Scene Perception using Pareidolia of Faces and Expressions of Emotion

Kenny Hong and Stephan K. Chalup
School of Electrical Engineering and Computer Science
The University of Newcastle, Australia
kenny.hong@uon.edu.au
stephan.chalup@newcastle.edu.au

Robert A.R. King
School of Mathematical and Physical Sciences
The University of Newcastle, Australia
robert.king@newcastle.edu.au

Michael J. Ostwald
School of Architecture and Built Environment
The University of Newcastle, Australia
michael.ostwald@newcastle.edu.au

Abstract—The aim of this study is to simulate the pareidolia capability of humans to produce an emotional response to a scene using analysis of facial expressions associated with abstract face-like patterns. We developed a system that uses a holistic face detector and a facial expression classifier. The $v$ and SVDD One-Class Support Vector Machines (SVM) were evaluated for creating a holistic face detector, which looks for faces that can vary from natural faces to minimal face-like patterns. A Pairwise Adaptive $C$ and $v$-SVM ($pa$-SVM) were evaluated for creating the facial expression classifier. In both scenarios, a dataset of human faces and facial expressions was used to produce a number of preprocessed images (grayscale, histogram equalised grayscale; and their respective Sobel and Canny edges) at a number of resolutions for analysis. A Gaussian and a degree two polynomial kernel were used with the SVM methods and the results were obtained using a 10 fold cross validation technique. A concern with the face detectors is verifying that they can look for minimal face-like patterns empirically. To address this concern, we created cartoon faces of the human face dataset and degraded these cartoon faces to produce an array of minimal face-like patterns. We then evaluated the face detectors and facial expression classifiers with the best model parameters on these cartoon faces. The outcome is a holistic system with the potential to describe a scene by producing an array of emotion scores corresponding to Ekman’s seven Universal Facial Expressions of Emotion.

Scene perception is defined as how we see a scene and perceive its meaning. From the literature this meaning is known as the gist and it refers mostly to our ability to categorise a scene. For instance, vague and hybrid images of a hallway, a city and a street are used to measure the speed that we can determine its category [22]. Our interest with scene perception is the idea of using faces and facial expressions to produce an emotional response. So far, facial expression and scene perception literature analyses facial expressions in the context of emotional scenes [13, 16, 27, 28]. In our previous work we hypothesised that the expressions of faces perceived in house designs through pareidolia can produce an emotional response in the observer [4, 5]. In this study, we refine our software system for pareidolia and extend our work into scene perception where a broader application can be evaluated (i.e. faces in cars, clouds, trees, etc).

A study of newborns recognising faces and facial expressions [15] implies that our brain may be hard-wired to look for faces from birth, and the use of schematic face-like patterns in the study suggests that newborns are capable of perceiving minimal face-like patterns. The idea of seeing faces from minimal face-like patterns is known as pareidolia. It is defined as the ability to perceive significance from random and vague stimuli. It is an innate ability which is commonly practised – especially for faces [31]. For instance, seeing the face of Jesus Christ on a potato chip [31] and the turf of Crystal Palace FC at London [7]; Mother Teresa on a cinnamon bun [7]; Virgin Mary on a grilled cheese sandwich [19, 31]; Satan in the smoke of 9/11 [19]; the face on Mars [19]; and a devil in the Queen’s hair of a 1954 Canadian banknote [19]. There are indications that examples of pareidolia of faces from minimal face-like patterns activate the same brain regions as real faces [9, 20].

