Understanding emotional impact of images using Bayesian multiple kernel learning

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

Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Espoo, Finland

ARTICLE INFO

Article history:
Received 23 February 2014
Received in revised form
15 August 2014
Accepted 9 October 2014
Available online 4 April 2015

Keywords:
Image emotions
Multiple kernel learning
Multiview learning
Variational approximation
Low-level image features

ABSTRACT

Affective classification and retrieval of multimedia such as audio, image, and video have become emerging research areas in recent years. The previous research focused on designing features and developing feature extraction methods. Generally, a multimedia content can be represented with different feature representations (i.e., views). However, the most suitable feature representation related to people’s emotions is usually not known a priori. We propose here a novel Bayesian multiple kernel learning algorithm for affective classification and retrieval tasks. The proposed method can make use of different representations simultaneously (i.e., multiview learning) to obtain a better prediction performance than using a single feature representation (i.e., single-view learning) or a subset of features, with the advantage of automatic feature selections. In particular, our algorithm has been implemented within a multilabel setup to capture the correlation between emotions, and the Bayesian formulation enables our method to produce probabilistic outputs for measuring a set of emotions triggered by a single image. As a case study, we perform classification and retrieval experiments with our algorithm for predicting people’s emotional states evoked by images, using generic low-level image features. The empirical results with our approach on the widely-used International Affective Picture System (IAPS) data set outperform several existing methods in terms of classification performance and results interpretability.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Affective computing [1] 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. Among the emotional stimuli, affective image classification and retrieval have attracted increasing research attention in recent years, due to the rapid expansion of the digital visual libraries on the Web. While most of the current Content-Based Image Retrieval (CBIR) systems [2] are designed for recognizing objects and scenes such as plants, animals and outdoor places, an Emotional Semantic Image Retrieval (ESIR) system [3] aims at incorporating the user’s affective states to enable queries like “beautiful flowers”, “cute dogs”, “exciting games”, etc.

Though emotions are highly subjective human factors, still they have stability and generality across different people and cultures [5]. As an example, Fig. 1 shows two pictures taken from a photo sharing site [ArtPhoto [4]]. The class labels of “Amusement” and “Fear” are determined by the emotion that has received the most votes from people. Intuitively, an “Amusement” picture usually makes people feel pleasant or induces high valence, whereas a “Fear” picture may induce low valence but high arousal to the viewer.

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” [6]. Concerning the studies related to image affect recognition, three major challenges can be identified: (a) the modeling of affect, (b) the extraction of image features to reflect affective states, and (c) the building of classifiers to bridge the “affective gap”.

Most of the current works (e.g., [5,7,4,8]) use descriptive words (e.g., the scenario in Fig. 1) to represent affective space. To obtain the ground truth label for learning, each image is assigned with a single emotional label among various emotional categories based on the maximum votes from the viewers. However, an image can usually evoke a mixture of affective feelings in people rather than a single one. Furthermore, the emotions often conceptually correlate with each other in the affective space. For example, the two paintings shown in Fig. 2 are labeled as “Excitement” and “Sad (ness)” according to the maximum votes (from the web survey in [4]). Nevertheless, by examining the votes from the viewers, each image actually has evoked a distribution of emotions rather than a
single one. Moreover, the correlations can be observed between certain emotions. For example, “Amusement” is closely associated with “Excitement”, and “Fear” often comes with “Sadness”.

Feature extraction is a prerequisite step for image classification and retrieval tasks [2], especially for the recognition of emotions induced by pictures or artworks. In the literature, much effort has been spent on designing features specific to image affect recognition (e.g., [9,7,4,10,11]). Other works (e.g., [12–14,8]) used the generic low-level color, shape, and texture features for detecting the image emotions. Concerning the inference, supervised learning has been used more often than unsupervised learning for inferring the image emotions. Among the classifiers, Support Vector Machines (SVMs) have been adopted by most of the works (e.g., [13,15,7,16,8]). Since the most suitable feature representation or subset related to people’s emotions is not known a priori, feature selection has to be done for better prediction performance prior to the final prediction, which increases the computational complexity. Instead of using a single representation or view, we can also make use of different representations or views at the same time. This implies that multiview learning [17] is preferred to single-view learning. Multiview learning with kernel-based methods belongs to the framework of Multiple Kernel Learning (MKL), which is a principled way of combining kernels calculated on different views to obtain a better prediction performance than single-view learning methods (see [18] for a recent survey).

In this paper, we propose a novel Bayesian multiple kernel learning algorithm for affective classification and retrieval tasks with multiple outputs and feature representations. Thanks to the MKL framework, our method can learn the feature representation weights by itself according to the data and task at hand 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 values for measuring the intensities of a set of emotions triggered by a single image. As a case study, we conduct classification and retrieval experiments with our proposed approach for predicting people’s emotional states evoked by images, using conventional low-level color, shape, and texture image features. The experimental results on the widely-used International Affective Picture System (IAPS) data set show that our proposed Bayesian MKL approach outperforms other existing methods in terms of classification performance, feature selection capacity, and results interpretability.

