Synthesising emotion-driven images

Konrad Cybulski

27807460

Literature Review - Semester 1, May 2019

Supervisor: Jon McCormack

Faculty of Information Technology
Monash University
Australia

Contents

1 Introduction

For centuries artists have been extremely talented at creating pieces of artwork that convey a range of emotions to those who view them. Extensive research has been conducted into how visual features affect humans emotionally and how these can be used to predict and detect the emotional content of images and text (Machajdik and Hanbury, 2010; Zhao et al., 2014). Due to the subjective and qualitative nature of human emotion, assigning a quantitative measure of emotion to an image is no easy task. Furthermore the ability to computationally recognise the emotional content of an image has wide-ranging applications from classifying posts on social media, to the creation of images, text, and even physical spaces in an emotionally quantifiable way.

This research aims to investigate quantitative representations of emotion and its effect on image emotion classification. Primarily focusing on comparing categorical (happy, sad, etc.) and continuous (valence-arousal-dominance) representations for emotion in the context of image classification, and extending the use of such a classifier in image synthesis. Through the generative process, gain a better understanding of visual patterns learnt by such a classifier by maximising classification confidence through the use of conditional generative adversarial networks (Gauthier, 2014), and evolutionary techniques explored by (Nguyen et al., 2015a).

Methods for quantifiably representing emotion have been explored thoroughly in the domain of psychology, with continuous multi-dimensional models being used in lieu of a single emotional label. The circumplex model of emotion introduced a two-dimensional space characterised by valence, and arousal; respectively representing positivity or negativity, and the level of excitement associated with it (Russell, 1980). Such a continuous model is not without flaw, failing to accurately capture more complex emotional states, often those that represent concurrent conflicting sides of a given axis (Larsen and Diener, 1992). The complexity of such a continuous representation of emotion has been investigated in depth and extended in various ways, including through the addition of a third dimension: dominance (Bradley and Lang, 1994); which is of particular interest within social dynamics.

The domain of emotion classification has had a particular focus on facial expressions and text (Cambria, 2016; Warriner et al., 2013). Image emotion classifiers have been explored (Kim et al., 2018; Machajdik and Hanbury, 2010; Chen et al., 2015, 2014) yet their use has been limited. Humans are able to fluently recognise, label, and discuss the emotional affect of an image, yet computers currently lack this ability. Ongoing research is further enabling and advancing the computational classification of image affect, and understanding the underlying patterns associated with image emotion.

As with the use of deep neural network image classifiers such as ResNet, AlexNet, and Inception, the shapes and patterns learnt by deep learning systems are difficult to extract. Recent work has looked at the use of quality-diverse generative algorithms with deep classifiers (Nguyen et al., 2015a,b). The underlying patterns learned by the classifier can be surfaced by synthesising images

that maximise various desirable features. This method allowed the exploration of images to which the classifier assigns a given label such as *school bus* or *lighthouse*, enabling a deeper understanding of the representative visual characteristics of each class. This process however does not limit itself to the use of evolutionary algorithms for image synthesis, with other methods that lend themselves to this process, including generative adversarial networks (GANs) (Goodfellow et al., 2014).

2 Aims and Scope

The primary focus of the proposed research is to produce an image emotion classifier, and to use this for emotion-driven image synthesis. Additionally to explore the patterns learnt by the classifier using techniques introduced by Nguyen et al. (2015a,b). This literature review aims to explore and understand the existing research in areas of quantitative representations of emotion, image classification techniques with a focus on neural networks, and image synthesis techniques. Furthermore to understand the impacts of emotion representation on image classification tasks, and processes for gaining a deeper understanding of the information learnt by deep classifiers depending on the representation.

3 Models of affect

Affect is a psychological term used to describe the experience or feeling of emotion. Affect is fundamentally different to both emotion and feeling. Feeling represents the label used to denote the inner processes triggered by situation and context (Shouse, 2005). The term emotion represents the label attached to the resulting display of a feeling, often attributed to actions such as facial gestures, variation in desires, and changes in brain activity (Sloman et al., 2001). Affect however represents the underlying internal and external causal relationships that result in variations of feeling and emotion or otherwise (Russell, 2003). Affect as such represents the greatly more abstract underlying processes that are linked most closely to valence: positivity or negativity; and arousal: states of excitement or lack thereof (Russell, 2003).

The field of affect classification has surged in recent years given a popularity and rising interest in the use of facial recognition, gesture, voice, and body language for emotion recognition. Some research has focused on exploring ways in which emotion can be recognised in images, text, and more abstract content. Underlying the area of classifying and recognising the affect of content are the methods for representing affect and emotion in both a meaningful, and accurate way. This section will discuss representations in the context of computational classification, in addition to their psychological underpinnings.

3.1 Affect in psychology

Given the abstract nature of affect, we often measure and describe it through introspection of emotion and feeling, or observing their outward displays and projections. As a result, shallow representations of emotion become widely used models of affect due to their simplicity. Shallow models generally involve the translation of outward displays of emotion such as facial expressions, gestures, and brain activation, to their corresponding labels of emotion: happy, sad, etc (Sloman et al., 2001). Aiming to capture more complex and deeper representations of emotion, Sloman et al. (2001) introduces the interpretation of emotion as the composition of information processing architectures. In the field of artificial intelligence, and computing as a whole, models of affect are lacking in their ability to describe complex and semantically rich interactions between actors in addition to the effect of their environment (Sloman et al., 2001).

