Activity Classification using MHI

Jijun Hu

I. Introduction

In General, the MHI method uses the idea that the video is actually a sequence of frames captured over time, meaning itself is a function of space over time, or we can write as I (x, y, t).

The approach showed here can also be considered as a two components version of a temporal template, which based motion-energy image – A image that represents where motion has occurred in an image sequence and motion-history image – A scalar-valued image where intensity is a function of recency of motion. [1]

II. Methodology

2.1 Data samples

For this project, I'm using data from http://www.nada.kth.se/cvap/actions/ both for data training and validation. A brief introduction of these data, they basically are made of by 6 actions by different persons, and are around 10 – 30 seconds each. I used person 15 as my training sample, and later pick different actions from different person to create validation sets.

2.2 Create binary images: Background subtraction

In order to recognize the activity, we need to find the object to detect at first. The way we are doing here is to use Background subtraction which can be expressed as below:

$$B_t(\text{x, y, t}) = \begin{cases} 1 & if\left|I(x,y,t)-I(x,y,t-1)\right|\geq\theta \\ 0 & \quad if\ otherwise \end{cases}$$

The idea here is comparing two consecutive frames, and see the absolute difference, if there is any motion or movement, we will observe the difference, and we put theta here as a threshold, meaning we threat those minimum difference as noise, and we can filter out. θ is a parameter to tune when training the model, 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 [2]. Compared using the adjacent frame difference and average frame difference, and the latter is better. After this, we still can observe quite a few noisy points in the Binary Image, so I did some research and try to use morphology method introduced by [3], and that leads to a much better Binary Image set.

2.3 Create temporal-templates

This is a major step in the process, creating temporal-templates, since this is the image that captures motions. Here, we are trying to calculate two images – MHI/MEI images.

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

$$MEI: E_\tau(x, y, t) = \text{¿} i=0 \text{¿} \tau-1 D(x,y,t-i)$$

Motion energy images (MEI) represents the spatial accumulation of motion, and these images can be used to suggest both the movement occurring and the viewing condition. While Motion history images (MHI) is a scalar-valued-image where more recently moving pixels are brighter. Obviously, MEI can be generated by thresholding the MHI above 0. To point out, like the Binary image, we do have a parameter τ here in both equation, that is something we are going to tune later when we training the model, τ is the frame number we are going to select to represent one action.

2.4 Image Hu moments

The two-dimensional (p+q) th order moments of a density distribution function $\rho(x,y)$ are defined as below:

$$\mu_{pq}=\iint_{-\infty}^{\infty} (x-\acute{x})^p (y-\acute{y})^q \rho(x,y)\, d(x-\acute{x})\, d(y-\acute{y})^{\square}$$

By normalize it, we can calculate all the normalized centralized moments:

$$\eta_{pq}=\frac{\mu_{pq}}{(\mu_{00})^\gamma}$$

where $\gamma=\frac{p+q}{2}+1 \wedge p+q\geq 2$

Finally, I calculate seven orientation invariant Hu Moments defined in [4].

2.5 Build a classifier

The method finally chosen to build the classifier is the KNN, this is a very straight-forward classification algorithm that very robust in low-dimensional problem. Given we are only doing classification based on descriptive statistics (7 Hu Moments for each MEI & MHI) – 14 in total, KNN is an effective and easy to implement way for this practice.