Our contributions are thus two-fold:

1. Instead of single view representation, a multiview learning with kernel-based method has been applied to emotional image recognition, with the advantages of better prediction performance, automatic feature selection, and interpretation of image emotional impact.
2. A novel Bayesian multiple kernel learning algorithm with multiple outputs and feature representations has been proposed for affective classification and retrieval tasks. Our method is able to
capture the correlations between emotions and give probabilistic outputs for measuring the intensities of a distribution of emotions triggered by an image.

We start in the following section with a concise review on the related work. Section 3 gives the mathematical details of the proposed method. In Section 4, the experimental results on affective image classification and retrieval are reported. Finally, the conclusions and future work are presented in Section 5.

2. Related work

In this section, we review the works related to image affect recognition, with an emphasis on affective modeling, feature extraction, and classifier construction.

Affect has been conceptualized in psychology [19]. There are two primary ways to modeling affect: the dimensional approach and the discrete approach. The dimensional approach [20] describes affect within a 3D continuous space along Valence, Arousal, and Dominance. Valence is typically characterized as the affective states ranging from pleasant or positive to unpleasant or negative. Arousal is characterized as the state of being awake or reactive to stimuli, ranging from calm to excited. Dominance denotes power and influence over others, ranging from no control to full control. The discrete approach describes affect with a list of descriptive or adjectival words (as the example given in Fig. 2). A popular example is Ekman’s six basic emotion categories, namely, happiness, sadness, fear, anger, disgust, and surprise. Most of the current works [12,7,4,10,8] related to image affect recognition focus on recognizing the discrete emotions extended from these basic emotions. For example, positive emotions may include amusement, awe, contentment, and excitement, while the negative emotions consist of anger, disgust, fear, and sadness [21]. For our work, we adopt the discrete approach as well.

Features specific to affective image classification have been developed in [9,7,4,10,11]. For example, the authors in [9] and [4] designed color features based on Iriens’ contrast theory. Specifically, the authors in [9] exploited semiotic principles to represent the visual content at the expressive level, while the authors in [4] used the composition features such as the low depth-of-field indicators, rule of thirds, and proportion and skin pixels in images, which have been found useful for aesthetics. The luminance-warm-cool and saturation-warm-cool color histograms were derived in [7] based on the fuzzy theory. In [10], the authors investigated the relationship between shape and emotions. They found that roundness and complexity of shapes are fundamental to understanding emotions. On the contrary, the conventional low-level image features have been adopted in [12–14,8]. For example, a large set of generic color, shape, and texture image features have been used in [14,8]. These low-level features were extracted from both the raw images and compound image transforms such as color transform and edge transform, which were found highly effective earlier in face recognition and the classification of painters and schools of art. In our work, we also use the conventional low-level image features, and we show later in the experiments that the proposed method can learn well enough to predict image emotions by using low-level features.

As for classifiers, SVM [22] is the most favorite one and has been used in [13,15,7,16,8]. Others include the naïve Bayes classifier [23] used in [11,4,10] and the regression trees [24] used in [15]. In this paper, we follow the Bayesian approach. As a methodological contribution, the proposed algorithm is the first multiple kernel learning algorithm that combines multiview learning and multi-label learning with full Bayesian treatment. There are existing Bayesian MKL algorithms and multilabel learning methods applied to image classification problems (e.g., [25]) but there is no previous study on a coupled approach. In this case, our method has the advantage of utilizing the emotional correlations in image affect recognition.

3. Proposed method

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

Fig. 3 illustrates the proposed probabilistic model for multilabel binary classification with a graphical model. We extended the model presented in [26] by trying to capture the correlation between the class labels with the help of shared kernel weights. The kernel matrices \( \{K_1, \ldots, K_p\} \) are used to calculate intermediate outputs using the weight matrix \( A \). The intermediate outputs capture the correlations between emotions and give probabilistic outputs for measuring the intensities of a distribution of emotions triggered by an image.

