



Human Activity Recognition in Video Sequences – A Survey

Kinjal V. Joshi

Computer Engineering Department
G H Patel College of Engineering and Technology,
Vallabh Vidyanagar, Anand, Gujarat 388120, INDIA.

Abstract: *Human Activity Recognition (HAR) is one of the active research areas in computer vision field because of its various applications like security surveillance, healthcare, human computer interaction, automatic video indexing and retrieval, and content based video retrieval. Recognizing human activities from video sequences or quiet images is a puzzling task due to problems, such as background disorder, partial occlusion, changes in scale, illumination, and appearance. In this paper a detailed review of recent research in the area of HAR is provided. The review covers various sensing technologies and methodologies used in HAR.*

Keywords: *human activity recognition, sensors, neural network, feature vector, support vector machine*

I. Introduction

Human Activity Recognition is active open research area in computer vision [1]–[3]. Human activity recognition is an ability to interpret human body gesture or motion via sensors and determine human activity or action [4]. HAR is used in surveillance systems installed at public places i.e. banks or airports [5], [6]. Ryoo [7] introduced a new paradigm of human activity prediction to prevent crimes and dangerous activities from occurring at public places. For prediction of next activity current activity must be recognized. HAR is widely employed in healthcare systems installed in residential environment, hospitals and rehabilitation centers. HAR is used for monitoring the activities of elderly people staying in rehabilitation centers for chronic disease management and disease prevention [8]. HAR is also integrated into smart homes for tracking the elderly people's daily activities [9], [10]. Besides, HAR is used to encourage physical exercises for patients. In the field of human computer interaction, HAR has been applied quite commonly in gaming. Through HAR, human body gestures are recognized to instruct the machine to complete dedicated tasks. In this paper, the literature related various sensors and methodologies used for human activity recognition is summarized. The rest of the paper is organized as follows. Section II contains information about different sensing technologies. Section III contains brief description about various methodologies used for human activity recognition and conclusion is presented in section IV.

II. Sensing Technologies

Generally, the sensors in a conventional HAR play an important role in recognizing human activity. The sensors capture the information acquired from human body gesture and the recognition engine analyses the information and determines the type of activity has been performed. Different sensing technologies are used in HAR. These technologies are classified as RGB camera-based, depth sensor-based and wearable-based [11]. The various research papers indicate both depth sensor and wearable sensor technologies are gaining more popularity in HAR research recently. On the other hand, RGB camera has obtained less emphasis in HAR research, because of its imitation in capturing the scene and human motions in 3D space [11]. The depth sensors have become popular due to low cost, high sample rate and capability of combining visual and depth information [11]. However some common visual based issue and challenges still persist for depth sensor such as occlusion and limitation of sensor viewpoint [11]. The wearable sensor systems can address occlusion and viewpoint limitation challenges [11]. The wearable sensors are well known being flexible in providing location independent and seamless human monitoring without affecting their daily lifestyle i.e. privacy issue [11]. It is quite cheap, compact and consumes low power. The main drawback against wearable sensor is the recognition accuracy. Usually a wearable-based HAR system requires the subject to wear or attached with multiple sensors on various body parts [11]. This is so troublesome, intrusive and inconvenient for the subjects [11]. The wearable-based HAR could not work effectively as there is a tendency where the human subject can forget to put on or displace the dedicated sensor [11].

III. HAR Methodologies

A. Adaboost Method

Paul Viola et al. [12] have used adaboost method for pedestrian detection by which walking activity can be recognized. This system works directly with images extracting short term patterns of motion, as well as appearance

information, to detect all instances of potential objects. There is no need of separate mechanisms for tracking, segmentation, alignment, and registration which each involves parameters and adjustment. One needs only select a feature set, a scale for the training data, and the scales used for detection. All remaining tuning and adjustment happens automatically during the training process.

