Activity Recognition using Motion History Imaging (MHI)

SAHIL DHINGRA

CS-6476 Computer Vision
Project
Fall - 2019

1. Introduction

In this article, we would discuss the implementation activity classification algorithm using Motion History Images (MHI) and Hu moments for human activity classification. The six types of human actions to be classified are – walking, jogging, running, boxing, hand-clapping and hand-waving.

2. Implementation

KTH dataset was used for training for this particular problem. The video database contains six types of human actions (walking, jogging, running, boxing, hand-clapping and hand-waving) performed several times by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3 and indoors s4 as illustrated below. The database contains 2391 sequences, all taken over homogeneous backgrounds with a static camera with 25fps frame rate, spatial resolution of 160x120 pixels and a length of four seconds on average.

2.1. MHI

From each video, Motion History Images (MHIs) are constructed to represent motion between two frames and generate scaled and unscaled moments which are then used in the classification model. First, binary difference images $B_t(x,y,t)$ are generated by taking difference between two successive frames and highlighting pixels where change in intensity is greater than a threshold value, θ.

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

Binary images are then converted to MHIs $M_\tau(x,y,t)$ by setting the values for highlighted pixels to τ, and with temporally diminishing values for places where motion was detected in earlier MHIs. This way, MHI images have maximum intensity where a change has been detected in the current image and temporally decreasing intensities cutting off at τ time steps. In other words, maximum weight is given to pixels where motion is detected in the current frame and

lowest weight to pixels where motion was detected τ frames ago, and no weights for frames before τ.

$$M_\tau(x,y,t) = \begin{cases} \tau & \text{if } B_t(x,y) = 1 \\ \max(\ M_\tau(x,y,t-1) - 1,\ 0\ ) & \text{if } B_t(x,y) = 0 \end{cases}$$

Using these MHIs, central moments, which are statistical measures to summarize the patterns of intensity along x and y axes, are computed. Scaled and unscaled versions of these moments are computed. From these moments, Hu moments are calculated, which would eventually be used as features in the classification model.

2.2. Model training

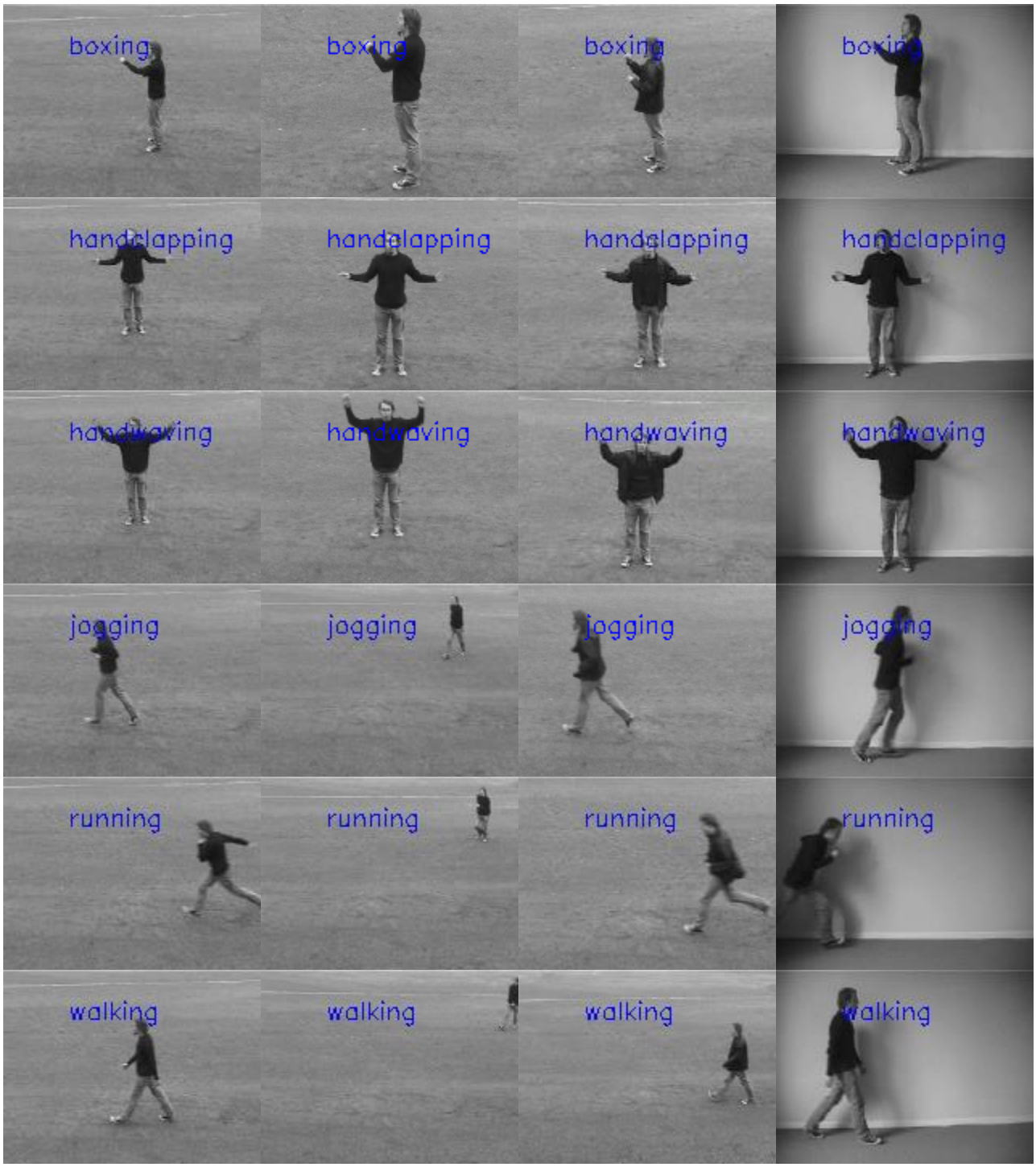
1. **Dataset** – The dataset was divided into three sets: train (for model training), validation (for hyper-parameter tuning and validation) and test (for final model validation). Training and validation sets consist of all activities performed by a set of 8 persons each and 9 persons for test set. After validation, the model is trained on entire train + validation set and scores on test set are reported.

