EMOTIONAL VALENCE CATEGORIZATION USING HOLISTIC IMAGE FEATURES

V. Yanulevskaya,[*] J.C. van Gemert,[*] K. Roth,[#] A.K. Herbold,[#] N. Sebe,[*] J.M. Geusebroek[*]

[*]University of Amsterdam
Informatics Institute
Amsterdam, The Netherlands

[#]University Clinic of Bonn
Department of Medical Psychology
Bonn, Germany

ABSTRACT

Can a machine learn to perceive emotions as evoked by an artwork? Here we propose an emotion categorization system, trained by ground truth from psychology studies. The training data contains emotional valences scored by human subjects on the International Affective Picture System (IAPS), a standard emotion evoking image set in psychology. Our approach is based on the assessment of local image statistics which are learned per emotional category using support vector machines. We show results for our system on the IAPS dataset, and for a collection of masterpieces. Although the results are preliminary, they demonstrate the potential of machines to elicit realistic emotions when considering masterpieces.

***Index Terms*—** Emotion categorization, scene categorization, natural image statistics.

1. INTRODUCTION

One of the main intentions of a master is to capture the scene or subject such that the final masterpiece will evoke a strong emotional response. Each stroke of an artist's brush against the canvas brings not only depth to the painting itself but also to the emotional expression that the painting will convey in the end. While the emotional feelings may vary from one person to another, there is some common ground, as many people tend to experience similar emotions when provoked by famous artworks. However, to mimic emotional feelings is a non trivial task for a machine. One could argue that emotions, among other stimuli, may be derived from visual content. Considering the current advances of scene categorization techniques, we believe that it is possible to achieve machine prediction of emotions as evoked by visual scenes for humans.

While the recognition of emotions as expressed by humans, for example by facial expressions, has matured to a state where robust emotion recognition software is available [1] the perception of emotions as evoked by visual scenes is an almost untouched area of research [2, 3]. In this paper, we consider whether we can train a visual categorization algorithm to mimic the emotions as perceived by humans when looking at master paintings, see Fig. 1. We have chosen the domain of masterpieces as the masters are well known for

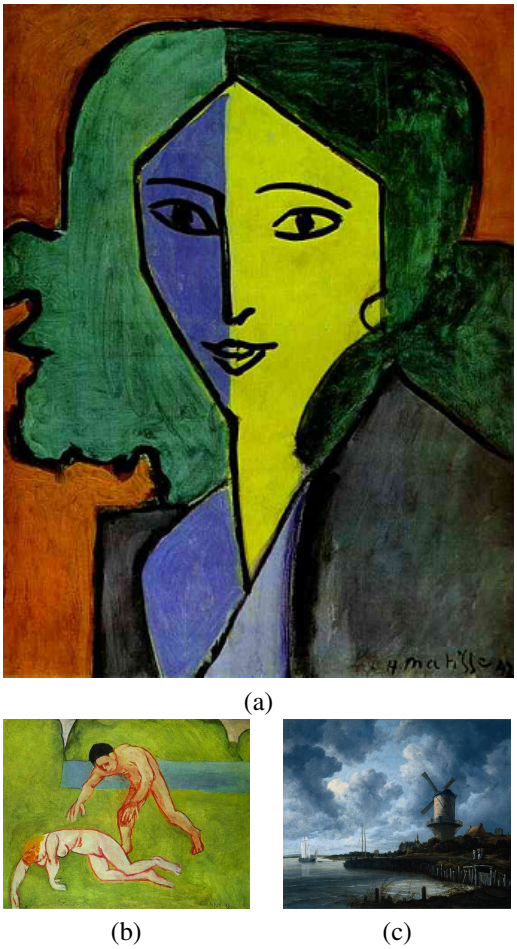

Fig. 1. Which emotions do these masterpieces evoke? Can a machine learn to perceive these emotions? (a) Portrait of Lydia Delectorskaya, the Artist's Secretary by Matisse; (b) Satyr and Nymph by Matisse, Both at the Hermitage, St. Petersburg, Russia; (c) The Mill at Wijk bij Duurstede by Jacob van Ruisdael, Rijksmuseum, Amsterdam, The Netherlands.

This work has been funded by the EU-NEST project PERCEPT.
Corresponding author: J.M. Geusebroek, mark@science.uva.nl.

their accurate and consistent arousal of emotions from their paintings.

2. MACHINE EMOTION PERCEPTION FROM IMAGES

We use scene analysis and machine learning techniques to learn to differentiate between pictures from various emotion evoking categories. The training set is the International Affective Picture System (IAPS) dataset extended with subject annotations to obtain ground truth categories.