Motion information is extracted from difference between pairs of images. Two filters are used in cascaded manner one is motion filter for motion information and the other is appearance filter for appearance information. Different filters are used for static and dynamic pedestrian detection. By setting proper threshold good accuracy can be achieved. This method takes about 0.25 seconds to detect all pedestrians in a 360 x 240 pixel image on a 2.8 GHz P4 processor.

B. ConfAdaBoost Algorithm

This algorithm is proposed by Attila Reiss et al [13].

It is tested on publicly available dataset of daily human activities. The dataset was recorded with 30 subjects, performing 6 different activities: walking, ascending stairs, descending stairs, sitting, standing and lying.

The sensor signals of both the accelerometer and the gyroscope are pre-processed, then sampled in fixed-width sliding windows of 2.56 seconds and 50% overlap. A feature vector is extracted from each window by computing variables in both time and frequency domain, on different components of the acceleration and angular velocity signals. The dataset consists of 561 features per activity instance. However, these features have different importance in the task of recognizing physical activities. Two subsets of the entire feature set were defined: 'Small' feature set: this set only uses features extracted from acceleration data. It includes the time domain features mean, standard deviation, and correlation; the frequency domain features energy, entropy, mean, and maximum frequency; and inclination. In total, this feature set contains 26 out of the original 561 features. 'Large' feature set: in addition to the 'small' feature set, this set also includes features extracted from gyroscope data. Moreover, some additional features are included such as skewness, kurtosis, and the energy in different frequency bands. This feature set contains 128 out of the original 561 features. These two feature subsets are compared to the entire set of features.

ConfAdaBoost.M1 (cf. Algorithm 1) is a confidence-based extension of the AdaBoost.M1 algorithm. It is a direct multiclass classification technique, using the information about how confident the weak learners are in the prediction of the instance's classification. Here the confidence information is used in both the training and prediction steps. The main idea of ConfAdaBoost.M1 can be described as follows. In the training part of the algorithm the weak learner's confidence of the classification is returned for each instance, and is then used to compute the new weight of that instance: the more confident the weak learner is in a correct classification or misclassification the more the weight will be reduced or increased, respectively. Moreover, the confidence values are used in the prediction part of the algorithm: the more confident the weak learner is in a new instance's prediction the more it counts in the output of the combined classifier. This method gives more accuracy i.e. 99.29% compared to simple adaboost method that is tested on small, large and full dataset.

C. Convolutional Neural Network and Recurrent Neural Network

This method is proposed by Alberto Montes et al [14].

Authors have proposed a simple pipeline to classify and temporally localize activities in untrimmed videos. As shown in figure 1 this system uses features from a 3D Convolutional Neural Network (C3D) as input to train a recurrent neural network (RNN) that learns to classify video clips of 16 frames. A network is designed to process a sequence of C3D-f6 features from a video, and returns a sequence of class probabilities for each 16-frames clip. They use LSTM layers, trained with dropout with probability $p = 0.5$ and a fully connected layer with a softmax activation. The system has the following architecture: input (4096) – dropout (0.5) – $N_lstm(c)$ – dropout(0.5) – softmax($K+1$) where K is the number of activity classes at the dataset.

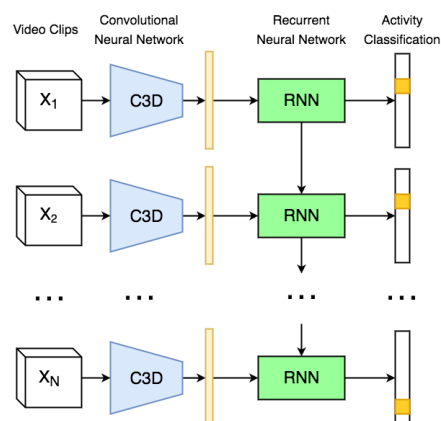


Fig.1 HAR using CNN and RNN [14]

Given a video, the prediction of proposed model is sequence of class probabilities for each 16-frame video clip. This output is post-processed to predict the activity class and temporally localize it. First, to obtain the activity prediction for the whole video, the average of the class probabilities over all video clips is computed. The class with maximum predicted probability is considered as the predicted class. To obtain the temporal localization of the predicted activity class, a mean filter of k samples is applied to the predicted sequence to smooth the values through time. Then, the probability of activity is predicted for each 16-frames clip, being the activity probability the sum of all probabilities of activity classes, and the no activity probability, the one assigned to the background class. Finally, only those clips with an activity probability over a threshold are kept and labeled with the previously predicted class.

