COMP3007 COMPUTER VISION COURSEWORK

Callum Gooding - psycg2 - 14298750

The University of Nottingham

ABSTRACT

Facial recognition has been a problem space for computer vision researchers since the little-published work of Woody Bledsoe in 1964. It is an example of a task performed effortlessly by humans who have evolved an innate affinity for detecting and differentiating between faces. However, it presents many challenges to computers due to the high variability of facial images due to head tilt, facial expression, and differences in lighting.

1. INTRODUCTION

This report presents a baseline method, along with two additional methods, for classifying images of faces given a set of training images representing each class. The set of training images contains only one image per class, all images are 600 by 600 pixels and some of the images are full colour and some are grayscale.

The methods are, broadly, facial similarity algorithms in that they take pairs of images of faces and provide a score of their similarity. To classify any given test image, the similarity score is found for that image and each training image. The class of the training image with the highest score is chosen as the predicted class of the test image.

2. METHODOLOGY

2.1. Baseline method

This method uses the full size of image, only grayscaled, as a feature vector. These vectors are then zero-mean normalised. To compare these vectors to determine a similarity score, the dot product is found between each training image and each test image. The label of the training image with the highest dot product with each test image is used as the label for the test image.

2.2. Additional method 1

This method uses Histograms of Oriented Gradient (HOG) descriptors to represent each image. These descriptors are used by Dalal and Triggs [1] to perform human detection. Additionally, a Gaussian filter is passed over each image before feature extraction in order to remove high frequency noise. To compare features after extraction, the dot product is found in the same fashion as the baseline method.

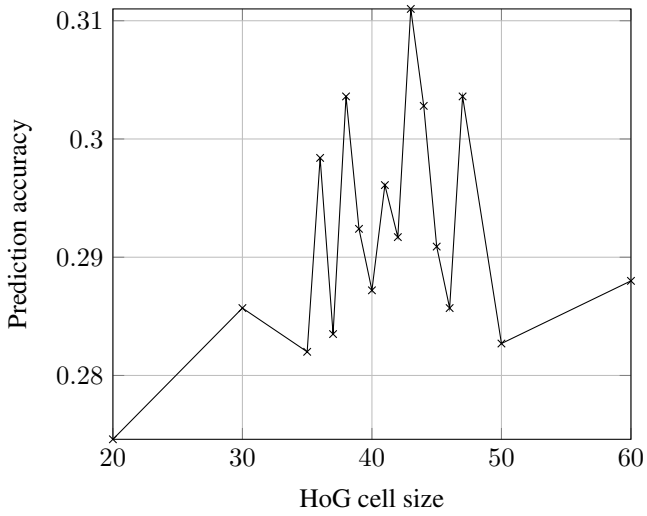

Figure 1: Graph showing HoG cell size against prediction accuracy

2.2.1. Cell size selection

An optimal cell size ensures detection of features at a scale relevant only to facial features. Figure 1 plots how the prediction accuracy changes with changes in HoG cell size. Note all input images are first passed through a Gaussian filter where $\sigma = 0.5$. As a result of this, a cell size of 43 was chosen as it gives a prediction accuracy of 31.1%. Figure 2 shows a visualisation of the HoG features that are compared between each image.

2.2.2. Gaussian σ value selection

In their report on discerning the presence of a hot dog [2], Banas, Jin, Lafayette and Marsh use a low pass filter before extracting HoG features. This filtering removes non-essential noise whilst preserving the relevant information. The σ value for the filter was obtained in the same fashion as the HoG cell size. Figure 3 plots this value against pre-

Figure 2: Visualisation of HoG features with a cell size of 43 overlaid on the source image