Supporting the pareidolia effect, a study on rapid scene categorisation suggests that the presence of the target category can be perceived outside the focus of spatial attention; even when the scenes were task-irrelevant [24]. And, a recent study using noisy images and fMRI suggests that faces can be perceived with minimal face-like patterns even when these features are not centred in our visual focus [26]. The existence of brain regions that are particular for scenes and faces suggests that priority is given to processing these features. The pathways that link these brain regions and the speed at which face-like information is processed suggest a subconscious and a conscious process [6, 30]; and where a stimuli is considered emotional, the amygdala may be involved in modulating the threshold of subconsciousness and consciousness [21]. ‘How does the subconscious relate to our hypothesis?’ We hypothesise that when we perceive a scene, we may lack the ability to fully describe our emotional response to it as the answers in our conscious mind are limited by our ability to question our subconscious. The rapid rate at which the subconscious is processed suggests that it is rich in information regarding the emotion of scene perception, and the ability for a machine to relay this information allows our emotional response to a scene to become transparent. The supporting studies on pareidolia for rapid face detection and scene perception mentioned previously [24, 26] suggest a complex interplay between the conscious and the subconscious.

Several machine learning methods are available to realise
our perceptive machine. However, from the suggestion of our brain being hard-wired to look for faces from birth we hypothesised that the encoded knowledge would primarily be influenced by face data (of the positive class) rather than non-face data (of the negative class). Our hypothesis directs our attention to one-class classification methods. Two comparative studies on general tasks have shown that one-class classification using SVMs has outperformed its competitors [8, 33]. In one study where the one-class SVM did not achieve the overall best performance, it did show more robustness than other methods [25]. Narrowing our review for face detection, there is literature of one-class SVMs being used [3, 14], however, (with the exception of our previous work [4, 5]) those studies are specific to finding human faces.

In this study we investigate a number of image preprocessing techniques, resolutions and different configurations of classifiers to understand which parameters and classifiers are best to detect a range of faces from natural faces to minimal face-like patterns. We propose a two stage solution. The first uses a 10 fold cross validation to isolate the parameters and classifiers that achieve the best one-class classification accuracy. Then we produce a set of cartoon faces using our human face training data, and degrade these cartoon faces across several resolutions to produce an array of minimal face-like patterns. A window search is then carried out where a degraded cartoon face is placed in the center of a white image to isolate those parameters and classifiers that produce the best detection rate. The purpose of the degraded cartoon faces and window search is to measure the robustness of the best parameters and classifiers empirically.

Following our previous publication [4, 5], we use SVMs for creating and evaluating our expression classifiers. A Pairwise Adaptive SVM (pa-SVM) is used to evaluate a number of resolutions and image preprocessing techniques, and is compared to a standard multiclass SVMs [11]. For face detection, a one-class SVDD and ν SVM is evaluated. A Gaussian and a degree 2 polynomial kernel are used for our one class and multiclass SVMs.

A system overview is presented in the next section, followed by a description of the dataset and how the degraded cartoon faces were generated and the image preprocessing techniques that were used.

For both processes the datasets are partitioned into classes and divided into training and validation sets using a k-fold cross validation algorithm (where k is 10) – 7 classes for the expression classifiers and 1 class for the face detectors. This is then followed by obtaining the SVM and kernel grid parameters for the experiments.

Once the best parameters are determined for the face detectors, we discard from further consideration the face detectors that see faces in nothingness. Then we perform a window search on a set of test images where a degraded cartoon face is placed in the center of a white image. The cartoon faces are generated using the manually inserted feature points; and are degraded across several resolutions to produce an array of minimal face-like patterns. The detected faces from the window search are evaluated with the best expression classifiers to determine which one is optimal. Details of further steps are explained in subsequent sections.

I. SYSTEM OVERVIEW

Our system consists of several steps as illustrated in Figure 1. The first step was to align the face dataset using manually inserted feature points. Six different preprocessed images were used. Grayscale, histogram equalised grayscale, and their respective Sobel and Canny edges. Once the datasets are normalised between the values of -1 and +1 they are split into two processes. One process uses the dataset to create face detectors for detecting a range of faces, from natural faces to minimal face-like patterns. The other process uses the dataset to create expression classifiers, which are later evaluated on the detected faces.

