Literature Review: Synthesising emotion-driven images

Konrad Cybulski

May 2019

1 Aims and Scope

The primary focus of the proposed research is to produce an image emotion classifier, and to further leverage it in the process of emotion-driven image synthesis. The two core themes explored are image generation techniques, and processes for classifying and generating emotional content. As a result this literature review aims to explore and understand the progression of knowledge in the fields of emotion representation and classification, computational image synthesis, and affective content synthesis. All of which represent the multiple facets of the proposed body of research with which an investigation into the use of image emotion classifiers for emotion-driven image generation will be explored.

2 Emotion representation

The field of emotion classification has surged in recent years given a popularity and rising interest in facial emotion recognition. Other research has focused on exploring ways in which emotion can be recognised in images, text, and more abstract content. Underlying these two areas of emotion recognition is the methods for quantitatively representing emotion in both a meaningful, and accurate way. This section will discuss the evolution of these aspects of emotion classification, in addition to their commonalities, and differences.

The computational recognition of emotion has been explored in countless studies and projects, in both the context of image emotion recognition [21, 42, 15] and facial emotion classification [25]. However a core component and key difference between a large number of such bodies of research is the method by which emotion is represented. Two of the most common representations represent a 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 [21, 1, 37, 24] however the number of emotion labels explored varies greatly. An example of the size of emotion label subsets used in studies of image emotion classification

using such a discrete model are 7 [1], 8 [21], 11 [37], and even 19 [24]. Due to this lack of consistency in emotion classification targets, such studies often resort to performing similar data gathering and classifier training methods. Despite the consequences of inconsistency in the emotion labels chosen, due to the simplicity of discrete categorical emotion assignment, data gathering can be performed with ease compared to continuous methods of emotion representation. In order to create large image datasets of labelled images can be completed with greater ease when the number of such labels is reduced. However there remains difficulty in labelling images with respect to emotion given the inherent variation in the emotion felt by someone when both viewing an abstract image, or accurately determining the emotion expressed by a facial expression.

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. While proposed representations vary, the most recognised basis for many continuous models is the circumplex model of emotion [33]. The circumplex model of emotion introduced by Russell [33] 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 [17]. The primary limitations of such a model involve it's likelihood of misinterpretation particularly when considering the labels of each dimension, and any information loss arising from reducing more complex emotions to a two-dimensional space. One way in which the circumplex model of affect introduced by Russell [33] has been extended to address its weaknesses is through the introduction of a third-dimension: Dominance, the amount of control associated with an emotion. Despite its limitations, the circumplex model has been used extensively in the field of psychology [3, 38], and in computational emotion classification to a limited extent [44].

While both models for the representation of emotion 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 [22]; this is becomes increasingly evident when attempting to classify the affective emotion embodied and expressed by 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 [14]. Valence-arousal values assigned to images often vary when considering a group of people. To address this potential for error in reporting the subjective emotion imparted by content, the self-assessment manikin (SAM) was introduced [16]. SAM, as can be seen in Figure 1, was developed to aid in the evaluation of emotion, particu-

larly 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 [3].

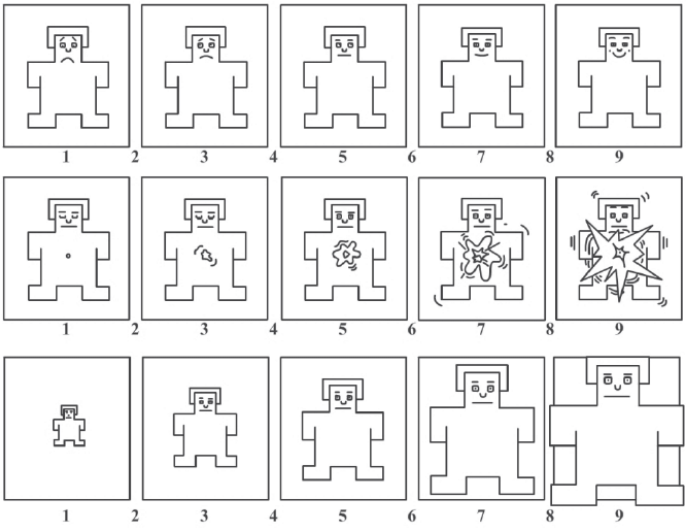

Figure 1: The self-assessment manikin: a guide for reporting individual emotional affect in the three dimensions valence (top), arousal (middle), and dominance (bottom) as introduced by Lang [16]

While 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 way in which the data gathering process for content emotion classification is performed. These two representations of emotion can be translated between, since all categorical emotions can be converted to the continuous circumplex model, and the inverse. However 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.

3 Emotion classification

Emotion classification has been researched in depth particularly with respect to two sub-domains: facial emotion recognition, and content emotion classification (image, text, sound). Facial emotion recognition has been more heavily researched in this field when compared to content emotion classification.