This system achieves competitive results on both tasks. The sequence to sequence nature of the proposed network offers flexibility to extend it to face more challenging tasks in video processing, e.g. where more than a single activity is present in the video.

D. Support Vector Machine

Mitja Lustrek and Bostjan Kaluza [15] have used machine learning. In this approach, the attributes characterizing the user's behavior are selected. The considered attributes are the locations of body parts in the reference coordinate system (fixed with respect to the environment), the locations of body parts in a body coordinate system (affixed to the user's body) and the angles between adjacent body parts. Using these attributes, feature vector is generated and by training, data model is created and using testing data classification accuracy is measured i.e. 95%. Authors have tested various machine learning algorithms and results are compared. SVM gives highest accuracy.

IV. CONCLUSION

In this paper overview of various sensing technologies for acquiring data is represented. Using single sensor or combination of different sensors data can be acquired. Here various approaches for human activity recognition are denoted briefly. Still any system does not match the capability of the human vision faultlessly. The real time cost-effective and fully automatic Human Activity Recognition might have a long way from the final prototype yet.

REFERENCES

- [1] Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 93, no. 6, pp. 1445–1457, Jun. 2013.
- [2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Underst.*, vol. 115, no. 2, pp. 224–241, Feb. 2011.
- [3] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.
- [4] J. Yang, J. Lee, and J. Choi, "Activity Recognition Based on RFID Object Usage for Smart Mobile Devices," *J. Comput. Sci. Technol.*, vol. 26, no. 2, pp. 239–246, Mar. 2011.
- [5] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien, "Gait Recognition with Kinect," in *Workshop on Kinect in Pervasive Computing at Pervasive 2012*, 2012.
- [6] V. Tamas, "Human Behavior Recognition In Video Sequences," Technical University of CLUJ-NAPOCA, 2013.
- [7] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *2011 International Conference on Computer Vision*, 2011, no. Iccv, pp. 1036–1043.
- [8] O. Banos, M. Damas, H. Pomares, A. Prieto, and I. Rojas, "Daily living activity recognition based on statistical feature quality group selection," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 8013–8021, Jul. 2012.
- [9] L. Chen, C. D. Nugent, and H. Wang, "A Knowledge-Driven Approach to Activity Recognition in Smart Homes," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 961–974, Jun. 2012.
- [10] A. Jalal, Z. Uddin, J. T. Kim, and T. Kim, "Recognition of Human Home Activities via Depth Silhouettes and A Transformation for Smart Homes," no. February, pp. 467–475, 2011.
- [11] Ong Chin Ann and Lau Bee Theng, "Human Activity Recognition: A Review", IEEE International Conference on Control System, Computing and Engineering, 28 - 30 November 2014, Penang, Malaysia
- [12] Paul Viola, Michael Jones and Daniel Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance", Proceedings of the International Conference on Computer Vision (ICCV), October 13, 2003, Nice, France.
- [13] Attila Reiss, Gustaf Hendeby and Didier Stricker, "A Competitive Approach for Human Activity Recognition on Smartphones", European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013) proceedings. Bruges (Belgium) 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
- [14] Alberto Montes, Amaia Salvador, Santiago Pascual, Xavier Giro-i-Nieto, "Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks" 1st NIPS Workshop on Large Scale Computer Vision Systems (LSCVS 2016), Barcelona, Catalonia/Spain.
- [15] Mitja Lustrek and Bostjan Kaluza, "Fall Detection and Activity Recognition with Machine Learning" Informatica vol 33 No.3(2009) pp 205-212