II. DATASET AND IMAGE PREPROCESSING

In this section we describe our face dataset, how the images were aligned, how the degraded cartoon faces were generated and the image preprocessing techniques that were used.
A. Dataset preprocessing

Our face datasets were generated using the JACFEE and JACNEUF datasets which are certified by the Paul Ekman Group LLC [23]. We specifically chose this dataset because it has discrete expression classes (the 7 universal facial expressions of emotion as well as neutral faces). The dataset consists of 280 coloured face images, including 140 neutral faces and 20 images of 7 different facial expressions: happy, sad, surprised, fearful, disgusted, angry and contemptuous. We consider the neutral face separate from the 7 universal facial expressions of emotion; because of this, it was not used for expression analysis – however, we used it for face detection.

A total of 60 feature points were manually inserted for each face image along the brows, eyes and mouth – as shown in Figure 2. We use 2 feature points of the inner eye corners (a.k.a lacrimal caruncle) for aligning our face dataset. Once aligned the images are cropped and resized across 10 resolutions with a height and width of 15, 19, 24, 30, 38, 47, 59, 73, 92 and 114 pixels – roughly a 1.25 step up using 24 as the base.

The face model defined by the manually inserted feature points was proposed in our previous publication [12], which compared a component approach for classifying discrete facial expressions to a holistic approach. As mentioned the feature points of the face model are used for aligning the faces in the dataset and for creating cartoon faces which are then degraded across a number of resolutions – i.e. we take an image of \((m,m)\) pixels, resize it to a lower \((n,n)\) resolution, and then resize it back to the original resolution (this normalisation step is necessary to set a standard size).

Specifically, 16 degraded cartoon faces of 7 facial expressions were used for our array of minimal face-like patterns. At one end we start with a clear and sharp cartoon face, which is generated by averaging the \(xy\) positions of the manually inserted feature points of the 20 faces per expression, and drawing lines between these averaged \(xy\) positions. The resolution of the clear and sharp cartoon face is set at \(114 \times 114\) pixels. The array of degraded cartoon faces is produced by resizing each face to a height and width of 92, 73, 59, 47, 38, 30, 24, 19, 15, 10, 8, 6, 5, 4 and 3 pixels, then back to \(114 \times 114\) – and then placing each in the center of a white image for assessing the best parameters and best classifiers empirically.

B. Image preprocessing

We describe a number of preprocessing techniques used: the grayscale, histogram equalised grayscale; and their respective Sobel and Canny edges.

The grayscale uses an ITU-R 601-2 luma transform on RGB (red green blue) colours [1]:

\[
\text{grayscale} = \text{red} \times 0.299 + \text{green} \times 0.587 + \text{blue} \times 0.114
\]

The purpose of histogram equalisation is to stretch the intensity values of the grayscale to the full range of available intensities. It is a popular technique for improving contrast in an image [2]. The edge operators are then performed on the non-equalised and equalised grayscale.

The Sobel and Canny edge operators have several parameters that impact the final edge image. Selection of the Sobel and Canny parameters was by visual evaluation for ideal face edges. The chosen parameters are as follows: aperture size of 3 and blur of 1 for Sobel; aperture size of 3, lower threshold of 127 and upper threshold of 255 for Canny.

Figure 3 shows sample images of the 7 universal face expressions of emotion in grayscale, Sobel, and Canny. Corresponding cartoon images are also shown alongside, and Figure 4 shows a subset of the 7 cartoon faces used for the window search. We can see that at low resolutions the facial expressions become difficult to distinguish.

III. SUPPORT VECTOR MACHINES (SVM) AND KERNELS

As mentioned in the introduction, we use a SVDD and \(\nu\) one-class SVM for face detection and two pa-SVM variants of \(C\) and \(\nu\)-SVM [11] for expression analysis. Our choice of the \(C\)-SVM and \(\nu\)-SVM with a Gaussian and a degree 2 polynomial has some parameters to tune. They are \((C, \gamma)\) for the \(C\)-SVM with 2 kernels and \((\nu, \gamma)\) for the \(\nu\)-SVM with 2 kernels – where \(C\) is a tuning parameter for \(C\)-SVM, \(\nu\) is a tuning parameter for \(\nu\)-SVM; and \(\gamma\) is a tuning parameter for the Gaussian and polynomial kernel.

