Predicting Emotional States of Images Using Bayesian Multiple Kernel Learning

He Zhang, Mehmet Gönen, Zhirong Yang, and Erkki Oja

Department of Information and Computer Science
Aalto University School of Science, Espoo, Finland
{he.zhang,mehmet.gonen,zhirong.yang,erkki.oja}@aalto.fi

Abstract. Images usually convey information that can influence people's emotional states. Such affective information can be used by search engines and social networks for better understanding the user's preferences. We propose here a novel Bayesian multiple kernel learning method for predicting the emotions evoked by images. The proposed method can make use of different image features simultaneously to obtain a better prediction performance, with the advantage of automatically selecting important features. Specifically, our method has been implemented within a multilabel setup in order to capture the correlations between emotions. Due to its probabilistic nature, our method is also able to produce probabilistic outputs for measuring a distribution of emotional intensities. The experimental results on the `International Affective Picture System (IAPS)` dataset show that the proposed approach achieves a bette classification performance and provides a more interpretable feature selection capability than the state-of-the-art methods.

Keywords: Image emotion, low-level image features, multiview learning, multiple kernel learning, variational approximation.

1 Introduction

Affective computing [11] aims to help people communicate, understand, and respond better to affective information such as audio, image, and video in a way that takes into account the user's emotional states. Affective image classification has attracted increasing research attention in recent years, due to the rapid expansion of the digital visual libraries on the Web. In analogy to the concept of "semantic gap" that implies the limitations of image recognition techniques, the "affective gap" can be defined as "the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal" [5].

The previous research (e.g., [9,8,13]) has focused on designing features that are specific to image affect detection, after which a general-purpose classifier such as SVM [3] is used to project an image to a certain emotional category. However, the most suitable feature representation or subset related to people's emotions is not

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 274–282, 2013.

known a priori, and feature selection has to be done first for a better prediction performance in final predictions, which increases the computational complexity. Besides, an image often evokes mixed feelings in people rather than a single one, and the ground-truth labels or emotions usually conceptually correlate with each other in the affective space. In such cases, it makes more sense to assign an image several emotional labels than a single one.

In this paper, we propose a novel Bayesian Multiple Kernel Learning (MKL) method for affective image classification using low-level color, shape and texture image features. An image can be represented by different feature representations or views. MKL combines kernels calculated on different views to obtain a better prediction performance than single-view learning methods (see [4] for a recent survey). Thanks to the MKL framework, our method can learn the image feature representation weights by itself without an explicit feature selection step, which makes the interpretation easy and straightforward. Our method has been implemented within a multilabel setup in order to capture the correlations between emotions. Due to its probabilistic nature, our method is able to produce probabilistic outputs to reflect a distribution of emotional intensities for an image. The experimental results on the `International Affective Picture System (IAPS)` dataset show that the proposed Bayesian MKL approach outperforms the state-of-the-art methods in terms of classification performance, feature selection, and result interpretation.

Section 2 introduces the image features used in this paper. Section 3 gives the mathematical details of the proposed method. In Section 4, the experimental results on affective image classification are reported. Finally, the conclusions and future work are presented in Section 5.

2 Image Features

We have used a set of ten low-level color, shape, and texture features to represent each image. The features are extracted both globally and locally. Note that the features calculated for five zones employ a tiling mask, where the image area is divided into four tiles by the two diagonals of the image, on top of which a circular center tile is overlaid [12]. Table 1 gives a summary of these features. All the features are extracted using PicSOM system [6].

Four of the features are standard MPEG-7 descriptors: Scalable Color, Dominant Color, Color Layout, and Edge Histogram. 5Zone-Color is defined as the average RGB values of all the pixels within the zone. 5Zone-Colm denotes the three central moments of HSV color distribution. Edge Fourier is calculated as the magnitude of the 16×16 FFT of Sobel edge image. 5Zone-Edgehist is the histogram of four Sobel edge directions. 5Zone-Edgecoocc is the co-occurrence matrix of four Sobel edge directions. Finally, 5Zone-Texture is defined as the histogram of relative brightness of neighboring pixels.

