GEORGIA TECH'S CS 6476 COMPUTER VISION

Final Project: Activity Classification using MHI

Ho Tin Kong

hkong43@gatech.edu

Abstract

This report explores the utilization of Motion History Image (MHI) in activity classification, integrating learning algorithms such as Random Forest (RF), k-nearest neighbors (kNN), and Support Vector Machines (SVM) for precise activity categorization. The dataset used, the KTH Actions dataset, comprises six activities performed by 25 participants, with the Hu moment and MHI computed for each frame using OpenCV. Following data preprocessing, a hierarchical classifier is introduced to distinguish between movement and gesture groups, enhancing classification accuracy. Results demonstrate that the hierarchical classifier with the random forest algorithm exhibited the highest performance at 94.5%. Challenges of classification are identified, suggesting strategies for future improvement. This report underscores the efficacy of MHI-based hierarchical classification in enhancing activity recognition accuracy.

Introduction

Activity recognition has been one of the most popular topics in computer vision. This topic relates to different industrial and commercial usages such as human-computer interaction, surveillance systems, and robotics [1]. Among different approaches, the Motion History Image (MHI) is one of the most widely used techniques for capturing temporal information within video sequences. The MHI can be further used for activity recognition by three approaches: template matching, state space approaches, and semantic description of human behaviors [2].

This report delves into the application of MHI in activity classification, utilizing advanced machine learning algorithms such as Random Forest (RF), k nearest neighbor (kNN), and Support Vector Machines (SVM) for accurate activity categorization. Different parameters such as the centroid of image moments and τ value of MHI will also be analyzed and optimized. Through the fusion of MHI with these classification methods, a robust framework is established to discern and classify various activities based on their temporal dynamics captured within video data.

LITERATURE REVIEW

The concept of motion energy image and motion history image is first introduced by Bobick and Davis. Their temporal template representation is represented as static vector images, encoding motion properties at spatial locations in image sequences. They demonstrated the effectiveness of a simple template structure by using 7 Hu moments as the moment-based features [3].

Alp and Keles (2017) proposed a method for action recognition using Hu moments derived from modified Motion History Images (MHIs) in conjunction with Hidden Markov Models (HMMs). Their approach achieved a notable 99% classification accuracy on the Weizmann dataset, showcasing its effectiveness in addressing challenges related to variations in appearance and environmental conditions [1].

Recently, Raj and Kos presented an enhanced human activity recognition technique leveraging Convolutional Neural Networks (CNNs). They introduce a CNN-based model that utilizes sensor input sequences to classify human activities, achieving a significant accuracy rate of 97.20% on the WISDM dataset. The study highlights the potential of deep learning methods for human activity classification. [4]

Methodology

Background isolation

The foreground extraction function is done by frame subtraction with a median filter.

The median filter is obtained by:

$$B(x, y, t) = median(x, y, i)$$

Where $1 \leq i \leq T$ and T is the number of frames.

Hence, the foreground filter is obtained using this equation:

$$|I(x, y, t) - B(x, y, t)| > th$$

Where th is the intensity threshold, which is set to 35 for this project.

Motion history image

The motion history image (MHI) is done with the frame difference sequence of the frames after foreground extraction I_f [2].

$$\psi(x, y, t) = \begin{cases} 1, if \left| I_{f,t}(x, y) - I_{f,t-1}(x, y) \right| \geq \theta \\ 0, otherwise \end{cases}$$

After applying the frame difference sequence with $\theta = 5$, noise reduction is done through dilation and erosion with

a 3x3 kernel. Hence, the MHI is calculated using the formula according to Bobick and Davis [3]:

$$M_t(x,y,t) = \begin{cases} \tau, if\ \psi_t(x,y) = 1 \\ \max(M_t(x,y,t-1) - 1, 0), if\ \psi_t(x,y) = 0 \end{cases}$$

To get better performance for calculating the image moment, the intensity is further increased for motion-related pixels.

$$M_t(x,y,t) = M_t(x,y,t) + (255 - \tau), \quad if\ M_t(x,y,t) \neq 0$$

Image moment

The concept of image moment is learned from the physics equation of moment:

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y)$$

A good parameter for activity classification needs to be translational, scale, and rotation irrelevant. Since the image moment is with respect to the origin (0,0) To obtain a translational invariance value, the central moment is required.

