UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

UNIVERSITAT ROVIRA i VIRGILI

Artificial Vision and Pattern Recognition
Assignment No. 2

Jason Klimack

Barcelona, Master in Artificial Intelligence

December 2, 2020

Contents

1 Introduction

Image classification is an important task in the field of computer vision. The main goal of image classification is to assign a label to an image in order to categorize the set of images as belonging to specific groups. For example, one possible task is to classify the scene in an image as belonging to one of the following categories: building, nature, vehicle, or person.

In this report, I perform image classification on a set of still images representing different human actions (figure 1 shows example images). The goal of the model is to be able to correctly classify the human action that is being performed in the images. The different categories/actions are:

1. Interacting with computer
2. Photographing
3. Playing Instrument
4. Riding Bike
5. Riding Horse
6. Running
7. Walking

The main components of the human action classification algorithm include: preprocessing the images, extracting features from the images, and finally, training a classifier model using the extracted features as input. In section 2, I explain briefly different approaches that can be performed for each of these sections, as well as discuss the results of the experiments that I performed.

2 Experiments

In this section, I discuss the experiments that I performed, as well as the results of each experiment. I performed an ablation study for both the preprocessing of the input images, as well as the feature extraction step.

All of the code was implemented in Matlab version R2018b, and run on a PC with an Intel Core i7-5500 processor that runs at 2.4GHz with 8GB of RAM.

2.1 Preprocessing

In this section, I compare a select number of different image preprocessing techniques in order to determine which method provides the best results for feature extraction and image classification. To do this, I fix the feature extraction and classifier over all of the experiments in this section, so that the results correctly capture the behaviour of the preprocessing step alone. HOG features will be used for the feature extraction, and an SVM will be used as the classifier. Table 1 shows the results.

First, I begin by resizing all of the images so that they all have the exact same dimensions. As the HOG feature extraction method divides the image into sub-regions with a specified cell size, together we find the best HOG cell size with the associated image size. As shown in the first section of table 1, the best image size was 256x256 while using HOG cell size of 12x12, which had an accuracy of 0.46. Figure 2.b shows the results of converting the original image into a grayscale image, followed

