Activity Classification using MHI

Parin Patel - ppatel480@gatech.edu

Georgia Institute of Technology, United States

Abstract

As the focus of computer vision is shifting from images to sequence of images, one of the tasks is to determine the activity an object is performing in the video (sequence of images). This report focuses on exploring the use of Motion History Images and Hu moments to classify different actions on KTH actions dataset[1]. By using very simple machine learning algorithm with hu moments as the features, it was possible to obtain 89.1% validation accuracy.

Keywords: Motion Energy Images, Motion History Images, Hu Moments, KNN

1. Introduction

Action recognition in videos have applications in different areas including video surveillance, human-computer interaction and video retrieval. There have been a lot of related research in recognizing human actions, reconstruction of human form from videos/image sequences using view-based methods. On an abstract level, action/activity recognition methods are based on the idea that the video is a sequence of images and thus a function of space (x,y) and time (t) $[I(x, y, t)]$.

Motion History Images based approach uses a temporal template to capture the motion itself. It is an approach for recognition and representation of action based on views, which is designed to support the direct recognition of the motion. The temporal template generated using MHI based algorithm represents both the at which portion the image is having motion as well as how much it has changed in intensity over time.

The two key challenges in using this approach are object detection and occlusion. As prerequisite to recognizing the motion/action in the image is to detect the object itself, object detection is an significant aspect of the problem. Another issue occurs if the motion or object is occluded during the video as it makes it difficult to identify the object and the correct motion representation.

2. Related Work

There are various approaches to activity recognition other than the use of temporal templates, like using CNN to get dynamic images from pooling layer[2] or using different methods of classification like Linear Discriminant Analysis, Quadratic Discriminant Analysis and so on. There are mainly two groups of approaches - vision-based and sensor-based. Dohnàlek et al.[3] describes different approaches in the paper. A recent paper has used MHI with Hidden Markov Model for prediction.[4] Another paper uses latest deep learning techniques like transfer learning combined with MHI and spatial information of the scene to predict activities[5].

The work based on which this project is created, is that of Davis et al.[6] This uses creation of Motion History Images and Motion Energy Images as temporal templates to recognize actions.

Another work in the same field as opposed to using temporal templates, Schuldt et al[1] uses space time features and a support vector machine as the classifier. The space-time features are basically space features of interest which occur at different times resulting in a pattern over space-time. It uses image gradients and Gaussian Convolution Kernel to determine points of interests as the local maximas.

3. Implementation

The generalized algorithm implemented in this project is as follows:

1. Create MEI: Create a binary image (motion energy image) by subtracting background image to detect the object.
2. Create MHI: Create a motion history image based on different size of sequences for recognizing different actions.
3. Create Moments: Create image moments, central moments, scale invariant moments and Hu moments for MHIs.
4. Append the activity class label to the data for each MHI to generate the classification data.
5. Train K-nearest neighbors classifier.
6. Get validation accuracy and confusion matrix.
7. Test the classifier using various videos.

3.1. Temporal Templates

As described in Davis et al. [6], creating temporal template using MHI is a two step method, in which first it creates binary motion-energy image(MEI), which is a representation of where the motion occured in an image sequence. The next step is to generate motion history image(MHI), which is a scalar-valued image where intensity is a function of recency.[6]

Motion Energy Images are binary cumulative motion images. Binary image sequence, which represents portions of images moved can be obtained using subtracting the previous image from the current image, in turn eliminating background in the process and the shape of moved object will become prominent in the foreground of that binary image. Let the binary image sequence be represented by $D(x, y, t) = I(x, y, t) - I(x, y, t-1)$, where $I(x, y, t)$ represents the image at time t. Motion Energy Image is then defined as

$$E_\tau(x, y, t) = \cup_{i=0}^{\tau-1} D(x, y, t-i) \tag{1}$$

MEI represents where the motion is in the image, whereas MHI represents how motion is in the image sequence. In MHI H_τ, pixel intensity is a function of temporal history of motion at that point. Motion History Image is defined as follows in terms of a decay operator: [6]

$$H_\tau = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max(0,\ \mathrm{H}_\tau(x, y, t-1)) & \text{otherwise} \end{cases}$$