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

Where

$$\bar{x} = \frac{M_{10}}{M_{00}}, \bar{y} = \frac{M_{01}}{M_{00}}$$

To improve data quality, instead of using the MHI, only the motion of current frame is used to calculate the image centroid. This will widen the data boundaries between walking, running, and jogging.

$$I_{centroid}(x,y,t) = \begin{cases} 255, if\ \left|I_{f,t}(x,y) - I_{f,t-1}(x,y)\right| \geq \theta \\ 0, otherwise \end{cases}$$

To achieve scale invariance, the moment needs to be further normalized.

$$v_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1+\frac{p+q}{2}}}$$

Finally, scale and rotational irrelevant parameters can be obtained using the Hu moment equations [5]. Note that the last Hu moment H_7 is reflection antisymmetric.

$$H_1 = v_{20} + v_{02}$$

$$H_2 = (v_{20} - v_{02})^2 + 4v_{11}^2$$

$$H_3 = (v_{30} - 3v_{12})^2 + (3v_{21} - v_{03})^2$$

$$H_4 = (v_{30} + v_{12})^2 + (v_{21} + v_{03})^2$$

$$H_5 = (v_{30} - 3v_{12})(v_{30} + v_{12})((v_{30} + 3v_{12})^2 - 3(v_{21} + v_{03})^2) + (3v_{21} - v_{03})(v_{21} + v_{03})((v_{30} + v_{12})^2 + 3(v_{21} + v_{03})^2)$$

$$H_6 = (v_{20} - v_{02})((v_{30} + v_{12})^2 - (v_{21} + v_{03})^2) + 4v_{11}(v_{30} + v_{12})(v_{21} + v_{03})$$

$$H_7 = (3v_{21} - v_{03})(v_{30} + v_{12})((v_{30} + v_{12})^2 - 3(v_{21} + v_{03})) + (v_{30} - 3v_{12})(v_{21} + v_{03})(3(v_{30} + v_{12})^2 - (v_{21} - v_{03})^2)$$

Data preparation

The dataset used for training and testing is KTH Actions dataset. There are six types of activities, which are walking, jogging, running, hand clapping, hand waving, and boxing, performed by 25 participants with a total of 2391 sequences [6].

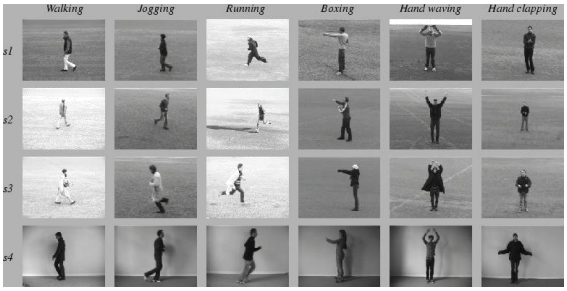

Fig. 1.1 example of KTH action dataset

The Hu moment and MHI are calculated for each frame of each video with OpenCV. Only the frames with motion and the object that is not occluded will be used as training and testing datasets. Images with any Hu moment equals to zero are not considered. Extreme data with values larger than 1.8 times of standard deviation and some ineffective background reduction frames are also excluded. The following are some examples of extracted MHI.

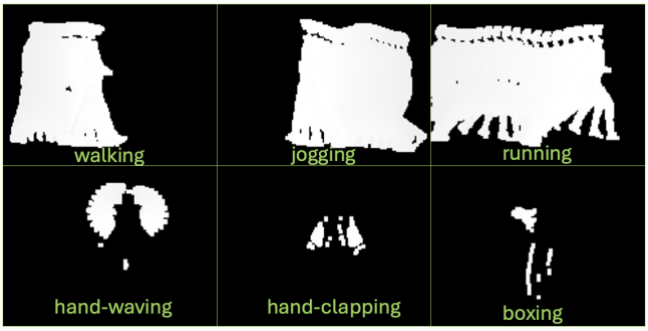

Fig. 1.2 example of MHI extraction

The following table shows the summary of the dataset, 80% of the data is used as the training set and the remaining 20% is the testing set.

Motion	Walking	Jogging	Running	Hand-clapping	Hand-waving	Boxing	Total
No. of samples	991	687	393	1186	1171	718	5146
	Movement sub-total: 2071			Gesture sub-total: 3075			

Classification

The dataset with H_1 to H_6 as input is normalized using the Min-max scaler. H_7 has not been used for training and classification due to its reflection asymmetric property. Principal component analysis (PCA) is done to analyze the data quality. Hence, K-nearest-neighbour (KNN) with 3 nearest neighbors, support vector machine (SVM), and random forest (RF) algorithms are used to train and evaluate the classification performance.

