Distinct Actions Classification Using Human Action Tracker Technique in Sports Videos

Kanimozhi S^{a,1}, Anbarasi S^b and Mythili M^a

^a*Department of Information Science and Technology, College of Engineering, Anna University, Chennai, India.*

^b*Department of Computer Science, Meenakshi College for Women, Chennai, India.*

Abstract. Recognizing human action in sports is difficult task as various sequences of activities involved in every scene. Identifying each action individually without overlapping of movements is a tedious process due to continuous change of frames within short duration. So proper tracking of human movements for each action is important. Hence new structure-based human action recognition and tracker technique (HART) is proposed. It uses joint trajectory images and visual feature to design each human action. At first, a structural based method employed to extract human skeleton data points from RGB (Red Green Blue) videos. Next, a Multitude Object Tracker (MOT) is proposed which uses the trajectory of human skeleton joints in an image space for identification of actions. Then, Histogram of Oriented Gradients (HOG) combined with Support Vector Machine (SVM) is applied to extract physical body shape and action information. Finally, the action label and interconnected keypoints in humans is jointly detected as end result. The proposed HART technique effectively performed well with the accuracy of about 82% over the other activity recognition methods.

Keywords. Histogram of Oriented Gradients, Support Vector Machine, Openpose, Human Action Recognition and Tracker (HART).

1. Introduction

Human activities have an inherent hierarchical structure, which can be considered as a three-level categorization. First, for the bottom level, there is an atomic element in which more complex human activities are constitutes as action primitives. After that, the detection of action/activity comes as second level. Finally the top level is complex interactions, which refers to the human activities that consist of involvement of more than two persons and objects. This three-level categorization varies and maintains a consistent theme. Action primitives are those atomic sports actions at the limb level, such as “stretching the left arm,” and “raising the right leg”. There are two challenges arising that make state-of-the-art deep network-based action recognition algorithms ineffective.

¹Kanimozhi S, Department of IST, College of Engineering, Anna University, Chennai.
E-mail: kanimozhiist21@gmail.com.

First problem is to extract boundary based shape features to overcome the occlusion and noise. Second, wide field-of-view (FoV) considering in the point of omnidirectional cameras results and not as multiview (actor style, clothing style, ethnicity) in the sports field. In a practical scenario where the camera is installed at a place with large pedestrian volume, the videos are likely to capture many people performing actions at the same time. Since it is computationally intensive to analyse each person individually, an efficient method is needed to simultaneously recognize actions performed by multiple people. Furthermore, from the perspective of curating training data, it is both expensive and more time-consuming to extensively annotate each person's position (i.e., bounding box) and action. On the other hand, it is much easier to acquire annotation only for the video-level action labels without linking each action to a specific person. The main contribution of the proposed HART model:

- i) Specify distinction among various object present in a frame,
- ii) Track the variations in targeted object individually and notify them,
- iii) Spot on labelling each sports action using target object motion pattern separately.

2. Related works

2.1. Vision based Human Activity Recognition

Set of actions that are performed by specific part of the human body is called as atomic actions [1]. Like “stretching the left arm,” and “raising the right leg” etc. are the actions that results due to the movements of hands, arms, or upper body part. An actions primitive [2,3] in terms of temporal sequence of a single person interaction (not more number of objects, humans) is considered for this approach. As it works based on space-time interest points (STIPs) [4] in which the variations among the image vectors is shown in terms of both time and space. Meanwhile, another method based on local descriptors such as histogram of oriented gradients (HOG) [5] and histogram of optical flow (HOF) is used for recognizing object. This method is widely used for 3D in HAR area.

2.2. Multi-Person Action Recognition

To monitor multi person actions the camera along with viewing angle plays an important role. When camera is placed at a maximum height (top view) omnidirectional cameras [6] can be more convenient while comparing with other multi conventional cameras. It provides high quality 4k videos at lower cost. Installation also simple as it needs an algorithm for entire working. This will avoids the synchronization problem, coordination among multiple conventional cameras, and reduces security risks of privacy attack. Still there are two challenges in recognizing of humans. First is to identify the appearance of distorted [7] humans. Next to recognize the occluded human's structure. To overcome these issues a deep network-based action recognition [8,9] method is proposed. In which an omni-directional video is transformed into a panoramic video where people stand upright. Another method which is similar to [8] called as multiple instances learning (MIL) [10] is mostly used for predicting drug activity. By incorporating deep network with this MIL, weakly supervised classification can be done in an effective manner. Extended version of this approach is used for multi-label learning, which can recognise multiple instances at a time.