3.1 Facial emotion recognition

==========
==========

Why is facial emotion recognition used and is of importance? How is it done (methods)? i.e. neural networks vs feature decomposition vs computer vision etc? Any current issues ==========
==========

Humans are known to be more accurate in classifying facial emotion of another when they are culturally similar due to their in-group advantage. This was also found to be the case with computational facial emotion classifiers when trained more heavily on one culture over another [6].

3.2 Image emotion classification

==========
==========

Brief history, applications, why it's used

Methods: feature decomposition vs CV vs neural networks

Existing problems and shortcomings of current methods. ==========
==========

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 [21]; and more recently using techniques such as deep neural networks [4, 15]. Feature extraction and analysis has been used for various applications such as measuring aesthetic appeal [7, 10, 8] and as an emotional feature vector for sentiment classification [21]. Due to the artistic and psychological underpinnings used by Machajdik and Hanbury [21], 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. [42], which uses a comparably small feature vector to Machajdik and Hanbury [21], resulting however in a 5% classification increase to state-of-the-art approaches at the time.

Deep neural networks in this domain 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 singular emotion labels, and adjective-noun pairs (ANP) have been used for the training of deep neural network classifiers [5, 39] with up to 200% performance gains over support vector machine classifiers.

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. Techniques used by image emotion recognition have been used extensively in domains such as image sentiment analysis [40], and image-to-text synthesis [36]. 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 can and have been used in the domain of image emotion [15, 37].

The use of continuous emotion representations, particularly relating to the circumplex model have been explored in image emotion recognition tasks [15, 44, 43]. 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 [15]. Recent datasets produced for image emotion recognition have used the valence-arousal-dominance model due to its continuity [44]. 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. [15] and Zhao et al. [44] in image emotion recognition over categorical classification.

4 Computational image synthesis

Computer-generated images (CGI), has 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 and model generation. However generative systems and methods for both supervised, and unsupervised image synthesis have evolved in recent years with the increased use of evolutionary algorithms, and in more recent times, neural networks. While extensive research has been conducted into the generation of visually aesthetic images, applications of image synthesis extend to the synthesis of text to describe a given image [23], as well as the generation of an image according to a target caption [31, 41].

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 [34] and Machado and Cardoso [20] involving evolutionary techniques. *NEvAr* [20] was one of the first such image synthesis systems able to produce greatly impressive images through evolutionary techniques. Evolutionary art leveraged methods introduced and exemplified by Sims [34], producing images such as those shown in Figure **??**. Sims [34] 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 through evolutionary techniques [20, 34, 8, 9, 32]. However the way in which NEvAr and Sims evolved images involved the manual process of selecting individuals in the population they deemed to be of higher fitness than the rest.

Despite the slow nature of the interactive process, Sims [34] and Machado

and Cardoso [20] were able to produce images with visually striking characteristics. Ross et al. [32] investigated measures of aesthetics for fitness evaluation in artificially evolving images. This research primarily used observations by Ralph [30], 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 [34] or Machado and Cardoso [20], it represented a self-contained system able to generate appealing imagery without human interaction.

The introduction of the generative adversarial network architecture (GAN) by Goodfellow et al. [13] 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. [2] where images have been synthesised to fine-detailed target labels such as bird species' and actors. Zhang et al. [41] and Reed et al. [31] 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. [35] 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 resemble closely those in the training dataset, an advantage when similarity and realism to existing data is desired, and a detriment when generative creativity is a target attribute.

Recent work by Nguyen et al. [28] and Nguyen et al. [27] 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) [26] and Novelty Search [18, 19] have been developed to address the need for a high quality, yet diverse solution space in related optimisation domains. The use of such QD algorithms has shown great promise in its efficiency and accuracy on a number of hard optimisation problems [29] such as maze navigation [19]. Nguyen et al. [27] and Nguyen et al. [28] 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, and as such the generated images allow the exploration of label representative patterns and shapes learnt by the classifier. Nguyen et al. [27] 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. [28] 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. [28] and Nguyen et al. [27] 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 Affective content synthesis

The application of generative systems in the domain of affective computing, particularly with regard to emotion and content synthesis, is limited. Sentiment-driven examples of generative systems include image captioning according to target sentiment [23]. The task of describing an image was extended from a traditional GAN approach through the addition of a sentiment target input. The method used to train such a generative system involved the conditional GAN architecture as described by Gauthier [12]. A similar technique was used in style-driven image captioning (factual, romantic, humorous) in combination with a long short-term memory (LSTM) neural network model Gan et al. [11]. In the context of image-to-image synthesis, *emotion transfer* was explored by Ali and Ali [1] which involved the transformation of an image's colour and style with the aim of altering its conveyed emotion.

6 Conclusion

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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*.

[6] Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J., and Cottrell, G. W. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion*, 10(6):874.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

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

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

[20] 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.

[21] 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.

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

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

[24] 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)*.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] 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*.

[32] 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.

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

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

[35] 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.

[36] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

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

[38] 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.

[39] 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.

[40] 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*.

[41] 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.

[42] 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.

[43] 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.

[44] 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.