$$\hat{Y}(x)=\frac{1}{k}\sum_{xi\in Nk(x)} y_i$$

2.6 Action recognition

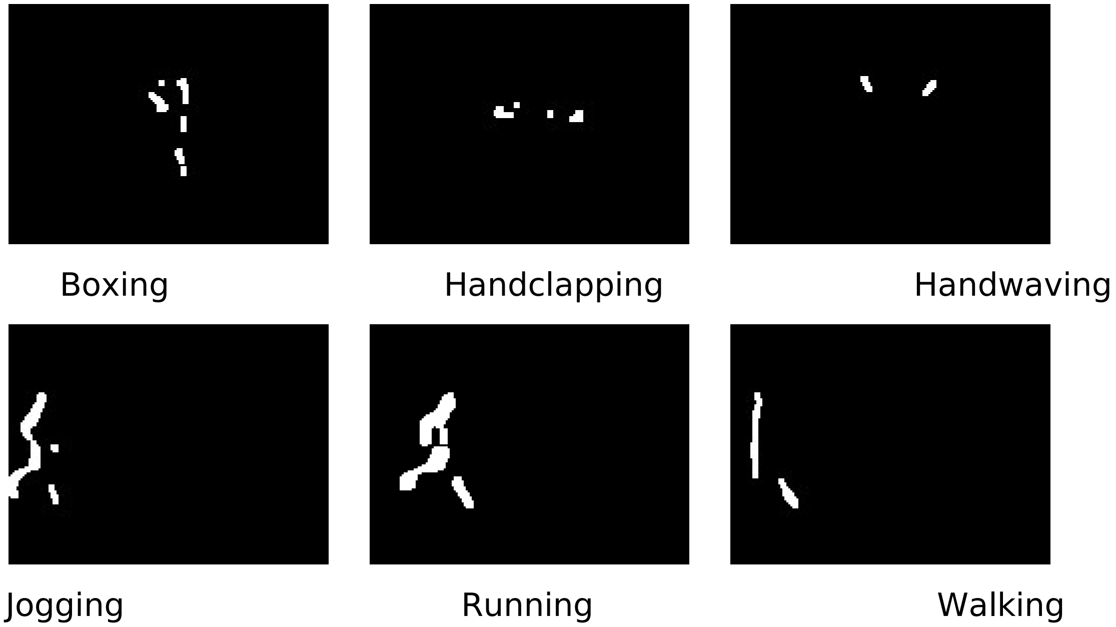
Once we have the Model trained, then we can apply it to do the prediction, I combined some of the new data we haven’t used before as our validation set, and too couple of frames try to identify their activity to see if the method described above works or not.

III. 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.

The binary images in Fig 1. below described each action, and the background was subtracted by using the frame differencing. As we can see, the binary images were able to capture the interested action scene and removed the background.



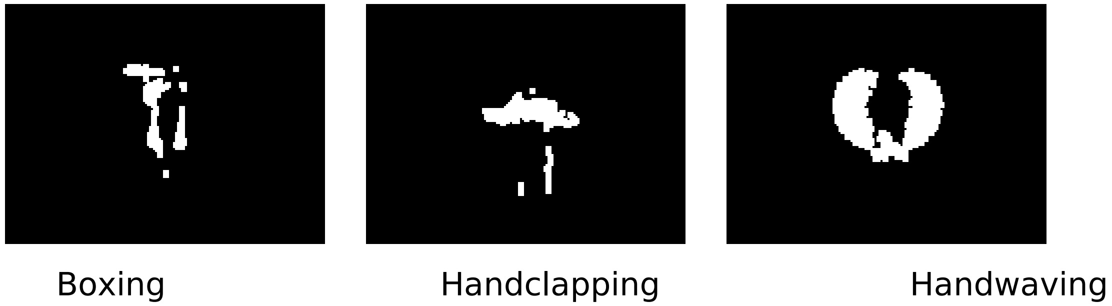
Boxing Handclapping Handwaving

Jogging Running Walking

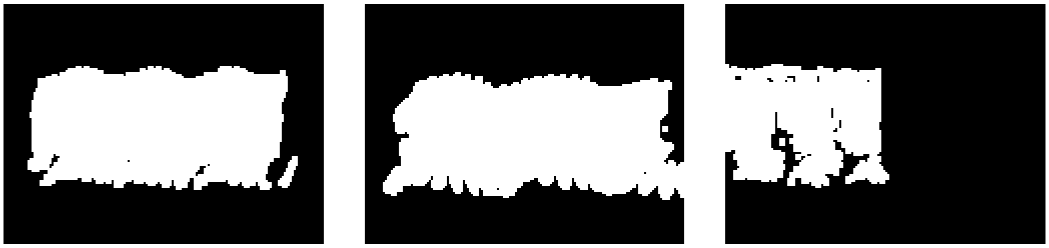
Figure 1: Binary images from different actions in the training.

3.2 The MEI/MHI from training

Below figures are the MEI/MHI from each action.