![Graphical model for Bayesian multilabel multiple kernel learning.](image)

Table 1

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>Number of training instances</td>
</tr>
<tr>
<td>( p )</td>
<td>Number of kernels</td>
</tr>
<tr>
<td>( L )</td>
<td>Number of output labels</td>
</tr>
<tr>
<td>( {K_1, \ldots, K_p} \in \mathbb{R}^{N \times N} )</td>
<td>Kernel matrices</td>
</tr>
<tr>
<td>( A \in \mathbb{R}^{N \times K} )</td>
<td>Weight matrix</td>
</tr>
<tr>
<td>( \Lambda \in \mathbb{R}^{K \times K} )</td>
<td>Priors for weight matrix</td>
</tr>
<tr>
<td>( {G_1, \ldots, G_q} \in \mathbb{R}^{P \times N} )</td>
<td>Intermediate outputs</td>
</tr>
<tr>
<td>( e \in \mathbb{R}^{P \times N} )</td>
<td>Kernel weight vector</td>
</tr>
<tr>
<td>( \omega \in \mathbb{R}^{K \times N} )</td>
<td>Priors for kernel weight vector</td>
</tr>
<tr>
<td>( b \in \mathbb{R}^{L} )</td>
<td>Bias vector</td>
</tr>
<tr>
<td>( \gamma \in \mathbb{R}^{G \times N} )</td>
<td>Priors for bias vector</td>
</tr>
<tr>
<td>( F \in \mathbb{R}^{G \times N} )</td>
<td>Auxiliary matrix</td>
</tr>
<tr>
<td>( Y \in {\pm 1}^{N \times L} )</td>
<td>Label matrix</td>
</tr>
</tbody>
</table>
(G_1, ..., G_L), kernel weights e, and bias parameters b are used to calculate the classification scores. Finally, the given class labels Y are generated from the auxiliary matrix F, which is introduced to make the inference procedures efficient [27]. 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

\[ \lambda_o^j \sim \mathcal{G}(\lambda_o^j; \alpha_j, \beta_j) \quad \forall (i, o) \]
\[ \alpha_o^j \sim \mathcal{N}(\alpha_o^j; 0, (\lambda_o^j)^{-1}) \quad \forall (i, o) \]
\[ g_o^m \mid (a_o, k_m) \sim \mathcal{N}(g_o^m; a_o^m, k_m, 1) \quad \forall (o, m, i) \]
\[ \gamma_o \sim \mathcal{G}(\gamma_o; \alpha_{\gamma}, \beta_{\gamma}) \quad \forall o \]
\[ b_o \sim \mathcal{N}(b_o; 0, \gamma_o^{-1}) \quad \forall o \]
\[ \omega_m \sim \mathcal{G}(\omega_m; a_{\omega}, b_{\omega}) \quad \forall m \]
\[ f_o^j \mid b_o, e, g_o \sim \mathcal{N}(f_o^j; e^T g_o + b_o, 1) \quad \forall (o, i) \]
\[ y_o^j \sim \mathcal{D}(y_o^j > \nu) \quad \forall (o, i) \]

where the margin parameter \( \nu \) 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 [28]. As short-hand notations, all priors in the model are denoted by \( \Xi = \{ \gamma, \Lambda, \omega \} \), where the remaining variables by \( \Theta = (A, b, e, F, G_1, ..., G_L) \) and the hyper-parameters by \( \xi = \{ \alpha_j, \beta_j, \alpha_{\omega}, \beta_{\omega}, \alpha_{\gamma}, \beta_{\gamma} \} \). Dependence on \( \xi \) is omitted for clarity throughout the paper.

The variational methods use a lower bound on the marginal likelihood using an ensemble of factored posteriors to find the joint parameter distribution [29]. Assuming independence between the approximate posteriors in the factorable ensemble can be justified because there is not a strong coupling between our model parameters. We can write the factorable ensemble approximation of the required posterior as

\[ p(\Theta, \Xi | (K_{m=1})^p, Y) = q(\Theta, \Xi) \]
\[ = q(\Lambda)q(\Theta)q(Z)q((G_o^T)_{o=1}^L)q(\gamma)q(\omega)q(b, e)q(F) \]

and define each factor in the ensemble just like its full conditional distribution:

\[ q(\Lambda) = \prod_{i=1}^N \prod_{l=1}^L \mathcal{G}(\lambda_o^j; \alpha_{\lambda_o^j}, \beta_{\lambda_o^j}) \]
\[ q(\Theta) = \prod_{o=1}^L \mathcal{N}(a_o; \mu(a_o), \Sigma(a_o)) \]
\[ q((G_o^T)_{o=1}^L) = \prod_{o=1}^L \prod_{m=1}^N \mathcal{N}(g_o^m; \mu(g_o^m), \Sigma(g_o^m)) \]
\[ q(\gamma) = \prod_{o=1}^L \mathcal{G}(\gamma_o; \alpha_{\gamma}, \beta_{\gamma}) \]
\[ q(\omega) = \prod_{m=1}^L \mathcal{G}(\omega_m; \alpha_{\omega}, \beta_{\omega}) \]
\[ q(b, e) = \mathcal{N}(b, e; \mu(b, e), \Sigma(b, e)) \]
\[ q(F) = \prod_{o=1}^N \prod_{m=1}^N \mathcal{D}(f_o^j > \nu) \]

where \( \alpha(\cdot), \beta(\cdot), \mu(\cdot), \) and \( \Sigma(\cdot) \) denote the shape parameter, the scale parameter, the mean vector, and the covariance matrix for their arguments, respectively. \( \mathcal{D}(f_o^j > \nu) \) denotes the truncated normal distribution with the mean vector \( \mu \), the covariance matrix \( \Sigma \), and the truncation rule \( \rho(\cdot) \) such that \( TN(\cdot; \mu, \Sigma, \rho(\cdot)) \), if \( \rho(\cdot) \) is true and \( TN(\cdot; \mu, \Sigma, 0) \) otherwise.

