Activity Classification using MHI

Georgia Institute of Technology

Written by

Zhi Zhang

December 2, 2017

1 Introduction

At a high-level, the activity recognition methods has existed over many years, it uses the idea that the video as a sequence of frames captured over time, and the image data is a function of space $I(x, y, t)$, so the video can be thought as an 'image stack'.

MHI based approach is to use a static image or temporal template where the value at each point in the image is a function of the motion properties at the corresponding spatial location in the sequence of frames. The static image can describe where the motion present and also how the motion is moving. The key challenge with using this temporal template is on object detection, the goal of activity recognition is to recognize the activity of the objects, so finding the objects are important, however, the temporal template can fail when two people are in the field of view and even worse when one occludes another. Also if the motion part of the body is not specified can cause many variability. So it needs to mask away regions of this type of motion.

The 3D extension of the 2D MHI approach which used the Motion History Volume (MHV) [1] is developed on visual hull for viewpoint-independent action recognition, and it's view-invariant MHI representation [2]. In the MHV approach, image pixels are replaced with voxels, and the standard image differencing function $D(x, y, t)$ is substituted with the space occupancy function $D(x, y, z, t)$.

The dynamic image based approach is to use dynamic image which is obtained by using the rank pooling CNN layer on the raw image pixels of video [3]. It utilizes the CNN idea that deep neural networks allow massive information and long term dynamics patterns from images to be represented.

2 Methods

2.1 Initial Investigation

My initial investigation focused on applying the existed activity recognition approach, which used the approach published in [4]. This phase was an experiment to determine which condition gave the best results. My goal was to quickly find a single, high-performing algorithm that could then be optimized. I used the MHI static image approach instead of the 3D view-invariant or the dynamic pattern because my training data has only one person and action is not complex, also due to its simplicity and ease to implement with good performance.

2.2 Data Handling

For this project, I had available 6 actions video data. These data were supplied as six $10 - 30$ seconds video. For each action, I obtained three section of frame sequences that covered the whole range of the action. Then each section of frame sequences is used for future processing.

2.3 Create binary images: Background subtraction

In order to recognize the activity, the object must be detected. The object is the person, so the background must be subtracted $(I(x, y, t) - B(x, y, z))$. The background is estimated to be as the previous frame. The theta is a threshold used for the subtraction. The theta is then used a parameter that I will tune later. The higher value of theta would result in more difference between the frame and the background. The background can also be the mean of the previous n frames [5]. Compared using the adjacent frame difference and average frame difference, and the later is better. And there is also the mix of gaussians to be used to model the background [6].

2.4 Create temporal-templates

In the next step I created static image that can represent the frame sequence pattern. The motion energy images and motion history images are used. Motion energy images represents the spatial accumulation of motion, and collapse over specific time window. The motion history images are a different function of temporal volume, and it's created by

$$\begin{aligned} I_\tau(x,y,t) &= \tau \quad if\,moving \\ I_\tau(x,y,t) &= max(I_\tau(x,y,t-1)-1,0) \quad Otherwise \end{aligned} \tag{1}$$

Then the temporal-templates are equal to the $MEI + MHI$

2.5 Calculate the image moments

The next step I calculated the image moments which summarize the shape of a given image $I(x,y)$, the Hu moments are set of 8 moments that applied to motion history image for global space-time shape descriptor, and it's translation and rotation and scale invariant.

2.6 Build a classifier

I then used the K nearest neighbors algorithm to build models of each class of action, the training data is the hu moments for each sample, I used the leave-one-out cross-validation which is essentially an estimate of the generalization performance of a model trained on n−1 samples of data, and is generally a slightly pessimistic estimate of the performance of a model trained on n samples.

In detail, I fit the model to all of the data, I iterate all of the samples to train the model, with one sample out to used as test, and used the trained classifier to predicate the label in the test, built the confusion matrices based on the predicated result, which is the Figure 5.

2.7 Action recognition

I used the classifier built to test a multi-actions video which contains the actions recorded by myself, the classifier was able to recognize most of the actions in the video, result is presented in the next session.

3 Results

3.1 The binary images from training

Below figures show that binary images obtained from six actions, and each action has three sequence frames, the original frames for each one was also provided as for comparison.

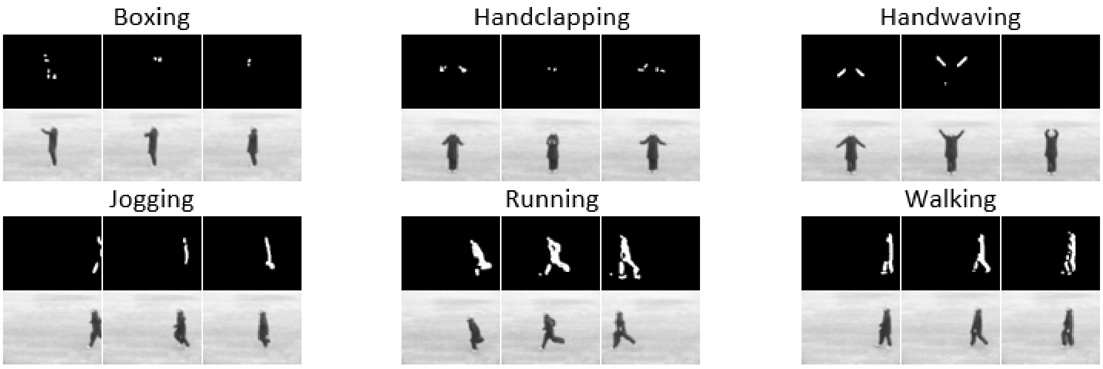

Figure 1: Binary images from different actions in the training.