2.1. Emotional valences ground truth

IAPS is a common stimulus set frequently used in emotion research. It consists of 716 natural colored pictures taken by professional photographers. They depict complex scenes containing objects, people, and landscapes (Fig. 2). A large data set already exists that characterizes those pictures in their reliability to elicit specific emotions. All pictures are categorized in emotional valence (positive, negative, no emotion; [4]). The images used as ground truth in our experiment, a subset of 396 of the IAPS images which are categorized in distinct emotions by Mikels et al. [5], in anger, awe, disgust, fear, sadness, excitement, contentment, and amusement, see Fig. 2. The categorization was made in two steps: A pilot study (20 subjects) with an open answering format has revealed eight frequently named types of emotions. In the main study (60 subjects) participants had to label each picture concerning these eight categories on a seven-point scale. Using this method 396 pictures were labeled either as one specific emotion or as a mixture of several emotions, see [5] for more details. Note that single pictures can belong to different emotional categories.

2.2. Holistic features from image statistics

We follow the scene categorization method put forward in van Gemert et al. [6]. This method has proven itself in realistic scenarios like the visual categorization task of TRECVID [7]. We aim to decompose complex scenes according to an annotated vocabulary. The visual words in this vocabulary provide a first step to automatic access to image content [8]. Given a fixed vocabulary, we assign a similarity score to all words for each region in an image. Different combinations of a similarity histogram of visual words provide a sufficient characterization of a complex scene.

In contrast to common codebook approaches [9, 10, 11, 12, 8], we use the similarity to all vocabulary elements [6]. A codebook approach uses the single, best matching vocabulary element to represent an image patch. For example, given a blue area, the codebook approach must choose between water and sky, leaving no room for uncertainty. Following [6], we use the similarity to all vocabulary elements. Hence, we

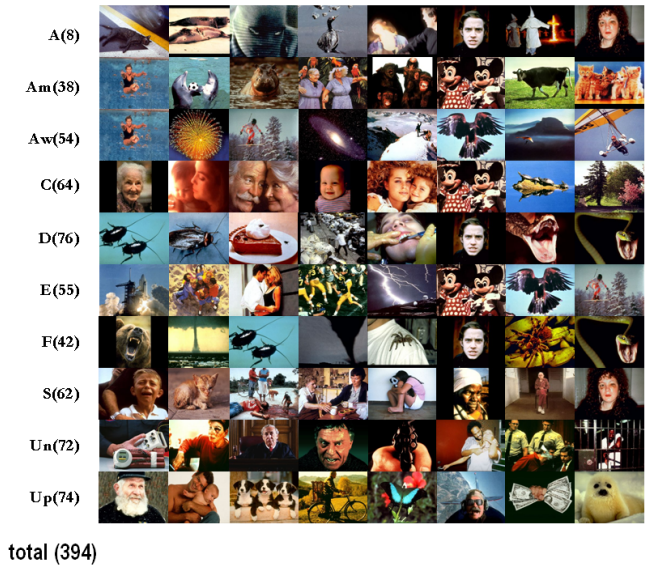

Fig. 2. International Affective Picture System (IAPS) [4]. The following emotional classes are distinguished in the dataset [5]: A - anger, Am - amusement, Aw - awe, C - contentment, D - disgust, E - excitement, F - fear, S - sadness, Un - undifferentiated negative, Up - undifferentiated positive.

model the uncertainty of assigning an image patch to each vocabulary elements. By using similarities to the whole vocabulary, our approach is able to model scenes that consist of elements not present in the codebook vocabulary.

We extract visual features for each sub-region of an image, densely sampled on a regular grid. The grid is constructed by dividing an image in $n \times n$ overlapping rectangular regions. The overlap between regions is one half of the region size. The number of regions is governed by a parameter r, that indicates the number of regions per dimension, where the two dimensions in the image are the width and height. In our experiments, we use a coarse sampling of the image with $r = 4$ and a fine sampling of the image using $r = 17$.