2.3. Joint Trajectory Character Recognition for Human Action Recognition

Estimating the human posture based on visual features [11] is carried out by considering Red Green Blue images (RGB) and its depth (RGB-D). As it provide the information in terms of position, direction and scale information of different human parts in the images, which helps a lot for human action recognition. Skeleton structure based human pose estimation [12] is another method where joints are plotted as individual points and they are interconnected [13] along with its position. This process is done in every video frame which creates a trajectory. Finally, all the trajectory images are combinedly used for recognizing the actions. For lengthy video field coordinating the trajectory is again a big problem. So a Coordinate-based action [14] representation is used for different co-ordinate systems. It works on optical character recognition method like joining of hand-written characters [15], the joining trajectory is taken place. This transforms the action recognition problem into characters detection that being drawn from the skeleton joint trajectory. Finally, a graph theory [16] based approach is proposed, where common count of joint point is made over a 2D image. To meet different needs, Openpose [17] method is used as it provides different real-time detection points. By using this method 18 body joint point is generated on the Microsoft COCO data set. Using these joint points the trajectory is constructed and tracked to finalize the action performed in that video.

3. Proposed Human Action Recognition and Tracker (HART) model

Figure 1. Proposed architecture for HART model

In the proposed method (Figure 1), sports scene is given as input where frames are extracted and denoising is done for all segregated frames. Once the noise removal [18,13] is done, cleaned data is given as input for the VGG-16 model [19] in which filter is used for convolution [20] taken care by its hyper parameter. It has dual fully connected layer (FC) followed by a softmax layer results with a refined feature vector value for each class. The 16 layers in the VGG-16 model helps in identifying the various set of objects present in

each frame along with humans (as object) for further processing. Object tracking [21] based on not just distance, velocity but also what that person looks like. Multitude Object (MOT) tracker algorithm in table-1 allows us to add this feature by computing deep features for every bounding box and using the similarity between deep features also factor into the tracking logic.

Kalman filter [22] allows us to model tracking based on the position along with velocity of an object and predict where it is likely to be. It models future position and velocity using Gaussians [12]. When it receives a new reading, it can use probability to assign the measurement to its prediction and update itself. It is light in memory and fast to run. And since it uses both position and velocity of motion, it has better results than the centroid based tracking.

Algorithm 1 Multitude Object Tracker (MOT) algorithm.

Input: Sports video scene

Output: Feature value for each object

- 1: Use KNN to find the distance- both partial fit and distance
 - 2: Generate bounding box , confidence score, feature matrix
 - 3: Apply NMS for the detected confidence score less than 0.7
 - 4: Track the predicted object using mean() and covariance()
 - 5: Match the detected value with partial fit and distance in step1
 - 6: Update track with corresponding detection
 - 7: **if** track is mismatched|updated time is too long **then**
 - 8: Delete Track
 - 9: **for** unmatched detection and mismatch **do**
 - 10: Obtain latest track list, save only confirmed and tentative
 - 11: Update distance metrics and collect track ID of confirmed status
 - 12: Join and match track ID with features
 - 13: Update feature set value
 - 14: **Return** feature value
-

For each object present in the frame multiple predicted boxes is generated. So it is necessary to provide a bounding box that exactly fit in over the object. It can be done with the help of non-maxima suppression (NMS) techniques [23] which select box based on the two metrics. One is using the objectiveness score given by the model. Second is the comparison with the IoU (Intersection over Union) score of the box. Once these two processes completed removal of boxes is done with a threshold of IoU score value below 50% among overall score. At last the bounding box that satisfy the highest score value is taken in account. Next is to recognise the actions with the help of identified object along with the relationship among them. For that the OpenPose [14] technique is used. It is constructed based on two branch multi stage VGG-16 convolution neural network, in which top branch predicts the confidence maps of various locations of human body parts. Bottom branch is used to predict the Part Affinity Fields (PAF) of each level. During the first stage the initial set of confidence map (M) and affinity fields (A) is identified and in subsequent stages previous results from both the branches along with image feature (L) is get concatenated . In the part of set (affinity fields), correlation information of each point is described. Finally, the points are transformed into confidence maps and 18 two-dimensional body joints of all people in the image are obtained by reasoning and analy-

