Action Recognition Using Temporal Templates with a Random Forest Classifier

Andrew Parmar (GT ID: 903389515)

CS6476 Computer Vision - Spring 2020 - Final Project

Abstract. This report showcases results from an implementation of human action recognition from image sequences using a Random Forest Classifier trained on features from temporal templates. Using frame-by-frame Motion History Images (MHI) and absolute Motion Energy Images (MEI) over a look-back window, a set of vector representations of human actions are created from a labeled action video database. Calculating scale and translation invariant image-moments for the motion image pairs, vectors are created that form the basis of the feature set to train a Random Forest Classifier, which can then be used to predict human actions given a new temporal template based vector. Finally, the classifier is augmented to use a buffer and frequency counter to predict and label actions in video.

1 Introduction

Recognition of human actions is an important field of study within computer vision and is the basis of many applications including, but not limited to video surveillance, health care, and human-computer interaction. [1] As such, various techniques have been proposed over the years as described in [2]. Ahad et al. [3], categorized action recognition into three groups: (i) template matching, (ii) state-space approaches and (iii) semantic description of human behaviors.

This study focuses on the area of action recognition via template matching. In particular, the concept of Motion History Images and Temporal Templates are explored as presented by Bobick and Davis [4] with the aim of re-implementing their proposed methods, while building up on their findings using machine learning classification techniques. The building blocks of temporal-templates [5] are implemented which comprise of binary Motion Energy Images (MEI) and scalar Motion History Images(MHI) which together form a vector based representation of human actions. Also examined are methods to convert these temporal images into scale invariant vectors that are then used to train a Random Forest classifier.

2 Related work

Much work has been done with template matching techniques for action recognition, particularly the use of MHIs. Keles and Alp [6] showed a modified version of the MHI (MMHI) that uses an exponential decay factor for τ as compared to the linear method first introduced by Bobick and Davis [4]. They combine this with an advanced Hidden Markov Model model to gain up to 99% classification accuracy as applied to the Wiezmann dataset. Rosales and Rómer [7] used generic MHIs and showed that the Hu moments could be processed via Principal Components Analysis to reduce the dimensionality of this representation, and then train K-nearest neighbor, Gaussian, and Gaussian mixture classifiers to perform template matching. Likewise, Meng et al. [8] used a support-vector machine as a classifier for MHI template matching.

3 Method

This study aims to bypass the Mahalanobis distance template matching methodology by Bobick and Davis [4] in favor of a fully supervised Random Forest classifier and compare the outcomes.

3.1 Creation of temporal images

The creation of temporal images is a sequence of three sequential steps. Images are first processed to undergo background subtraction. This results is a binary image indicating pixels where motion is detected. For the binary image creation, image-differencing is used as defined in Eq.1. Other techniques such as median and mode frame subtractions are more effective, and create cleaner binary images, however, they rely on a buffer of frames prior to outputting results. This results in loosing the first n frames of a sequence to the buffer, which is not desirable. Eq.2 and Eq.3 are next used to generate the MHI and MEI images with τ set between a range of 10 and 30 for fastest and slowest actions, respectively.

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

$$M_t(x,y,t) = \begin{cases} \tau & \text{if } B_t(x,y,t) = 1 \\ \max(M_t(x,y,t-1) - 1, 0) & \text{if } B_t(x,y,t) = 0 \end{cases} \tag{2}$$

$$E_t(x,y,t) = \begin{cases} 1 & \text{if } M_t(x,y,t) > 0 \\ 0 & \text{if otherwise} \end{cases} \tag{3}$$

Fig. 1 shows examples of the three stages of temporal templates for walking and hand-waving. The lack of detail in Fig1a for hand-waving, is contrasted with the details seen in Fig1e