In this case, for each action tau is different and is as shown in the following table:

Action	Tau
Walking	7
Running	11
Handclapping	13
Handwaving	17
Jogging	21
Walking	35

3.2. Hu Moments

In order to quantify the details of MHI and MEIs, moments like Image moments, unscaled and scaled central moments and second as well as third degree moments were used. Moments are defined based on density distribution functions. Regular Image moments are defined as $M_{ij} = \sum_x \sum_y x^i y^j I(x, y)$. Now as defined in the problem description, $\bar{x} = \frac{M_{10}}{M_{00}}$ and $\bar{y} = \frac{M_{01}}{M_{00}}$.

The unscaled central moments are defined as

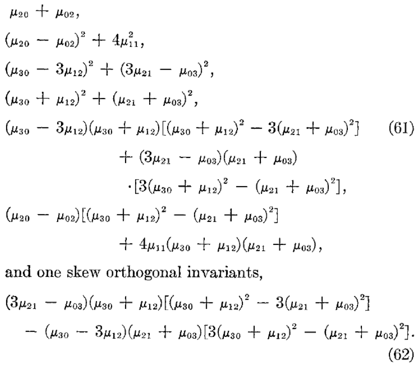
$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y) \tag{2}$$

The scaled central moments are defined as

$$\upsilon_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(1+\frac{p+q}{2})}} \tag{3}$$

The 7 other second and third degree moments are defined in the following figure. 6 of them are absolute orthogonal invariants and one is skew orthogonal invariants as described in Figure 1.

Figure 1: Second and third order moments[7]



$$\mu_{20} + \mu_{02},$$

$$(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2,$$

$$(\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2,$$

$$(\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2,$$

$$\begin{aligned}(\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\ + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\ \cdot [3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2],\end{aligned} \tag{61}$$

$$\begin{aligned}(\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\ + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}),\end{aligned}$$

and one skew orthogonal invariants,

$$\begin{aligned}(3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\ - (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2].\end{aligned} \tag{62}$$

3.3. Machine Learning Model

Here, K-nearest neighbor model is used for classification. The attributes for classification are Hu moments, scaled and unscaled central moments.Three different prediction techniques were implemented:

1. For each Tau in the table if any action gets predicted, return the action.
2. Take sequence length corresponding to each tau, and predict action for all the sub-sequences in the video. For each frame in a sub sequence the prediction will be the same. So, there will be 6 predictions per frame and the final prediction for each frame will be the prediction class which appeared maximum number of times.
3. Do the entire part 2, with the assumption that the entire video has one action. So, out of all the predictions for frames, the action which appeared the most will be the video prediction.

For single action detection, option 3 worked the best. And for more than one action prediction option 2 worked better.

4. Experiment Results

4.1. MEI and MHI

As described in section 3.1, different actions will have different MHIs and MEIs associated with it. Preprocessing like grayscaling, Gaussian Blurring and morphing was done on the images before generating MEI and MHI. Here are some example images from the KTH dataset of MHIs. Figure 2 shows details of MEI, MHI and real images for each iteration of Tau in the algorithm. The Last image in the third row of images is the final MHI of 7 frames of boxing.

Figure 3 shows Motion History Images for all 6 actions. As you can observe from the images, for older times the images are darker and newer images are brighter, which shows that it also records the recency of images.