sis of Hungarian matching algorithm [24]. At first the skeleton trajectory is extracted by HOG (Histogram of Oriented Gradients), which is a histogram of directional gradients. It is used to extract human body shape and action information. The generation of HOG feature vectors is split into five steps. Initially, the image is normalized by colour and Gamma. Then the gradient of the image is calculated by first-order differential within the vertical or horizontal direction. Then, the direction weight projection supported the gradient magnitude is carried out, and therefore the weight of every cell's gradient direction is calculated by using the oblong HOG model. Gradient of the image is calculated by first-order differential in the vertical or horizontal direction. Then, the direction weight projection based on the gradient magnitude (Eq. (1)) is carried out. The weight of each cell's gradient direction (Eq. (2)) is calculated by using the rectangular HOG model [25].

$$G = \sqrt{G_u^2 + G_v^2} \quad (1)$$

$$\theta = \arctan \frac{G_v}{G_u} \quad (2)$$

Where G is the magnitude of the gradient along u and v point, is the direction of gradient along three channels. Finally, the gradient histogram of the cell is obtained and the L2-norm function is used to normalize the feature vectors. The final feature vectors of HOG containing the location information of many skeleton joints are obtained.

4. Experimental Results

UCF101 dataset which is a human action video data set used in our proposed work. It contains 101 action categories. For each different action category, videos are divided into 25 groups. Video clips in the same group may share some common features, similar backgrounds or similar perspectives. Because the number of recognizable videos provided by different action categories in UCF 101 data sets is different, the 4 action sets used in the experiment are filtered [26] according to the amount of data. For each action category, we selected 4 groups as the experimental data. During training phase, the detection function of 18 body joints based on UCF-101 data set provided by OpenPose is applied. OpenPose provides different real-time human body detection points in a text format. Using that text data, we first generate a CSV file using those points. CSV file contains the (x,y) co-ordinates of all 18 body parts and binary data of each class. Next step of action recognition is training the model using that CSV file. Trained data will be stored in the h5 file. Next is to label the data by identifying the connected components in an image and assigning each one a unique label. The resulting matrix is called a label matrix. Each video in the data set is belonging to a particular class. Class label is mapped to a specific class according to the binary data contained in the CSV file. In Figure 2 the output of the proposed system is shown for various actions like kick, run, bowl and jump. As video scenes in the dataset consist of single player performing various sports actions, the skeleton is generated focusing on individual person but in concern from various orientations.

Figure 2. Resulted recognised action labels from UCF-101 dataset

Finally, recognition is done based on the given input data and the trained model. Generating skeleton joint [17] position matrix and skeleton joint trajectory in the input frame and the order of frame sequence in the video is then applied into the trained model. Humans are mapped in a bounding box and the action label is placed on top of that. The performance evaluation for the proposed model is done using the confusion matrix where x –axis represent the predicted actions and y-axis is the true label shown in Figure 3. From the matrix it is evident that correctly prediction rate of HART model is about

Figure 3. Confusion matrix for the proposed model

96% for all the considered human actions.

The performance of the proposed system is compared with other state-of-art methods which clearly shown in Table 1. The HART model outperforms other methods on an average of about 4% higher than corresponding methods.

Table 1. Performance comparison with state-of-art method

Methods	Accuracy
SVM	89%
CNN	91.3%
ResNet	92.5%
Proposed HART	96%

5. Conclusion

Core motivation of the proposed work is to recognise and track human actions individually in sports videos. The VGG-16 model is used to extract the image feature vector. The HART model uses MOT algorithm in table-1, where every object present in the frame is highlighted. Then filtering is made so that only targeted object along with its actions is tracked individually. Human actions are divided into three levels including action primitives, actions/activities, and interactions. At last summarization of the classic and representative approaches to activity representation is done. The proposed model results with 96% of performance improvement in recognising and tracking of human actions. In future, the work is focused towards detection of multi target and movements in an entire sports video. This can be more helpful in preparing sports commentary.

