Visual Sentiment Classification via Low-Rank Regularization and Label Relaxation

Xiao Jin, *Member, IEEE*, Peiguang Jing, Jiesheng Wu, Jing Xu, *Member, IEEE*, and Yuting Su, *Member, IEEE*

***Abstract*—In the human cognitive system, the emotional feeling is a complicated process. Visual sentiment classification aims to predict the human emotions evoked by different images. In this article, we proposed a novel visual sentiment classification algorithm by modeling this task as a low-rank subspace learning problem. To reduce the discrepancy between global and local features, image features of relevant regions are selected from the whole image by sparse encoding. The label relaxation item is employed for alleviating the label ambiguity caused by subjective evaluation. We develop an alternative iterative method to optimize the proposed objective function. This model can be naturally extended for online learning, which improves efficiency. We conduct extensive experiments on three publicly available data sets. Compared with several state-of-the-art methods, we achieve better performance.**

***Index Terms*—Dictionary learning, low-rank regularization, subspace learning, visual sentiment classification.**

I. Introduction

RECENT advances in social media platforms have led to massive visual contents on the Internet. People use pictures and videos to share their emotions of daily lives. Visual sentiment analysis aims to associate image contents with emotions that they arouse in humans [1]–[4]. Previous research on visual sentiment analysis can be summarized into categorical and dimensional approaches [5]. The categorical approaches classify emotions evoked by visual contents into several types. The dimensional methods map sentiments into a multidimensional space, e.g., valance–arousal. In this article, we mainly focus on categorical approaches.

A. Motivations

Although many strategies have been designed for this task, there still exist several challenges.

1) Affective and Semantic Gap: As we all know, there is an obvious gap between low-level image features and high-level emotions it caused. In the process of human cognition, the emotion corresponding to a certain image is very subjective and abstract. The inconsistency between the image features and the expected sentiment tags is the main challenge in visual emotion analysis. Thus, low-rank subspace learning is employed to model this problem, since an enormous number of images belong to a few sentimental categories.

2) Global and Local Feature Discrepancy: When someone looks at an image, his emotion is not only influenced by the whole image but also evoked by some specific regions. Different areas have various effects on the viewer's affective states. The semantic information of some objects has strong correlations with image sentiments, while some global features are redundant and noisy for this task. Therefore, how to extract the most effective features becomes a critical issue in this problem. Motivated by this phenomenon, we seek a set of sparse representations that incorporates the most relevant and the most discriminative features [6].

3) Label Ambiguity Caused by Subjective Evaluation: Emotional feeling is a complicated process in the human recognition system. The same image leads to various emotions of different viewers, which is caused by the cultural background, personality, and social context [7]. Besides, the same person may have inconsistent emotions at different time points [8]. Therefore, the sentiment labels are vague, and outliers may exist in the training data set. Thus, it is important to investigate this problem from the label space and build a robust classifier against the outliers. To address this issue, we further consider the label relaxation strategy when exploring the label space information.

4) Strong Dependence on Large-Scale Labeled Data: The current performance of deep learning algorithms is highly dependent on the amount of training data. Small-scale data sets may cause overfitting and performance degradation. However, most data sets in this research topic contain less than thousands of pictures [9]. Though some data sets include more data, collection and annotation are labor-intensive and time-consuming. Therefore, it is difficult to obtain sufficient training data to build robust classifiers. Furthermore, it is difficult to contain all the emotional tags in a data set because of the subjectivity of human judgments. In practice, pictures are continuously uploaded to social media platforms, and new samples with new categories probably appear. It is crucial to handle this case while maintaining the efficiency of the algorithm. To solve this problem, we extend the proposed method to an online version.