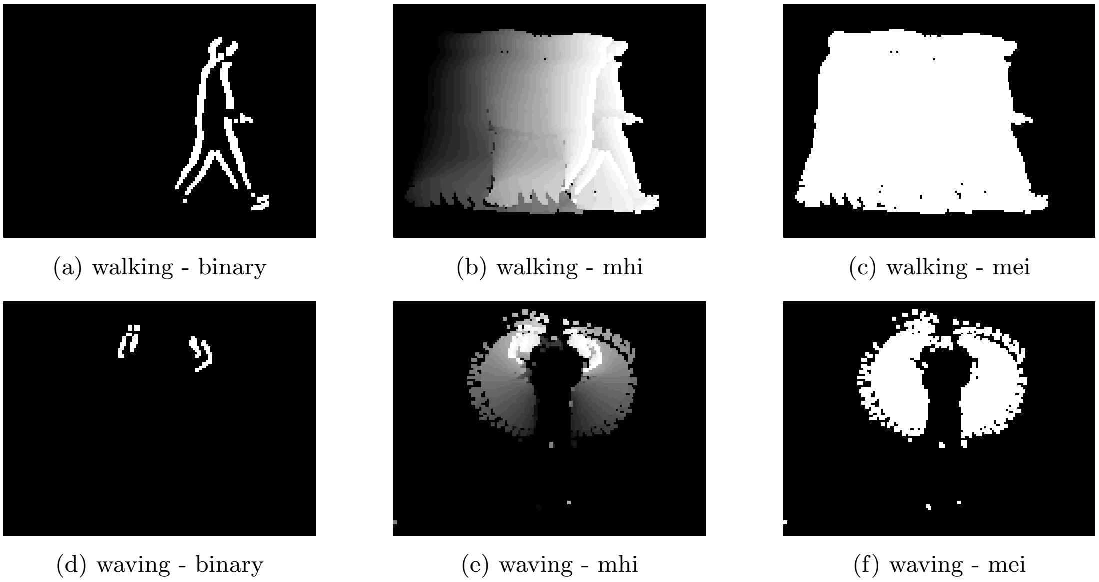
3.2 Calculation of Hu-moment based features

Using the MHI and MEI, scale and translation invariant moments for these images are created using the functions proposed by Hu [9]. Using seven scalar moments for each component in the MHI/MEI pair, a 14 dimensional vector representing an action sequence over τ image frames is created.

Fig2 shows plots of the MHI moments over the first 100 frames of video of various actions. Over the 100 frame window, each action shows a distinct Hu characteristic curve.

3.3 Random-forest classifier

Using 144 action videos from the Laptev et al. [10] video database [11], a total of 45599 Hu-moments vectors are generated. As each video action is known, a list of labels for each frame in the feature set is also created. Using Scikit-Learn's `RandomForestClassifer`, a model is trained using 100 estimators with bootstrapping for tree creation. $\sqrt{14}$ features are used for the split decision. These values were obtained via a grid search cross-validation subroutine.



(a) walking - binary

(b) walking - mhi

(c) walking - mei

(d) waving - binary

(e) waving - mhi

(f) waving - mei

Fig. 1: Stages of creation of the temporal template with $\tau = 20$ for the the top row (walking) and $\tau = 30$ for bottom row (hand-waving)

3.4 Creation of n-window lookback τ features-sets

The plots for running and jogging in Fig. 2 appear as time scaled versions of each other. This behaviour is consistent with intuition, as they are the same activity varying mostly in speed. This property of similar actions proposes a challenge to recognition as having no prior knowledge of the action in an image sequence, a known τ value cannot be used for temporal-template creation. Bobick and Davis [4] propose that comparisons must be run against all τ values that the classifier is trained on. Generating an MHI/MEI for the maximum value of τ and using Eq. 4, Hu moment vectors for all τ in the min and max range can be generated.

$$H_{t-\Delta\tau}(x,y,t) = \begin{cases} H_t(x,y,t-\Delta\tau) & \text{if } H_t(x,y,t) > \Delta\tau \\ 0 & \text{if otherwise} \end{cases} \tag{4}$$

4 Experiment

4.1 Naive comparison of classifier accuracy on training and test data

A comparison of classifier accuracy is shown in Fig. 3. Hu features are created for both the training and testing set of videos while accurately applying a known value of τ. This comparison is naive due to the reasons mentioned in Section3.4. The training set performs predictions with $\tilde{9}9.9\%$, while the testing data comes in at $\tilde{6}7.8\%$ cumulative accuracy. Fig. 3 clearly shows that the poor performance is primarily seen in the jogging, running and walking classes. This is not unexpected, as visual inspection of the various test videos showed a variation in running speeds between people. In comparison, picking a random label for each of the training set data points, obtains a $\tilde{1}2.6\%$ accuracy. Though a naive method, this validates the classifier's usefulness.