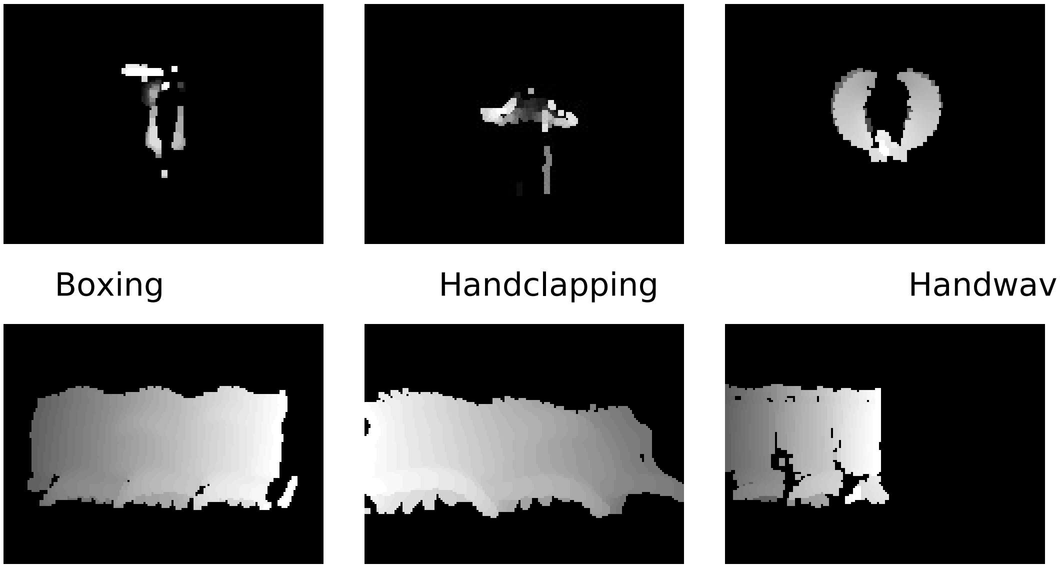
Boxing Handclapping Handwaving



Jogging Running Walking

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

Looking at both MEI above and MHI below, we can see, MEI gives a very good indication capturing what a person actually doing during a certain of time. While the motion history images, adding the time decay into the MEI image, you can tell the difference between image jogging, running and walking since it by putting more weights on latest frame, we can see the speed and direction of movement which we cannot achieve by looking at only MEI.



Boxing Handclapping Handwaving

Jogging Running Walking

Figure 3: Motion History Image from different actions in the training.

3.3 Model Training / Testing

Once we have all the MHI/MEI together with their invariant Hu Moment, we can start training the model. As we showed initially, there are couple of parameters we need to tune to see a decent performance. The training sample confusion matrix is showed below left in Figure 4a. While the confusion matrix for testing samples are shown below as Figure 4b.