It is important however to recognise the advantages of what are deemed to be *shallow* models of affect. Shallow models enable the definition of a broad landscape of affective descriptors despite their inability to effectively define their relationships (Sloman et al., 2001). More nuanced states of affect such as depression and embarrassment however represent states that a shallow and broad model still fails to capture (Gunes et al., 2011). Despite this, the use of categorical emotional states and labels represents the majority of research at the intersection of computing and affect (Gunes et al., 2011). More complex relationships between the processes involved with affect, feeling, and emotion are often represented as dichotomous dimensional models, such as the two-dimensional model of valence and arousal (discussed below) in which each dimension represents polar states of affect (Grandjean et al., 2008).

3.2 Categorical

The computational recognition of emotion has been explored in countless studies and projects, in both the context of image emotion recognition (Machajdik and Hanbury, 2010; Zhao et al., 2014; Kim et al., 2018) and facial emotion classification (Mollahosseini et al., 2016). However a core component and key difference between a large number of such bodies of research is the method for representing emotion. Two of the most common representations are the discrete, and continuous approach.

Discrete emotion representations generally involve the categorisation of emotion to a series of labels. Discrete approaches represent the method used in a large number of papers (Machajdik and Hanbury, 2010; Ali and Ali, 2017; Wang and Lewis, 2017; Mohammad and Kiritchenko, 2018) however the number of emotion labels used varies greatly. An example of the size of the category set used in studies of image emotion classification with this discrete model are 7 (Ali and Ali, 2017), 8 (Machajdik and Hanbury, 2010), 11 (Wang and Lewis, 2017), and even 19 (Mohammad and Kiritchenko, 2018). Due to this lack of consistency in emotion classification label sets, studies often repeat similar data gathering and classifier training methods again for their own use. Despite the

consequences of inconsistency in the emotion labels chosen, due to the simplicity of the categorical model, data gathering can be performed with ease compared to continuous methods of emotion representation. The creation of large labelled image datasets can be completed with greater ease when the number of labels is reduced. However inherent difficulty remains in labelling images with their respective affect. This difficulty may arise as a result of variation between individuals in the affect attributed to an image or a facial expression.

The categorical representation is the simplest method for emotion classification, due not only to the consequent ease with which data can be gathered, but due to the extensive research surrounding methods categorical classification. Many techniques for prediction, both for regression and classification rely on sufficiently large datasets of labelled data. It is often difficult to determine the label of a datapoint computationally, since this ability would imply that such a predictive model already exists. As a result dataset creation often involves a manual process of labelling datapoints.

Research in psychology has understood the increased error associated with continuous measurement tasked performed by humans in comparison to categorical classification (Harnad, 2003). With the human brain's attempt to sort perceived objects and situations into learned categories, it has been shown to warp continuous variables and scales to do so. As a result there exists an increased human error associated with the absolute measurements of continuous, compared to categorical variables. However such absolute error can be mitigated by introducing comparative measurements in combination with a standard ranking algorithm such as Glickman (2012, 1995). Furthermore the introduction of processes such as the self-assessment manikin (Lang, 1980) assist the assessment of emotional affect associated with the circumplex model of affect (valence, arousal, and dominance).

3.3 Continuous dimensional

The aforementioned difficulty associated with labelling images according to their respective emotional content is accentuated with the introduction of a dimensionally continuous representation of emotion. The most recognised and commonly used continuous model for emotional affect the circumplex model of emotion (Russell, 1980). The circumplex model of emotion introduced by Russell (1980) asserts that emotion can be measured in terms of two continuous variables: valence, and arousal. Valence measures the positivity or negativity associated with an emotion; and arousal measures the excitement associated with it. For example, the discrete emotion label of **happiness** could be represented in the continuous valence-arousal (VA) space with high-valence, medium-arousal; while the label **depressed** would translate to low-valence, low-arousal; and **relaxed** being medium-valence, low-arousal. While this continuous dimensional model allows for greater continuity in measuring emotions, it is not without fault (Larsen and Diener, 1992). The primary limitations of such a model pointed out by Larsen and Diener (1992) involve the likelihood of misinterpretation, particularly referring to the names of the respective dimensions. This issue of

misinterpretation was of particular interest as the circumplex model of emotional affect was being refined and compared to other models for emotional assessment. One extension on the circumplex model of affect is the introduction of a third-dimension: dominance, the amount of control associated with an emotion. The three-dimensional model of affect has been used in both the field of psychology and computing (Warriner et al., 2013; Zhao et al., 2016). Despite its limitations, both the two and three-dimensional circumplex model have been used extensively in the field of psychology (Bradley and Lang, 1994; Warriner et al., 2013), and to a limited extent in computational emotion classification (Zhao et al., 2016).