Manuscript received 3 August 2021; revised 15 October 2021; accepted 11 December 2021. Date of publication 21 December 2021; date of current version 9 December 2022. This work was supported in part by the Science and Technology Planning Project of Tianjin, China, under Grant 17JCZDJC30700 and Grant 18ZXZNGX00310; in part by the Tianjin Natural Science Foundation under Grant 19JCQNJC00300; and in part by the Fundamental Research Funds for the Central Universities of Nankai University under Grant 63201192 and Grant 63211116. *(Corresponding author: Jing Xu.)*

Xiao Jin, Jiesheng Wu, and Jing Xu are with the College of Artificial Intelligence, Nankai University, Tianjin 300350, China (e-mail: jinxiao@nankai.edu.cn; jasonwu@mail.nankai.edu.cn; xujing@nankai.edu.cn).

Peiguang Jing and Yuting Su are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: pgjing@tju.edu.cn; ytsu@tju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2021.3135948.

Digital Object Identifier 10.1109/TCDS.2021.3135948

The scattered samples with labels are utilized to incrementally learn a more robust model while reducing the running time without any retraining.

B. Contributions

To deal with the current challenges in visual sentiment classification, we introduce a novel method based on joint sparse and low-rank feature representation. The proposed model projects the feature space into a high-level semantic space by low-rank subspace learning and encodes the image features with sparse dictionary learning. We employ the label relaxation strategy to explicitly embed label information into our framework. Moreover, we also extend our method to an online version. The contributions of this article are listed as follows.

1) The visual sentiment classification is modeled as a low-rank subspace learning problem to bridge the gap between low-level image features and high-level sentiments they evoked. A set of sparse features is obtained by dictionary learning, which selects the most discriminative region representations from the whole images and reduces the discrepancy of global and local features. To enlarge the distance between different classes and alleviate the label ambiguity, the label relaxation is adopted in our framework.
2) Considering the practical settings for visual emotion classification, we also extend the proposed model to an online version to incorporate more data for better performance. This attempt builds a more robust classifier, while the computational complexity is much less than the supervised setting.
3) We develop an alternative iterative method to optimize the proposed objective function. The results of experiments conducted on three image sentiment data sets demonstrate that the classification performance of our approach is competitive to state-of-the-art methods.

The remainder of this article is organized as follows. In Section II, we review the related works. The proposed algorithm is detailed in Section III. We describe the experiment settings in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Visual Sentiment Classification

Visual information plays a significant role in the human cognitive system [10]–[15]. According to [5], previous research on visual sentiment analysis can be classified into categorical and dimensional approaches. The categorical methods aim at classifying emotions evoked by visual contents into sentiment categories. The dimensional approaches map sentiments into a multidimensional space, e.g., valance–arousal. This article mainly discusses the categorical approaches.

Many researchers have made great efforts concerning the classification of visual sentiments. Some previous studies have focused on handcrafted features. Sartori *et al.* [16] investigated art theory concepts and color combinations to identify the emotions aroused by images. Zhao *et al.* [17], [18] exploited multitask hypergraph learning to predict personalized emotions by different factors.

In recent years, CNN-based methods have been promisingly applied to analyze visual sentiments. Wu *et al.* [19] first discovered multiple relative local regions via visual attributes learned by multitask learning and then built the classifier upon these regions, obtaining a significant performance improvement. You *et al.* [20] built a large-scale data set and evaluated the performance of CNNs on image sentiment classification. Zhu *et al.* [21] fused the outputs from different layers in the CNN model by a bidirectional recurrent neural network to recognize the visual emotions. You *et al.* [22] adopted the progressive and training domain transfer strategy to deal with the weakly supervised nature of image emotion classification. Rao *et al.* [23] designed a multilevel region-based CNN framework to extract the features from both global and local views. To fully use the information of multiple source domain, Lin *et al.* [24] developed a domain adaptation method based on the generative adversarial network. She *et al.* [5] presented a weakly supervised coupled network that can accomplish detection and classification tasks simultaneously. In the work [9], a multiple kernel network (MKN) learning representation from strongly and weakly supervised CNNs is presented for visual sentiment analysis.

B. Dictionary Learning

For decades, one of the most popular ideas in representation learning research is dictionary learning, also called sparse representation or sparse coding [25]–[27]. The key benefit of these algorithms is that complex signals can be represented sparsely and concisely. Besides, these methods are powerful for representing images, since images admit naturally sparse representations. According to the way of dictionary training, it can be roughly classified into two categories: 1) offline learning and 2) online learning.