We can bound the marginal likelihood using Jensen’s inequality:

\[ \log p(Y | (K_{m=1})^p) \geq E_{q(\Theta, \Xi)}[\log p(Y, \Theta, \Xi | (K_{m=1})^p)] - E_{q(\Theta, \Xi)}[\log q(\Theta, \Xi)] \]

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

\[ q(\tau) = \exp \left[ E_{q(\Theta, \Xi)}[\log p(Y, \Theta, \Xi | (K_{m=1})^p)] \right] \]

For our model, thanks to the conjugacy, the resulting approximate posterior distribution of each factor follows the same distribution as the corresponding factor.

### 3.1. Inference details

The approximate posterior distribution of the priors of the precisions for the weight matrix can be found as a product of gamma distributions:

\[ q(\Lambda) = \prod_{i=1}^N \prod_{l=1}^L \mathcal{G}(\lambda_o^j; \alpha_{\lambda_o^j} + \frac{1}{2}, \frac{1}{2} (\frac{1}{\beta_{\lambda_o^j}^2} + \frac{\beta_{\lambda_o^j}^2}{2})^{-1}) \]

where the tilde notation denotes the posterior expectations as usual, i.e., \( b(\tau) = E_{q(\tau)}[b(\tau)] \). The approximate posterior distribution of the weight matrix is a product of multivariate normal distributions:

\[ q(\Theta) = \prod_{o=1}^L \mathcal{N}(a_o; \mu(a_o), \Sigma(a_o)) \left( \sum_{m=1}^L K_m g_o^m \right)^{-1} \]

The approximate posterior distribution of the projected instances can also be formulated as a product of multivariate normal distributions:

\[ \sum_{m=1}^L K_m g_o^m \]

where the kernel weights and the auxiliary variables defined for each label are used together.

The approximate posterior distributions of the priors on the biases and the kernel weights can be found as products of gamma distributions:

\[ q(b) = \prod_{o=1}^L \mathcal{G}(b_o; \alpha_{b_o} + \frac{1}{2}, \frac{1}{2} (\frac{1}{\beta_{b_o}^2} + \frac{\beta_{b_o}^2}{2})^{-1}) \]

\[ q(\omega) = \prod_{m=1}^L \mathcal{G}(\omega_m; \alpha_{\omega} + \frac{1}{2}, \frac{1}{2} (\frac{1}{\beta_{\omega}^2} + \frac{\beta_{\omega}^2}{2})^{-1}) \]

The approximate posterior distribution of the biases and the kernel weights is a product of multivariate normal distributions:

\[ q(b, e) = \mathcal{N}\left( b, e; \mu(b, e), \Sigma(b, e) \right) \]

\[ \mathcal{D}(f_o^j > \nu) \]
The predictive distribution of the auxiliary variables is a product of truncated normal distributions:

\[ q(F) = \prod_{o=1}^{N} \prod_{m=1}^{P} N(f_{o,m}^p, e^T g_{o,m} + b_o, 1.5 f_{o,m} > \nu) \]  

(7)

where we need to find the posterior expectations in order to update the approximate posterior distributions of the projected instances and the classification parameters. Fortunately, the truncated normal distribution has a closed-form formula for its expectation.

### 3.2. Complete algorithm

The complete inference algorithm is listed in Algorithm 1. The inference mechanism sequentially updates the approximate posterior distributions of the model parameters and the latent variables until convergence, which can be checked by monitoring the lower bound. The first term of the lower bound corresponds to the sum of exponential forms of the distributions in the joint likelihood. The second term is the sum of negative entropies of the approximate posteriors in the ensemble. The only nonstandard distribution in the second term is the truncated normal distributions of the auxiliary variables; nevertheless, the truncated normal distribution has a closed-form formula also for its entropy.