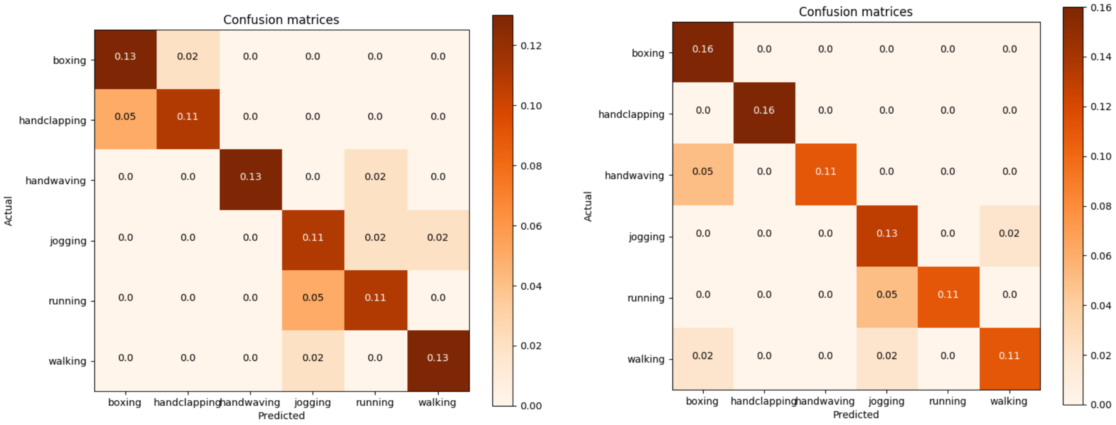


Figure 4a, 4b: Confusion Matrix for model fitting and validation

3.4 Performance statistics analysis-Confusion Matrices

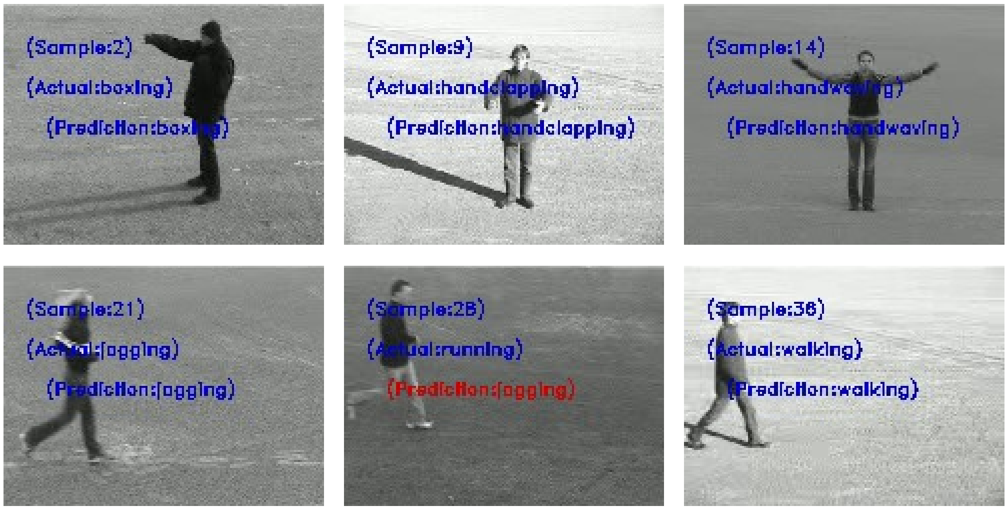


Figure 5: Actual validation Scene

In real testing, we took 6 scenes for each action to collect 36 data points both in the training and testing sets. From the confusion matrix in Figure 4, the algorithm performs pretty good in finding boxing and handclapping, but there are some misclassifications when looking at jogging/running/walking like what you can see in Figure 5, since when you think about all these three actions, the only difference may just be the speed, especially when you look at MHI in Figure 3. The differences between jogging and running are actually not very significant.

IV. Discussion

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

Generally speaking, this method works pretty well, as it achieves over 80% accuracy in test samples. When the action is pretty unique, this method works, while when looking at jogging/running/walking sample, since these actions are pretty similar and the only difference is speed, the method performs relatively bad. The better way to improve it, is to count exact information into the model, instead of describe it as $I(x, y, t)$, we add one more dimension into the model, such as add one more CNN layer on the raw image pixels of video [6], but the trade off is this is much computational expensive.

4.2 Comparison to the state-of-the-art methods

1. One of the differences 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 [7].
2. The other difference is that they also convert the MHI from view-based to view-invariant method, like motion history volumes and 3DMHI, 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.

4.3 Proposal on methods can be improved Some aspects that can improve my approach:

1. Maybe we can use the multi-camera system or splitting optical flow vectors or 3D view invariant model to deal with the self-occlusion.
2. The performance may be better if we can add an algorithm, instead of detecting, we try to track the locate the motion path.
3. Consider using dynamic background subtraction technique to better segment target moving foreground extraction.

V. References

[1] 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.

[2] 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.

[3] R. Haralick and L. Shapiro Computer and Robot Vision, Vol. 1, Addison-Wesley Publishing Company, 1992, Chap. 5, pp 174 - 185.

[4] H. Ming-Kuei, "Visual pattern recognition by moment invariants," Information Theory, IRE Transactions, vol. 8, pp. 179-187, 1962.

[5] 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.

[6] 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.

[7] 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.