Offline dictionary learning utilizes the category labels at hand to build a dictionary in the training stage, without any incoming unlabeled samples. Existing literature has explored the effectiveness of dictionary learning in research community [28], such as image fusion [29], image denoising [30], object recognition [31], visual tracking [32], and image classification [33], [34].

This field has gradually broadened as online dictionary learning in large-scale settings. Mairal *et al.* [35] designed an online optimization method for large-scale matrix factorization problem. They introduced stochastic approximations for handling millions of training samples, which can extend to matrix factorization and guarantee the convergence. Lu *et al.* [36] modeled the online updating process as minimizing several quadric functions iteratively, which takes time linear to the size of data and robust against outliers. Naderahmadian *et al.* [37] presented an adaptive update strategy by only considering the atoms in sparse learning of the new training data. Zhao *et al.* [38] incorporated both user–item relationship and item content features into a unified online social recommendation system. Adeli *et al.* [6] noticed that collaboratively learning both labeled and unlabeled samples can build better intrinsic geometry of the sample space and

proposed a semisupervised classifier to alleviate the problem of sample-outliers and feature-noise.

III. Proposed Method

In this section, we first present the notations and preliminaries. Then, we formulate our visual sentiment classification framework and propose an alternative iterative algorithm for minimizing the objective function. Finally, we extend this method to an online version to incrementally strengthen the proposed model. The pipeline of our method is illustrated in Fig. 1.

A. Notations and Preliminaries

To formulate our method, we first summarize the involved notations, dimensions, and operations. Except for specific definitions, we present a scalar with a Greek letter, e.g., α; a column vector with a lowercase bold letter, e.g., $\mathbf{x}$; and a matrix with an uppercase bold letter, e.g., $\mathbf{X}$. For convenience, we use $\mathbf{x}^i$ and $\mathbf{x}_i$ to denote the ith row and the ith column of the matrix $\mathbf{X}$, respectively. The ith row and jth components of matrix $\mathbf{X}$ are denoted as $\mathbf{X}_{ij}$. The $l_{p,q}$-norm of matrix $\mathbf{X} \in \mathbb{R}^{D\times N}$ is defined as

$$\|\mathbf{X}\|_{p,q} = \left[\sum_{i=1}^{D}\left(\sum_{j=1}^{N}|\mathbf{X}_{ij}|^p\right)^{q/p}\right]^{1/q} \tag{1}$$

where $\mathbf{X}_{ij}$ is the (i,j)th element of matrix $\mathbf{X}$. By changing the values of p and q, there are several types of norms defined as follows. When $p = q = 1$, the l_1-norm is defined as

$$\|\mathbf{X}\|_1 = \sum_{i=1}^{D}\|\mathbf{x}^i\|_1. \tag{2}$$

When $p = q = 2$, the Frobenius norm is defined as

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{D}\sum_{j=1}^{N}\|\mathbf{X}_{ij}\|^2}. \tag{3}$$

When $p = 2, q = 1$, the $l_{2,1}$-norm is defined as

$$\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{D}\|\mathbf{x}^i\|_2. \tag{4}$$

The nuclear norm of matrix $\mathbf{X}$ is defined as

$$\|\mathbf{X}\|_* = \sum_i \delta_i(\mathbf{X}) \tag{5}$$

where $\sum_i \delta_i(\mathbf{X})$ is the sum of singular values of matrix $\mathbf{X}$. The basic notations and descriptions are listed in Table I.

B. Problem Formulation

Without loss of generality, suppose that we have a collection of n samples in c classes. Each sample can be represented by a vector of d_{fea} dimension. A feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d_{\text{fea}}\times n}$ and a label matrix $\mathbf{L} \in \mathbb{R}^{c\times n}$ with sentiment labels are obtained after feature extraction. The main goal of our work is to learn a sparse matrix $\mathbf{A}$ that reveals the connection between the learned image features and the emotions they evokes.