**Algorithm 1.** Bayesian multilabel multiple kernel learning.

Require: \(|K_m| = 1, Y, \nu, \alpha, \beta, \alpha_*, \beta_*, \alpha_o, \beta_o, \) and \(\beta_w \)

1. Initialize \(q(\Lambda), q((g_o)^{T})_o, q(b, e), \) and \(q(F) \) randomly

2. repeat

3. Update \(q(\Lambda) \) and \(q(\Lambda) \) using (1) and (2)

4. Update \(q(g_o)^{T}_o \) using (3)

5. Update \(q(y), q(\omega), \) and \(q(b, e) \) using (4), (5), and (6)

6. Update \(q(F) \) using (7)

7. until convergence

8. return \(q(\Lambda) \) and \(q(b, e) \)

### 3.3. Prediction

In the prediction step, we can replace \(p(A|K_{m,1} = 1, Y) \) with its approximate posterior distribution \(q(A) \) and obtain the predictive distribution of the intermediate outputs \( \{g_{a,*}\}_{a=1} \) for a new data point as

\[
p(\{g_{a,*}\}_{a=1} | K_{m,1} = 1, Y) = \prod_{o=1}^{N} \prod_{m=1}^{P} N(g_{a,m}^p, \mu(a_o)^T k_{m,1} + 1 + k_{m,1} \Sigma(a_o) k_{m,1}).
\]

The predictive distribution of the auxiliary variables \(f_* \) can also be found by replacing \(p(b, e|K_{m,1} = 1, Y) \) with its approximate posterior distribution \(q(b, e)\):

\[
p(f_* | \{g_{a,*}\}_{a=1} = 1, Y) = \prod_{o=1}^{N} \left( f_o^p, \mu(b_o)^T 1 + [g_{a,*} \Sigma(b_o, e)] 1 \right)
\]

and the predictive distribution of the class label \(y_* \) can be formulated using the auxiliary variable distribution:

\[
p(y_* = 1|K_{m,1} = 1, Y) = (Z_*^{P})^{-1} \Phi \left( \frac{\mu(f_*^p) - \nu}{\Sigma(f_*^p)} \right) \forall \theta
\]

where \(Z_*^{P} \) is the normalization coefficient calculated for the test data point and \(\Phi(\cdot)\) is the standardized normal cumulative distribution function.

### 4. Experiments

In this section, we present the experimental results using our proposed Bayesian MKL algorithm in two different scenarios: affective image classification and affective image retrieval. 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. Data sets

Two affective image data sets have been used in the experiments, the International Affective Picture System (IAPS) [30] and the ArtPhoto [4].

The IAPS data set 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, including snakes, insects, animals, landscapes, babies, guns, and accidents, among others. Each image is evaluated by subjects (males and females) on three continuously varying scales from 1 to 9 for Valence, Arousal, and Dominance. A subset of 394 IAPS images have been grouped into 8 discrete emotional categories based on a psychophysical study [21]. Among the 8 emotions, Amusement, Awe, Contentment, and Excitement are considered as the positive class, whereas Anger, Disgust, Fear, and Sad are considered as the negative class. The ground truth label for each image was selected as the category that had majority of the votes. Both Machajdik and Hanbury [4] and Lu et al. [10] used this subset for emotional image classification, and we also used it in our experiment to compare with their results.

The ArtPhoto data set was originally collected by Machajdik et al. [4] and it contains 806 artistic photographs obtained using discrete emotional categories as search queries in a photo sharing site. The discrete categories are the same as those adopted in the above IAPS subset and the images cover a wide range of semantic contents as well. We used this data set in our image retrieval experiment.

![Fig. 4. The five-zone partitioning scheme [33.](image)](image)

4.2. Image features

We have used a set of ten low-level content descriptors for still images, including color, shape, and texture features. Four of them are standard MPEG-7 [31] visual descriptors: Scalable Color, Dominant Color, Color Layout, and Edge Histogram. These low-level features have been widely used in image classification and retrieval tasks, as well as in image affect detections [13,14,8].

When presented a new picture or painting, people tend to first get a holistic impression of it and then go into segments and details [32]. Therefore, our features are extracted both globally and locally from each image. For certain features, a five-zone image partitioning scheme (see Fig. 4) is applied prior to the feature extraction [33]. Similar to the rule of thirds used in photography, the central part of an image usually catches most of people’s attention. All the features have been extracted by using PicSOM system [34]. Table 2 gives a summary of these features.

<table>
<thead>
<tr>
<th>Index</th>
<th>Feature</th>
<th>Type</th>
<th>Zoning</th>
<th>Dims.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Scalable Color</td>
<td>Color</td>
<td>Global</td>
<td>256</td>
</tr>
<tr>
<td>F2</td>
<td>Dominant Color</td>
<td>Color</td>
<td>Global</td>
<td>6</td>
</tr>
<tr>
<td>F3</td>
<td>Color Layout</td>
<td>Color</td>
<td>8 x 8</td>
<td>12</td>
</tr>
<tr>
<td>F4</td>
<td>SZone-Color</td>
<td>Color</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>F5</td>
<td>SZone-Color</td>
<td>Color</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>F6</td>
<td>Edge Histogram</td>
<td>Shape</td>
<td>4 x 4</td>
<td>80</td>
</tr>
<tr>
<td>F7</td>
<td>Edge Fourier</td>
<td>Shape</td>
<td>Global</td>
<td>128</td>
</tr>
<tr>
<td>F8</td>
<td>SZone-Edgehist</td>
<td>Shape</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>F9</td>
<td>SZone-Edgecoocc</td>
<td>Shape</td>
<td>5</td>
<td>80</td>
</tr>
<tr>
<td>F10</td>
<td>SZone-Texture</td>
<td>Texture</td>
<td>5</td>
<td>40</td>
</tr>
</tbody>
</table>

4.2.1. Color features

Scalable Color: The descriptor is a 256-bin color histogram in HSV color space, which is encoded by a Haar transform.

