Automatic Human Activity Segmentation and Labeling in RGBD Videos

David Jardim[1,2,3,4], Luís Nunes[2,3], and Miguel Sales Dias[1,3,4]

[1] MLDC, Lisbon, Portugal,
[2] Instituto de Telecomunicações, Lisbon, Portugal,
[3] University Institute of Lisbon (ISCTE-IUL), Lisbon, Portugal,
[4] ISTAR-IUL, Lisbon, Portugal,
t_dajard@microsoft.com, luis.nunes@iscte.pt,
miguel.dias@microsoft.com

Abstract. Human activity recognition has become one of the most active research topics in image processing and pattern recognition. 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 health-care. This paper aims to recognize an action performed in a sequence of continuous actions recorded with a Kinect sensor based on the information about the position of the main skeleton joints. The typical approach is to use manually labeled data to perform supervised training. In this paper we propose a method to perform automatic temporal segmentation in order to separate the sequence in a set of actions. By measuring the amount of movement that occurs in each joint of the skeleton we are able to find temporal segments that represent the singular actions. We also proposed an automatic labeling method of human actions using a clustering algorithm on a subset of the available features.

Keywords: Human motion analysis, Motion-based recognition, Action recognition, Temporal segmentation, Clustering, K-means, Labeling, Kinect, Joints, Video sequences

1 Introduction

Human activity recognition is a classification problem in which events performed by humans are automatically recognized. 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 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 health-care [15, 13, 8, 10]. Ultimately, we are interested in recognizing high-level human activities and interactions between humans and objects. The main sub-tasks of this recognition are Usually achieved using manually labeled data to train classifiers to recognize a set of human activities. An interesting question is how far can we take the automatic labeling of human actions using unsupervised learning? From our experiments we have found that this labeling is possible, but still with a large margin for improvement.

2 Related Work

Human activity recognition is a classification problem in which events performed by humans are automatically recognized by a computer program. Some of the earliest work on extracting useful information through video analysis was performed by O'Rourke and Badler [9] 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 [16] did some work on understanding the motion of 2D points in which he was able to infer 3D position. Driven by application demands, this field has seen a relevant growth in the past decade. This research has been 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 [4]. Now that that problem is well explored, researchers are moving towards recognition of complex realistic human activities involving multiple persons and objects. In a recent review written by [1] an approach-based taxonomy was chosen to categorize the activity recognition methodologies which were divided into two categories. Single-layered approaches [2, 20, 18] 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 [1]. Hierarchical approaches can be seen as statistical, syntactic and description-based [3, 6, 8, 14, 17, 21].

The previous approaches all used computer vision (CV) techniques to extract meaningful features from the data. Motion capture data (MOCAP) has also been used in this field, a relevant approach found was [22] where they pose the problem of learning motion primitives (actions) as a temporal clustering one, and derive an unsupervised hierarchical bottom-up framework called hierarchical aligned cluster analysis (HACA). HACA finds a partition of a given multidimensional time series into m disjoint segments such that each segment belongs to one of k clusters representing an action. They 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 the cost in mind Microsoft released a sensor called Kinect which captures RGB-D data and is also capable of providing joint level information in a non-invasive way allowing the developers to abstract away from CV techniques. Using Kinect [11] the authors 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. Given a RGB-D video, they jointly model the human activities and object affordances as a Markov random field where the nodes represent objects and sub-activities, and the edges represent the relationships between object affordances, their relations with sub-activities, and their evolution over time. The learning problem is formulated using a structural support vector machine (SSVM) approach, where labelings over various alternate temporal segmentations are considered as latent variables. The method was tested on a dataset comprising 120 activity videos collected from 4 subjects, and obtained an accuracy of 79.4% for affordance, 63.4% for sub-activity and 75.0% for high-level activity labeling.