2.2.1. Wiccest Features

We rely on Wiccest features for image feature extraction on regular grids. Wiccest features [13] utilize natural image statistics to effectively model texture information. Texture is described by the distribution of edges in a certain image. Hence, a histogram of a Gaussian derivative filter is used to represent the edge statistics. It was shown in [14] that the complete range of image statistics in natural textures can be well modeled with an integrated Weibull distribution. This distribution is given by

$$f(r) = \frac{\gamma}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma(\frac{1}{\gamma})} \exp\left\{-\frac{1}{\gamma}\left|\frac{r-\mu}{\beta}\right|^{\gamma}\right\}, \qquad (1)$$

where r is the edge response to the Gaussian derivative filter and $\Gamma(\cdot)$ is the complete Gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-1}dt$. The parameter β denotes the width of the distribution, the parameter γ represents the 'peakness' of the distribution, and the parameter μ denotes the mode of the distribution. The position of the mode is influenced by uneven illumination and colored illumination. Hence, to achieve color constancy the values for μ is ignored.

The Wiccest features for an image region consist of the Weibull parameters for the color invariant [13] edges in the region. Thus, the β and γ values for the x-edges and y-edges of the three color channels yields a 12 dimensional descriptor. The similarity between two Wiccest features is given by the accumulated fraction between the respective β and γ parameters: $\sum \left(\frac{\min(\beta_F,\beta_G)}{\max(\beta_F,\beta_G)} \frac{\min(\gamma_F,\gamma_G)}{\max(\gamma_F,\gamma_G)} \right)$, where F and G are Wiccest features. We compute the similarity to 15 proto-concepts [6] for F and G. We divide an input frame into multiple overlapping regions, and compute for each region the similarity to 15 proto-concepts [6].

2.2.2. Gabor Features

In addition to the Wiccest features, we also rely on Gabor filters for regional image feature extraction. Gabor filters may be used to measure perceptual surface texture in an image [15]. Specifically, Gabor filters respond to regular patterns in a given orientation on a given scale and frequency. A 2D Gabor filter is given by:

$$\widetilde{G}(x,y) = G_\sigma(x,y) \exp\left\{ 2\pi i \begin{pmatrix} \Omega_{x_0} \\ \Omega_{y_0} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right\}, \quad i^2 = -1 \ , \tag{2}$$

where $G_\sigma(x,y)$ is a Gaussian with a scale σ, $\sqrt{\Omega_{x_0}^2 + \Omega_{y_0}^2}$ is the radial center frequency and $\tan^{-1}(\frac{\Omega_{y_0}}{\Omega_{x_0}})$ the orientation. Note that a zero-frequency Gabor filter reduces to a Gaussian.

In order to obtain an image region descriptor with Gabor filters we follow these three steps: 1) parameterize the Gabor filters 2) incorporate color invariance and 3) construct a histogram. First, the parameters of a Gabor filter consist of orientation, scale and frequency. We use four orientations, $0°, 45°, 90°, 135°$, and two (scale, frequency) pairs: (2.828, 0.720), (1.414, 2.094). Second, color responses are measured by filtering each color channel with a Gabor filter. The $\mathcal{W}$ color invariant is obtained by normalizing each Gabor filtered color channel by the intensity. Finally, a histogram is constructed for each Gabor filtered color channel, where we use histogram intersection as a similarity measure between histograms. Again, we divide an input frame into multiple overlapping regions, and compute for each region the similarity to 15 proto-concepts [6].

2.3. Machine learning of emotional categories

The extracted features for each ground truth image are used to train a classifier to distinguish between the various emotional valences. We use the popular Support Vector Machine (SVM) framework for supervised learning of emotion categories. Here we use the LIBSVM implementation with radial basis functions. We obtain good SVM parameter settings by using an iterative search on a large number of SVM parameter combinations. We optimize SVM C and γ parameters. Furthermore, we select the best features per class, being the Gabor features, the Wiccest features, or both. We estimate performance of all parameter and feature combinations based on 8-fold cross validation on the IAPS training set, and average 3 times to yield consistent performance indications.

3. RESULTS

In order to evaluate how well our method expresses emotions, we apply the trained system to the IAPS test set, and to a set of masterpieces from the Rijksmuseum, Amsterdam, website.

3.1. Evaluation on IAPS

Our system is trained on 70% of the images per IAPS category, and tested on the remaining 30%. As the IAPS dataset is relatively small, we repeat the training and testing 25 times. Performance is measured by the percentage of correctly classified images per category. Average performances are given in Fig. 3. Overall, the system performs a bit better than chance level (50% for one-versus-all), which can be expected for this challenging task approached with such a small set of training images. For anger, only 8 samples constituted the training set, making machine learning an undoable challenge. However, we obtain encouraging results for some categories, as can be derived from a detailed analysis of the performance. Awe and disgust can be identified by the color distribution of the input image. The emotions are linked to specific scene categories, like landscapes or insects. In future work, one could exploit the learning of smaller and more coherent sub categories to boost performance for these emotions. Similarly, sadness and undifferentiated positive are linked to textures in the scene.