Dominant Color: The descriptor is a subset from the original MPEG-7 XM descriptor and is composed of the LUV color system values of the first and second most dominant color. If the XM routine only found one dominant color, then it was duplicated.

Color Layout: The image area is divided into 8 x 8 non-overlapping blocks where the dominant colors are solved in YCbCr (6, 3, 3) color system. Discrete Cosine Transform (DCT) is then applied to the dominant colors in each channel and the coefficients of DCT used as a descriptor.

SZone-Color: This descriptor is a three-element vector that contains the average RGB values of all the pixels within each zone.

SZone-Colm: The color moments feature treats the HSV color channels from each zone as probability distributions, and calculates the first three moments (mean, variance, and skewness) for each distribution.

4.2.2. Shape features

Edge Histogram: The image is divided into 4 x 4 non-overlapping sub-images where the relative frequencies of five different edge types (vertical, horizontal, 45°, 135°, non-directional) are calculated using 2 x 2-sized edge detectors for the luminance of the pixels. The descriptor is obtained with a nonlinear discretization of the relative frequencies.

Edge Fourier: This descriptor calculates the magnitude of the 16 x 16 Fast Fourier Transform (FFT) of Sobel edge image.

5 Zone-Edgehist: The edge histogram feature is the histogram of four Sobel edge directions. It is not the same as the MPEG-7 descriptor with the same name.

4.3. Texture features

5 Zone-Texture: The texture neighborhood feature is calculated from the Y (luminance) component of the YIQ color representation of each zone pixels. The 8-neighborhood of each inner pixel is examined, and a probability estimate is calculated for the probabilities that the neighbor pixel in each surrounding relative position is brighter than the central pixel. The feature vector contains these eight probability estimates.

4.4. Affective image classification

4.4.1. Experimental setup

In this experiment, we evaluate the performance of the proposed Bayesian MKL algorithm within a classification framework and compare with the results in [4,10]. Note that for [10], we compared the result by using their proposed image shape features. The TAPS subset was used in this task. For training and testing, we used the same procedure as in [4,10]: we used 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. As for the free parameters, we manually set $\{\nu, \alpha, \beta, \gamma, \delta, \omega\}$ to be $\{1, 1, 1, 0.001, 0.001, 0.001, 0.001\}$, respectively, based on the cross-validation results from the training data set. Through our experiments, we found that the last four parameters ($\alpha, \beta, \omega, \delta$) need careful selections as they directly control the kernels or features sparsity, whereas the other three ones do not affect the final performance much on emotional image predictions.

4.4.2. Results

Fig. 5 shows the classification results (average of 8 classes). It is clear to see that our proposed algorithm is the best among the three. With rather generic low-level image features, our classifier can achieve better classification performance than methods of [4,10] which rely on the design of complicated domain-specific features. Table 3 shows the comparison result with four other existing MKL methods, including the RBMKL [35], GMKL [36], NLMKL [37], and GLMKL [38]. The same ten low-level image features described in this paper were utilized as the input for all the MKL methods. We can see that our method is slightly better than NLMKL and GLMKL, yet much better than RBMKL and GMKL.

To further demonstrate the advantage of multiple kernel (multi-view) learning over single kernel (single-view) learning, we trained and tested a single SVM classifier using each of the 10 features.
Table 4
The image features ranked by SVM classification accuracies.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dominant Color</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>Color Layout</td>
<td>0.22</td>
</tr>
<tr>
<td>3</td>
<td>Edge Fourier</td>
<td>0.22</td>
</tr>
<tr>
<td>4</td>
<td>SZone-Texture</td>
<td>0.21</td>
</tr>
<tr>
<td>5</td>
<td>SZone-Color</td>
<td>0.20</td>
</tr>
<tr>
<td>6</td>
<td>Scalable Color</td>
<td>0.20</td>
</tr>
<tr>
<td>7</td>
<td>SZone-Color</td>
<td>0.20</td>
</tr>
<tr>
<td>8</td>
<td>SZone-Edgecoccc</td>
<td>0.20</td>
</tr>
<tr>
<td>9</td>
<td>SZone-Edgehist</td>
<td>0.19</td>
</tr>
<tr>
<td>10</td>
<td>Edge Histogram</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Fig. 6. The average feature representation weights over 5-fold cross-validation for the multilabel multiple kernel learning scenario.

