Human Action Recognition with Extremities as Semantic Posture Representation

Ram Krishna
2016csb1053
Department of Computer Science Engineering,
IIT Ropar

Shailendra Kumar Gupta
2016csb1059
Department of Computer Science Engineering,
IIT Ropar

1 Abstract

This project presents a Recurrent neural Network(RNN) methodology for Human Action Recognition using star skeleton as a representative descriptor of human posture. Star skeleton is a fast skeletonization technique by connecting from centroid of target object to contour extremes. To use star skeleton as feature for action recognition, we clearly define the feature as a five-dimensional vector in star fashion because the head and four limbs are usually local extremes of human shape. In our project we assumed an action is composed of a series of star skeletons over time. Therefore, time-sequential images expressing human action are transformed into a feature vector sequence. Then the feature vector sequence must be transformed into symbol sequence so that RNN can model the action. We used RNN because the features extracted are time dependent.

2 Introduction and Motivation

Recognizing Human actions in videos is becoming one of the popular areas of research in Computer Vision. So we tried to implement the same in this project. It is motivated by a great deal of applications, such as automated surveillance system, smart home application, video indexing and browsing, virtual reality, human-computer interface and analysis of sports events. Unlike gesture and sign language, there is no rigid syntax and well-defined structure that can be used for action recognition. This makes human action recognition a more challenging task. Several human action recognition methods were proposed in the past few years by many researchers. Most of the previous methods can be classified into three classes: model-based methods, eigenspace technique and Hidden Markov Model. The model-based method has high computation cost and the rate of recognition of eigenspace technique can be improved. Hence implementing Hidden Markov Model is better compared to other methods.

But as HMM became an old method to apply on such problems, so we implemented recurrent neural network(RNN) for training our model and predicting it over the test video. The main motive behind using RNN is to make use of sequential features stored from the star skeltonization.

In a traditional neural network we assume that all inputs (and outputs) are independent of each other. But for many tasks thatâĂŹs a very bad idea. If we have to predict the next word in a sentence we better know which words came before it. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a âĂIJmemoryâĂİ which captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to look back only a few steps in the sequence.

3 Related Works

Human Action Recognition is the most useful topic in computer vision. A lot of researchers have proposed several methods. In [2] they also used star skeleton method and after that they chose the feature points to analyze cyclic motion of the human and then they calculated the angle between extremal points with vertical and then took the average of the angle as a parameters for recognizing actions.

Another paper [4] described the HMM model to use as human action recognition. To recognize an action, the HMM which best matches the action is chosen. It achieves high recognition rate by using distinguishable feature, and requires low process time. HMM transforms the problem of action recognition into the problem of pattern recognition. They used HMM to recognize six different tennis strokes among three players.

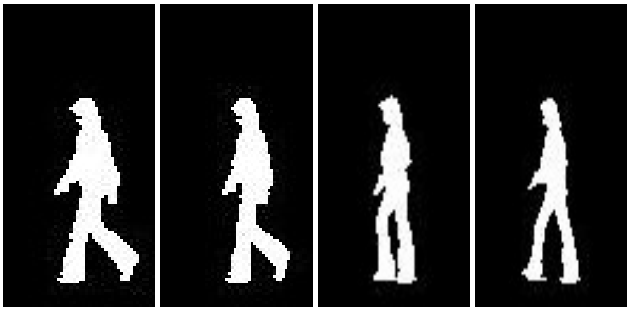

Figure 1: Human Silhouette.

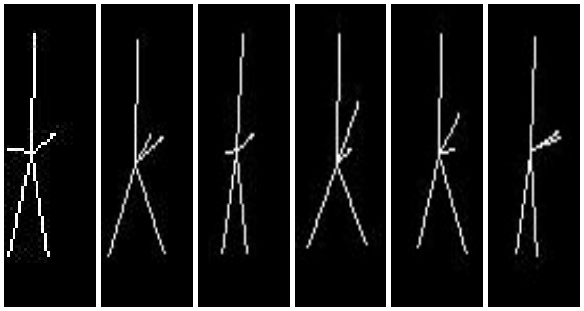

Figure 2: Star Skeleton of images.

Another paper [1] recognized the human action with the help of depth motion maps as features.It used a different approach for action recognition.

Now a days human action recognition is more or less a solved topic.

4 Methodology

After reading from different resources following are the different methods we have implemented in this project of Human Action Recognition. This section explains each of the step briefly. First it discusses about the method of extracting human body contour from a frame of the given video. Then it talks about the single star skeleton method to represent the human body extremities which will represent feature vectors for corresponding training examples while classification process. After that it discusses about the training model implemented for the classification of data.