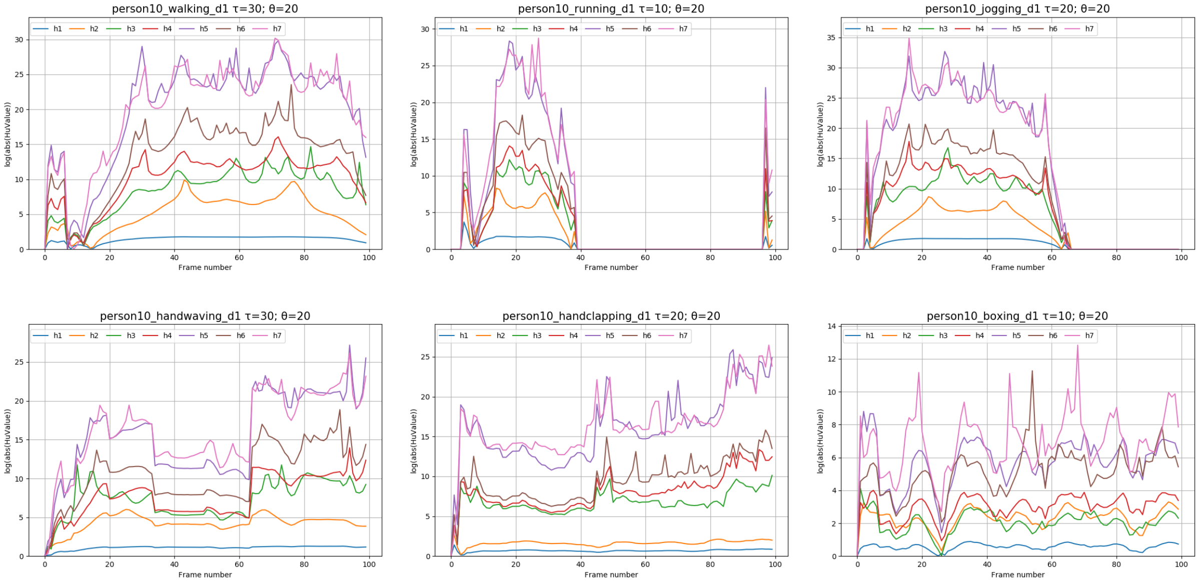
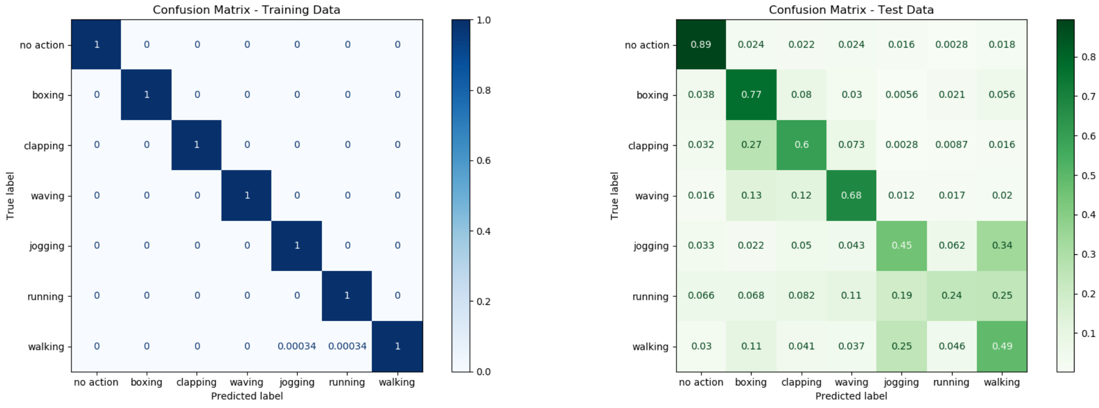