While both models for the representation of affect have trade-offs, both suffer from issues related to cross-cultural differences in emotion expression and recognition. Cultures inherently differ with respect to how emotions are both felt and expressed (Markus and Kitayama, 1991). This becomes increasingly evident when attempting to classify the affective emotion associated with more abstract content such as imagery and sound. The popular image-emotion dataset used for classification known as the International Affective Picture System (IAPS) was shown to have a significantly different valence-arousal assignment for up to 31.74% of images between Chinese and American young adults (Huang et al., 2015). Even with respect to the distribution of image valence-arousal values as assigned within a group of the same culture, these values often vary. This culture-independent variation can be reduced by introducing valence-arousal assignment guidelines and benchmark comparisons. One of the most commonly used methods for emotional affect self-assessment is the self-assessment manikin (SAM) (Lang, 1980). SAM, as can be seen in Figure 1, was developed to aid in the evaluation of emotion, particularly its translation into the commonly used three dimensions of the circumplex model of affect: valence, arousal, dominance. It has proved itself as more accurate and effective than other methods of emotion self-assessment while being less complex (Bradley and Lang, 1994).

Both discrete and continuous methods of representing emotion have their respective benefits and trade-offs. The choice to use one over another often depends on the data gathering process for the classification task, in addition to the previous research on which the task is based. These two representations of emotion can be translated between, since all categorical emotions can be converted to the continuous circumplex model, and the inverse. This translation however does introduce error since the valence-arousal-dominance model captures states of affect that fail to be captured by a broad and shallow categorical model Due to the higher complexity nature of the continuous model, even with the use of SAM, dimensional representations are more difficult to gather on a large scale in comparison to categorical emotions.

4 Image classification

Emotion classification has been researched particularly with respect to two subdomains: facial emotion recognition, and content emotion classification (image,

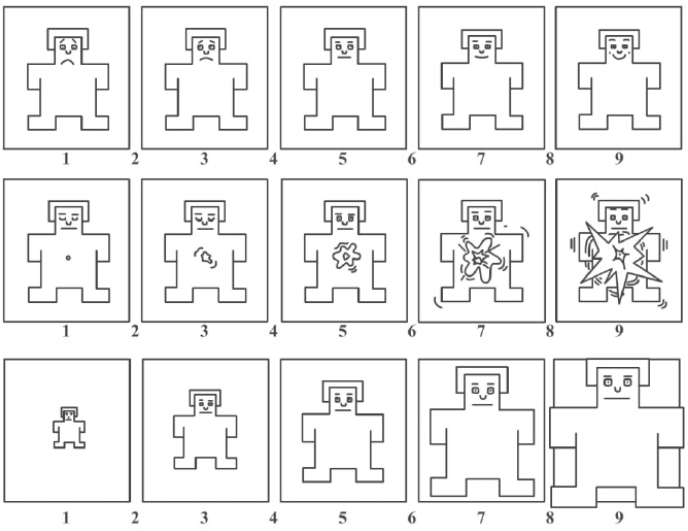

Figure 1: The self-assessment manikin: a guide for reporting individual emotional affect in the three dimensions valence (top), arousal (middle), and dominance (bottom) as introduced by Lang (1980)

text, sound). While the applications of content emotion classifications seem few in comparison to that of facial recognition, emotion classification of text, image, and sound has seen increased research interest. Image emotion classification has itself been a domain built upon the foundation of image classification techniques. From hand-crafted feature decomposition, to neural network architectures, the emotional content of images and text rely heavily on findings in other domains such as image object classification and text sentiment analysis.

4.1 Image object classification

As the basis for numerous computer vision applications, image object classification has been a primary focus of research in the field of machine learning. The techniques associated with such tasks have for the most part in recent years involved the use of deep convolutional neural networks (DCNN). DCNN have proven to be state-of-the-art classifiers, with several high performing architectures such as ResNet, AlexNet, and Inception being used as staple image classifiers in other domains (Pan and Yang, 2009). While neural networks and their use date back to the 1980s, the higher availability and lower cost of single-instruction multiple-data (SIMD) processing architectures such as graphics processing units (GPU) has enabled their widespread use (Rawat and Wang, 2017).

Before popular use of neural network image classifiers, the majority of image classification tasks lent on manually deriving image features for use in other

more traditional classification models (regression, support vector machines, etc). As a result the limiting factor of such classifiers were predictors derived from extracted image features. The primary advantage of extracted image features lies particularly in relation to such features being more interpretable and human-understandable when manually predefined. The features extracted from neural network classifiers are often abstract and may lack any human-comprehensible parallels despite their comparative predictive advantage. While it is often seen to be a black-box, the features learnt by a deep convolutional neural network, or any neural network architecture for that matter, do represent a repeatable and content-dependent image feature representation. As such the ability to reuse pre-trained neural networks and their feature extraction capabilities has been thoroughly explored in various bodies of work (Kim et al., 2018; Pan and Yang, 2009; You et al., 2015). This practice is known as transfer learning.

4.2 Neural networks

This section will give a brief overview of neural networks and their application in image and content classification. The principles underlying neural networks are simple, yet their ability to capture and learn highly complex non-linear representations of information has been key in their rise to popularity. The fundamental building block of neural networks is the neuron, which itself is a parameterised mathematical function with which a range of inputs can be mapped to a single, or multiple outputs. Examples of such mappings range from simple mathematical functions such linear or logistic regression, to more complex applications involving convolution and functions of input sequences.