TABLE I
Notations Used in This Article and Their Descriptions

Notations	Descriptions
$\mathbf{X} \in \mathbb{R}^{d_{fea}\times n}$	feature matrix
$\mathbf{P} \in \mathbb{R}^{d_{com}\times d_{fea}}$	projection matrix (transformation matrix)
$\mathbf{D} \in \mathbb{R}^{d_{com}\times d_{dict}}$	dictionary matrix
$\mathbf{A} \in \mathbb{R}^{d_{dict}\times n}$	sparse representation
$\mathbf{L} \in \mathbb{R}^{c\times n}$	label matrix
$\mathbf{H} \in \mathbb{R}^{c\times n}$	luxury matrix
$\mathbf{M} \in \mathbb{R}^{c\times n}$	label relaxation matrix (nonnegative)
$\mathbf{E} \in \mathbb{R}^{d_{com}\times n}$	error matrix
$\mathbf{Z} \in \mathbb{R}^{n\times n}$	reconstruction matrix
d_{fea}	dimension of input feature
d_{com}	dimension of common space
d_{dict}	dimension of dictionary space
n	number of samples
c	number of categories
α, β, γ	regularization parameters
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_1$	l_1-norm
$\|\cdot\|_{2,1}$	$l_{2,1}$-norm
$\|\cdot\|_*$	nuclear norm
$\odot$	Hadamard product operation

1) Low-Rank Representation: In some other image classification tasks, samples belonging to the same class are usually correlated and reside in the same low-dimensional subspace. The representation for samples from one class should be reasonable of low rank [39]. Since a large number of pictures correspond to a small number of emotional tags, visual sentiment classification can be naturally modeled as a low-rank subspace learning problem [40]–[44].

The low-rank representation can be expressed as follows:

$$\min_{\mathbf{Z}} \operatorname{rank}(\mathbf{Z}), \quad \text{s.t.} \quad \mathbf{X} = \Gamma\mathbf{Z} \tag{6}$$

where $\operatorname{rank}(\cdot)$ is the rank of a matrix, $\mathbf{X}$ is the input matrix, Γ is used to represent the inputs, and $\mathbf{Z}$ is the reconstruction matrix. In practical application, noise is inevitable. The above model can be rewritten as

$$\min_{\mathbf{Z},\mathbf{E}} \operatorname{rank}(\mathbf{Z}) + \|\mathbf{E}\|_l, \quad \text{s.t.} \quad \mathbf{X} = \Gamma\mathbf{Z} + \mathbf{E} \tag{7}$$

where $\|\cdot\|_l$ is the regularization item of noise and $\mathbf{E}$ is the error matrix. Following a common practice in rank minimization, we substitute the rank function with the nuclear norm, resulting in this optimization problem:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \|\mathbf{E}\|_l, \quad \text{s.t.} \quad \mathbf{X} = \Gamma\mathbf{Z} + \mathbf{E}. \tag{8}$$

By choosing the data matrix $\mathbf{X}$ itself to replace Γ [44], we have

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \|\mathbf{E}\|_l, \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}. \tag{9}$$

Based on the above analysis, in this article, we utilize the low-rank subspace learning to bridge the semantic gap between low-level image features and high-level sentiment categories. Each feature vector corresponding to each image is modeled as a sample, while each sentiment category is modeled as a subspace. Assuming $\mathbf{P}$, $\mathbf{Z}$, and $\mathbf{E}$ are the project matrix, the subspace reconstruction matrix, and the error matrix, respectively, this problem can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E},\mathbf{P}} \quad & \|\mathbf{E}\|_{2,1} + \|\mathbf{Z}\|_* \\ \text{s.t.} \quad & \mathbf{P}\mathbf{X} = \mathbf{P}\mathbf{X}\mathbf{Z} + \mathbf{E} \end{aligned} \tag{10}$$