Table 1. The set of low-level image features used

Index	Feature	Type	Zoning	Dims.
F1	Scalable Color	Color	Global	256
F2	Dominant Color	Color	Global	6
F3	Color Layout	Color	8×8	12
F4	5Zone-Color	Color	5	15
F5	5Zone-Colm	Color	5	45
F6	Edge Histogram	Shape	4×4	80
F7	Edge Fourier	Shape	Global	128
F8	5Zone-Edgehist	Shape	5	20
F9	5Zone-Edgecoocc	Shape	5	80
F10	5Zone-Texture	Texture	5	40

3 Methods

In order to benefit from the correlation between the class labels in a multilabel learning scenario, we assume a common set of kernel weights and perform classification for all labels with these weights but using a distinct set of classification parameters for each label. This approach can also be interpreted as using a common similarity measure by sharing the kernel weights between the labels.

The notation we use throughout the manuscript is given in Table 2. The superscripts index the rows of matrices, whereas the subscripts index the columns of matrices and the entries of vectors. $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{G}(\cdot; \alpha, \beta)$ denotes the gamma distribution with the shape parameter α and the scale parameter β. $\delta(\cdot)$ denotes the Kronecker delta function that returns 1 if its argument is true and 0 otherwise.

Figure 1 illustrates the proposed probabilistic model for multilabel binary classification with a graphical model. The kernel matrices $\{\mathbf{K}_1, \ldots, \mathbf{K}_P\}$ are used to calculate intermediate outputs using the weight matrix $\mathbf{A}$. The intermediate outputs $\{\mathbf{G}_1, \ldots, \mathbf{G}_L\}$, kernel weights $\boldsymbol{e}$, and bias parameters $\boldsymbol{b}$ are used to calculate the classification scores. Finally, the given class labels $\mathbf{Y}$ are generated from the auxiliary matrix $\mathbf{F}$, which is introduced to make the inference procedures efficient [1]. We formulated a variational approximation procedure for inference in order to have a computationally efficient algorithm.

The distributional assumptions of our proposed model are defined as

$$
\begin{aligned}
\lambda_o^i &\sim \mathcal{G}(\lambda_o^i; \alpha_\lambda, \beta_\lambda) && \forall(i,o) \\
a_o^i | \lambda_o^i &\sim \mathcal{N}(a_o^i; 0, (\lambda_o^i)^{-1}) && \forall(i,o) \\
g_{o,i}^m | \boldsymbol{a}_o, \boldsymbol{k}_{m,i} &\sim \mathcal{N}(g_{o,i}^m; \boldsymbol{a}_o^\top \boldsymbol{k}_{m,i}, 1) && \forall(o,m,i) \\
\gamma_o &\sim \mathcal{G}(\gamma_o; \alpha_\gamma, \beta_\gamma) && \forall o \\
b_o | \gamma_o &\sim \mathcal{N}(b_o; 0, \gamma_o^{-1}) && \forall o \\
\omega_m &\sim \mathcal{G}(\omega_m; \alpha_\omega, \beta_\omega) && \forall m
\end{aligned}
$$

$$
\begin{aligned}
e_m|\omega_m &\sim \mathcal{N}(e_m; 0, \omega_m^{-1}) && \forall m \\
f_i^o|b_o, \boldsymbol{e}, \boldsymbol{g}_{o,i} &\sim \mathcal{N}(f_i^o; \boldsymbol{e}^\top \boldsymbol{g}_{o,i} + b_o, 1) && \forall (o,i) \\
y_i^o|f_i^o &\sim \delta(f_i^o y_i^o > \nu) && \forall (o,i)
\end{aligned}
$$

where the margin parameter ν is introduced to resolve the scaling ambiguity issue and to place a low-density region between two classes, similar to the margin idea in SVMs, which is generally used for semi-supervised learning [7]. As shorthand notations, all priors in the model are denoted by $\boldsymbol{\Xi} = \{\boldsymbol{\gamma}, \boldsymbol{\Lambda}, \boldsymbol{\omega}\}$, where the remaining variables by $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{b}, \boldsymbol{e}, \mathbf{F}, \mathbf{G}_1, \ldots, \mathbf{G}_L\}$ and the hyper-parameters by $\boldsymbol{\zeta} = \{\alpha_\gamma, \beta_\gamma, \alpha_\lambda, \beta_\lambda, \alpha_\omega, \beta_\omega\}$. Dependence on $\boldsymbol{\zeta}$ is omitted for clarity through-