2. **Frames** – From each video, only frames within start frame and end frame, as per the .txt file, were used for model training and testing purposes i.e. frames with a person were kept and rest were discarded. This is important as frames with no motion are disregarded and only relevant frames are used for training.

3. **Machine learning model and hyper-parameters** – A gradient boosting model was used for multi-class classification. The model was developed with the following parameters:
   - Number of trees: 1200
   - Subsampling rate: 70%
   - Minimum samples in leaf: 10
   - Maximum depth: 5

4. **MHI parameter selection** – For tuning θ and τ, a manual approach was tried, so that only the important changes in the images were captured to detect motion and a relevant history of motion retained in MHIs. However, this did not result in the most optimal parameters and there was difficulty in separating apart certain combinations of actions. To deal with this, different values of θ and τ were tested and the one with optimal validation score was used.

3. Results


boxing
boxing
boxing
boxing
handclapping
handclapping
handclapping
handclapping
handwaving
handwaving
handwaving
handwaving
jogging
jogging
jogging
jogging
running
running
running
running
walking
walking
walking
walking

4. Performance statistics

The model has an overall **precision of 87.8%** and a **recall of 86.5%**. Precision is a more appropriate metric since it is important for us that our predictions are mostly correct, wherever made. We can avoid making a prediction where we are not very confident. However, this criterion of making a prediction only beyond a certain level of confidence has not been taken into implementation here, but it could help with an improvement in precision.

	Walk	Jog	Run	Box	Hwav	Hclap
Walk	**90**	2	1	1	11	1
Jog	0	**92**	3	0	0	7
Run	0	1	**88**	0	0	7
Box	4	0	0	**93**	8	0
Hwav	6	0	0	5	**81**	2
Hclap	0	5	8	0	0	**83**

	Walk	Jog	Run	Box	Hwav	Hclap
Walk	**83**	1	0	1	14	1
Jog	0	**86**	2	0	0	12
Run	0	1	**83**	0	0	16
Box	4	0	0	**84**	11	0
Hwav	4	0	0	4	**90**	2
Hclap	0	3	3	0	0	**93**

Fig. 1 Confusion matrices for Precision and Recall respectively, actual values on the rows and predictions along columns

5. Analysis of results

5.1. Analysis on why this works on some images but not on others –

1. The θ and τ values have been carefully tuned towards the characteristics of activities performed by a small sample of 16 people, in the presence of noise, change in lighting and scale. Therefore, this model might misclassify subjects with different activity characteristics, such as faster walking pace. Since this model is quite sensitive to MHI parameters, it can fail at higher level of noise due to the θ value.
2. Since this solution deals only with sequences with single subject with no distractions and no noise, it would not perform well on sequences with multiple subjects or other distractions with moving backgrounds.

5.2. Other methods published in recent research – Other solutions for motion detection have been developed since, some of them resulting in much better accuracies.

1. Histograms or PCA of optical flow or spatio-temporal gradient vectors
2. Local descriptors on interest points in space-time – Local descriptors are more useful for activity recognition as these extract feature descriptors for multiple space-time points of interest in the image, unlike global descriptors like Hu moments which describe the whole image
3. CNNs with LSTMs

5.3. State of the art methods – Methods such as the ones published in recent research, as mentioned above, can be very accurate, with accuracy >95%

5.4. Proposal on how the performance could be improved –

1. Shrinking the MHI/MEI images into moments (mainly providing statistical summarization of images) results in loss of information. Utilizing the full MHIs, with algorithms such as CNNs, would result in lesser information loss and improvement in the performance.
2. Additionally, if conditions for application of the model are known, a less generalized model or multiple models could be developed – with θ and τ chosen for specific conditions – parameters that work better for each scenario but not for others.
3. Rather than using a fixed θ for a complete activity, a statistical version of θ could instead be chosen for constructing the binary difference images. In the statistical distribution of pixel intensity change between two frames, a statistical threshold could be chosen, such as μ, or $\mu + \sigma$.
4. A criterion for making a prediction only beyond a certain level of confidence has not been implemented, which could help with a better precision score. Even though recall would drop, precision is a more relevant metric for us here.

References

J. Davis and A. Bobick, The representation and recognition of action using temporal templates, In Proc. CVPR, pages 928–934, 1997.

Ivan Laptev and Tony Lindeberg, Local Descriptors for Spatio-Temporal Recognition, CVAP