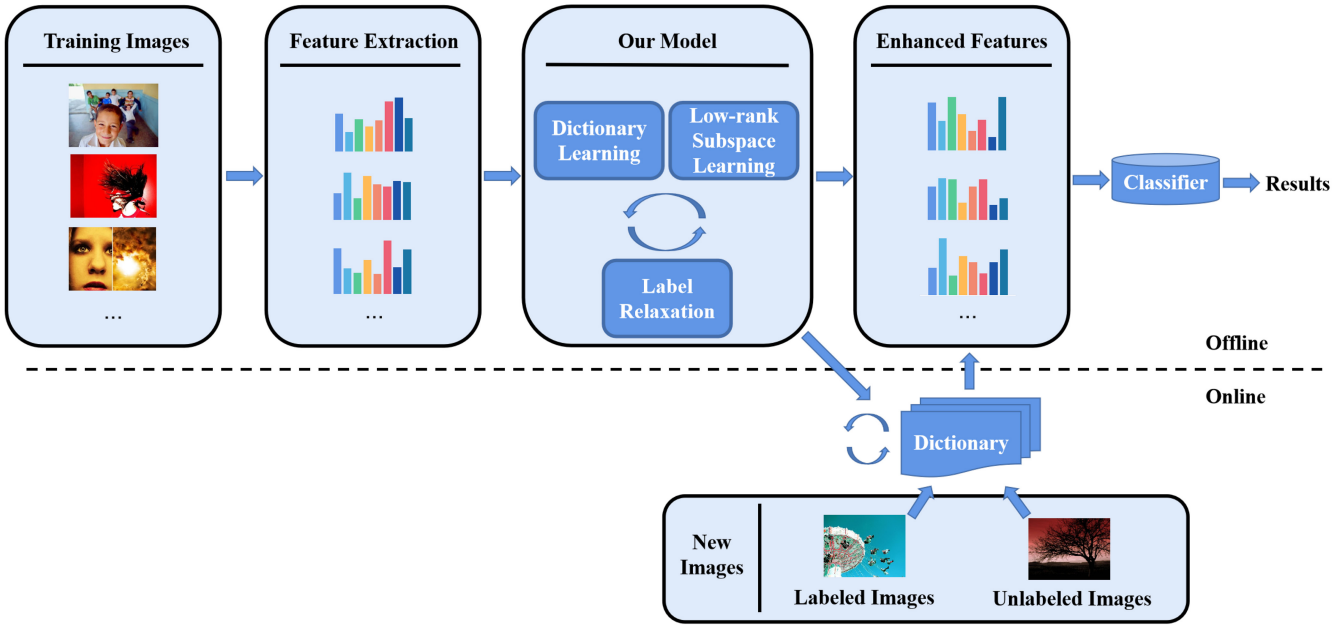

Fig. 1. Pipeline of the proposed method.

where the $l_{2,1}$-norm is employed to model the sample-specific corruptions and outliers in the error matrix $\mathbf{E}$. Compared with the common low-rank subspace learning formulation, the transformation matrix $\mathbf{P}$ is introduced for bridging the semantic gap.

Low-rank representation mainly captures the global structure of the data by imposing a low-rank constraint on the data representation matrix [44]. To construct a robust model, we also need the local structure of the data. Thus, we analyze this aspect by sparse dictionary learning in the following.

2) Sparse Dictionary Learning: To build a model that can capture global and local information simultaneously, a common way is to apply low-rank and sparse constraints on the representation features [41], [45]. As mentioned in Section I-A, visual sentiment classification still suffers from global and local feature discrepancy, outliers caused by label ambiguity, and strong dependence on large-scale labeled data. Inspired by recent work [6], we alleviate these problems by encoding image features with sparse dictionary learning

$$\min_{\mathbf{D},\mathbf{A},\mathbf{P}} \|\mathbf{PX} - \mathbf{DA}\|_F + \|\mathbf{A}\|_1 \tag{11}$$

where $\mathbf{D}$ is the dictionary matrix and $\mathbf{A}$ is the sparse representation matrix. The advantages of such formulation are as follows.