Table 2. List of notation

N	Number of training instances
P	Number of kernels
L	Number of output labels
$\{\mathbf{K}_1, \ldots, \mathbf{K}_P\} \in \mathbb{R}^{N \times N}$	Kernel matrices
$\mathbf{A} \in \mathbb{R}^{N \times L}$	Weight matrix
$\boldsymbol{\Lambda} \in \mathbb{R}^{N \times L}$	Priors for weight matrix
$\{\mathbf{G}_1, \ldots, \mathbf{G}_L\} \in \mathbb{R}^{P \times N}$	Intermediate outputs
$\boldsymbol{e} \in \mathbb{R}^P$	Kernel weight vector
$\boldsymbol{\omega} \in \mathbb{R}^P$	Priors for kernel weight vector
$\boldsymbol{b} \in \mathbb{R}^L$	Bias vector
$\boldsymbol{\gamma} \in \mathbb{R}^L$	Priors for bias vector
$\mathbf{F} \in \mathbb{R}^{L \times N}$	Auxiliary matrix
$\mathbf{Y} \in \{\pm 1\}^{L \times N}$	Label matrix

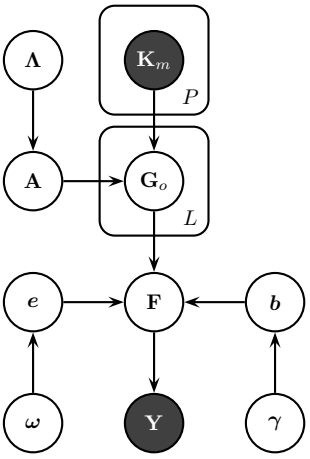


Fig. 1. Graphical model for Bayesian multilabel multiple kernel learning

out the manuscript. The variational methods use a lower bound on the marginal likelihood using an ensemble of factored posteriors to find the joint parameter distribution [2]. We can write the factorable ensemble approximation of the required posterior as

$$p(\boldsymbol{\Theta}, \boldsymbol{\Xi} | \{\mathbf{K}_m\}_{m=1}^P, \mathbf{Y}) \approx q(\boldsymbol{\Theta}, \boldsymbol{\Xi}) = q(\boldsymbol{\Lambda})q(\mathbf{A})q(\mathbf{Z})q(\{\mathbf{G}_o\}_{o=1}^L)q(\boldsymbol{\gamma})q(\boldsymbol{\omega})q(\boldsymbol{b}, \boldsymbol{e})q(\mathbf{F})$$

and define each factor in the ensemble just like its full conditional distribution. We can bound the marginal likelihood using Jensen's inequality:

$$\log p(\mathbf{Y} | \{\mathbf{K}_m\}_{m=1}^P) \geq \mathrm{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\Xi})}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi} | \{\mathbf{K}_m\}_{m=1}^P)] - \mathrm{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\Xi})}[\log q(\boldsymbol{\Theta}, \boldsymbol{\Xi})]$$

and optimize this bound by optimizing with respect to each factor separately until convergence. The approximate posterior distribution of a specific factor $\boldsymbol{\tau}$ can be found as

$$q(\boldsymbol{\tau}) \propto \exp\big(\mathrm{E}_{q(\{\boldsymbol{\Theta}, \boldsymbol{\Xi}\} \setminus \boldsymbol{\tau})}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi} | \{\mathbf{K}_m\}_{m=1}^P)]\big).$$

For our model, thanks to the conjugacy, the resulting approximate posterior distribution of each factor follows the same distribution as the corresponding factor. The exact inference details are omitted due to the space limit.

4 Experiments

In this section, we present the experimental results using our proposed Bayesian MKL method for affective image classification. We implemented our method in Matlab and took 200 variational iterations for inference with non-informative priors. We calculated the standard Gaussian kernel on each feature representation separately and picked the kernel width as $2\sqrt{D}_m$, where D_m is the dimensionality of corresponding feature representation.

4.1 Dataset and Comparison Methods

The `IAPS` dataset is a widely-used stimulus set in emotion-related studies. It contains altogether 1182 color images that cover contents across a large variety of semantic categories. A subset of 394 `IAPS` images have been grouped into 8 discrete emotional categories based on a psychophysical study [10], including Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and Sad(ness). The ground truth label for each image was selected as the category that had majority of the votes. Both Machajdik *et al.* [9] and Lu *et al.* [8] used this subset for image emotion classification, hence we used it to compare with their results in [9,8].

4.2 Experimental Setup

We used the same training and testing procedure (80% samples for training, 20% for testing) as in [9,8]: we ran 5-fold Cross-Validation (CV) and calculated the average classification accuracy. As a baseline method, the standard SVM (with Gaussian kernel and 5-fold CV) was also implemented for comparison, where each feature was taken separately for training a single classifier.