To improve the classification performance, a hierarchical classifier is introduced. The motions are divided into two main groups, the movement group with walking, jogging, and running, and the gesture group with clapping, waving, and boxing. A first-layer classifier will classify which groups the motion belongs to, and then sub-layer classifiers will classify the specific action. For the Hu moment of the first-layer and movement layer, the number of frames to track in MHI τ is set to 6. For the gesture layer, τ is set to 15 to keep more details. The performance of each approach can be found in the result section.

Results

Performance of hierarchical classifier versus single classifier

Using Knn classifier as reference, the Principal component analysis (PCA) graph, confusion matrix of testing data, and accuracy of using single classifier versus hierarchical classifier is evaluated. Note that these results are based on the improved centroid method (centroid using current frame motion) and all the confusion matrices are using the testing set as data.

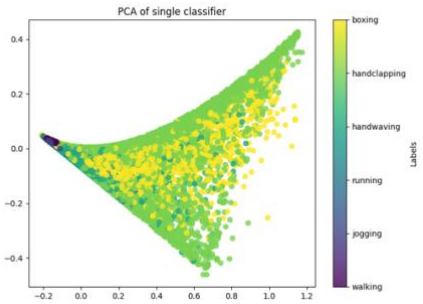

Fig.2.1 PCA result of single classifier

Confusion matrix of single layer kNN classifier

	Walking	Jogging	Running	Waving	Clapping	Boxing
Walking	803	143	12	20	2	28
Jogging	221	365	76	36	2	8
Running	51	141	209	12	2	11
Waving	72	47	14	1034	140	76
Clapping	17	6	4	307	716	108
Boxing	107	24	4	227	151	330

Total accuracy: 62.6%

Table 1 – Confusion matrix of single classifier

Prediction result of hierarchical classifier

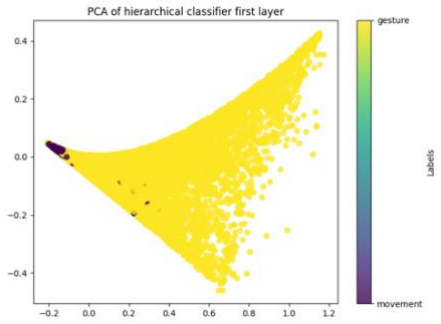

Fig.2.2 PCA result of hierarchical classifier first layer

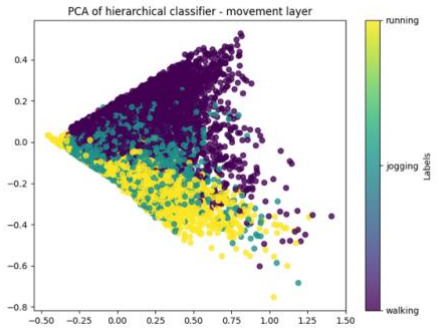

Fig.2.3 PCA result of hierarchical classifier movement layer

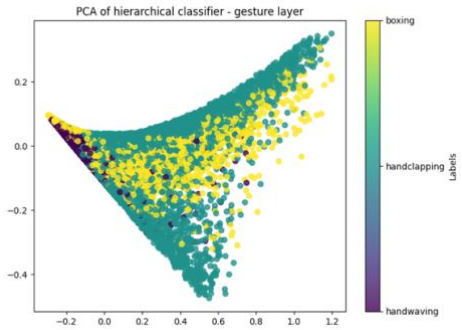

Fig.2.4 PCA result of hierarchical classifier gesture layer

Confusion matrix of first layer (kNN)

	Movement	Gesture
Movement	1986	156
Gesture	230	3154

Accuracy: 93%

Table 2 – Confusion matrix of first layer

Confusion matrix of movement layer (kNN)

	walking	jogging	running
walking	824	162	5
jogging	186	453	48
running	22	71	300

Accuracy:76.1%

Table 3 – Confusion matrix of movement layer

Confusion matrix of Gesture layer (kNN)

	waving	clapping	boxing
waving	1052	70	64
clapping	187	884	100
boxing	146	155	417

Accuracy:76.5%

Table 4 – Confusion matrix of gesture layer

Overall accuracy: 71.0%

The result shows that using a hierarchical classifier successfully improved the total prediction accuracy by 8.4%. PCA plot also shows a less chaotic pattern for the hierarchical classification.