Specifically, we use the same \((C, \gamma)\) search parameters as [11] in our multiclass SVM experiments:

\[
C \in \{2^{-5}, 2^{-4}, ..., 2^{14}\}
\]

and

\[
\gamma \in \{2^{-17}, 2^{-16}, ..., 2^{2}\}
\]

And likewise, for the sequences of \((\nu, \gamma)\) we grow \(\nu\) by a linear step of 0.05 up to 1.0; while \(\gamma\) remains the same. i.e.

\[
\nu \in \{0.05, 0.10, ..., 0.95, 1.00\}
\]

For our one-class SVM experiments that use only \(\nu\) and \(\gamma\), we use the same \((\nu, \gamma)\) values as for the multiclass SVM experiments.
IV. NULL TEST FOR DISCARDING FACE DETECTORS THAT SEE FACES IN NOTHINGNESS

A concern with one class classification is if the boundary decision encompasses a known negative. In our case, a known negative is the detection of faces in nothingness. Two instances of this are black or white blank images; and any face detector that sees a positive face in these instances is considered undesirable and should be discarded from further processing. To discover which face detectors to discard from further processing we ran our face detectors over a series of black and white blank images. The face detectors that pass this test are used in the window search described next.

V. WINDOW SEARCH AND A BOUNDING BOX FOR ESTIMATING THE TRUE POSITIVES RATE

A concern with one-class face detectors is verifying that they can look for minimal face-like patterns empirically. To address this concern, a window search is performed on the 16 degraded cartoon faces with the 7 expressions; where each degraded cartoon face is placed at the center of a $342 \times 342$ test image – which is 3 times the size of the centred image. An evaluation procedure is undertaken by sliding a $114 \times 114$ sized window across the test image. At each step the window image is resized and preprocessed to reflect the dimension and the image preprocessing technique used for training the candidate face detector. Not all possible positions in the test image need to be searched – this is because the null test would have thrown out those face detectors that see faces in nothingness.

To measure the true positives rate we devised a number of bounding boxes around the center of the test image. This can best be explained using Figure 5. If a face is detected with its center within the red box it is counted as a true positive; if not, it is a false positive. We define the red box as a 25% bound from the center of the image – it is a good balance for the true positives rate after a pilot study comparing a 50% and 10% bound. The 100% bound is the dark green box, which considers all candidate face positive. The light coloured boxes represent the outer edges of faces detected on the boundary of the dark coloured boxes.

VI. RESULTS AND DISCUSSION

Because this investigation covers a number of image preprocessing techniques, resolutions and different configurations of classifiers, the interpretation of our results can be complex. However, as we progress through this section we do arrive with a set of best parameters and best classifiers, which have shown to be robust and have passed our empirical tests of degraded
cartoon faces. The use of degraded cartoon faces was one of many possible options. However, in using them we are able to show empirically that our face detectors are able to perceive faces and minimal face-like patterns that are different to their training data. These differences (notably sharp lines to blurred tones against a white backdrop) mean that particular image preprocessing techniques used for training the candidate face detector would produce better results.

Our presentation of results and discussion follows the structure of our system overview diagram in Figure 1. We begin by presenting the results and provide appropriate discussions with regard to face detection, followed by facial expression analysis, and we then evaluate their robustness when both are integrated. However, the foremost issue to address is the artificial edges, which are produced when the edge operators operate over histogram equalised grayscale.

A. Edge operators over histogram equalised grayscale produces artificial edges

As described in section II, the purpose of histogram equalisation is to increase the contrast of an image by stretching the intensity values of the grayscale to the full range of available intensities. By stretching the intensity values artificial edges appear, which is an undesirable effect. An illustration of these artificial edges can be seen in Figure 6 when comparing the non-equalised version of the first row with the equalised version of the second row. The edge operator parameter values remained the same in both cases.