3.2. Generalization to masterpieces

To see whether we can express realistic emotions for masterpieces, we use all of the IAPS images as training data, and tried the system on a set of master paintings. Figure 4 shows some typical results, and a few failure cases. Note that in the case of masterpieces the painting techniques as well as the colors chosen by the artist contribute significantly to the emotional effect and therefore our low-level features perform better than in the case of the IAPS dataset.

4. CONCLUSIONS

We have shown initial results for a scene categorization system aiming to distinct between emotional categories. Our system is trained on the IAPS dataset, and we applied it to a collection of masterpieces. Although the results are preliminary, they demonstrate the potential of machines to elicit realistic emotions as can be derived from visual scenes.

5. ACKNOWLEDGEMENTS

We are grateful to Dr. Lang for providing us with the IAPS dataset, and to Dr. Joe Mikels for providing the ground truth emotion categories. We acknowledge the Rijksmuseum, Amsterdam, and the Hermitage, St. Petersburg for their kind permission to use the artwork pictures.

6. REFERENCES

[1] R. Valenti, N. Sebe, and T. Gevers, "Facial expression recognition: A fully integrated approach," in *Int. Workshop on Visual and Multimedia Digital Libraries*, 2007.

[2] A.B. Eder, B. Hommel, and J. De Houwer, "How Distinctive is Affective Processing?," *Cognition & Emotion*, vol. 21, 2007.

[3] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE Multimedia*, vol. 6, pp. 38–53, 1999.

[4] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," Tech. Rep., Gainesville, Centre for Research in Psychophysiology, 1999.

[5] J. A. Mikels et al., "Emotional category data on images from the international affective picture system," *Behavior Research Methods*, vol. 37, pp. 626–630, 2005.

[6] J.C. van Gemert, J.M. Geusebroek, C.J. Veenman, C.G.M. Snoek, and A.W.M. Smeulders, ""robust scene categorization by learning image statistics in context"," in *CVPR Workshop on Semantic Learning Applications in Multimedia (SLAM)*, 2006.

[7] C. G. M. Snoek et al., "The MediaMill TRECVID 2005/2006/2007 semantic video search engine," in *Proceedings of the 3rd,4th,5th TRECVID Workshop*, Gaithersburg, USA, November 2005,2006,2007.

[8] J. Vogel and B. Schiele, "Natural scene retrieval based on a semantic modeling step," in *ICVR*, Dublin, Ireland, July 2004.

[9] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.

[10] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *ICCV*, 2005.

[11] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky, "Describing visual scenes using transformed dirichlet processes," in *NIPS*, 2005.

[12] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image classification for content-based indexing," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 117–130, 2001.

[13] J.M. Geusebroek, "Compact object descriptors from local colour invariant histograms," in *British Machine Vision Conference*, 2006, vol. 3, pp. 1029–1038.

[14] J. M. Geusebroek and A. W. M. Smeulders, "A six-stimulus theory for stochastic texture," *Int. J. Comput. Vision*, vol. 62, pp. 7–16, 2005.

[15] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, 1990.

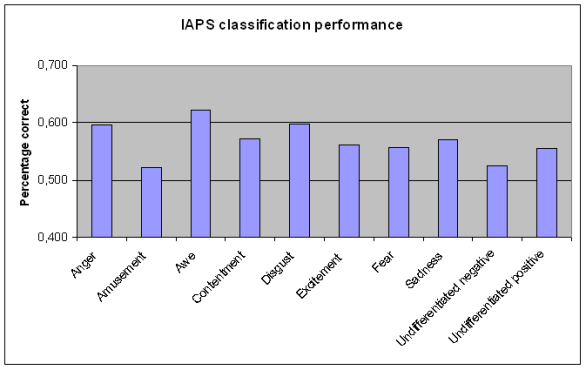

Fig. 3. Performance of emotional valence classification on the IAPS set. Average performance over 25 training repetitions using 70% of the images for training, and 30% for testing.

(a) Awe, Excitement

(b) Undifferentiated positive

(c) Undifferentiated negative, Undifferentiated positive

Fig. 4. Emotion categorization results for (a) Explosion of Dutch Gunboat by Schouman; (b) Still life with fruit and oysters by Mignon; (c) The Celebration of the Peace of Münster by Bartholomeus van der Helst. All at the Rijksmuseum, Amsterdam, The Netherlands - (c) in long term loan from the council of Amsterdam.