4.3 Results

Figure 2 shows the classification results. It is clear to see that our proposed approach is the best among the three. With rather generic low-level image features, our classifier can achieve very good classification performance. Note that the compared methods [9,8] utilize complicated domain-specific features.

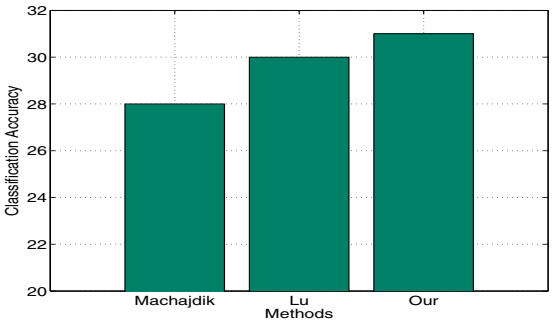


Fig. 2. The classification results of the compared methods

To further demonstrate the advantage of multiple kernel (multiview) learning over single kernel (single-view) learning, we trained and tested a single SVM classifier using each of the 10 features separately (with the same partition as MKL setup). Table 3 lists the classification accuracies. The best SVM classifier (trained with Dominant Color) can only achieve an accuracy of 0.22, which is about 9 percent lower than that of our method. And an SVM using all 10 features can give an accuracy of 0.25. This demonstrates the advantage of multiview learning over single-view learning. It also validates the strength of our proposed classifier in terms of mapping low-level image features to high-level emotional responses.

Another advantage of our MKL method is that it can select features automatically without explicit feature extraction and selection procedures. Figure 3 shows the average feature representation weights (i.e., kernel weights) in the range $[0, 1]$ based on 5-fold CV for the multiple kernel learning scenario. We clearly see that, among the ten image feature representations, Edge Histogram (F6) ranks first, followed by Scalable Color (F1), 5Zone-Colm (F5), and Edge

Table 3. The image features ranked by SVM classification accuracies

Rank	Feature	Accuracy
1	Dominant Color	0.22
2	Color Layout	0.22
3	Edge Fourier	0.22
4	5Zone-Texture	0.21
5	5Zone-Colm	0.21
6	Scalable Color	0.20
7	5Zone-Color	0.20
8	5Zone-Edgecoocc	0.20
9	5Zone-Edgehist	0.19
10	Edge Histogram	0.18

Fourier (F7) etc. This reveals that colors and edges of an image are the most informative features for emotions recognition, which is in agreement with the studies in [9] and [8]. This also shows that multiple kernel learning helps to identify the relative importances of feature representations using a common set of kernel weights.

It is worth emphasizing that an image can evoke mixed emotions instead of a single emotion. Our Bayesian classifier is capable of producing multiple probabilistic outputs simultaneously for an image, which allows us to give the image a "soft" class assignment instead of a "hard" one. This characteristic is particularly useful for detecting emotion distribution evoked by an image. Figure 4 gives some examples. One can see that the probabilistic outputs of our Bayesian classifier generally agree well with the real human votes for certain images.

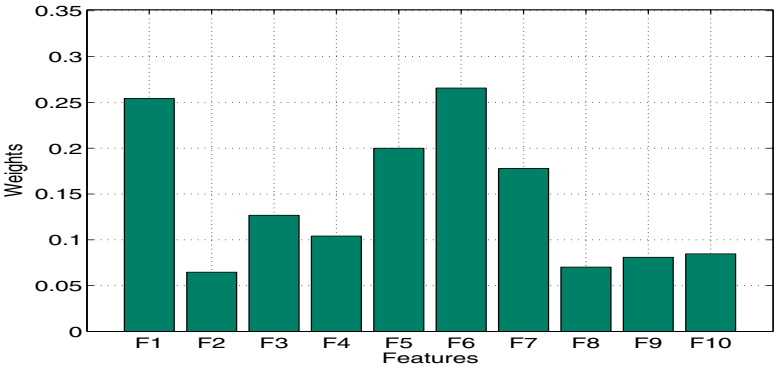
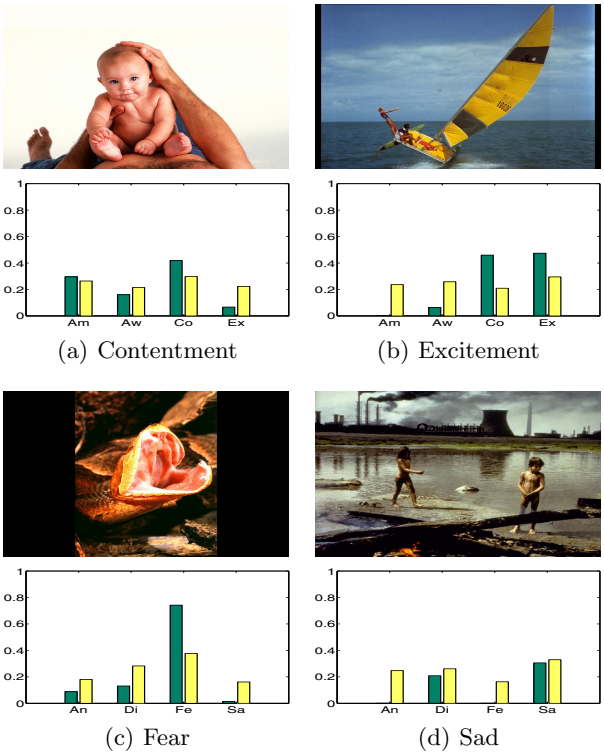