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, which allows us to give a “soft” class assignment instead of a “hard” one. This characteristic is particularly useful for detecting emotion distribution evoked by an image. Fig. 7 gives some examples. One can see that the probabilistic outputs of our Bayesian algorithm generally agree well with the real human votes for certain images.

4.5. Affective image retrieval

4.5.1. Experimental setup

Also, we have designed an experiment for affective image retrieval based on our proposed Bayesian MKL algorithm. Firstly, we define the dissimilarity measure (the Euclidean distance in the implicit feature space) between a query image \( q \) and a retrieved image \( r \) as

\[
d_q(q, r) = \sqrt{k(q, q) + k(r, r) - 2k(q, r)}
\]

\[
k_q(q, q) = \sum_{m=1}^{p} e_m k_m(q, q)
\]

\[
k_r(r, r) = \sum_{m=1}^{p} e_m k_m(r, r)
\]

\[
k_q(q, r) = \sum_{m=1}^{p} e_m k_m(q, r)
\]

where \( k_m(\cdot, \cdot) \) denotes the kernel function calculated on the \( m \)th feature representation and \( e_m \) is the weight for the corresponding kernel learned by our algorithm. Therefore, given a query image \( q \), our aim is to find those images with the smallest \( d_q(q, r) \) values. In essence, the smaller \( d_q(q, r) \) is, the more probable that the retrieved image \( r \) evokes similar emotional states in people. We selected query images from the ArtPhoto data set and let the algorithm retrieve images from the IAPS data set. Both data sets use the same emotional categories. The kernel weights \( \{e_m\}_{m=1}^{p} \) were selected by training on the whole IAPS data set. Note that neither of the compared methods [4,10] had explored image emotions from a retrieval perspective as their focus was on feature design only.

4.5.2. Results

Fig. 8 gives some query-return examples from the results of image retrieval experiments. For the “Contentment” image, our algorithm successfully finds three other contentment images as its nearest neighbors. Similar query-return patterns can be seen from the “Disgust” and “Fear” query images. An interesting phenomenon is that both the “Amusement” and “Excitement” query images have retrieved the “Awe” image, and both the “Anger” and “Sad” query images have

separately (with the same partition as MKL setup). Table 4 lists the classification accuracies. The best SVM classifier (trained with Dominant Color) can only achieve an accuracy of 22%, which is about 9% lower than that of our algorithm. And an SVM using all 10 features can give an accuracy of 25%. This demonstrates the advantage of kernel learning helps to identify the relative importance of feature representations using a common set of kernel weights.

Also, we have designed an experiment for affective image retrieval based on our proposed Bayesian MKL algorithm. Firstly, we define the dissimilarity measure (the Euclidean distance in the implicit feature space) between a query image \( q \) and a retrieved image \( r \) as

\[
d_q(q, r) = \sqrt{k(q, q) + k(r, r) - 2k(q, r)}
\]

\[
k_q(q, q) = \sum_{m=1}^{p} e_m k_m(q, q)
\]

\[
k_r(r, r) = \sum_{m=1}^{p} e_m k_m(r, r)
\]

\[
k_q(q, r) = \sum_{m=1}^{p} e_m k_m(q, r)
\]

where \( k_m(\cdot, \cdot) \) denotes the kernel function calculated on the \( m \)th feature representation and \( e_m \) is the weight for the corresponding kernel learned by our algorithm. Therefore, given a query image \( q \), our aim is to find those images with the smallest \( d_q(q, r) \) values. In essence, the smaller \( d_q(q, r) \) is, the more probable that the retrieved image \( r \) evokes similar emotional states in people. We selected query images from the ArtPhoto data set and let the algorithm retrieve images from the IAPS data set. Both data sets use the same emotional categories. The kernel weights \( \{e_m\}_{m=1}^{p} \) were selected by training on the whole IAPS data set. Note that neither of the compared methods [4,10] had explored image emotions from a retrieval perspective as their focus was on feature design only.

4.5.2. Results

Fig. 8 gives some query-return examples from the results of image retrieval experiments. For the “Contentment” image, our algorithm successfully finds three other contentment images as its nearest neighbors. Similar query-return patterns can be seen from the “Disgust” and “Fear” query images. An interesting phenomenon is that both the “Amusement” and “Excitement” query images have retrieved the “Awe” image, and both the “Anger” and “Sad” query images have...
found the “Fear” image among their top candidates. This is meaningful in that the former three emotions belong to the positive class which usually induces high valence, while the latter three emotions belong to the negative class which usually induces low valence but high arousal. Besides, the retrieval result again reveals the fact that an image often evokes multiple emotional states that are correlated with each other. For example, an amusement image usually elicits partial feeling of awe, and the feeling of sadness is closely connected with the feeling of fear. To a certain extent, our algorithm has detected such correlations that exist among emotions using rather low-level image features.