4.1 Human Silhouette Extraction

The first target in this project is to extract human body contour from given image. To achieve so we first need to extract the human body from the given frame. We have two types of videos in our data-set. First set of videos are those in which the person stays at the same place and performs the action like bending, doing jacks, waving with single or double hands, etc(we call **inFrameVideos**). The second set includes those videos in which the human enters the video in one frame and exits from video in coming frames i.e. the whole human body is moving from one place to another(we call **outFrameVideos**). So according to the type of video we used two methods to extract human body from a frame of the video.

- **In Frame Videos**: For this type of videos we used direct differencing method to extract the human body i.e. we took direct difference between the background and the current frame to obtain the human body. We had one background image corresponding to each video. The method gave good results for human silhouette extraction.

- **Out Frame Videos**: For this type of videos we used the inbuilt Gaussian Mixture Model based Foreground detection method to extract the human body from frames of the videos.

4.2 Human Contour Extraction

Now we will use the extracted human silhouette to contruct the human contour. First we have applied used morphological operations to fill the gaps if exists in the human silhouette. After that we applied Canny edge detector to obtain the edge image of the human silhouette. Edge detector gives us the border of the human body in the frame. Next using this edge image we have extracted the contour points of the border of the human body in the frame. For this task we have an inbuilt function of MATLAB which takes a point of the border of an object as input and returns the collection of boundary points of the object in cyclic order. This collection of boundary points of the human body is called contour points.

4.3 Star Skeletonization

The concept of star skeleton is to connect from centroid to gross extremities of a human contour. To find the gross extremities of human contour, the distances from the centroid to each border point are processed in a clockwise or counter-clockwise order. Extremities can be located in representative local maximum of the distance function. Since noise increases the difficulty of locating gross extremes, the distance signal must be smoothed by using smoothing filter or low pass filter in the frequency domain. Local maximum are detected by finding zero-crossings of the smoothed difference function. The star skeleton is constructed by connecting these points to the target centroid.

4.3.1 Star skeleton Algorithm

Below are the brief steps of star skeleton method. The algorithm takes human contour as input and produces a skeleton in star fashion as the output of the algorithm.

1. Determine the centroid of the target image border.
2. Calculate the distances from the centroid to each border point.
3. Smooth the distance signal for noise reduction by using linear smoothing filter or low pass filter in the frequency domain.
4. Take local maximum of the distance signal as extremal points, and construct the star skeleton by connecting them to the centroid. Local maximum are detected by finding zero-crossings of the difference function.

As a feature, the dimension of the star skeleton must be fixed. The feature vector is then defined as a five dimensional vectors from centroid to shape extremes because head, two hands, two legs are usually local maximum. The final cut-off frequency of star skeleton is determined automatically. The cut-off frequency is first set to a higher frequency, and gradually decreases until the dimension of star skeleton is within five. For postures with more than five contour extremes, we adjust the low pass filter to lower the dimension of star skeleton to five. On the other hand, zero vectors are added for postures with less than five extremes.

Since the used feature is vector, its absolute value varies for people with different size and shape, normalization must be made to get relative distribution of the feature vector. This can be achieved by dividing vectors on x-coordinate by human width, vectors on y-coordinate by human height.

4.4 Feature Extraction

Now these star skeleton of each videos will be used as its features for training the model. There will good number of frames for each videos, therefore it will be difficult to use the whole skeleton image of each frame as the feature of a single video. So somehow we will reduce the number of features representing a video. Now as discussed we have got five points for each frame which will be used as features for that frame. Directly we cannot use these points as the features so what we did is, we calculated the angle of line joining the points from the centroid and used these angles as the features for that frame. Also we have used the distance of the centroid from one of the vertical boundary of the image to keep track of the position of human body from center of the image, which will help us to differentiate from pjump action(in place jump) and jump action(jumping from end of the image to other end of the image). In this way every frame is represented with 6 dimensional vector out of which 5 represents the angle of each point and 6th represents the distance of human body from the center of image.

4.5 Training the Model using RNN

A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. Therefore we preffered to use recurrent neural network for our classification problem. We used tensorflow framework for rnn networks.

5 Experimental Settings

The dataset for this project has been taken from the below sources:

- Action as Space-Time Shapes by M. Blank et at.

  (`http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`).

Our dataset consists of 10 different types of actions. They are

- Bend.
- Jack.
- Jump in place.
- Waving with one hand.
- Waving with two hand.
- Run.
- Walk.
- Side Run.
- Skip.
- Jumping from one end of image to another.

It contains approximately 90 videos. Each action has 9 videos. We have taken 7 videos for training and 2 videos from each action as testing. We trained the features extracted from star skeleton with RNN model.