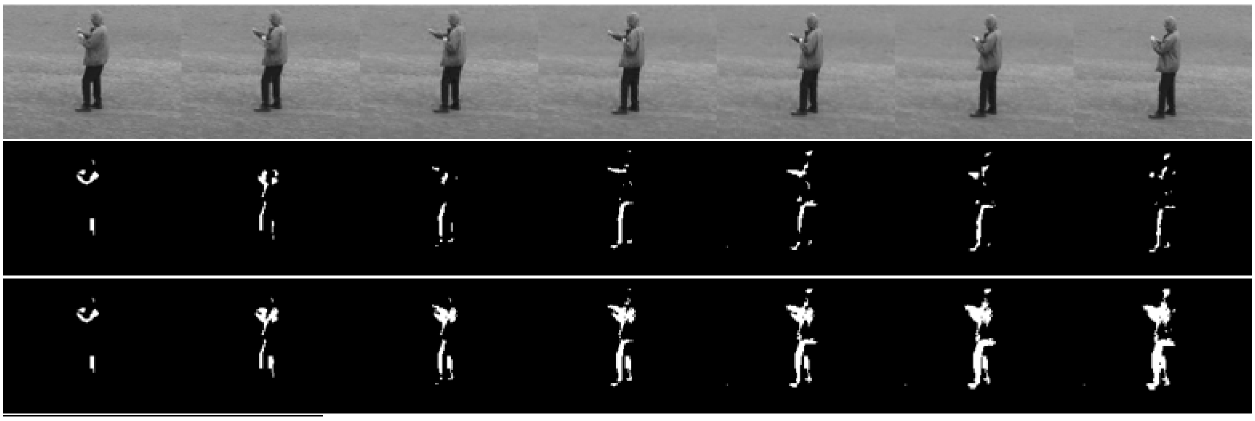
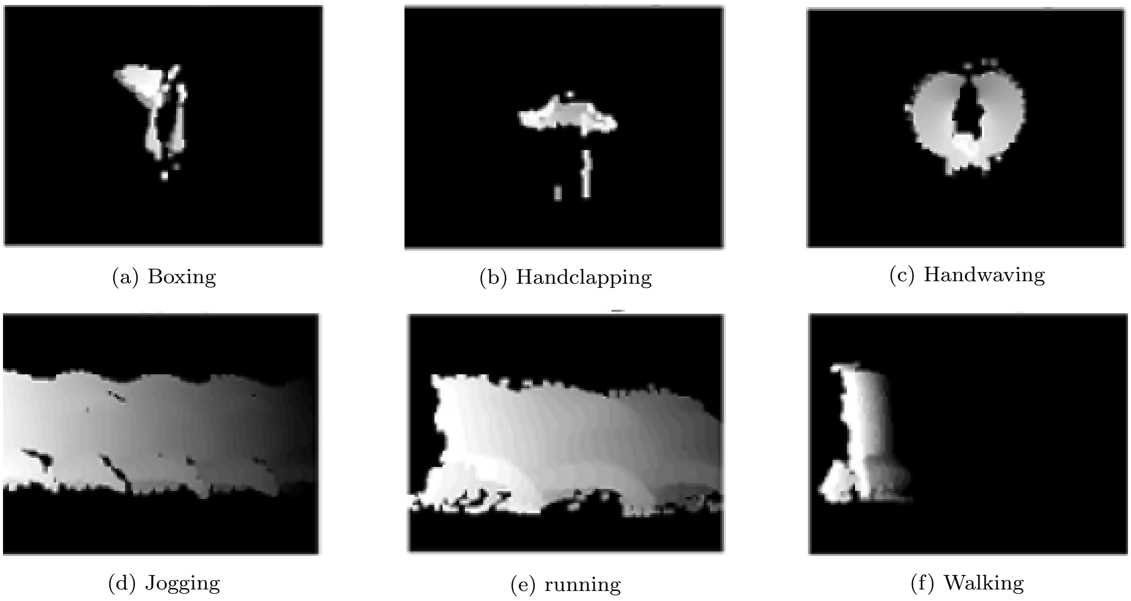


Figure 2: Real image, MEI and MHI respectively for boxing with tau=7



(a) Boxing

(b) Handclapping

(c) Handwaving

(d) Jogging

(e) running

(f) Walking

Figure 3: Motion History Images of all 6 actions

4.2. Predictions

For verification of all the predictions, I have used 80% of the dataset for training and 20% of the dataset for testing. The videos were chosen randomly in the train-test split. For understanding what are the accuracies for different actions Figure 4 shows the confusion matrix for the testing dataset. Please note that the accuracy is described with 6 orthogonal second and third order moments along with scaled and unscaled central moments as input features.