In [7] the covariance matrix for skeleton joint locations over time is used as a discriminative descriptor for a sequence of actions. To encode the relationship between joint movement and time, multiple covariance matrices are deployed over subsequences in a hierarchical fashion. The descriptor has a fixed length that is independent from the length of the described sequence. Their experiments show that using the covariance descriptor with an off-the-shelf classification algorithm one can obtain an accuracy of 90.53% in action recognition on multiple datasets.

In a parallel work [5] authors propose a descriptor for 2D trajectories: Histogram of Oriented Displacements (HOD). Each displacement in the trajectory votes with its length in a histogram of orientation angles. 3D trajectories are described by the HOD of their three projections. HOD is used to describe the 3D trajectories of body joints to recognize human actions. The descriptor is fixed-length, scale-invariant and speed-invariant. Experiments on several datasets show that this approach can achieve a classification accuracy of 91.26%.

Recently [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. Kintense uses a combination of: (1) an array of supervised learners to recognize a predefined set of aggressive actions, (2) an unsupervised learner to discover new aggressive actions or refine existing actions, and (3) human feedback to reduce false alarms and to label potential aggressive actions. The system is 11% - 16% more accurate and 10% - 54% more robust to changes in distance, body orientation, speed, and subject, when compared to standard techniques such as dynamic time warping (DTW) and posture based gesture recognizers. In two multi-person households it achieves up to 90% accuracy in action detection.

3 Temporal Segmentation

This research is framed in the context of a doctoral program where the final objective is to predict the next most likely action that will occur in a sequence of actions. In order to solve this problem we divided it in two parts, recognition and prediction. This paper will only refer to the recognition problem. Human activity can be categorized into four different levels: gestures, actions, interactions and group activities. We are interested in the actions and interactions category.

An initial research was conducted to analyze several datasets from different sources like LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information) dataset

[19], CMU (Carnegie Mellon University) MoCap dataset[5], MSR-Action3D and MSR-DailyActivity3D dataset [13] and verify it's suitability to our problem. All these datasets contain only isolated actions, and for our task we require sequences of actions. We saw this as an opportunity to create a new dataset that contains sequences of actions.

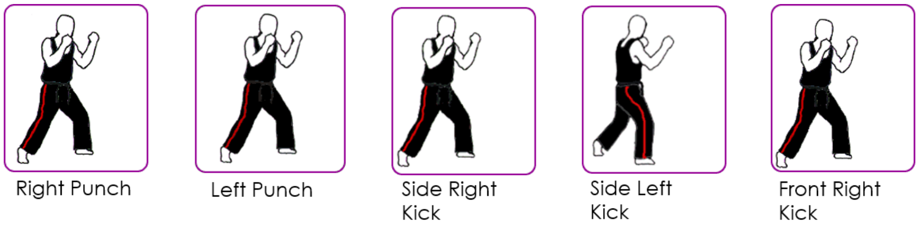

Fig. 1: Example of a sequence of actions in the dataset

We used Kinect to record the dataset which contains 8 aggressive actions like punching and kicking, 6 distinct sequences (each sequence contains 5 actions). Recorded 12 subjects, each subject performed 6 sequences. Total of 72 sequences, 360 actions. An example of a recorded sequence is illustrated on Figure 1. Kinect captures data at 30 frames per second. The data is recorded in .xed files which contains RBG, depth and skeleton information, and also a light version in .csv format containing only the skeleton data. We expect to make the dataset available to public in a near future on a dedicated website.