The above binary images described each action, and the background was subtracted by using the frame differencing. The binary images were able to capture the interested action scene and removed the background.

3.2 The MHI from training

Below figures are the motion history images from each action,

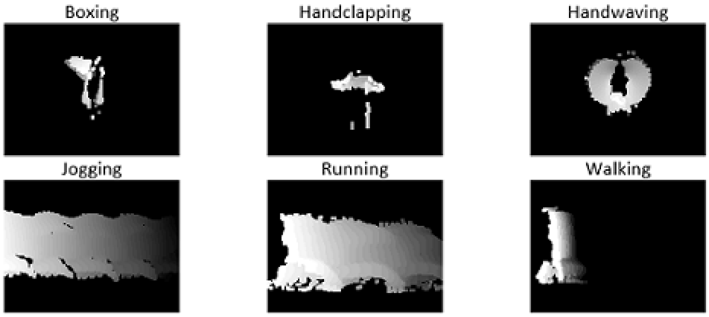

Figure 2: Motion History Image from different actions in the training.

The motion history images for each action represent the motion decay, each action has different time window length, the motion history image can reflect this difference. The running had more decay than the jogging than the walking, and the handwaving had more decay than the handclapping and boxing.

3.3 Test real video

I have recorded a multiple actions video and used the trained classifier obtained to recognize the action. There are total six different actions in the video, including boxing, handwaving, handclapping, walking, running, and jogging. The video was loaded and converted to binary value indicating the presence of motion, and then applied with motion history images to indicate the recency of motion in a sequence. Since the actions are performed at varying speeds than the training, I need to choose the right tau for the computation of the MEI and the MHI. Below is motion history images obtained.

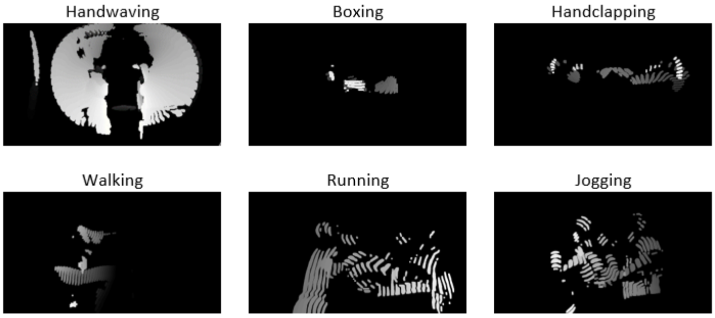

Figure 3: Motion History Image from different actions of the tested video.

I then computed the Hu moments on each action for each image. Then the pretrained KNN classifier was applied to the Hu moments. Since the KNN checks the Euclidean distance of the MHI and MEI parameters against the known action, the action found to be with the smallest distance would be the predicated action. Below are the figures of sample frames including their action label.

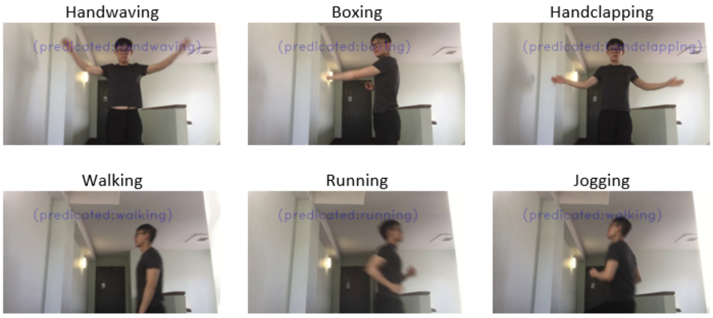

Figure 4: Recognized Actions from different actions of the tested self-recorded video.