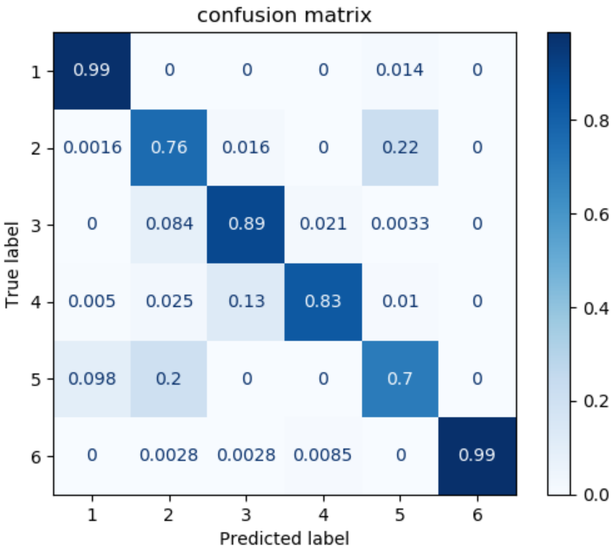
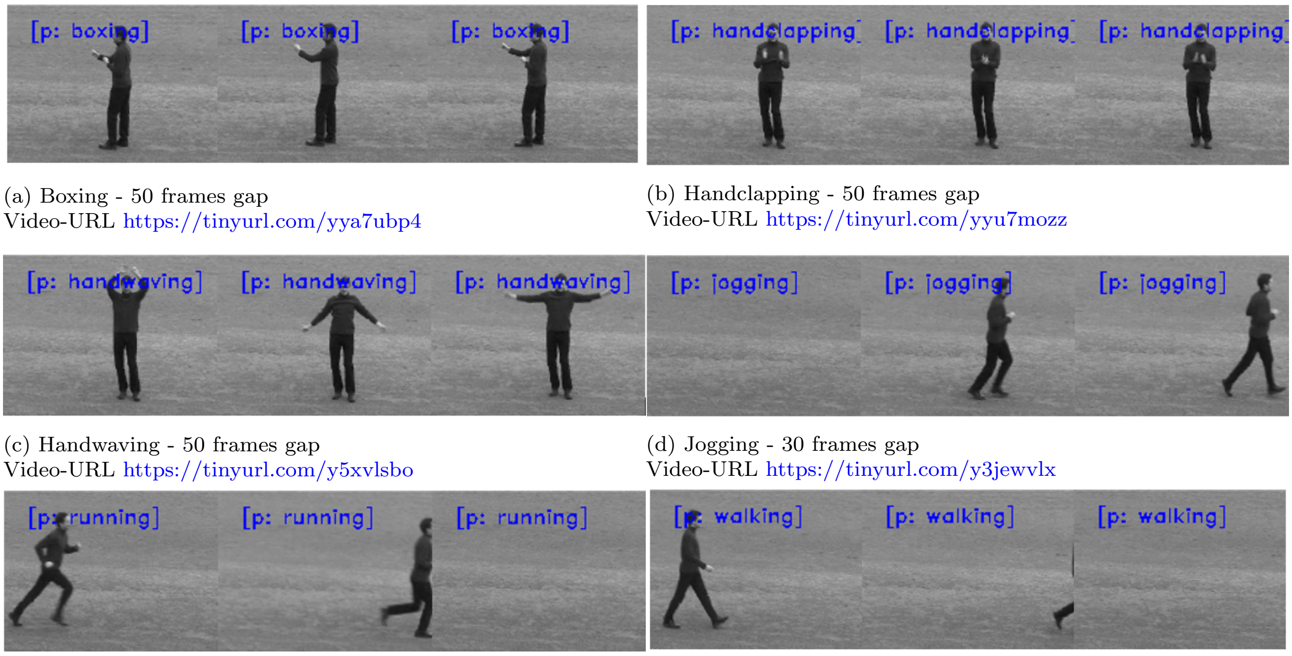