Fig. 2: Plots of log(abs(Hu-moments)) of the six different types of actions in the video dataset

4.2 Comparison of Random Forest classifier with other Scikit-learn classifiers.

The Random Forest classifier is also compared to two more generic classification techniques, namely Naive Bayes, and K-Nearest Neighbors. In doing this comparison no additional tuning was conducted on the estimators, they are used with their default settings from the Scikit-learn library. As can be see in Fig. 3 the Random Forest shows a prediction record of true positives compared to the other two.



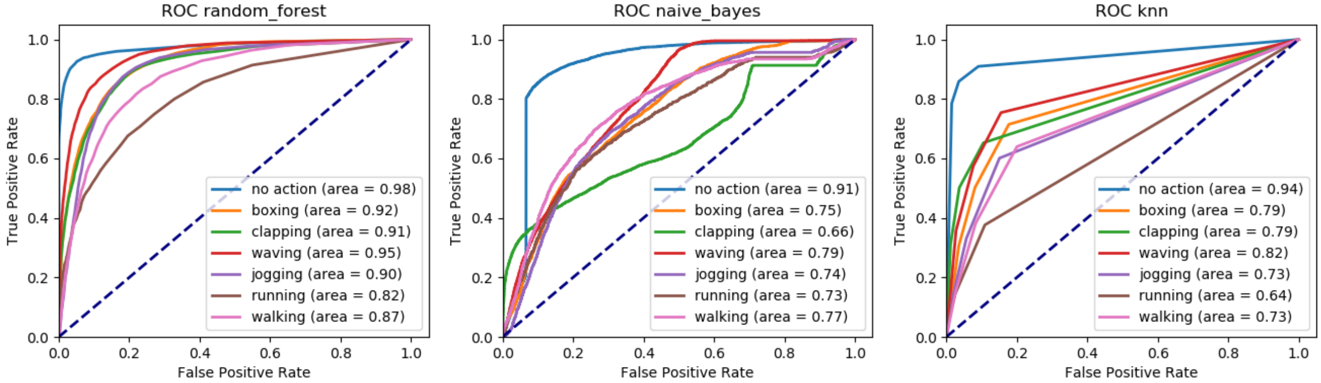
Confusion Matrix - Training Data

	no action	boxing	clapping	waving	jogging	running	walking
no action	1	0	0	0	0	0	0
boxing	0	1	0	0	0	0	0
clapping	0	0	1	0	0	0	0
waving	0	0	0	1	0	0	0
jogging	0	0	0	0	1	0	0
running	0	0	0	0	0	1	0
walking	0	0	0	0	0.00034	0.00034	1

Confusion Matrix - Test Data

	no action	boxing	clapping	waving	jogging	running	walking
no action	0.89	0.024	0.022	0.024	0.016	0.0028	0.018
boxing	0.038	0.77	0.08	0.03	0.0056	0.021	0.056
clapping	0.032	0.27	0.6	0.073	0.0028	0.0087	0.016
waving	0.016	0.13	0.12	0.68	0.012	0.017	0.02
jogging	0.033	0.022	0.05	0.043	0.45	0.062	0.34
running	0.066	0.068	0.082	0.11	0.19	0.24	0.25
walking	0.03	0.11	0.041	0.037	0.25	0.046	0.49