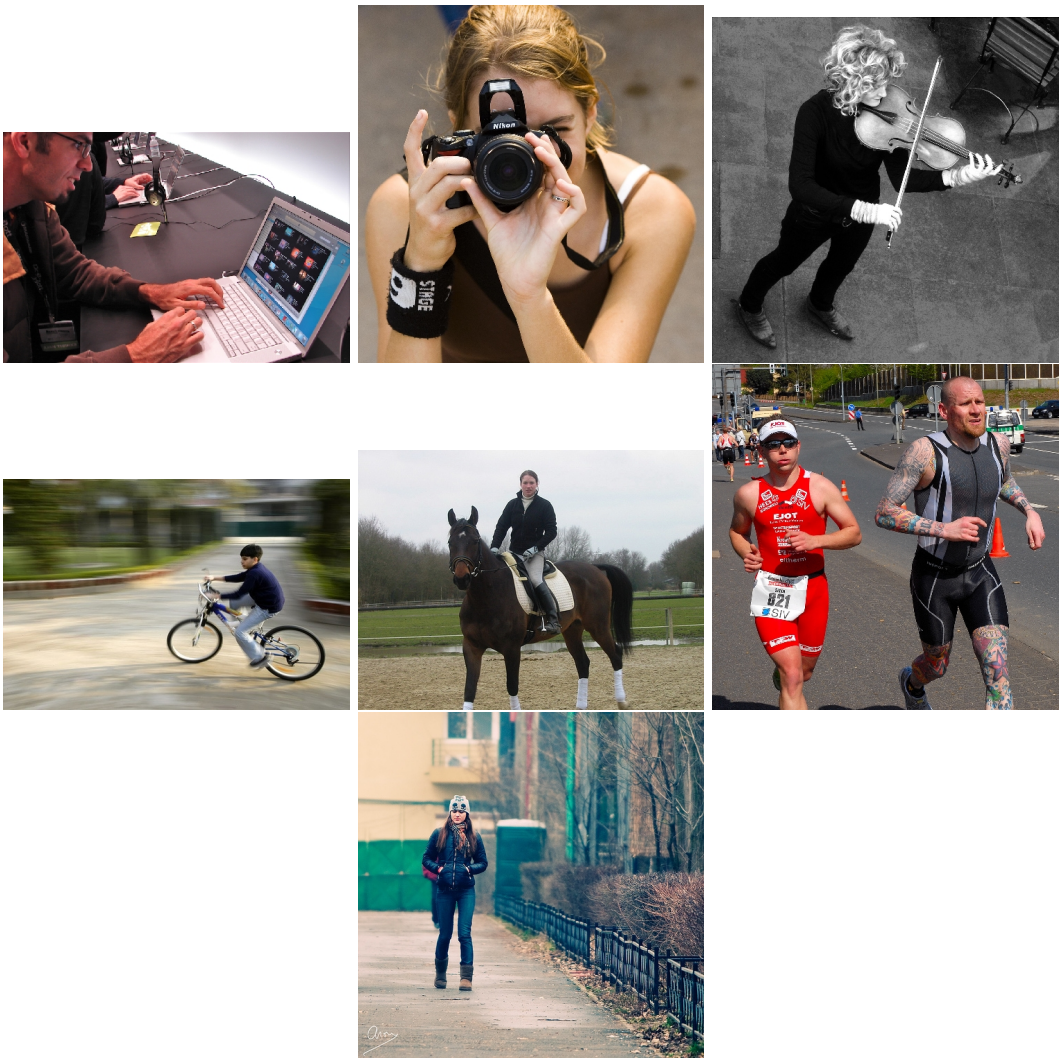

Figure 1: Example images from each of the different categories. In left-right, top-bottom order, the categories of the images shown are: Interacting with computer, Photographing, Playing Instrument, Riding Bike, Riding Horse, Running, Walking.

HOG cell	Image Size	Filter	HE	Accuracy
12x12	512x512	None	None	0.37
12x12	256x256	None	None	0.46
12x12	128x128	None	None	0.40
12x12	64x64	None	None	0.40
8x8	512x512	None	None	0.39
8x8	256x256	None	None	0.41
8x8	128x128	None	None	0.43
8x8	64x64	None	None	0.39
4x4	512x512	None	None	0.36
4x4	256x256	None	None	0.39
4x4	128x128	None	None	0.36
4x4	64x64	None	None	0.43
12x12	256x256	None	YES	0.43
12x12	256x256	Gauss($\sigma = 2$)	None	0.43
12x12	256x256	Gauss($\sigma = 4$)	None	0.39
12x12	256x256	Median	None	0.39
12x12	256x256	Gauss($\sigma = 2$)	YES	0.39

Table 1: Results of the preprocessing ablation study.

by resizing the dimensions of the image.

The next step is to perform histogram equalization (HE) on the image, in order to determine whether the illumination changes help/hinder the classification. The second section of table 1 shows the results. As indicated, the histogram equalization had a negative impact on the classification results. Figure 2.c shows the resulting image after applying HE.

The third step of the image preprocessing experiments was to compare different image smoothing techniques on the image, to determine which had the best results, if any. To do this, I applied two different image filters: Gaussian filter, and Median filter. The Gaussian filter was applied with two different values of sigma: 2, and 4. In total, there were three trials performed for this experiment. Figure 2.d and 2.e show the effects that the two different smoothing functions have on the image. The results of the experiments are shown in the third section of table 1. Clearly, the Gaussian filter with σ=2 performed the best, with an accuracy of 0.43, but the performance is still worse than without using any filter where the accuracy of the model was 0.46.

Finally, I combine the Gaussian filter with HE in order to see the combined effect of both approaches. Figure 2.f shows the resulting image after applying these techniques, and the final row of table 1 shows the accuracy of the model as 0.39.

Taking into consideration the results of these preprocessing experiments, I will only be applying the conversion to grayscale and image resizing to each of the input images. The smoothin and HE only decreased the results of the model, hence these techniques will be omitted from the final pipeline.

2.2 Feature Extraction

There are two different feature extraction algorithms that I tested for this application: Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG). In this section, I compare the results of both algorithms applied to the human action detection problem.