Convolutional neural networks represent current state-of-the-art when using images or text for classification or regression. First introduced by Krizhevsky et al. (2012) as part of the ImageNet image classification competition, it exceeded the state-of-the-art at the time in classification accuracy. The way in which such an architecture processes image input data is through convolutional layers. As depicted in figure 2, each neuron in a convolutional layer takes as input, a spatial subset of the layer before it. The result is layers of neurons that have a greater deal of local spatial awareness when compared to the most commonly used fully connected layers. The spatial constraints associated with these layers lends itself to the domain of image processing particularly given the highly spatial relationship of image pixels.

4.3 Transfer learning

Given the large amount of work conducted into image object classification, producing neural networks such as AlexNet, ResNet, and Inception, transfer learning has been heavily relied upon in various domains of classification and synthesis of images (Pan and Yang, 2009). Each of these commonly used object classification architectures have been originally trained on common baseline image object classification datasets such as CIFAR-10, CIFAR-100, and ImageNet. These are three of the most common and widely-used image object classification

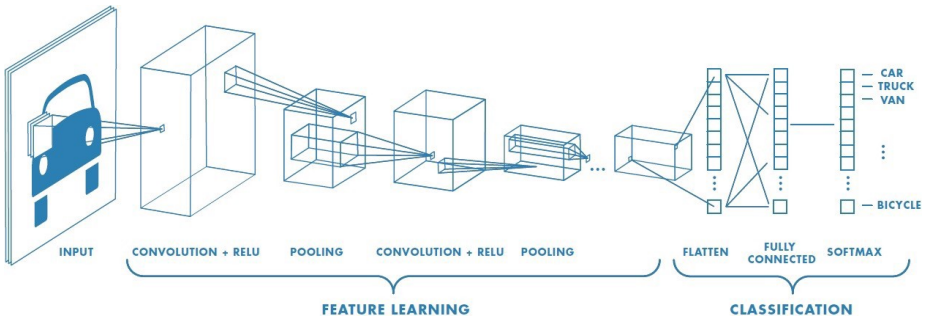

Figure 2: Visual representation of most convolutional neural network image object classifier structures as introduced by Krizhevsky et al. (2012)

datasets, in which images fall into one of 10, 100, and 1000 categories respectively. As a result, the number of outputs of neural networks trained on these datasets is the total number of predictable categories. Each of these pre-trained image classifiers can be used in other domains of machine learning by omitting the last layer of their architectures and feeding the previous layer into a new architecture. This last layer represents the object classification layer, while the previous feature layer represents a tensor of extracted image features. In doing so, the information held within the neural network can be repurposed, and the learnings reused in other domains and applications.

The use of transfer learning in numerous domains has increased the general ability to produce highly accurate networks even with comparatively small datasets. Not only do the feature layer dimensions of each architecture differs, the features they represent vary. As a result, using stacked predictor architectures, in which multiple feature vectors from multiple neural network classifiers are used as input to a further predictor network (Kim et al., 2018). Using such a stacked predictor architecture enabled Kim et al. (2018) to generate a highly accurate image emotion classifier using a relatively small dataset of 10766 images.

The subjectivity and human-dependent nature of emotion has naturally resulted in only small datasets of emotion-labelled images. Through transfer learning, pre-trained image object classifiers reduce both the computational and dataset size requirements (Pan and Yang, 2009) for creating new predictive models. Particularly in the domain of image emotion, transfer learning has enabled accurate classifiers to be built with relatively small image datasets (e.g. 10766 (Kim et al., 2018)) by extracting the image features derived by image object classifiers such as ResNet and AlexNet. Small datasets such as this however can benefit from other techniques such as data augmentation (Perez and Wang, 2017). This process involves transforming images from the dataset by rotation, cropping, blurring, colour distortions, and combinations of these. In doing so, the effective number of images available for training is increased since each augmented image is sufficiently different from the original. This technique,

in combination with transfer learning is the foundation of research by Wang and Lewis (2017) to create an emotion and theme classifier of art. Then further exploring simple variations on the transfer architecture by changing the network structure, in addition to which pre-trained image object classifier , the feature vector of VGG and ResNet object classifiers. This differs from the stacked feature extraction approach by Kim et al. (2018) with considerable success despite.

Transfer learning enables the training of classifier architectures that can base themselves on image object classifiers that have already been tirelessly trained (Krizhevsky et al., 2012; Pan and Yang, 2009). This allows classifiers to be trained in domains that suffer from a lack of data. While opportunities exist to maximise accuracy through image classifier stacking for feature extraction (Kim et al., 2018), simpler architectures involving a single image object classifier and less than three hidden layers to process these features (Wang and Lewis, 2017) have shown to provide considerably good results.

4.4 Image emotion classification

The area of image emotion and sentiment classification has been explored in a number of ways, primarily through image feature analysis derived from art and psychological factors (Machajdik and Hanbury, 2010); and more recently using techniques such as deep neural networks (Chen et al., 2015; Kim et al., 2018). Feature extraction and analysis has been used for various applications such as measuring aesthetic appeal (den Heijer and Eiben, 2010a,b, 2011) and as an emotional feature vector for sentiment classification (Machajdik and Hanbury, 2010). Due to the artistic and psychological underpinnings used by Machajdik and Hanbury (2010), the low-level features extracted from images can be understood at a high level. The relationship between an image's emotion and its core artistic components such as balance, harmony, and variety was further explored by Zhao et al. (2014). Zhao et al. (2014) used a comparably small feature vector to Machajdik and Hanbury (2010), resulting however in a 5% classification increase to state-of-the-art approaches at the time.