B. Cross validation of face detection shows non-equalised grayscale as the best and most stable performer

The one-class cross validation results of Figure 7 show non-equalised grayscale as our best and most stable performer; followed by Sobel and then equalised or non-equalised Canny. To clarify, when we say equalised Canny or Sobel we mean the edge operators over histogram equalised grayscale; while non-equalised Canny or Sobel means the edge operators over non-equalised grayscale.

C. Null test shows Canny detects faces in nothingness – which is undesirable

Before we present the results of our window search, we analyse the results of our null test, which uses all the best face detectors from the one-class cross validation method in our previous section and tests them on black and white blank images. Our analysis shows that all SVM types and kernels for Canny do detect faces in nothingness. To a lesser extent, a number of non-equalised grayscale classifiers at low resolutions and a number of non-equalised Sobel classifiers at high resolutions are affected – both these classifiers use a \( \nu \)-SVM with the degree 2 polynomial kernel.

D. Window search shows Sobel detected faces to be robust

As described in section V, the purpose of a window search is to assess the robustness of the one-class face detectors. The graphs in Figure 8 show only Sobel (green) to be robust. To determine which particular classifier resolution was robust we...
examined 2 scenarios. The first was the average detection rate, followed by the number of degraded cartoon faces that were successfully detected by the classifiers. Examining the two scenarios, we isolated for non-equalised Sobel the $30 \times 30$ and $38 \times 38$ classifier as the best; for equalised Sobel the $114 \times 114$ classifier was identified as the best.

The poor performance of grayscale does not necessarily mean that it is a bad choice for a perceptive machine. Firstly, it passed the null test, and secondly (as we have mentioned) the cartoon faces are set against a white backdrop where there are no tonal shades between the eyes and mouth, which is evident in the grayscale training data; the white backdrop is also not part of the grayscale training data. Once the tonal shades are made less relevant using edge operators such as Sobel, good performance is seen as expected.

Fig. 8: Window search shows only Sobel to be robust. Using a cut-off average detection rate of 80% for the top graphs we show that for non-equalised Sobel detected faces there are two best Sobel classifier resolutions: $30 \times 30$ and $38 \times 38$; and for equalised Sobel detected faces there is only one Sobel classifier resolution that seems best: $114 \times 114$ – this is because the graph below shows the classifier detecting most of the degraded cartoon faces (identified by the vertical scale). The average detection rate in the top graphs is averaged over the rates from the number of degraded cartoon faces and the 7 expressions. For the bottom graphs, a more saturated colour tone means a higher number of expressions (of the 7) that passed our chosen cut off rate.

In the next section we present the cross validation results of our expression classifiers. Then we run the best expression classifiers over the Sobel detected faces from our window search in section VI-F to show which expression classifiers are robust.

E. Cross validation of expression analysis shows equalised grayscale as the best performer

The multiclass cross validation results of Figure 9 show equalised grayscale as the best performer; followed by Sobel then Canny. The plotted rates are based on our new pa-SVM method as they are higher than the standard multiclass SVM [11]. The clear gap between the grayscale and Sobel rates in the non-equalised face detectors of Figure 7 in section VI-B is now reflected here in the equalised results of our expression classifiers.

Similar to the point expressed in section VI-B, the accuracy rates may be high in both graphs of Figure 9, however, these accuracies do not reflect robustness. To show robustness we run the best expression classifiers over the Sobel detected faces from our window search, which is described below.

Fig. 9: Results of expression classifiers show equalised grayscale as best performer. The rates plotted use our new pa-SVM method and thus they are higher than the standard multiclass SVM.

F. Sobel expression classifiers over Sobel detected faces shows best performance

In this section we show two sets of expression classifier results over the Sobel detected faces from our window search. The first set in Figure 10 iterates over the Sobel detected faces from our window search – the rate displayed is a count of the number of candidate faces within the 25% bound where the classifier chooses the correct expression. The second set in Figure 10 serves as a relative comparison to the ground truth of how well the expression classifiers could perform if we classified over the exact face location of the Sobel detected faces (i.e. the face located on the exact centre of the test image).