Figure 4: Confusion Matrix

Here labels 1 through 6 are boxing, handclapping, handwaving, jogging, running and walking respectively. The average accuracy achieved was 89.1%. As it is evident from the confusion matrix, the prediction accuracy for walking and boxing are better then other actions. From observing the MHIs, boxing, handclapping and handwaving have more similar MHIs and jogging, running and walking have similar MHIs. So if the actions are divided in two portions - related to boxing and related to walking, the distributions looks quite similar in the confusion matrix. Thus, we can derive that there are some images where the classifier is not able to distinguish prominently between the fine actions, but is able to distinguish between generalized actions.

Figure 6 below shows some of the predicted frames and links to URL of video predictions of single action.



(a) Boxing - 50 frames gap
Video-URL https://tinyurl.com/yya7ubp4

(b) Handclapping - 50 frames gap
Video-URL https://tinyurl.com/yyu7mozz

(c) Handwaving - 50 frames gap
Video-URL https://tinyurl.com/y5xvlsbo

(d) Jogging - 30 frames gap
Video-URL https://tinyurl.com/y3jewvlx

(e) running - 15 frames gap
Video-URL https://tinyurl.com/y6eenwla

(f) Walking - 50 frames gap
Video-URL https://tinyurl.com/y3vgt4gl

Figure 5: Sample frames from predicted outputs

5. Improvements

5.1. what went wrong

In some of the videos it was not able to predict the correct action. The algorithm mainly was giving wrong predictions between fine action differences like jogging and running or running and walking in some cases. Figure 7 is a screenshot of one such wrong classification.

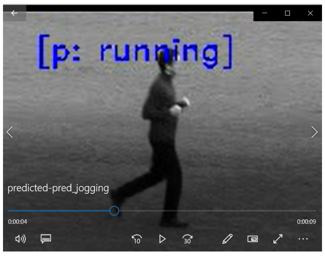


Figure 6: Predicted jogging as running

5.2. what can be improved

One of the reasons for wrong prediction in this video was that the person was jogging fast and thus it would have almost overlapped with tau for running and that is why resulting into the wrong prediction. Following are some more suggestions on how the algorithm can be improved:

1. It would be better if an algorithm was used to determine the best Tau for each action.
2. Filter sizes for image processing were chosen manually as well, it could be chosen according to the image frame size instead.
3. Use of Hidden Markov Model would have been better to predict future frames.
4. Using dynamic background subtraction techniques or using median filtering would have increased the scene understanding.
5. More advanced techniques like usage of bounding boxes, trajectory prediction algorithms or using stick figure representations would have been helpful.

References

[1] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proc. Int. Conf. Pattern Recognition (ICPR'04), Cambridge, U.K.

[2] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, S. Gould, Dynamic image networks for action recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3034–3042.

[3] P. Dohnálek, P. Gajdos, T. Peterek, V. Snáel, An overview of classification techniques for human activity recognition.

[4] E. C. Alp, H. Y. Keles, Action recognition using mhi based hu moments with hmms, in: IEEE EUROCON 2017 -17th International Conference on Smart Technologies, pp. 212–216.

[5] S. Zebhi, Human activity recognition by using mhis of frame sequences, TURKISH JOURNAL OF ELECTRICAL ENGINEERING COMPUTER SCIENCES 28 (2020) 1716–1730.

[6] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 257–267.

[7] Ming-Kuei Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory 8 (1962) 179–187.