References

- [1] Ullah A, Muhammad K, Del Ser J, Baik SW, de Albuquerque VH. Activity recognition using temporal optical flow convolutional features and multilayer LSTM. *IEEE Transactions on Industrial Electronics*. 2018 Nov 22;66(12):9692-702.
- [2] Zhang S, Wei Z, Nie J, Huang L, Wang S, Li Z. A review on human activity recognition using vision-based method. *Journal of healthcare engineering*. 2017 Jul 20;2017.
- [3] Li X. Influence of sandstorm on air pollution based on Gaussian mixture model and sports detection. *Arab J Geosci*. 2021;14(15):1-3.
- [4] Sagayam KM, Hemanth DJ. A probabilistic model for state sequence analysis in hidden Markov model for hand gesture recognition. *Computational Intelligence*. 2019 Feb;35(1):59-81.
- [5] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) 2005 Jun 20 (Vol. 1, pp. 886-893). IEEE.
- [6] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015* (pp. 4305-4314).
- [7] Bu X. Human Motion Gesture Recognition Algorithm in Video Based on Convolutional Neural Features of Training Images. *IEEE Access*. 2020 Aug 28;8:160025-39.
- [8] Huang Z, Huang L, Gong Y, Huang C, Wang X. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019* (pp. 6409-6418).
- [9] Nagrath P, Jain R, Madan A, Arora R, Kataria P, Hemanth J. SSDMNv2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. *Sustainable cities and society*. 2021 Mar 1;66:1-11.
- [10] Zucker JD, Chevaleyre Y. Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem (Doctoral dissertation, LIP6).
- [11] Khurana R, Kushwaha AK. Deep Learning Approaches for Human Activity Recognition in Video Surveillance-A Survey. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) 2018 Dec 15 (pp. 542-544). IEEE.
- [12] Singh A, Agarwal S, Nagrath P, Saxena A, Thakur N. Human pose estimation using convolutional neural networks. In 2019 Amity International Conference on Artificial Intelligence (AICAI) 2019 Feb 4 (pp. 946-952). IEEE.

- [13] Sharieff AH, Sabena S, Sathiyavathi V, SaiRamesh L. Intelligent framework for joint data hiding and compression using SMVQ and fast local image in-painting. *Int. J. Sci Technol. Res.* 2020;9(2):2267-71.
- [14] Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence.* 2019 Jul 17;43(1):172-86.
- [15] Sagayam KM, Hemanth DJ. ABC algorithm based optimization of 1-D hidden Markov model for hand gesture recognition applications. *Computers in Industry.* 2018 Aug 1;99:313-23.
- [16] Shen L, Yeung S, Hoffman J, Mori G, Fei-Fei L. Scaling human-object interaction recognition through zero-shot learning. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) 2018 Mar 12 (pp. 1568-1576). IEEE.
- [17] Liang X, Zhang HB, Zhang YX, Huang JL. JTCCR: Joint Trajectory Character Recognition for human action recognition. In 2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE) 2019 Oct 3 (pp. 350-353). IEEE.
- [18] Ren H, El-Khamy M, Lee J. Dn-resnet: Efficient deep residual network for image denoising. In Asian Conference on Computer Vision 2018 Dec 2 (pp. 215-230). Springer, Cham.
- [19] Li J, Liu J, Yongkang W, Nishimura S, Kankanhalli M. Weakly-Supervised Multi-Person Action Recognition in 360° Videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2020 (pp. 508-516).
- [20] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) 2017 Aug 21 (pp. 1-6). IEEE.
- [21] Salhi A, Ghazzi F, Fakhfakh A. Estimation for motion in tracking and detection objects with Kalman Filter. In Dynamic Data Assimilation-Beating the Uncertainties 2020 Oct 8. IntechOpen.
- [22] Vinodhini CA, Sabena S, Ramesh LS. A Robust and Fast Fundus Image Enhancement by Dehazing. In International Conference On Computational Vision and Bio Inspired Computing 2018 Nov 29 (pp. 1111-1119). Springer, Cham.
- [23] Shepley A, Falzon G, Kwan P. Confluence: A robust non-IoU alternative to non-maxima suppression in object detection. *arXiv preprint arXiv:2012.00257.* 2020 Dec 1.
- [24] Yang H, Yuan C, Xing J, Hu W. SCNN: Sequential convolutional neural network for human action recognition in videos. In 2017 IEEE International Conference on Image Processing (ICIP) 2017 Sep 17 (pp. 355-359). IEEE.
- [25] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 4694-4702).
- [26] Alippi C, Disabato S, Roveri M. Moving convolutional neural networks to embedded systems: the alexnet and VGG-16 case. In 2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN) 2018 Apr 11 (pp. 212-223). IEEE.