Deep neural networks provide less transparency to the process with which emotions and sentiment are classified compared to feature analysis. The emotional content of an image can be decomposed in various ways. Image databases with categorical emotion labels, or adjective-noun pairs (ANP) have been used for the training of deep neural network classifiers (Chen et al., 2014; Yang et al., 2018) with up to 200% performance gains over support vector machine classifiers. Compared to feature decomposition approaches used by Machajdik and Hanbury (2010), convolutional neural network architectures have become more popular and used in combination with transfer learning. This is largely due to their learning ability, particularly in recognising "hierarchical representations" (Lipton et al., 2015) of image features in a hands-off manner. This results in greatly improved classification accuracy when compared with manually crafted metrics derived through image decomposition.

The use of continuous emotion representations, particularly relating to the circumplex model have been explored in image emotion recognition tasks (Kim

et al., 2018; Zhao et al., 2016, 2017). Regression models produced to predict the valence-arousal (VA) values of given images have shown high accuracy on various datasets. In leveraging pre-trained image classification networks through transfer learning, even smaller datasets (10,000 images) can have high accuracy classification results (Kim et al., 2018). Despite the aforementioned difficulty associated with labelling images on continuous dimensions such as the circumplex model of affect (Russell, 1980), Zhao et al. (2016) have used image descriptions and comments to generate an image's respective valence, arousal, and dominance values. Zhao et al. (2016) further use this dataset with hypergraph learning techniques for the personalised prediction of an image's emotional affect. While categorical classification is more easily verified by humans, training predictive models with data that has an element of noise and uncertainty benefits from both continuity and volume. This is a key advantage of the circumplex model as applied by Kim et al. (2018) and Zhao et al. (2016) over categorical classification.

5 Computational image synthesis

Computer-generated images (CGI), have been a widely used in the film industry, as well as in countless other domains such as gaming, simulation, and art. CGI has, for the most part, involved human interaction and human-controlled image generation. Extensive research has been conducted into the generation of visually aesthetic images and methods for measuring aesthetics (den Heijer and Eiben, 2010b,a, 2011, 2014). A popular application of generative neural network architectures has been the generating text that accurately describes an image (Mathews et al., 2016), and recent work has been able to generate an image according to a target caption (Reed et al., 2016; Zhang et al., 2017).

5.1 Evolutionary computing

Some of the first methods for image generation focused on the synthesis of visually appealing images. While often using human-in-the-loop systems, visually striking and aesthetic images were the goal of methods introduced by Sims (1993) and Machado and Cardoso (2000) involving evolutionary techniques. The *NEvAr* (Machado and Cardoso, 2000) system for the interactive evolution of images exemplifies the range of visual outputs possible with these techniques. Evolutionary art leveraged methods introduced and exhibited by Sims (1993), producing images such as those shown in Figure 3. Sims (1993) proposed using *Lisp* expressions for genotype definitions, which map a coordinate (x, y) into a grayscale or RGB value. This genotype representation leveraged extensive research done into the use of evolutionary computing for optimisation problems. This genotype expression has been used in numerous further research of both supervised and unsupervised image synthesis with evolutionary techniques (Machado and Cardoso, 2000; Sims, 1993; den Heijer and Eiben, 2011, 2013; Ross et al., 2006). However the way in which NEvAr and Sims evolved im-

ages involved the manual process of selecting individuals in the population they deemed to be of higher fitness than the rest.

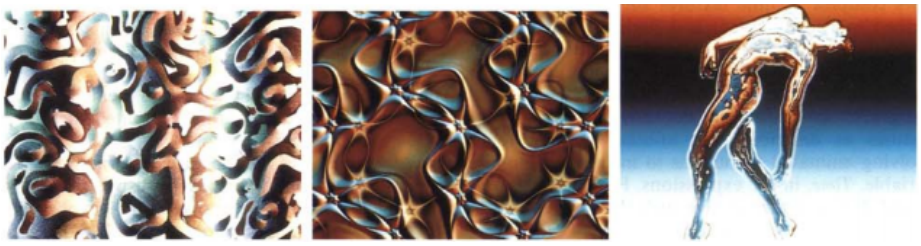

Figure 3: Images generated through the process of interactive evolution introduced by Sims (1993)

Evolutionary computing techniques for image synthesis have been able to produce increasingly interesting and appealing imagery and artwork. While the genotype representation introduced by Sims (1993) has formed the basis for numerous studies in the field of evolutionary image synthesis (den Heijer and Eiben, 2011, 2010a; Machado and Cardoso, 2000; den Heijer and Eiben, 2012, 2013), other genotype representations for image synthesis have been explored. While common generative representations have involved the use of expression trees (Sims, 1993; Machado and Cardoso, 2000; Ross et al., 2006), other methods used include direct pixel encoding (Nguyen et al., 2015a), and the more novel technique of line-drawing (Annunziato, 1998; McCormack and Bown, 2009). Line-drawing as exemplified by McCormack and Bown (2009) has built upon evolutionary techniques, using a collection of individuals interacting in real-time to synthesise artwork. In a traditional well-mixed (Sims, 1993) or distributed population (den Heijer and Eiben, 2013) each individual represents an image, or a model for generating one. The particle swarm technique employed by McCormack and Bown (2009) generates images through the interaction of actors all of whom draw on the same canvas. Despite producing visually interesting images, this real-time evolutionary image drawing mechanism has not been explored with regard to its use in other application domains.