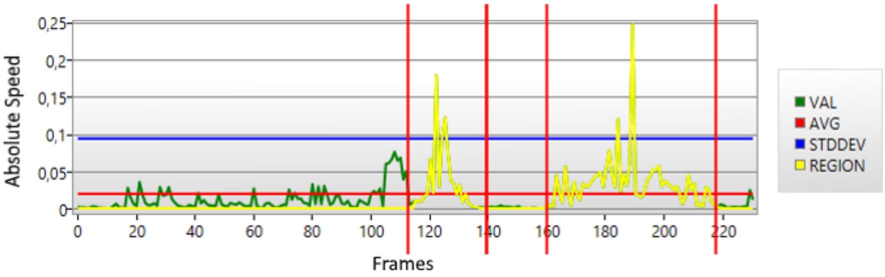

Fig. 2: Absolute speed of the right ankle while performing actions and regions of interest found

Since our dataset contains sequences of actions our very first task was to automatically decompose the sequence in temporal segments where each segment represents an isolated action. We went for a very simple approach. From visual observation we've

[5] http://mocap.cs.cmu.edu/

noticed that during each action there were joints that moved more than others. If we could measure that movement and compare it with other joints we could be able to tell which joint is predominant in a certain action and then assign a temporal segment to a joint. Figure 2 shows a timeline which represents the movement of the right ankle. It is perfectly visible that there are two regions where that joint has a significant higher absolute speed. These two regions represent moments in time where an action was performed that involved mainly the right leg.

Our first step was to create these regions which we called regions of interest. This was achieved by selecting frames in which the absolute speed value was above the standard deviation multiplied by a factor of two. Then we selected all the neighboring frames that were above the average value with a tolerance of 3 frames below of the average. This data was collected for four different joints: right and left ankle, right and left wrist. Then we searched for overlapping regions. While the user performs a kick the rest of his body moves, specially the hands to maintain the body's balance. Overlapping regions were removed by considering only the joint moving at a higher average speed in each frame. Figure 3 illustrates an example result of our automatic segmentation method. Each color of the plot represents a temporal segment to which we assigned a joint as being the dominant joint for that action. We obtained 5 temporal segments which successfully correspond to the number of actions that the sequence contains, in this case: right-punch; left-punch; front-right-kick; front-left-kick; side-right-kick.

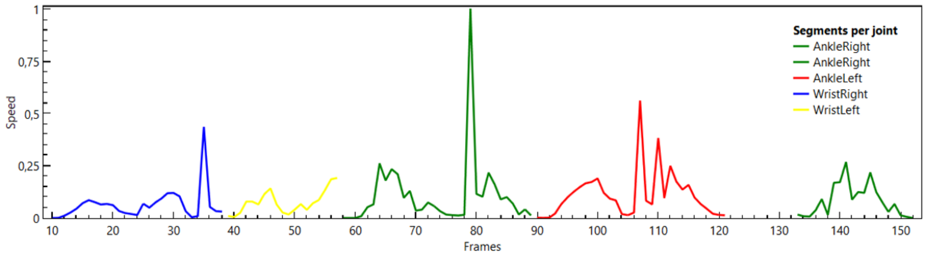

Fig. 3: Visual representation of our action segmentation method

Table 1 shows that the automatic segmentation can be improved. These results reflect the measurements between the frames of the annotated data and the frames of our automatic temporal segments. Overall the segmentation is satisfactory and we believe that the segments have the most important part of the actions. This method might be revisited in the future to improve the overall performance of our system.

Table 1: Automatic temporal segmentation per sequence

Sequence	1	2	3	4	5	6	Average
Segmentation Accuracy	94,23%	89,66%	73,55%	89,44%	76,31%	75,03%	83,04%

4 Action Labeling

In most cases, as seen in [12], action labeling is achieved by manually labeling the segments obtained by the segmentation algorithm and then use that data to train one classifier per action. Those classifiers would then be used in an application capable of recognizing actions in real-time. Instead we thought that it would be more interesting if we could automatically label equal actions performed by different subjects. For example a right-punch performed by subject 1 should be very similar to a right-punch performed by subject 2. This process is composed of the following stages:

1: Automatically find temporal segments that represent the actions of the sequence
2: Sample the dataset based on the previously found temporal segments
3: Extract meaningful features for each segment
4: Use clustering to automatically group similar actions and thus label them

4.1 Sampling

To sample the data for the clustering algorithm the program automatically selects the automatically found temporal segments which ideally should be 5 per sequence, which corresponds to the number of actions that compose the sequence. The most active joint is assigned to that segment. Based on the window-frame of the segment found for a specific joint we create new temporal segments for the remaining joints on the same exact window-frame. This can be portrayed has stacking the joints timeline one on top of another and making vertical slices to extract samples of data that correspond to temporal segments where an action has occurred.