5. Conclusions

In this paper, we have presented a novel Bayesian multiple kernel learning algorithm for affective image classification and retrieval tasks 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 algorithm is able to predict a set of emotions evoked by an image rather than a single one. 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.

It is worth emphasizing that our method is not confined to the image emotions recognition, but can be easily extended to other affective stimuli such as audio and video data. Due to the varying subjectivity in humans and the limit of the available affective databases, it is of course not guaranteed that our method can make a perfect classification or retrieval for every single image. Eventually, the development in this interdisciplinary area relies on the joint efforts from artificial intelligence, computer vision, pattern recognition, cognitive science, psychology, as well as color and art theories.

Fig. 7. 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) and (b)): Amusement, Awe, Contentment, Excitement, and negative emotions ((c) and (d)): Anger, Disgust, Fear, Sad. The y-axis shows the agreement in the range [0, 1]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)
Fig. 8. The image retrieval results using the ArtPhoto images as queries. The first column corresponds to the query images from the ArtPhoto data set, and the last three columns correspond to the top three retrieved images from the IAPS emotional subset ranked by distance. The ground-truth label is given under each image.
Acknowledgments

This work was financially supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, Grant no. 251170).

References


He Zhang

He received his Bachelor and Master degrees from Jinan University College of Communication Engineering, China, in 2004 and 2007, respectively. He obtained his Ph.D. degree on Machine Learning and Data Mining from the Department of Information and Computer Science at Aalto University Finland in 2014. His research area focuses on Nonnegative Matrix Factorization (NMF) for data clustering, and affective multimedia classification and retrieval. He is a member of the European Neural Network Society (ENNS) and is a PC member of the 2014 International Conference on Artificial Neural Networks (ICANN). He has already started his industrial career as a (client-side) Project Data Analyst in a Finnish data analytics corporation in the Helsinki Metropolitan area.

Mehmet Gönen

received the B.Sc. degree in industrial engineering, the M.Sc. and the Ph.D. degrees in computer engineering from Bogaziçi University, Istanbul, Turkey, in 2003, 2005, and 2010, respectively. He did his postdoctoral work at the Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland. He then was a Senior Research Scientist at Sage Bionetworks, Seattle, WA, USA. He is currently a Research Assistant Professor at the Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA. His current research interests include cancer biology, personalized medicine, kernel methods, and Bayesian methods.

Zhirong Yang

received his Bachelor and Master degrees from Sun Yat-Sen University, China in 1999 and 2002, and Doctoral degree from Helsinki University of Technology, Finland in 2008, respectively. Presently he is a Docent (Adjunct professor) in the Department of Information and Computer Science at Aalto University. He is currently a researcher in The Finnish Centre of Excellence in Computational Inference Research (COIN). His current research interests include machine learning, information visualization, computer vision and information retrieval, particularly focus on nonnegative learning, clustering analysis, nonlinear dimensionality reduction and learning representations. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the International Neural Network Society (INNS).

Erikki Oja

received the D.Sc. degree from Helsinki University of Technology in 1977. He is Director of the Computational Inference Research Centre and Distinguished Professor at the Department of Information and Computer Science, Aalto University (former Helsinki University of Technology). Finland. He holds an honorary doctorate from Uppsala University, Sweden, and from two Finnish universities. He has been a research associate at Brown University, Providence, RI, and a visiting professor at the Tokyo Institute of Technology and the Universite Pantheon-Sorbonne, Paris. He is the author or coauthor of more than 330 articles and book chapters on pattern recognition,
computer vision, and neural computing, and three books: “Subspace Methods of Pattern Recognition” (New York: Research Studies Press and Wiley, 1983), which has been translated into Chinese and Japanese; “Kohonen Maps” (Amsterdam, The Netherlands: Elsevier, 1999), and “Independent Component Analysis” (New York: Wiley, 2001; also translated into Chinese and Japanese). His research interests are in the study of principal component and independent component analysis, self-organization, statistical pattern recognition, and machine learning. Oja is a member of the Finnish Academy of Sciences and the Academia Europaea, Fellow of the Institute of Electrical and Electronics Engineers (IEEE), Founding Fellow of the International Association of Pattern Recognition (IAPR), Past President of the European Neural Network Society (ENNS), and Fellow of the International Neural Network Society (INNS). He is a member of the editorial boards of several journals and has been in the program committees of several recent conferences. He was general chair of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP) and the 2011 International Conference on Artificial Neural Networks (ICANN). Prof. Oja is the recipient of the 2004 IAPR Pierre Devijver Award, the 2006 IEEE Computational Intelligence Society Neural Networks Pioneer Award, and the 2010 INNS Hebb Award.