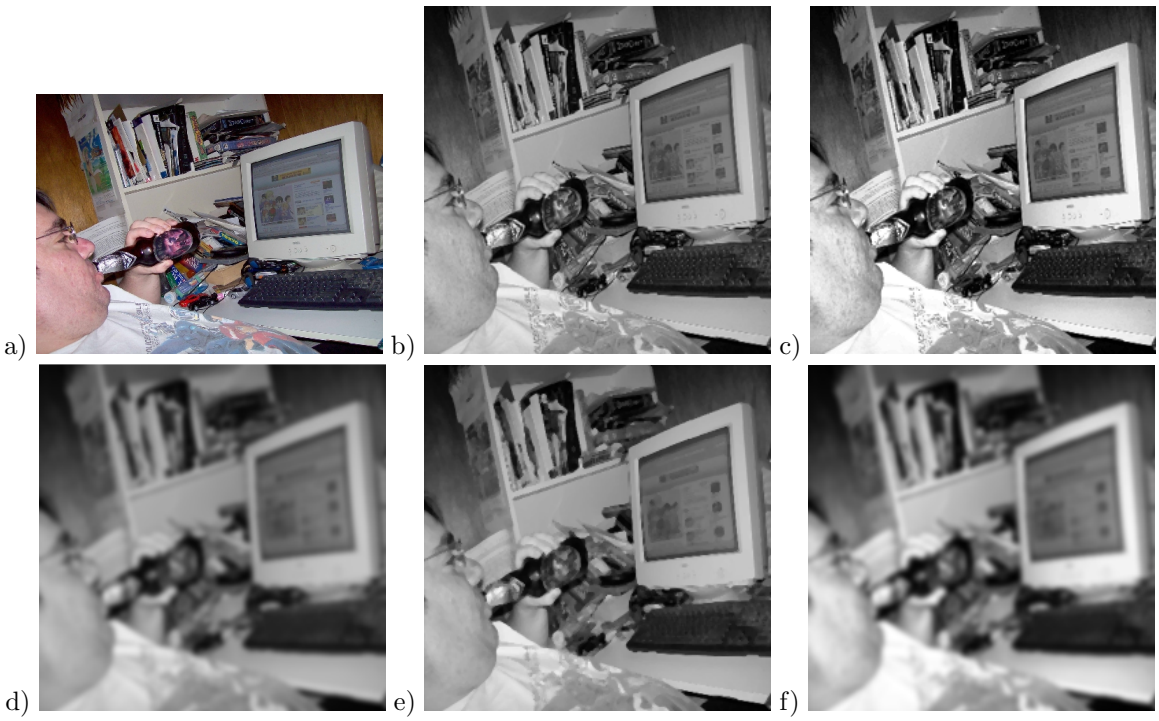

Figure 2: Different preprocessing techniques that were applied to the images. a) original image, b) grayscale and resize, c) HE, d) Gaussian filter (σ=2), e) Median filter, and f) Both HE and Gaussian filter (σ=2).

The first algorithm that I explored was the Local Binary Pattern (LBP) algorithm. It consists of determining the difference in intensities between each pixel in the image, and its neighbouring pixels. The result for each pixel is stored as a binary vector, which is then used to compute the global features for the image. The number of features used to represent the image is determined by the image size, cell size, and number of neighbours. Originally, I began the experiments with only one cell that had dimensions equal to the size of the image. The results were not very promising (around 25 %). Next, I tried different sizes of cells ranging from 8x8 to 24x24, as well as different number of neighbours ranging from 8-32. Finally, I also tried a few experiments regarding the value of the radius of the neighbourhood circle. Increasing the size of the radius had the smallest impact on the results, and the impact was always negative (decreasing the accuracy). Thus, the best model was to use a radius of 1. Overall, the best performance for LBP occurred when radius=1, number of neighbours=8, and the cell size was either 12x12 or 16x16. The results are shown in table 2.a.

The second algorithm for feature extraction from images that I used was Histogram of Oriented Gradients (HOG). HOG involves the calculation of the image gradients in both the horizontal and vertical directions, followed by dividing the image into a series of cells. For each cell, a histogram of N bins is then computed where each pixel in the cell contributes to one of the N angles calculated from the gradients, and thus is added to the bin associated with that direction. The result is that each cell represents a histogram of oriented gradients. The global feature of the image can then be calculated by flattening all of the computed histograms into one single feature vector. Thus, the number of features used to represent the image is dependent on the size of the image, the size of the cells, and the number of bins in each cell.

In the experiments that I performed for the HOG feature extraction method, I looked at different cell sizes, different sizes of blocks, and different number of bins for the histogram. The results are

shown in table 2.b, except for the different sized cells, which is shown in the first block of table 1. The different cell sizes that were tested were 4x4, 8x8, and 12x12. Clearly, the results show that 12x12 had the best performance with an accuracy of 46 percent. The different block sizes that were tested were 2x2 and 4x4. The 4x4 block size had fewer features than the 2x2, however was not as good at classifying the images (ie. lower accuracy). The number of bins that were tested were 9 and 12. When the block size was 2x2, there was no difference in the accuracy of the model simply by changing the number of bins between 9 and 12. However, there was a slight change when the block size was 4x4, where the 9 bin model outperformed the 16 bins.

a)