4.2 Feature Extraction

An action can be seen as a sequence of poses over time. Each pose respects certain relative positions and orientation of joints of the skeleton. Based on the positions and orientations of the joints we extracted several features that will be used to model the movements performed by the subjects. We have experimented with several features (speed; absolute speed; speed per axis; joint flexion in angles; bone orientation). After a comparison of these different approaches (to be published) we selected the angles of the elbows and the knees *a1,a2,a3,a4* and the relative position of the wrists and ankles *s1,s2,s3,s4* (Figure 4) and used these to calculate other features like relative speed of each joint. Different subsets of these features combined will constitute the feature vectors that will be used by the clustering algorithm.

4.3 Clustering Experiments

As previously mentioned, the objective is to cluster similar actions performed by different subjects (or by the same subject in different recordings). For that purpose we will use k-means which is one of the simplest unsupervised learning algorithms. We made several experiments with different combinations of features.

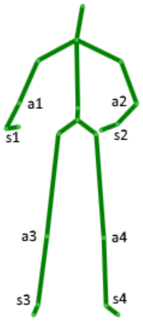

Fig. 4: Visual representation of body relative features used

Table 2: Clustering results for Sequence 1 using average speed as a feature

Action	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Right punch	8,33%	91,67%	0,0%	0,0%	0,0%
Left punch	0,0%	0,0%	0,0%	100,0%	0,0%
Front right kick	91,67%	0,0%	8,33%	0,0%	0,0%
Side right kick	33,33%	0,0%	66,67%	0,0%	0,0%
Side left kick	8,33%	8,33%	0,0%	0,0%	83,33%

Our initial experiments using simply the average speed of each joint over the whole segment as a feature. Results of this experiment are shown in Table 2. Clustering all the segments of the same sequence of actions being performed by different subjects brought interesting results. All the actions were correctly labeled except for the side-right-kick. As shown in the table this action was classified as a front-right-kick 33,33% of the time. These two actions are similar and originate from the same body part. These results lead us to believe that maybe more features could help distinguish these movements more clearly. Table 3 shows the results of clustering using also the angles of the knees and the elbows. Surprisingly the results are worst. The right-punch and the left-punch even when they are from different arms are labeled with the same cluster-label, the same happened to the front-right-kick and the side-right-kick.

Table 3: Clustering results for Sequence 1 using average speed and angles of the knees and elbows as a feature

Action	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Right punch	0,0%	0,0%	0,0%	100,0%	0,0%
Left punch	0,0%	0,0%	0,0%	100,0%	0,0%
Front right kick	100,0%	0,0%	0,0%	0,0%	0,0%
Side right kick	100,0%	0,0%	0,0%	0,0%	0,0%
Side left kick	0,0%	16,67%	50,0%	0,0%	33,33%

This can be explained by the angle features becoming more relevant than the speed features. Given that angles are less discriminative of these movements, this results in more miss-classifications. When considering the amplitude of the movements of the lower body members the differences between a right-punch and a left-punch become minor. To prove this a simple experiment was performed using only the temporal segments originated by an action from the upper part of the body. Table 4 shows that using only the upper body k-means is perfectly capable of distinguishing the actions of the right arm from the actions of the left arm using the same features as in Table 3.

Table 5 shows the results using only the angles of the knees and elbows as features. In this case the kicking actions are diluted amongst several clusters. So using only the angles as features has proven insufficient to correctly label the actions.