5.2 Measures of aesthetics

Despite the slow nature of the interactive process, Sims (1993) and Machado and Cardoso (2000) were able to produce images with visually striking characteristics. Ross et al. (2006) investigated measures of aesthetics for fitness evaluation in artificially evolving images. This research primarily used observations by Ralph (2006), that the distribution of colour gradients in fine art tend toward normal. While the images produced through this method did not meet the level of intricacy and detail as the results of Sims (1993) or Machado and Cardoso (2000), it represented a self-contained system able to generate appealing imagery without human interaction.

Measures of aesthetics have been explored and multiple have been derived using information about fine art (Ralph, 2006), measures of symmetry, and even complexity measures based on image compression ratios (den Heijer and Eiben, 2010a). Work by den Heijer and Eiben (2014) performed a comparison of seven measures of aesthetics, comparing even some of the most popular metrics such as the Ralph bell curve (Ralph, 2006). The primary finding of this work showed that the visual styles of the images generated depended heavily on the given aesthetic metric used to determine fitness. The process by which images are generated is an optimisation problem, maximising the aesthetic value as measured by the given metric. It was found that combining various pairs of metrics were correlated, resulting in images with highly similar characteristics when using one metric or the other. den Heijer and Eiben (2014) also further explored the multi-objective optimisation problem of image synthesis when using combinations of aesthetic measures. The multi-objective optimisation variant of this research showed an increased aesthetic appeal of the images produced particularly with combinations of non-correlated metrics.

With goals of synthesising *interesting* images through the use of such aesthetic measures, current state-of-the-art metrics include measures of compression complexity and fractal dimension (Johnson et al., 2019). The perceived beauty of an image can be predicted with considerable accuracy using these measures of aesthetics, particularly the fractal dimension of the image (Forsythe et al., 2011). However recent works has investigated the accuracy of such metrics with respect to user-reported image complexity and aesthetic appeal (Johnson et al., 2019). Even in controlled environments, the evaluations of images made by study participants can vary substantially. This variation depends heavily on the context in which such evaluations are made, and the data gathering process can be greatly impacted when performed in more uncontrolled environments such as online.

The mathematical functions for the measure of visual aesthetic value explored by den Heijer and Eiben (2014) and Forsythe et al. (2011) can provide useful optimisation targets in applications of image synthesis. Another popular technique used in this domain of application is the use of an image collection used as a seed in the synthesis process. This collection of images, known as an *inspiring set* (Johnson et al., 2019), forms a set on which a synthesis process is run, often aiming to replicate style and form while remaining novel from the inspiring set. This approach, while omitting targets such as fractal dimension and compression complexity, uses measures of colour distribution or geometrical analysis (Johnson et al., 2019). Kim et al. (2018) utilises exactly this method in its image-to-image *emotion transfer*, looking particularly at the transference of colour distributions based on an inspiring set determined by the target emotion.

5.3 Quality-diverse algorithms

Recent work by Nguyen et al. (2015b) and Nguyen et al. (2015a) investigated the use of quality-diverse algorithms for image generation particularly to better understand the patterns learned by deep neural network image classifiers. Quality-

diverse (QD) evolutionary algorithms such as Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) (Mouret and Clune, 2015) and Novelty Search (Lehman and Stanley, 2008, 2011) have been developed to address the need for a high quality, yet diverse solution space in related optimisation domains. The type of problems QD algorithms aim to address include primarily those in which a multitude of solutions exist within a multi-objective space, however the degree to which each objective is desired may vary. Thus algorithms such as MAP-Elites aims to maximise a given fitness function, while maintaining an N-dimensional feature space, where each dimensional represents the feature-specific fitness of a solution. The MAP-Elites algorithm results in not just a single or set of high fitness solutions, but a collection of high fitness solutions spatially distributed over the desired feature space.

The use of QD algorithms has shown great promise in its efficiency and accuracy on a number of hard optimisation problems (Pugh et al., 2016) such as maze navigation (Lehman and Stanley, 2011). Nguyen et al. (2015a) and Nguyen et al. (2015b) use MAP-Elites in conjunction with a pre-trained deep neural network (DNN) image classifier; assigning individual image fitness according to the accuracy with which it is classified. Using the MAP-Elites framework in this context, each dimension of the feature-space represents a classification label. The generated images show the exploration of representative patterns and shapes learnt by the classifier. Nguyen et al. (2015a) leverages such an architecture to show the shallowness with which an image classifier recognises images. Assigning the label of *school bus* to alternating yellow and black lines is a prime example of the way such a network has learnt to differentiate one class from the others. Thus enabling exploration into the inner workings of the DNN classifier by uncovering features that maximise the separation of one label to another. In contrast, Nguyen et al. (2015b) uses the same architecture to explore the novelty-driven evolutionary path taken by generated images and the potential for such a system in the field of content synthesis. While the conclusions derived from Nguyen et al. (2015b) and Nguyen et al. (2015a) contrast greatly, the quality-diverse generative method used to understand the visual components learnt by the classifier show such an architecture's exploratory abilities. This technique for understanding the patterns learnt by such a classifier has not been explored in the context of regression.