A presentation result video was also provided at `https://youtu.be/uzkIGALgEQ0`

3.4 Performance statistics analysis-Confusion Matrices

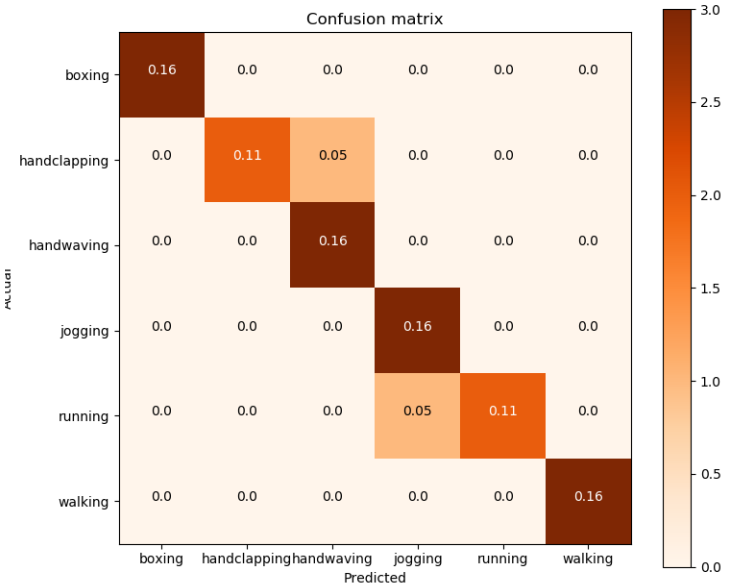

Figure 5: Confusion matrices in the training.

The above plot is the confusion matrices obtained from the classifier, there are six actions ('boxing', 'handclapping', 'handwaving', 'jogging', 'running', 'walking') used in the training, the y-axis is the true label, and the x-axis is the predicated label. The color bar on the right indicates the strength of actual label is equal to the predicated label.

The training data has six actions, and for each action, I extracted three different frames which covered the whole action the person performed, that's why the color bar tick ranges from 0.0 to 3.0. 3.0 means all of three predictions are correct, and 0.0 means none of them are correct. The values in the matrices indicate the experiment discrete probability distribution value, the sum of all the values in the matrices is equal to 1.0. There are total 6 actions, 18 samples or experiments, so each action can have maximum value as 1/6 which is 0.16, minmum value as 0.0.

From the Figure 5, the overall predication accuracy for the training data is about 88.9%, there are 2 wrong predictions, as you can see, there is one 'handclapping' but predicated as 'handwaving', there is one 'running' but predicated as 'jogging'.

4 Discussion

4.1 Analysis on why methods work on some images and not on others

My approach was able to perform well overall, it can recognize most of the actions and can achieve 88.9% accuracy in the training period using the leave-one-out cross validation. The overall model is robust with every action been correctly predicated generally, and it was able to differentiate the actions in the multi-actions video. Errors exist in the similar actions, it mislabeled the jogging as walking in the self-recorded video. These actions are similar each other except they have different speed. The 3D-MHI model which is called the Volumetric Motion History Image [7], can be used for the analysis of irregularities in human action and to detect unusual behavior [8]. This approach is found to be invariant to motion self-occlusion, speed variability in action, and in variable-length motion sequences. Also, a hierarchical extension was used to the original MHI framework which introduces a second parameter , the decay factor, in order to vary the length of the captured history of movement [9]. The pyramid of MHIs allows for recovering to a certain extent motions of varying speed by exploiting spatial gradient information.

4.2 Comparison to the state of the art methods

1. One of the difference is state of the art methods use of a combination of feature types to represent the human action, optical flow was used as large-scale features and SURF descriptors as local patch features to represent complex actions [10]. 2. The other difference is that they also convert the MHI from view-based to view-invariant method, like motion history volumes and 3D-MHI, namely each action is represented as a unique curve in a 3D invariance-space, surrounded by an acceptance volume, which can deal with self-occlusion, speed variation. 3. Another difference is the MHI I used can have overwriting or self-occlusion, such as in the running, the person run from left to right, and then right to left, but MHI has no clue that the person ran from left to right earlier. Current approach solved this by considering multi-camera system or multi-level history images or splitting optical flow vectors.

4.3 Proposals on methods can be improved

Some aspects that can improve my approach: 1. Consider using the multi-camera system or splitting optical flow vectors or 3D view invariant model to deal with the self-occlusion. 2. Consider using dynamic background subtraction technique to better segment target moving foreground extraction. 3. Consider using a tracking bounding box that aims to isolate the relevant motions if two or more people in the field of view. 4. Consider adding a tracking algorithm to better locate the trajectory of the motion path.

References

[1] D. Weinland, R. Ronfard, and E. Boyer, "Automatic discovery of action taxonomies from multiple views," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 1639–1645, IEEE, 2006.

[2] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[3] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3034–3042, 2016.

[4] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 928–934, IEEE, 1997.

[5] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, pp. 255–261, IEEE, 1999.

[6] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, pp. 246–252, IEEE, 1999.

[7] A. B. Albu and T. Beugeling, "A three-dimensional spatiotemporal template for interactive human motion analysis.," *Journal of Multimedia*, vol. 2, no. 4, 2007.

[8] A. B. Albu, T. Beugeling, N. Virji-Babul, and C. Beach, "Analysis of irregularities in human actions with volumetric motion history images," in *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*, pp. 16–16, IEEE, 2007.

[9] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on*, pp. 39–46, IEEE, 2001.

[10] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.