Algorithm	Cell Size	NumNeighbors	Radius	NumFeatures	Accuracy
LBP	size(Image)	8	1	59	0.25
LBP	size(Image)	16	1	243	0.26
LBP	8x8	8	1	60,416	0.39
LBP	12x12	8	1	26,019	0.41
LBP	12x12	16	1	107,163	0.39
LBP	12x12	16	2	107,163	0.36
LBP	16x16	8	1	15,104	0.41
LBP	24x24	8	1	5900	0.39
LBP	24x24	32	1	99,500	0.30
LBP	24x24	32	4	99,500	0.27

b)

Algorithm	Cell Size	Block Size	Num Bins	NumFeatures	Accuracy
HOG	12x12	2x2	9	14,400	0.46
HOG	12x12	4x4	9	11,664	0.43
HOG	12x12	2x2	16	25,600	0.46
HOG	12x12	4x4	16	20,736	0.40

Table 2: Results of tuning the different hyperparameters for the LBP and HOG feature extraction algorithms are shown in tables (a) and (b), respectively. HOG cell size was determined in table 1

.

2.3 Classifier

In this work, I used a support vector machine (SVM) to classify the feature vectors as belonging to one of the seven categories listed in section 1. In this section I document the results of the final model.

Thus far, I have looked at pre-processing and feature extraction. The best pre-processing technique was to simply convert the image to grayscale, and then resize the image to 256x256. The best feature extraction algorithm was to use HOG with cell size of 12x12, block size 2x2, and bin count of 9. The resulting feature vector has 14,400 dimensions, which is used as input to the SVM.

As shown in table 2.b, the accuracy of the model is 46 percent. Figure 3 shows the confusion matrix of the prediction. As you can see, a large portion of the error comes from the running class (number 6). Much of this error is a result of the running and walking categories being very similar. Another class that has lots of error is the photography category. One quarter of the images that have a ground truth of belonging to photography were classified as belonging to the running category. Likely, the stance that a photographers position themselves has resemblance to a person running, which is why these images are misclassified.

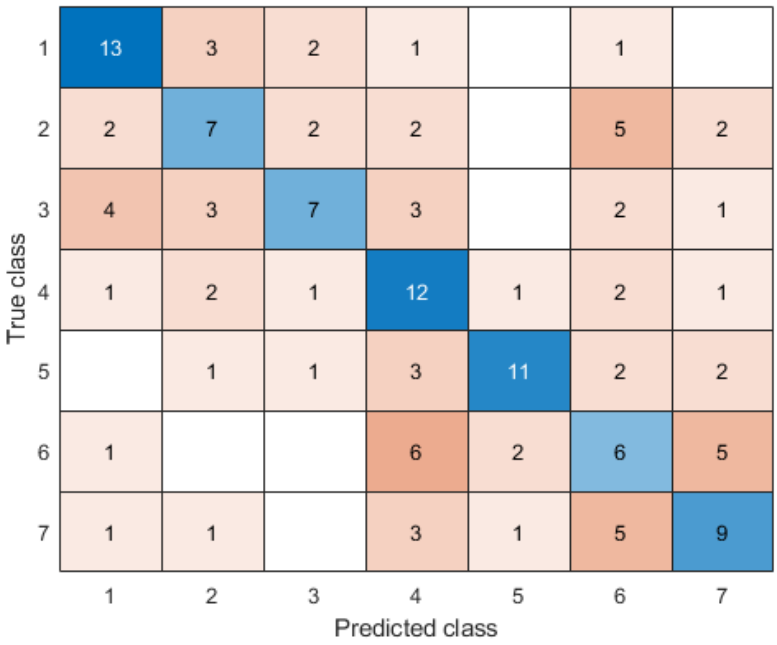

Figure 3: Confusion matrix after predicting the categories of the images in the test set on the final pipeline. The numerical values shown for the categories correspond to the associated class listed in section 1.

3 Conclusion

In conclusion, I reviewed and experimented with different techniques for performing human action classification.

Image smoothing and histogram equalization were shown to be counterproductive, as the classification accuracy decreased when these techniques were used. HOG and LBP images feature extraction techniques had similar performance, however the HOG approach had a slightly higher accuracy. The final classifier model was trained using an SVM, with the HOG feature vectors as input for each image.

The final result of the model was poor for the application, as it only had an accuracy of 46 percent. There are a number of factors which can contribute to this performance: the dataset is simply too complex, not enough samples to sufficiently train the model, the feature extraction and preprocessing techniques need more refining, or a different classification algorithm other than SVM may provide higher results.

There have been many recent advances in deep learning using convolutional neural networks (CNN) that have high performance results in comparison to hand crafted techniques. Thus, some future work may include applying a CNN model to the human action image dataset in order to achieve higher accuracy performance.