5.4 Generative adversarial neural networks

Neural networks have been applied and researched extensively with regard to prediction and classification, as discussed in Section 4.2, and recently shown exceedingly interesting and even visually realistic results with the generative adversarial architecture. Their use is limited with regard to image synthesis using more traditional feed-forward network architectures due to the difficulty in converting such networks into image generators. Limited work has been performed using image classifiers in conjunction with evolutionary computing techniques by Nguyen et al. (2015b) as discussed previously.

The introduction of the generative adversarial network architecture (GAN)

by Goodfellow et al. (2014) allowed the process of image generation to depend only on collecting a sufficiently large dataset. Common GAN application has involved the generation of realistic images, including work by Bao et al. (2017) where images have been synthesised to fine-detailed target labels such as bird species' and actors. Zhang et al. (2017) and Reed et al. (2016) have recently explored text to image synthesis, in which detailed descriptions of birds and flowers have been converted into photo-realistic images using the GAN model. Such an architecture has been applied to the area of art synthesis by Tan et al. (2017) in which images were generated according to a target genre and artist. Learning from a dataset of countless artworks in various categories, styles, and artists, it was hugely successful in generating images that were stylistically similar to existing art of the target artist/genre. Due to the competitive relationship of the generator and discriminator networks, the patterns learned by the discriminator propagate through the generator network. The discriminator network of the GAN architecture aims to learn patterns and styles from the dataset on which it is trained in order to discriminate between the generated and existing images. As a result, the images generated tend to closely resemble those in the training dataset. This is advantageous when similarity to existing data or realism is desired, and detrimental when generative creativity is a target attribute.

The GAN architecture has been built on in various ways (Han et al., 2018). The conditional GAN introduced by Gauthier (2014) represents the base architecture of conditioned generative networks (Mathews et al., 2016; Tan et al., 2017). This method involves bases itself on the normal GAN, however with the target condition being used as additional input to both the generator and discriminator networks. Thus when the discriminator aims to determine the training image, both images *should* satisfy the target condition. The conditional GAN has formed a basis for the generation of sentiment-driven, and affect-driven content synthesis.

5.5 Affective image synthesis

The application of generative systems in the domain of affective computing, particularly with regard to emotion and content synthesis, is limited. The domain of affective content synthesis, involves the generation of text, image, sound, and other types of content according to a target emotional affect. Work in this domain has largely involved the generation of content conditioning on sentiment (Cambria, 2016; Gunes et al., 2011; Mathews et al., 2016). The task of captioning or describing an image conditioning on sentiment has often been performed using a conditional GAN architecture by (Gauthier, 2014). Sentiment is comparatively simpler to represent and classify than emotion. Often represented categorically as negative, neutral, and positive being valued -1, 0, and 1 respectively or continuously between -1 and 1 following a similar pattern (Mathews et al., 2016; Gunes et al., 2011; Zhao et al., 2016). The comparative complexity of emotional representation self-evident as the complexity of categorical classification increases, and the one-dimensional continuous model is replaced by up to three dimensions (Zhao et al., 2016).

In a similar domain is the synthesis of factual image captions according to target textual themes (Gan et al., 2017). This differed from the conditional GAN approach being used by Mathews et al. (2016), instead using a long short-term memory (LSTM) neural network model in style-driven image captioning: factual, romantic, humorous. Despite the target being simply one of three themes, each had a fundamentally different underlying emotion. The use of a categorical emotion representation is applied to affective image-to-image transformations by Ali and Ali (2017). Here Ali and Ali (2017) introduces *emotion transfer*, involving the transformation of an image's colour distribution with the aim of altering its affective emotion. The target emotional profile is represented here as proportions of seven emotions. Then performs colour transformations based on an inspirational set of images with known affect and their respective colour distributions. This differs from the practice of *style transfer* in which images are altered to stylistically match another image (Gatys et al., 2016). While style transfer aims to find a middle-ground between maintaining the content of the image and imparting the style of its target, the emotion transfer technique leaves the image content unaltered varying only its distribution of colour.

6 Conclusion

There exist numerous valuable applications of image affect and emotion classification, in addition to its use in image synthesis. While emotion and affect often represent abstract and fluid concepts, applications in computing ultimately require methods for their representation. There exist two main quantitative representations of affect most commonly used both in the field of computing and psychology. Categorical labels of expressed emotion (e.g. happy, sad, angry) represent the most commonly used representation for various reasons generally due to the ease with which labelled datasets can be produced. The second representation is dimensionally continuous, most commonly represented by the two dimensions valence and arousal, with a third, dominance, often added. When used in classification tasks, the decision to use one representation over another is often arbitrary or reflective only of past research. The deep understanding of their respective advantages and disadvantages in psychology has not been tested or understood in the domain of content affect classification.

Generative processes using neural networks or evolutionary computing have shown great promise in synthesising content. Adversarial approaches to generative architectures enable synthesis of increasingly realistic images, and stylistically accurate artworks. Such architectures tend to learn the form and style of the images shown to them during the training process. In the domain of affective content synthesis however, particularly images, there is limited work.

Such generative architectures have enabled exploration into the visual patterns learnt by image object classifiers such as ResNet and Inception. While often seen as a black box, neural networks and the information learnt by them can be better understood through such processes. This process however has not been widely used to explore other domains involving image classification.