Fig. 3: Naive classifier performance comparison



ROC random_forest

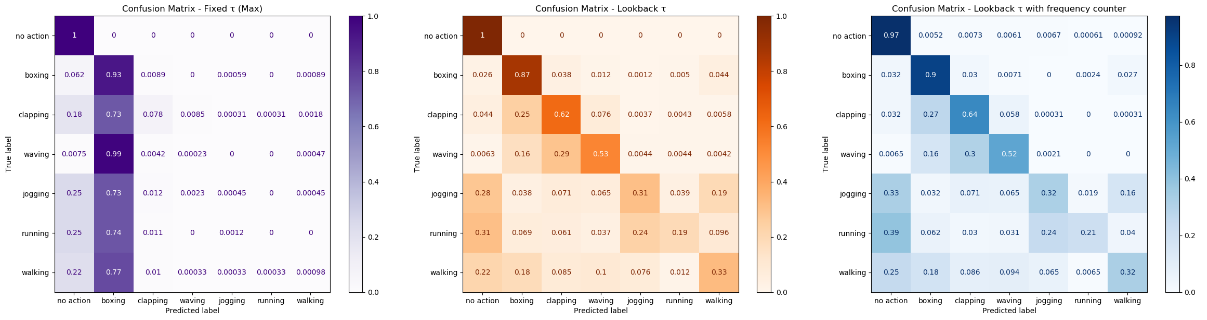
ROC naive_bayes

ROC knn

Fig. 4: ROC Curve comparison over different classifiers

4.3 Comparison of lookback τ vs fixed τ classifiers

As introduced in Sec. 3.4 the MHI for hand-waving performed at slow speed, could resemble that of clapping at normal speed, as only a small number of frames will be captured in the temporal template. A lookback τ predictor is a workaround to this problem. Wrapping the estimator's prediction methods with logic to select the highest confidence prediction from an n-window sized features-set, it is seen that the backward looking-tau produces a higher overall prediction accuracy compared to using a fixed τ. For this experiment, $\tau = 20$ was chosen as it was the midpoint of the range used for training, and would have the most overlap with each action's duration. A fixed τ in-fact does quite poorly as seen in Fig. 5



Confusion Matrix - Fixed τ (Max)

True label \ Predicted label	no action	boxing	clapping	waving	jogging	running	walking
no action	1	0	0	0	0	0	0
boxing	0.062	0.93	0.0089	0	0.00059	0	0.00089
clapping	0.18	0.73	0.078	0.0085	0.00031	0.00031	0.0018
waving	0.0075	0.99	0.0042	0.00023	0	0	0.00047
jogging	0.25	0.73	0.012	0.0023	0.00045	0	0.00045
running	0.25	0.74	0.011	0	0.0012	0	0
walking	0.22	0.77	0.01	0.00033	0.00033	0.00033	0.00098

Confusion Matrix - Lookback τ

True label \ Predicted label	no action	boxing	clapping	waving	jogging	running	walking
no action	1	0	0	0	0	0	0
boxing	0.026	0.87	0.038	0.012	0.0012	0.005	0.044
clapping	0.044	0.25	0.62	0.076	0.0037	0.0043	0.0058
waving	0.0063	0.16	0.29	0.53	0.0044	0.0044	0.0042
jogging	0.28	0.038	0.071	0.065	0.31	0.039	0.19
running	0.31	0.069	0.061	0.037	0.24	0.19	0.096
walking	0.22	0.18	0.085	0.1	0.076	0.012	0.33

Confusion Matrix - Lookback τ with frequency counter

True label \ Predicted label	no action	boxing	clapping	waving	jogging	running	walking
no action	0.97	0.0052	0.0073	0.0061	0.0067	0.00061	0.00092
boxing	0.032	0.9	0.03	0.0071	0	0.0024	0.027
clapping	0.032	0.27	0.64	0.058	0.00031	0	0.00031
waving	0.0065	0.16	0.3	0.52	0.0021	0	0
jogging	0.33	0.032	0.071	0.065	0.32	0.019	0.16
running	0.39	0.062	0.03	0.031	0.24	0.21	0.04
walking	0.25	0.18	0.086	0.094	0.065	0.0065	0.32

Fig. 5: Lookback τ performance: fixed τ(left), lookback τ(middle), lookback + buffert(right)

The scores and run-times for each of the three prediction methodologies are listed in Table 1. Using a lookback τ doubles the prediction score, however it is significantly slower. As the lookback predictor is just a wrapper around a normal prediction method, it is not very efficient. The results however prove the validity of this technique, and confirm the benefits of investing effort in optimizing this algorithm.