An examination of these graphs shows that Sobel expression classifiers return the best performance in our test configuration. In particular, the graphs show that certain expression classifiers are sensitive to the exact face location and the resolution used for detection. This suggests that a well chosen face detector resolution is required for robust classification of expression.

G. Low resolution for detection and slightly higher resolution for expression analysis

Our window search results using non-equalised images in section VI-D show that for face detection there were no clear benefits from having a higher resolution since the best face detectors were a $30 \times 30$ and $38 \times 38$ classifier. From our literature review, most studies show improved detection rates at higher resolutions. However their higher resolutions can still be considered as very low. For instance, [32] showed improved detection rate from $6 \times 6$ to $24 \times 24$ resolution using Modified Census Transform (MCT) and Boosting on colour
images (MCT behaves similar to histogram equalisation where the aim is to compensate for lighting variations); [10] showed an improved performance from $6 \times 6$ to $24 \times 24$ resolution using an AdaBoost based face detector.

Compared to studies in literature, our study investigated resolutions much greater than those reported. It shows that moving from very low to low resolutions (e.g. $15 \times 15$ to $24 \times 24$) improves the detection rate, however moving from low to high resolutions (e.g. $24 \times 24$ to $114 \times 114$) reduces the detection rate. The idea of returning a better performance at higher resolutions can only be seen for Canny in both equalised and non-equalised face detector results (Figure 7 of section VI-B).

Similarly, our results of expression classifiers show that there was no clear benefit for having a high resolution expression classifier. This is because most of the low resolution classifiers produce the same results as the high resolution classifiers. In our literature review, most results show a reduced detection rate when compared to a higher resolution classifier. For instance, [18], [29], and [17] returned reduced performance when the pixel resolutions fell below $64 \times 64$, $110 \times 150$ and $72 \times 96$ respectively. Looking at the cross validation results for our expression classifiers (Figure 9 of section VI-E), some specific SVM types and image preprocessing techniques do exhibit this trend. However, the highest cross validation rate can be seen just at the start of the resolution axis.

H. Do the decision values reflect confidence of faces detected?

A common consensus in the SVM learning community assumes that the decision values reflect the confidence of the classifier. To verify this consensus, we gathered all the decision values from the Sobel detected faces and averaged the decision values over the number of samples. For the assumption to be true, the decision values near the centre of the test image should be the highest. The plots shown in Figure 12 agree with this assumption.

Fig. 12: Decision values of Sobel detected faces do reflect the confidence of faces detected. A more saturated colour tone is a higher decision value. The very tiny circles are decision values that are very close to zero when averaged over the number of Sobel detected faces.

I. A framework for finding the face detectors of different face types

We gathered a large sample of different types of non-human faces from the creative commons flickr community, and tested our face detectors to see which face detectors are tuned for a particular face type. A sample of these faces is shown in Figure 13; the labels indicate the face detectors that returned a positive for that image - ‘G’ = grayscale, ‘S’ = Sobel. The Canny face detectors were not tested as they failed our null test in section VI-C.

VII. Conclusion

In this paper we have introduced a new method for perceiving a scene using pareidolia of faces and expressions of emotion. The use of pareidolia of abstract faces and facial
expressions to produce an emotion was first proposed in previous work [4, 5] where this technique was applied to house designs. A concern with this idea was showing that we can empirically perceive faces and expressions. By using a set of feature points in our training data, we were able to develop a set of cartoon faces, which we degrade to provide a set of test images. Using a number of image preprocessing techniques, resolutions and different configurations of classifiers for face detection and expression analysis, we were able to show a set of parameters and classifiers that work best for our test case. Essentially, we have contributed a framework where other image preprocessing techniques and a different set of test images can be substituted such that the robustness of the face detectors and expression classifiers can be empirically shown using a null test and a window search. With this simple procedure, we can pass in an image of a scene, and using the set of best parameters and classifiers, a description of this scene is produced as an array of emotional scores corresponding to a set of facial expressions of emotion.

REFERENCES