6 Results and Discussion

We used different methods for silhouette extraction. First method uses the thresholding on the saturation part of the HSV image. The second method is implemented from the paper by Jong-Wook et. al.[5]. Next we used direct background subtraction which gave the better results compared to others shown in figures.

We trained the model with RNN. Our dataset has only 7 videos per action for training the model. The F-measure on training and test data achieved by our model is 62.85% and 47.36% for 250 epochs. As the number of epochs are increased the model tries to overfit the training data hence the accuracy on testing data reduces. The figure shows the confusion matrix for the rnn model trained. The model confused more to predict run as this action is similar to skip and jump action. Also pjump action was confusing to predict as its star skeleton is similar to jack action.

Training Data Confusion Matrix

Pred-> / True	BEND	JACK	PJUMP	WAVE1	WAVE2	RUN	WALK	SIDE	SKIP	JUMP
BEND	6	1	0	0	0	0	0	0	0	0
JACK	0	6	0	1	0	0	0	0	0	0
PJUMP	2	2	2	1	0	0	0	0	0	0
WAVE1	0	0	1	6	0	0	0	0	0	0
WAVE2	0	0	0	1	6	0	0	0	0	0
RUN	0	0	0	0	0	2	0	0	3	2
WALK	0	0	0	0	0	0	4	2	0	1
SIDE	0	0	0	0	0	0	0	3	4	0
SKIP	0	0	0	0	0	3	0	0	3	1
JUMP	0	0	0	0	0	1	0	0	0	6

Figure 3: Confusion matrix for Training data.

Testing Data Confusion Matrix

Pred-> / True	BEND	JACK	PJUMP	WAVE1	WAVE2	RUN	WALK	SIDE	SKIP	JUMP
BEND	1	1	0	0	0	0	0	0	0	0
JACK	0	2	0	0	0	0	0	0	0	0
PJUMP	0	0	0	0	1	0	1	0	0	0
WAVE1	1	0	0	1	0	0	0	0	0	0
WAVE2	0	1	0	0	1	0	0	0	0	0
RUN	0	0	0	0	0	0	0	0	2	0
WALK	0	0	0	0	0	0	1	0	1	0
SIDE	0	0	0	0	0	0	0	2	0	0
SKIP	0	0	0	0	0	0	0	0	1	1
JUMP	1	0	0	0	0	0	0	0	0	0

Figure 4: Confusion matrix for Test data.

7 Challenges

It does not clearly recognize between two similar actions like side run and run. The basic difficulties found while doing experimentation are

- Finding human extremities due to the articulated human body, self occlusion, ambiguity from the absence of depth information, appearance variations caused by camera viewpoints, illumination and loose clothing.
- How better human silhouette is extracted as action will depend on human extremities which in turn will depend on how good is the human silhouette.
- It mostly works on uniform background.Our method does not work for non uniform background.

8 Extra Works

Apart from Human Action Recognition, we also used the star skeleton method to classify the different Yoga poses. The results for the yoga poses detection were not that much satisfactory as the star skeleton made for poses having similar hands and legs orientation like for standing pose and upside down pose star skeleton is almost similar to each other hence was difficult to distinguish between them. however the star skeleton method performed well for the poses having very dissimilar hands and legs orientation.

9 Summary

The star skeleton model is good for recognizing activities that are different from each other and there exists horizontal motion of human. It also gives good result for vertical motion but it fails or does not give satisfactory result when length of feature vector is less than 5.

One limitation is that the recognition is greatly affected by the extracted human silhouette. We used a uniform background to make the foreground segmentation easy in our experiments. To build a robust system, a strong mechanism of extracting correct foreground object contour must be developed. Also, the viewing direction is somewhat fixed. In real world, the view direction varied for different locations of the cameras. The implemented method should be improved because the human shape and extracted skeleton would change from different views.

10 References

[1] Nasser Kehtarnavaz Chen Chen, Kui Liu. Real-time human action recognition based on depth motion maps.

[2] A. J. Lipton. H. Fujiyoshi. Real-time human motion analysis by image skeletonization., 1998. IEEE Workshop on Applications of Computer Vision.

[3] Y.-W. Chen-S.-Y. Lee H.-S. Chen, H.-T. Chen. Human action recognition using star skeleton, 2006. In International Workshop on Visual Surveillance Sensor Networks.

[4] K. Ishii J. Yamato, J. Ohya. Recognizing human action in time-sequential images using hidden markov model. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.

[5] Jong-Wook Han Ki-Young Moon Jang-Hee Yoo, Doosung Hwang. Extracting human body based on background estimation in modified hls color space, 2009. International Journal of Electrical and Computer Engineering.

[6] M. A. Jack X. D. Huang, Y. Ariki. Hidden markov models for speech recognition, 1990. Edingurgh Univ. Press.