Table 1: Classifier performance comparison.

classifier-type	score (%)	run-time (sec)
mahalanobis dist [4]	$\tilde{8}$3	n/a
fixed-tau	31.5	0.137
lookback-tau	59.1	162.116
loobback-tau w/ freq	59.7	160.04

Table 2: Distribution of training and testing data classes

Category type	no action	boxing	clapping	waving	jogging	running	walking
Training	11802	6867	6652	8768	3428	2268	5814
Testing	3269	3387	3277	4288	2213	1663	3058

The confusion matrices in Fig. 5 show that boxing/clapping/waving do noticeably better than walking/jogging/running. The distribution of data-points shown in Table 2 reveals the training data used is not balanced. Videos with walking/jogging/running have many frames where the subject was out frame, as compared to the other actions. Balancing this data prior to classifier training is an area of enhancement and can provide improvements in the overall classification results.

4.4 Labeling videos

Testing the classifier against a video of images yields mixed results as shown in Fig 6. Frame differencing and a linear *tau* decay present phantom actions in the MHI images that throw the classifier off. One area of improvement is to use a more aggressive exponential decay as seen in [6]. The effects of this are seen primarily at the splice location of two different videos. The old action and new action mix to form a blob that falls under no particular action. As the classifier is trained heavily with boxing and waving data points, it tends to produce those labels for these cases. Using a buffer for prediction in the video also introduced a prediction delay as the buffer fills and overtakes the previous action in terms of frequency.

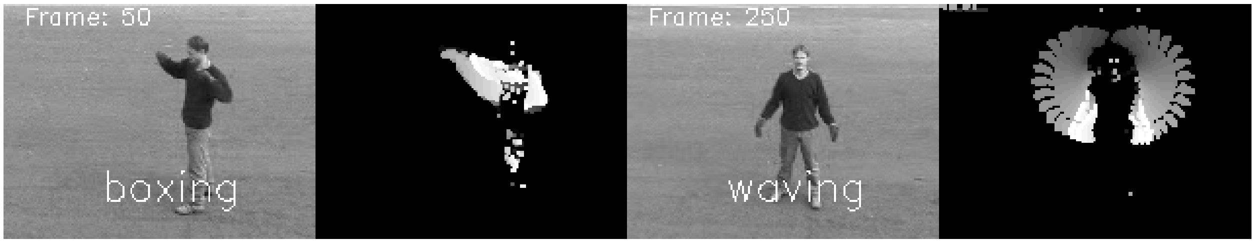


Fig. 6: Output of video processing showing label frame and equivalent MHI representation

References

1. Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., Li, Z.: A review on human activity recognition using vision-based method. Journal of Healthcare Engineering **2017** (07 2017) 1–31
2. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and AI **2** (2015) 28
3. Ahad, M.A.R., Jie, T., Kim, H., Ishikawa, S.: Motion history image: Its variants and applications. Machine Vision and Applications **23** (03 2010) 255–281
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3) (2001) 257–267
5. Bobick , A., Davis , J.: An appearance-based representation of action. In: Proceedings of 13th International Conference on Pattern Recognition. Volume 1. (1996) 307–312 vol.1
6. Alp, E.C., Keles, H.Y.: Action recognition using mhi based hu moments with hmms. In: IEEE EUROCON 2017 -17th International Conference on Smart Technologies. (2017) 212–216
7. Rosales, R.: Recognition of human action using moment-based feature (1998)
8. Meng, H., Pears, N., Freeman, M.J., Bailey, C.: Motion history histograms for human action recognition. (2009)
9. Ming-Kuei Hu : Visual pattern recognition by moment invariants. IRE Transactions on Information Theory **8**(2) (1962) 179–187
10. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Volume 3. (2004) 32–36 Vol.3
11. Schuldt, L., Caputo: Recognition of human actions video database proc. icpr'04, cambridge uk (2004)