given the current action and sequence predict the next action
Go back to action recognition

Figure 2: Action prediction module.

5 EXPECTED OUTCOME

This research intends to make the following contributions for advancing the state of the art of Human activity recognition and prediction:

- Develop an approach to this problem that can be implemented on top of currently available commercial hardware and software.

- Compare different approaches of action-division rules in 3D skeleton sequences.
- Compare the importance of different features extracted from the collected data, or calculated.
- Implement a cluster analysis method which will facilitate the task of recognizing actions by grouping a set of points representing the skeleton data when performing a specific action;
- Create an algorithm that, in an occurring sequence of actions, successfully detects patterns and predicts what will happen next. This prediction could be short, mid or long-term;
- Develop a functional prototype that, by using the previous methods, will be able to recognize and predict actions in real-time. This will be the main criteria for evaluation of the research results;
- Advance the state-of-the-art in the development of automated visual systems which have the task of recognizing and describing human actions and improve the performance of action prediction.

6 STAGE OF THE RESEARCH

The research is still at its early stages, the related work has been extensively researched, and we have already created a dataset recorded with Kinect that consists of skeleton data from 12 people, each performing 6 sequences containing 5 actions, with a total of 8 different actions. In total 72 sequences and 360 actions. This will increase in the future as we intend to use Kinect 2 as soon as it is released to the public. Also a framework capable of capturing and playing RGB-D videos has been developed. Our next step is to automatically partition and classify our sequences of action via a clustering algorithm.

ACKNOWLEDGMENTS

This work was partially funded by Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP), FCT: SFRH/BDE/52125/2013 and FCT: PTDC/EEI-SII/2312/2012.

REFERENCES

Aggarwal, J. K., and Ryoo, M. S. (2011). Human activity analysis. ACM Computing Surveys, 43(3), 1–43. doi:10.1145/1922649.1922653.

Allen, J. F., and Allen, J. F. (1983). Maintaining Knowledge about Temporal Intervals, 26(11), 832–843.

Bobick, A.F. Wilson, A.D., A state-based approach to the representation and recognition of gesture, Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.19, no.12, pp.1325-1337, Dec 1997 doi: 10.1109/34.643892.

C Wolf, J. Mille, L.E Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E Dellandrea, C.-E. Bichot, C. Garcia, B. Sankur, The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition, Technical Report RR-LIRIS-2012-004, LIRIS Laboratory, March 28th, 2012.

CMU Graphics Lab Motion Capture Database, Available: http://mocap.cs.cmu.edu/. Last accessed 14th August 2012.

Damen, D. Hogg, D., Recognizing linked events: Searching the space of feasible explanations, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on , vol., no., pp.927-934, 20-25 June 2009 doi: 10.1109/CVPR.2009.5206636.

Dariu M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding (CVIU) 73 (1) (1999) 82–92.

Gupta, A., Srinivasan, P., Jianbo Shi; Davis, L.S., Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on , vol., no., pp.2012-2019, 20-25 June 2009 doi: 10.1109/CVPR.2009.5206492.

Hawkins, K., and Vo, N. (2013). Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration. Proceedings of the IEEE.

Hoai, M., and De la Torre, F. (2013). Max-Margin Early Event Detectors. International Journal of Computer Vision, 107(2), 191–202. doi:10.1007/s11263-013-0683-3.

ICPR - HARL 2012 (Human activities recognition and localization competition), Available: http://liris.cnrs.fr/harl2012/. Last accessed 24th September 2012.

Intille, S. S., and Bobick, A. F. (1999). A Framework for Recognizing Multi-Agent Action from Visual Evidence, (489), 1–7.

Ivanov, Y.A. Bobick, A.F., Recognition of visual activities and interactions by stochastic parsing, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.22, no.8, pp.852-872, Aug 2000 doi: 10.1109/34.868686.

Kalman, R. E., and Bucy, R. S. (1961). New Results in Linear Filtering and Prediction Theory. Journal of Basic Engineering, 83(1), 95. doi:10.1115/1.3658902.

Keller, C. G., Dang, T., Fritz, H., Joos, A., Rabe, C., and Gavrila, D. M. (2011). Active Pedestrian Safety by Automatic Braking and Evasive Steering. IEEE

Transactions on Intelligent Transportation Systems, 12(4), 1292–1304. doi:10.1109/TITS.2011.2158424.

Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M. (n.d.). Activity Forecasting, 1–14.

Koppula, H. S. (2013). Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation, 28.

Koppula, H., and Saxena, A. (2013). Anticipating Human Activities using Object Affordances for Reactive Robotic Response. Robotics: Science and Systems.

Li, K., Hu, J., and Fu, Y. (2012). Modeling complex temporal composition of actionlets for activity prediction. Computer Vision–ECCV 2012, 286–299.

Liu, N., Lovell, B. C., Kootsookos, P. J., Davis, R. I. A., Imaging, I. R., and Group, S. I. (n.d.). Understanding HMM Training for Video Gesture Recognition School of Information Technology and Electrical Engineering, (Figure 2), 2–5.