Performance of MHI centroid versus single frame motion centroid

Similarly, taking the kNN classifier with the movement layer as a reference, the performance of using the MHI centroid method (taking the whole MHI to calculate the centroid) is shown below.

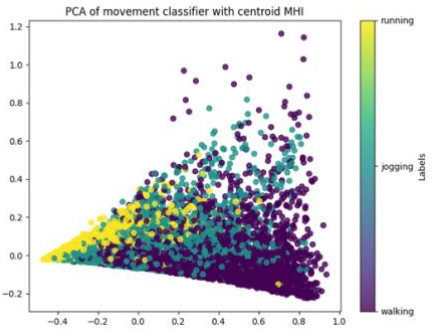

Fig. 2.5 PCA of centroid MHI movement layer

Confusion matrix of movement using MHI centroid

	walking	jogging	running
walking	775	144	6
jogging	213	444	55
running	27	89	302

Accuracy: 74%

Table 5 – Confusion matrix of MHI centroid

This result shows the performance of taking the centroid using current frame motion only (Fig. 2.3, table 3) performs better than the original MHI centroid method (Fig. 2.5, table 5) in movement classification.

Performance of different classifier types

Finally, the performance of different classifiers of all layers is evaluated. The data shows the classification performance of the gesture layer. For comparison, the performance of kNN is listed in table 4.

Confusion matrix of gesture layer (SVM)

	waving	clapping	boxing
waving	1101	68	17
clapping	206	923	42
boxing	248	227	243

Accuracy: 73.7%

Table 6 – Confusion matrix of SVM classifier

Confusion matrix of gesture layer (RF)

	waving	clapping	boxing
waving	1076	59	51
clapping	111	971	89
boxing	82	158	478

Accuracy: 82.1%

Table 7 – Confusion matrix of random forest classifier

The random forest (RF) classifier performs the best with 82.1% accuracy. k-NN classifier's accuracy is 76.5%, and the SVM classifier performs the worst, with an accuracy of 73.7%. Therefore, the random forest classifier is selected for the final classification model.

Final model

The final model has a hierarchical classifier structure, with one random forest classifier to classify the motion of movement or gesture, and two random forest classifiers to classify the specific motion. If the probability of prediction is less than a threshold, that frame will be considered as "no action". The frame kept for movement and layer classifier is 6 frames and the frame kept for gesture layer is 15. All Hu moments are calculated based on the current frame motion centroid instead of the whole MHI. Using this approach the testing data could achieve a 94.4% testing accuracy (Table 8).

Confusion matrix of final model

	Walking	Jogging	Running	Waving	Clapping	Boxing
Walking	744	9	0	0	0	2
Jogging	21	488	6	0	0	2
Running	3	4	224	0	0	0
Waving	24	24	5	1170	3	4
Clapping	5	5	6	9	893	5
Boxing	40	21	22	5	15	508