Fig. 3. The average feature representation weights over 5-fold cross-validation for the multilabel multiple kernel learning scenario



(a) Contentment (b) Excitement

(c) Fear (d) Sad

Fig. 4. The agreement of image emotion distribution between our predicted results (green bars) and the normalized human votes (yellow bars). The x-axis shows positive emotions ((a) & (b)): Amusement, Awe, Contentment, Excitement, and negative emotions ((c) & (d)) Anger, Disgust, Fear, Sad. The y-axis shows the agreement in the range $[0, 1]$.

5 Conclusions

In this paper, we have presented a novel Bayesian multiple kernel learning method for affective image classification with multiple outputs and feature representations. Instead of single feature (view) representation, our method adopts a kernel-based multiview learning approach for better prediction performance and interpretation, with the advantage of selecting or ranking features automatically. To capture the correlations between emotions, our method has been implemented within a multilabel setup. Due to its probabilistic nature, the proposed approach is able to produce probabilistic outputs for measuring the intensities of a distribution of emotions evoked by an image. More large-scale emotional datasets will be tested in the future. It is worth emphasizing that our method is not confined to the image recognition, but can be easily extended to other affective stimuli such as audio and video data.

Currently, only the conventional low-level image features are utilized, as our focus in this paper is not on the affective feature design. Rather, we would like to provide a new framework for better predicting people's emotional states, especially when an image evokes multiple affective feelings in people. Eventually, the development in this interdisciplinary area relies on the joint efforts from, for instance, artificial intelligence, computer vision, pattern recognition, cognitive science, psychology, and art theory.

Acknowledgements. This work has received funding from the Academy of Finland in the project Finnish Center of Excellence in Computational Inference Research (COIN). We gratefully acknowledge the *Center for the Study of Emotion & Attention* at University of Florida for providing the original `IAPS` image dataset.

References

1. Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88(422), 669–679 (1993)
2. Beal, M.J.: Variational Algorithms for Approximate Bayesian Inference. Ph.D. thesis, The Gatsby Computational Neuroscience Unit, University College London (2003)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
4. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268 (2011)
5. Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Magazine 23(2), 90–100 (2006)
6. Laaksonen, J., Koskela, M., Oja, E.: PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. IEEE Transactions on Neural Networks 13(4), 841–853 (2002)
7. Lawrence, N.D., Jordan, M.I.: Semi-supervised learning via Gaussian processes. In: Advances in Neural Information Processing Systems 17, pp. 753–760 (2005)
8. Lu, X., Suryanarayan, P., Adams Jr., R.B., Li, J., Newman, M.G., Wang, J.Z.: On shape and the computability of emotions. In: Proceedings of the International Conference on Multimedia, pp. 229–238 (2012)
9. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the International Conference on Multimedia, pp. 83–92 (2010)
10. Mikels, J., Fredrickson, B., Larkin, G., Lindberg, C., Maglio, S., Reuter-Lorenz, P.: Emotional category data on images from the International Affective Picture System. Behavior Research Methods 37(4), 626–630 (2005)
11. Picard, R.: Affective Computing. MIT Press (1997)
12. Sjöberg, M., Muurinen, H., Laaksonen, J., Koskela, M.: PicSOM experiments in TRECVID 2006. In: Proceedings of the TRECVID 2006 Workshop (2006)
13. Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., Alene, H.: Analyzing emotional semantics of abstract art using low-level image features. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 413–423. Springer, Heidelberg (2011)