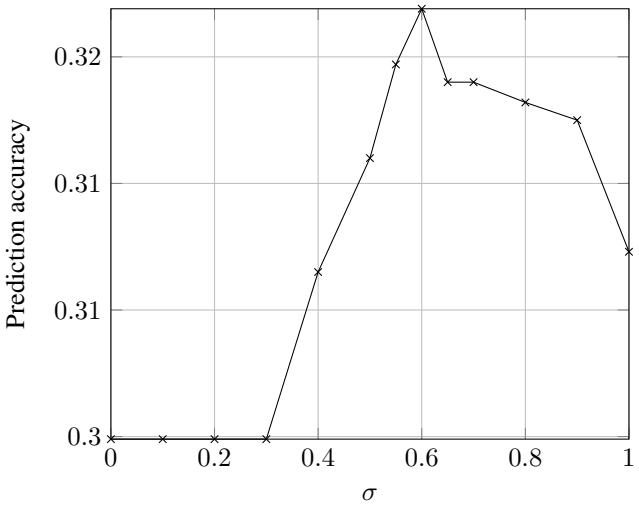

Figure 3: Graph showing Gaussian σ value against prediction accuracy

diction accuracy where the features are extracted with a cell size of 43. Where σ is shown as 0, no Gaussian filter is used. A value of 0.6 was consequently used as it results in a prediction accuracy of 32%.

2.3. Additional method 2

This method uses a neural network for feature extraction. It was trained as a Siamese network where the loss is defined as the contrastive loss between the reduced features of the input images. This results in a dimensionality reduction similar to that mentioned by Hasdell, Chopra and LeCun [3]. The network projects the high 64x64-dimensional vector representing the original image into a 10-dimensional manifold. The algorithm stores these feature vectors and uses Euclidean distance to measure the similarity between each training image and each test image.

The network was trained on the VGGFace2 dataset [4] as its feature space is similar to that of the problem set. The architecture of the network is as such:

1. Image input layer with dimensions 64x64x1 pixels
2. Convolution layer with 32 3x3 pixel filters, a stride of 1x1 and padding of 1x1x1x1
3. ReLU layer [5]
4. Convolution layer with 32 3x3 pixel filters, a stride of 1x1 and padding of 1x1x1x1
5. ReLU layer
6. Max pooling layer with a convolution size of 2x2, a stride of 2x2 and no padding
7. Convolution layer with 32 3x3 pixel filters, a stride of 1x1 and padding of 1x1x1x1
8. ReLU layer
9. Max pooling layer with a convolution size of 2x2, a stride of 2x2 and no padding
10. Convolution layer with 32 3x3 pixel filters, a stride of 1x1 and padding of 1x1x1x1
11. ReLU layer
12. Max pooling layer with a convolution size of 2x2, a stride of 2x2 and no padding
13. Fully connected layer with 1000 output neurons
14. Fully connected layer with 100 output neurons
15. Fully connected layer with 10 output neurons

The network was trained for 5000 iterations with an initial learning rate of 0.0002 which was decreased to 0.0001 for the last 2000 iterations. Imaged are pre-processed by first scaling down to 64x64 pixels, converted to grayscale and then each pixel is rescaled to lie within a range of 0 to 1.

3. METHOD EVALUATION

3.1. Baseline method

This method achieves a prediction accuracy of 25.37%. When comparing the algorithm's predictions against the truth, it can be seen that the algorithm matches high level features like noses and head position in the frame. This makes sense as these features make a large contribution to the brightness of the image in any given position due to the shadow of the node on the rest of the face and the fact that the surface of the face is generally lighter than the rest of the image. The algorithm seems more sensitive to these larger features than more localised features such as texture. As such, this algorithm best correlates image pairs that are differentiated by only changes in resolution, blurs and other effects whereas understanding of the relevant features of the face is not achieved.

3.2. Additional method 1

This method achieves a prediction accuracy of 31.69%, superior to that of the baseline method. Upon analysis of the predictions, it can be seen that this algorithm is very responsive to the overall shape of the face, which is the intended effect of the HoG features. Additionally, the outline of the face on the background is a large factor. This is understandable as this difference usually represents the most significant gradient on the image. The Gaussian blur helps with the extraction of these large features as it prevents small details from being misinterpreted as the larger gradient based features intended to be compared by the algorithm. Such small-scale details would most likely have a negative effect on the overall accuracy as many of the images in the problem set are degraded by factors such as noise and decreased resolution. These factors cannot be distinguished from relevant details such as facial texture and as such cannot be relied upon to correlate images.