Accuracy: 94.4%

Table 8 – Confusion matrix of final model

Discussion

Prediction of walking

Fig. 3.1 Person06_walking_d2, frame 306

Video link: output/person06_walking_d2_uncomp.avi

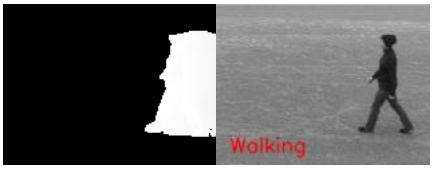

Fig. 3.2 Person19_walking_d1 frame 307

Video link: output/person19_walking_d1_uncomp.avi

Prediction of jogging

Fig. 3.3 Person10_jogging_d1 frame 163

Video link: output/person10_jogging_d1_uncomp.avi

Fig. 3.4 Person15_jogging_d2 frame 209

Video link: output/person15_jogging_d2_uncomp.avi

Prediction of running

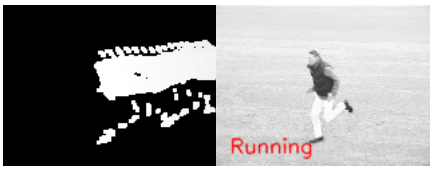

Fig. 3.5 Person11_running_d2 frame 319

Video link: output/person11_running_d2_uncomp.avi

Fig. 3.6 Person18_running_d1 frame 127

Video link: output/person18_running_d1_uncomp.avi

Prediction of hand-waving

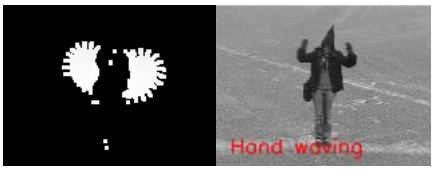

Fig. 3.7 Person08_handwaving_d3 frame 103

Video link: output/person08_handwaving_d3_uncomp.avi

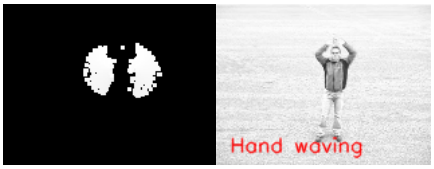

Fig. 3.8 Person12_handwaving_d1 frame 203

Video link: output/person12_handwaving_d1_uncomp.avi

Prediction of hand-clapping

Fig. 3.9 Person16_handclapping_d4 frame 130

Video link: output/person16_handclapping_d4_uncomp.avi

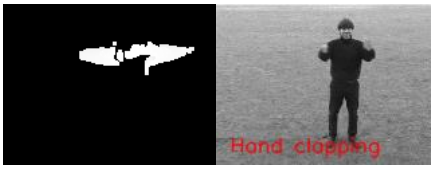

Fig. 3.10 Person21_handclapping_d1 frame 100

Video link: output/person21_handclapping_d1_uncomp.avi

Prediction of boxing

Fig. 3.11 Person07_boxing_d4 frame 79

Video link: output/person07_boxing_d4_uncomp.avi

Fig. 3.12 Person08_boxing_d1 frame 210

Video link: output/person08_boxing_d1_uncomp.avi

The prediction videos and images show that the classifier works effectively in most cases, including motion in different scales, positions, and orientations. However, some failure scenarios are observed.

Failure case

Shadow

Fig. 3.13 Person14_walking_d3, frame 43, shadow failure

MHI with shadows does not have similar characteristics to cases without shadows. Those data have not been handled in the current classifiers. Therefore, wrong classifications will occur for motion video with shadows. To improve, a new classifier can be trained with a new label for those cases with shadows.

Moving camera

Fig. 3.14 Person03_walking_d4, frame 35, moving camera failure

This failure occurs when the camera is moving, causing the failure of background subtraction. Low-quality MHI is produced in this case, leading to a wrong prediction. To deal with this issue, some better background subtraction methods such as CNN or Gaussian Mixture Models (GMM) can be used to improve MHI quality [7].

Low object contrast

Fig. 3.15 Person11_jogging_d3, frame 157, object contrast failure

This is also a failure of background subtraction. The contrast between the background and motion object is lower than the threshold, resulting in bad MHI extraction. Similarly, further tuning of parameters or using a more advanced background subtraction method could solve this failure.

Conclusion

In this study, a hierarchical classifier with MHI and Hu moment as input is trained to predict the motion. The MHI follows a modified equation to maximize the boundaries between different motions. The Hu moment calculation method is improved using current frame motion to calculate the image centroid. The performance of different types of classification algorithms is compared. The classifiers using the random forest algorithm have the best performance with a successful accuracy of 94.5%. Some failure cases of using this approach have been discussed, and possible future improvements are suggested.

References

[1] E. C. Alp and H. Y. Keles, "Action recognition using MHI based Hu moments with HMMs," IEEE EUROCON 2017 -17th International Conference on Smart Technologies, Ohrid, Macedonia, 2017, pp. 212-216, doi: 10.1109/EUROCON.2017.8011107.

[2] Ahad, M.A.R., Tan, J.K., Kim, H. et al. Motion history image: its variants and applications. Machine Vision and Applications 23, 255–281 (2012).

[3] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, March 2001, doi: 10.1109/34.910878.

[4] Raj, R., Kos, A. An improved human activity recognition technique based on convolutional neural network. Sci Rep 13, 22581 (2023)

[5] M. K. Hu, "Visual Pattern Recognition by Moment Invariants", IRE Trans. Info. Theory, vol. IT-8, pp.179–187, 1962

[6] P. M. Roth, T. Mauthner, I. Khan, and H. Bischof, "Efficient human action recognition by cascaded linear classification," in 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), Nov. 2009, pp. 546-553. doi: 10.1109/ICCVW.2009.5457655.

[7] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 1999, pp. 246-252 Vol. 2, doi: 10.1109/CVPR.1999.784637.