1) As mentioned in recent work [6], the l_1-norm can effectively select the most discriminative regions associated with the specific task from the whole image. This is important for the visual sentiment analysis, where the features from different regions of the image are extracted, but not all the regions are associated with emotions [23].
2) The l_1-norm is also robust to outliers, which alleviates the performance degradation caused by sentiment label ambiguity.
3) In addition, this format can easily extend to an online dictionary learning version. It can make full use of the available data while reducing the computational complexity. Assuming n is the number of offline training samples, $\tilde{n}$ is the number of online incoming samples, the online learning reduces the complexity from $\mathcal{O}((n+\tilde{n})^3)$ to $\mathcal{O}(n^3+\tilde{n})$. The detailed analysis of complexity is illustrated in Section IV-A.

Since sentiment classification is a subjective issue, it is hard to cover all the categories. In practical applications, there are few emotional categories in the existing data set. New categories and new samples that are not in the training set often appear. In this case, how to avoid the waste of computing resources caused by repeated training is critical.

3) Label Relaxation: Although label ambiguity in training data is a widely acknowledged challenge in visual sentiment classification, there is litter work to solve this problem from label space. To deal with the label ambiguity, we introduced a label relaxation item into the objective function.

The traditional binary label matrix lacks discriminative information between classes [46]. Both dictionaries and sparse representations are too rigid when input data is mapped into strict binary labels. For instance, there are three samples x_1, x_2, and x_3 belong to classes c_2, c_1, and c_3, respectively. Then, the label matrix is calculated as

$$\overline{\mathbf{L}} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

where columns represent samples and rows stand for labels. The distance between two samples from different classes is $\sqrt{2}$ in the label space. In this setting, the distances are equal for all the samples, while the characteristics for different data are ignored. Suppose we add a nonnegative label relaxation

matrix $\overline{\mathbf{M}}$ to $\overline{\mathbf{L}}$, obtaining a relaxed label matrix $\widehat{\mathbf{L}}$

$$\widehat{\mathbf{L}} = \begin{bmatrix} -\overline{\mathbf{M}}_{11} & 1+\overline{\mathbf{M}}_{12} & -\overline{\mathbf{M}}_{13} \\ 1+\overline{\mathbf{M}}_{21} & -\overline{\mathbf{M}}_{22} & -\overline{\mathbf{M}}_{23} \\ -\overline{\mathbf{M}}_{31} & -\overline{\mathbf{M}}_{32} & 1+\overline{\mathbf{M}}_{33} \end{bmatrix}$$
$$\overline{\mathbf{M}}_{ij} \geq 0, i \in \{1,2,3\}, j \in \{1,2,3\} \tag{13}$$

where $\overline{\mathbf{M}}_{ij}$ denotes the (i,j)th component in the matrix $\overline{\mathbf{M}}$. The distance between sample x_1 and x_2 is enlarged to

$$\sqrt{\left(\overline{\mathbf{M}}_{11}+1+\overline{\mathbf{M}}_{21}\right)^2+\left(1+\overline{\mathbf{M}}_{12}+\overline{\mathbf{M}}_{22}\right)^2+\left(\overline{\mathbf{M}}_{23}-\overline{\mathbf{M}}_{13}\right)^2} \geq \sqrt{2} \tag{14}$$

in the new label space.

Based on the above analysis, we use the nonnegative label relaxation matrix $\mathbf{M} \in \mathbb{R}^{c\times n}$ to relax the strict binary label matrix into a relaxed label matrix. The mathematical formulation is expressed as

$$\min_{\mathbf{M},\mathbf{P}} \ \|\mathbf{L}+\mathbf{H}\odot\mathbf{M}-\mathbf{PX}\|_F^2 \quad \text{s.t.} \ \ \mathbf{M}\geq 0 \tag{15}$$

where $\odot$ represents the Hadamard product operation, $\mathbf{M}$ is the label relaxation matrix, elements in the luxury matrix $\mathbf{H}$ are defined as $\mathbf{H}_{ij}=+1$ if $\mathbf{L}_{ij}=1$, -1 otherwise. This attempt can expand the distance between different classes as much as possible in the label space. Moreover, such settings encourage the dictionary $\mathbf{D}$ and the sparse representation $\mathbf{A}$ to become more flexible and more discriminative [47].