Visual exploration provides a novel avenue for understanding the effects of data representation and modelling on the information learnt by such a network.

References

Ali, A. R. and Ali, M. (2017). Emotional filters: Automatic image transformation for inducing affect. *arXiv preprint arXiv:1707.08148*.

Annunziato, M. (1998). The nagual experiment. In *Proceedings of the First International Conference on Generative Art*, pages 241–250.

Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. (2017). Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Chen, M., Zhang, L., and Allebach, J. P. (2015). Learning deep features for image emotion classification. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4491–4495. IEEE.

Chen, T., Borth, D., Darrell, T., and Chang, S.-F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

den Heijer, E. and Eiben, A. (2010a). Using aesthetic measures to evolve art. In *IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE.

den Heijer, E. and Eiben, A. (2011). Evolving art using multiple aesthetic measures. In *European Conference on the Applications of Evolutionary Computation*, pages 234–243. Springer.

den Heijer, E. and Eiben, A. (2012). Maintaining population diversity in evolutionary art. In *International Conference on Evolutionary and Biologically Inspired Music and Art*, pages 60–71. Springer.

den Heijer, E. and Eiben, A. (2013). Maintaining population diversity in evolutionary art using structured populations. In *2013 IEEE Congress on Evolutionary Computation*, pages 529–536. IEEE.

den Heijer, E. and Eiben, A. (2014). Investigating aesthetic measures for unsupervised evolutionary art. *Swarm and Evolutionary Computation*, 16:52–68.

den Heijer, E. and Eiben, A. E. (2010b). Comparing aesthetic measures for evolutionary art. In *European Conference on the Applications of Evolutionary Computation*, pages 311–320. Springer.

Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., and Sawey, M. (2011). Predicting beauty: fractal dimension and visual complexity in art. *British journal of psychology*, 102(1):49–70.

Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Gauthier, J. (2014). Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2.

Glickman, M. E. (1995). The glicko system. *Boston University*.

Glickman, M. E. (2012). Example of the glicko-2 system. *Boston University*, pages 1–6.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Grandjean, D., Sander, D., and Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, 17(2):484–495.

Gunes, H., Schuller, B., Pantic, M., and Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and Gesture 2011*, pages 827–834. IEEE.

Han, J., Zhang, Z., Cummins, N., and Schuller, B. (2018). Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *arXiv preprint arXiv:1809.08927*.

Harnad, S. (2003). Categorical perception.

Huang, J., Xu, D., Peterson, B. S., Hu, J., Cao, L., Wei, N., Zhang, Y., Xu, W., Xu, Y., and Hu, S. (2015). Affective reactions differ between chinese and american healthy young adults: a cross-cultural study using the international affective picture system. *BMC psychiatry*, 15(1):60.

Johnson, C. G., McCormack, J., Santos, I., and Romero, J. (2019). Understanding aesthetics and fitness measures in evolutionary art systems. *Complexity*, 2019.

Kim, H.-R., Kim, Y.-S., Kim, S. J., and Lee, I.-K. (2018). Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications.

Larsen, R. J. and Diener, E. (1992). Promises and problems with the circumplex model of emotion.

Lehman, J. and Stanley, K. O. (2008). Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, pages 329–336.

Lehman, J. and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223.

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Machado, P. and Cardoso, A. (2000). Nevar–the assessment of an evolutionary art tool. In *Proceedings of the AISB00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, Birmingham, UK*, volume 456.

Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92. ACM.

Markus, H. R. and Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review*, 98(2):224.

Mathews, A. P., Xie, L., and He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI Conference on Artificial Intelligence*.

McCormack, J. and Bown, O. (2009). Life's what you make: Niche construction and evolutionary art. In *Workshops on applications of evolutionary computation*, pages 528–537. Springer.

Mohammad, S. and Kiritchenko, S. (2018). Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

Mouret, J.-B. and Clune, J. (2015). Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*.

Nguyen, A., Yosinski, J., and Clune, J. (2015a). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Nguyen, A. M., Yosinski, J., and Clune, J. (2015b). Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pages 959–966. ACM.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Pugh, J. K., Soros, L. B., and Stanley, K. O. (2016). Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40.

Ralph, W. (2006). Painting the bell curve: The occurrence of the normal distribution in fine art.

Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Ross, B. J., Ralph, W., and Zong, H. (2006). Evolutionary image synthesis using a model of aesthetics. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1087–1094. IEEE.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Shouse, E. (2005). Feeling, emotion, affect. *M/c journal*, 8(6):26.

Sims, K. (1993). Interactive evolution of equations for procedural models. *The Visual Computer*, 9(8):466–476.

Sloman, A. et al. (2001). Beyond shallow models of emotion. *Cognitive Processing*, 2(1):177–198.

Tan, W. R., Chan, C. S., Aguirre, H. E., and Tanaka, K. (2017). Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE.

Wang, Y. and Lewis, M. (2017). Arttalk: Labeling images with thematic and emotional content.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Yang, J., She, D., Sun, M., Cheng, M.-M., Rosin, P. L., and Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525.

You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915.

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., and Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56. ACM.

Zhao, S., Yao, H., Gao, Y., Ji, R., and Ding, G. (2017). Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia*, 19(3):632–645.

Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X., and Chua, T.-S. (2016). Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1385–1394. ACM.