Lopes, P.F. Jardim, D. Alexandre, I.M. , Math4Kids, Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on , vol., no., pp.1-6, 15-18 June 2011.

Moore, D. (n.d.). Recognizing Multitasked Activities from Video using Stochastic Context-Free Grammar Introduction and Related Work Representation using SCFG The Earley-Stolcke Parsing AAAI-02, 770–776.

Nevatia, Ram Zhao, Tao Hongeng, Somboon, Hierarchical Language-based Representation of Events in Video Streams, Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on , vol.4, no., pp.39, 16-22 June 2003 doi: 10.1109/CVPRW.2003.10038.

Nguyen, N.T. Phung, D.Q. Venkatesh, S. Bui, H., Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on , vol.2, no., pp. 955- 960 vol. 2, 20-25 June 2005 doi: 10.1109/CVPR.2005.203.

Niu, W., Long, J., Han, D., Wang, Y., and Barbara, S. (n.d.). Human Activity Detection and Recognition for Video Surveillance, 1–4.

Oliver, N. Horvitz, E. Garg, A., Layered representations for human activity recognition, Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on, vol., no., pp. 3- 8, 2002 doi: 10.1109/ ICMI.2002.1166960.

O'Rourke, J. and N. I. Badler. 1980. Model-based image analysis of human motion using constraint propagation. IEEE PAMI, 2(4).

Pentland, A. and Liu, A. (1999). Modeling and prediction of human behavior. Neural computation, 11(1), 229–42. Retrieved from http://www.ncbi.nlm.nih.gov/ pubmed/9950731.

Pinhanez, C.S., Bobick, A.F., Human action detection using PNF propagation of temporal constraints, Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on, vol., no., pp.898-904, 23-25 Jun 1998 doi: 10.1109/CVPR.1998.698711.

Popa, M., Koc, A. K., Rothkrantz, L. J. M., Shan, C., and Wiggers, P. (n.d.). Kinect Sensing of Shopping related Actions.

Rashid, Rick. 1980. LIGHTS: a system for interpretation of moving light displays. Ph.D. thesis, University of Rochester Computer Science Department.

Ryoo, M. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. Computer Vision (ICCV), 2011 IEEE, (Iccv). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6126349.

Ryoo, M. S., and Aggarwal, J. K. (2008). Semantic Representation and Recognition of Continued and Recursive Human Activities. International Journal of Computer Vision, 82(1), 1–24. doi:10.1007/s11263-008-0181-1.

Ryoo, M.S. Aggarwal, J.K., Semantic Understanding of Continued and Recursive Human Activities, Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol.1, no., pp.379-378, 0-0 0 doi: 10.1109/ICPR.2006.1043.

Ryoo, M.S., Aggarwal, J.K. , Recognition of Composite Human Activities through Context-Free Grammar Based Representation, Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol.2, no., pp. 1709- 1718, 2006 doi: 10.1109/CVPR.2006.242.

Sinha, S. N., Frahm, J., Pollefeys, M., and Genc, Y. (2006). GPU-based Video Feature Tracking And Matching, 012(May), 1–15.

Starner, T. Pentland, A., Real-time American Sign Language recognition from video using hidden Markov models, Computer Vision, 1995. Proceedings., International Symposium on, vol., no., pp.265-270, 21-23 Nov 1995 doi: 10.1109/ISCV.1995.477012.

Uddin, M. Z., Byun, K., Cho, M., Lee, S., Khang, G., and Kim, T.-S. (2011). A Spanning Tree-Based Human Activity Prediction System Using Life Logs from Depth Silhouette-Based Human Activity Recognition. In P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch (Eds.), Computer Analysis of Images and Patterns (Vol. 6854, pp. 302–309). Springer Berlin Heidelberg. doi:10.1007/978-3-642-23672-3_37.

Vu, V., Bremond, F., Thonnat, M., Orion, P., Sophia, I. N. R. I. A., Cedex, B.-S. A., Vu, T., et al. (2004). Automatic Video Interpretation : A Novel Algorithm for Temporal Scenario Recognition, 1–6.

Wang, J. (n.d.). Mining Actionlet Ensemble for Action Recognition with Depth Cameras.

Yamato, J., Ohya, J. Ishii, K., Recognizing human action in time-sequential images using hidden Markov model, Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on , vol., no., pp.379-385, 15-18 Jun 1992 doi: 10.1109/CVPR.1992.223161.

Yu, E. Aggarwal, J.K., Detection of Fence Climbing from Monocular Video, Pattern Recognition, 2006. ICPR

2006. 18th International Conference on , vol.1, no., pp.375-378, 0-0 0doi: 10.1109/ICPR.2006.440.

Ziebart, B. D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J. A., Hebert, M., et al. (2009). Planning-based prediction for pedestrians. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 3931–3936. doi:10.1109/IROS.2009.5354147.