Much like the baseline method, this method relies heavily on large scale features but instead of being sensitive to large areas of similar colour, this method identifies similarities in shape. Additionally, neither method can account for differences in position or rotation of the face, demonstrating that they simply identify similarities in the image, and not the faces they represent and hence cannot extract any higher-level information from these images.

3.3. Additional method 2

This method attempts to learn from the underlying issues present in the baseline and first additional methods by understanding the faces themselves. However, it achieves an accuracy of only 23.88%. The most obvious factor in that whilst the dataset used to train the model is a reasonable approximation of the problem set, the faces contained within the images take up a significantly smaller portion of the image than in the problem set as they are taken from Google Image Search and are then only loosely cropped.

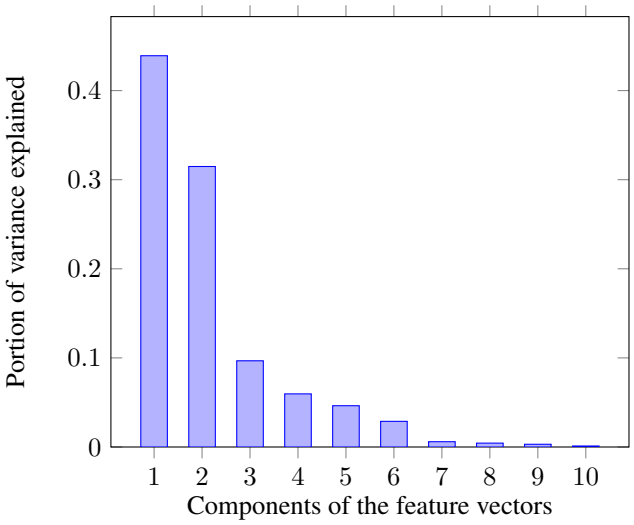

Figure 4: PCA analysis of the projections of the test data into the lower dimensional manifold

Upon inspection of the model's predictions, it can be seen that several domain specific features have been learnt such as hair style, facial structure, and facial texture. This results in many images that are "close" guesses in that the model's guesses share many higher-level features with the truth. This analysis raises the possibility that the accuracy may be greatly improved if multiple training images per class were used and hence this model's benefits would be more pronounced than those of the other methods presented.

In order to assess whether the feature space that the images are projected into it expressive enough, Principal Component Analysis [6] is used to show how much variance is explained by each component. Figure 4 shows that the model would not benefit from increasing the amount of output neurons in the final layer, increasing the dimensionality of the feature space. This would be redundant as 85% of the total variance within the manifold is explained by the three most principal components.

4. CONCLUSION

The methods presented by this report outline a promising foundation upon which to build more advanced methods. Method 1 reiterates the effectiveness of HoG feature extraction in facial comparison. Method 2 demonstrates promise in a neural network's ability to develop an understanding of higher-level features of the face and map them to a lower dimensional manifold to allow for more semantically aware

comparison.

5. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886–893 vol. 1.

[2] Joseph Lafayette Christopher Marsh Katherine Banas, Michael Jin, "Hot dog, or not dog? - an application to determine the presence of a hot dog in an image," 2018, Available at https:/gitlab.eecs.umich.edu/marshchr/notHotDog/-/blob/master/Final%20Report.pdf.

[3] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping.," in *CVPR (2)*. 2006, pp. 1735–1742, IEEE Computer Society.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[5] Vinod Nair and Geoffrey E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Madison, WI, USA, 2010, ICML'10, p. 807–814, Omnipress.

[6] Ian T Jolliffe, "Choosing a subset of principal components or variables," *Principal component analysis*, pp. 111–149, 2002.