Table 4: Clustering results for all the sequences using only upper body actions

Action	Cluster 1	Cluster 2
Right punch	2,08%	97,92%
Left punch	91,67%	8,33%
Back fist	2,08%	97,92%
Elbow strike	16,67%	83,33%

Table 5: Clustering results for Sequence 1 using angles as a feature

Action	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Right punch	0,0%	0,0%	0,0%	75,0%	25,0%
Left punch	0,0%	0,0%	0,0%	0,0%	100,0%
Front right kick	50,0%	0,0%	25,0%	0,0%	25,0%
Side right kick	58,33%	0,0%	41,67%	0,0%	0,0%
Side left kick	0,0%	25,0%	50,0%	0,0%	25,0%

Since a single value (average) is used to represent a temporal segment a loss in the granularity of information might be a problem. In the following experiments temporal segments were divided in equal parts to increase the feature vector, but the results were 30 to 40% lower. This can be explained in figure 5 where we show a comparison between the same movement performed by two different subjects. The curves are similar but they start at different frames, which if we divide the temporal segment in four parts, for subject 10 the first two parts will have a higher value and for subject 1 the last two will have a higher value. For this reason these two actions would probably be assigned different clusters. Overall the results are very similar to all the other sequences that we have in our dataset (total of 8). Due to space limitations we were unable to include the clustering results for each sequence. Our final experiment (Table 6) was to see how well k-means coped with all the sequences at the same time using only the average speed as

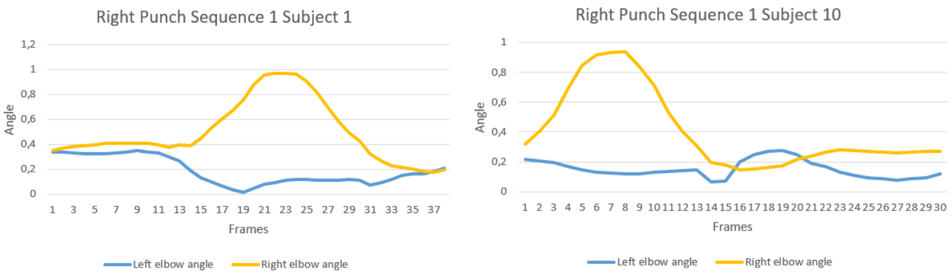

Fig. 5: Temporal segment of a right punch performed by subject 1 and subject 10

a feature since it was the feature that proved to have the best results. Again there is a clear separation from actions from the right and left side of the body. As for actions that are from the same part of the body there is room for improvement.

Table 6: Clustering results for all the sequences using speed as a feature

Action	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Right punch	2,08%	4,17%	0,0%	2,08%	0,0%	2,08%	75,0%	14,58%
Left punch	0,0%	0,0%	2,78%	0,0%	88,89%	2,78%	2,78%	2,78%
Front right kick	0,0%	0,0%	93,75%	0,0%	0,0%	4,17%	0,0%	2,08%
Side right kick	5,00%	0,0%	50,00%	0,0%	0,0%	45,00%	0,0%	0,0%
Front left kick	5,56%	0,0%	2,78%	86,11%	0,0%	0,0%	0,0%	5,56%
Side left kick	36,11%	0,0%	0,0%	38,89%	1,39%	2,78%	2,78%	18,06%
Back fist	0,0%	52,08%	0,0%	0,0%	2,08%	0,0%	45,83%	0,0%
Elbow strike	8,33%	16,67%	0,0%	0,0%	25,0%	0,0%	33,33%	16,67%

5 Conclusion

In this paper, we described a new dataset of sequences of actions recorded with Kinect, which is, to the best of our knowledge, the first to contain whole sequences. We proposed a method to achieve automatic temporal segmentation of a sequence of actions trough a simple filtering approach. We also proposed and evaluated an automatic labeling method of human actions using a clustering algorithm. In summary, our results show that, for the type of actions used, k-means is capable of grouping identical actions performed by different users. This is evident when the clustering is performed with all of the subjects performing the same sequence of actions. When all the sequences are used the accuracy decreases. This might be explained by the effect that the neighboring actions have on the current action. So for different neighboring actions, the same current action will have a different start and ending.

By using several features (absolute speed, absolute 3D speed, joint angle) we also show that the choice of features affects greatly the performance of k-means. The poor results achieved when using the angles of the knees and elbows, appear to be related to how the flexion angles are calculated using the law of cosines. In our next experiment Euler angles will be used which represent a sequence of three elemental rotations (rotation about X,Y,Z). We also think that we could improve the results if we applied dynamic time warping to the temporal segments. This technique is often used to cope with the different speed with which the subjects perform the actions.

Our study showed how clustering and filtering techniques can be combined to achieve unsupervised labeling of human actions recorded by a camera with a depth sensor which tracks skeleton key-points.

References

1. J. K. Aggarwal and M. S. Ryoo. Human Activity Analysis: A Review. *ACM Computing Surveys*, 43(3):1–43, 2011.
2. A. F. Bobick and A. D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1325–1337, 1997.
3. D. Damen and D. Hogg. Recognizing linked events: Searching the space of feasible explanations. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 927–934, 2009.
4. D. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
5. M. a. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban. Histogram of Oriented Displacements (HOD): Describing trajectories of human joints for action recognition. *International Joint Conference on Artificial Intelligence*, 25:1351–1357, 2013.
6. A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, pages 2012–2019, 2009.
7. M. E. Hussein, M. Torki, M. a. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. *International Joint Conference on Artificial Intelligence*, pages 2466–2472, 2013.
8. S. S. Intille and A. F. Bobick. A Framework for Recognizing Multi-agent Action from Visual Evidence. *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, (489):518–525, 1999.
9. J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):522–536, 1980.
10. C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila. Active pedestrian safety by automatic braking and evasive steering. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1292–1304, 2011.
11. H. Koppula, R. Gupta, and A. Saxena. Learning Human Activities and Object Affordances from {RGB-D} Videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.
12. S. Nirjon, C. Greenwood, C. Torres, S. Zhou, J. a. Stankovic, H. J. Yoon, H. K. Ra, C. Basaran, T. Park, and S. H. Son. Kintense: A robust, accurate, real-time and evolving

system for detecting aggressive actions from streaming 3D skeleton data. *2014 IEEE International Conference on Pervasive Computing and Communications, PerCom 2014*, pages 2–10, 2014.

13. W. Niu, J. Long, D. Han, and Y. F. Wang. Human activity detection and recognition for video surveillance. In *2004 IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXP (ICME), VOLS 1-3*, pages 719–722, 2004.
14. C. S. Pinhanez and A. F. Bobick. Human action detection using pnf propagation of temporal constraints. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998.*, pages 898–904. IEEE, 1998.
15. M. Popa, A. Kemal Koc, L. J. M. Rothkrantz, C. Shan, and P. Wiggers. Kinect sensing of shopping related actions. *Communications in Computer and Information Science*, 277 CCIS:91–100, 2012.
16. R. F. Rashid. Towards a system for the interpretation of moving light displays. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):574–581, 1980.
17. M. S. Ryoo and J. K. Aggarwal. Semantic Representation and Recognition of Continued andRecursive Human Activities. *International Journal of Computer Vision*, 82(1):1–24, 2009.
18. T. Starner, J. Weaver, and A. Pentland. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *Transactions on Pattern Analysis and Machine Intelligence*, 20(466):1371–1375, 1998.
19. C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014.
20. J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. *Computer Vision and Pattern Recognition*, pages 379–385, 1992.
21. E. Yu and J. K. Aggarwal. Detection of Fence Climbing from Monocular Video. *18th International Conference on Pattern Recognition (2006)*, 1:375–378, 2006.
22. F. Zhou, F. D. L. Torre, and J. Hodgins. Hierarchical Aligned Cluster Analysis (HACA) for Temporal Segmentation of Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):1–40, 2010.