4) Final Objective Function: Based on the above descriptions, our objective function has three parts: 1) low-rank subspace learning; 2) dictionary learning; and 3) label relaxation

$$\min_{\mathbf{D},\mathbf{A},\mathbf{M},\mathbf{P},\mathbf{E},\mathbf{Z}} \ \|\mathbf{PX}-\mathbf{DA}\|_F^2+\|\mathbf{A}\|_1 + \alpha\|\mathbf{L}+\mathbf{H}\odot\mathbf{M}-\mathbf{PX}\|_F^2+\beta\|\mathbf{E}\|_{2,1}+\gamma\|\mathbf{Z}\|_* \quad \text{s.t.} \ \ \mathbf{PX}=\mathbf{PXZ}+\mathbf{E}, \mathbf{M}\geq 0. \tag{16}$$

It is also worth noting that all the elements in $\mathbf{M}$ should be nonnegative. Without this restriction, there is no guarantee that the distance of samples from different categories in the relaxed label space becomes larger than that in the original label space. Furthermore, negative elements in $\mathbf{M}$ may cause some trivial solutions of the final objective function. For example, when $\mathbf{M}=-\mathbf{L}$, we obtain $\mathbf{P}=\mathbf{D}=\mathbf{A}=\mathbf{Z}=\mathbf{X}=\mathbf{0}$, which derives a zero loss and becomes an optimal solution. Additionally, we assume that the dimensionality of the common subspace d_{com} is the same as the number of categories c in this article.

C. Optimization Algorithm

The proposed objective function is multivariate. It is difficult to solve them at one step. Therefore, we design an alternative iterative algorithm to minimize this function, which divides a complex problem into a couple of separable subproblems. In each step, we optimize the objective function with respect to one unknown variable, while fixing the other variables. This procedure repeats until meeting the convergence condition or reaching the maximum number of iterations.

As a common approach for low-rank optimization, we rewrite (16) into (17) by introducing a relaxed variable $\mathbf{Z}_1$ to substitute $\mathbf{Z}$

$$\min_{\mathbf{D},\mathbf{A},\mathbf{M},\mathbf{P},\mathbf{E},\mathbf{Z},\mathbf{Z}_1} \ \|\mathbf{PX}-\mathbf{DA}\|_F^2+\|\mathbf{A}\|_1 +\alpha\|\mathbf{L}+\mathbf{H}\odot\mathbf{M}-\mathbf{PX}\|_F^2+\beta\|\mathbf{E}\|_{2,1}+\gamma\|\mathbf{Z}_1\|_* \quad \text{s.t.} \ \ \mathbf{PX}=\mathbf{PXZ}+\mathbf{E}, \mathbf{Z}=\mathbf{Z}_1, \mathbf{M}\geq 0. \tag{17}$$

We use the inexact augmented Lagrangian method [48] to optimize this problem. The augmented Lagrangian function of (17) can be written as $\mathcal{L}$

$$\begin{aligned}\mathcal{L}&(\mathbf{D},\mathbf{A},\mathbf{M},\mathbf{P},\mathbf{E},\mathbf{Z},\mathbf{Z}_1,\mathbf{Y}_1,\mathbf{Y}_2,\mu)\\ &= \|\mathbf{PX}-\mathbf{DA}\|_F^2+\|\mathbf{A}\|_1 \\ &\quad +\alpha\|\mathbf{L}+\mathbf{H}\odot\mathbf{M}-\mathbf{PX}\|_F^2+\beta\|\mathbf{E}\|_{2,1}+\gamma\|\mathbf{Z}_1\|_* \\ &\quad +\langle\mathbf{Y}_1,\mathbf{PX}-\mathbf{PXZ}-\mathbf{E}\rangle+\langle\mathbf{Y}_2,\mathbf{Z}-\mathbf{Z}_1\rangle \\ &\quad +\frac{\mu}{2}\|\mathbf{PX}-\mathbf{PXZ}-\mathbf{E}\|_F^2+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_1\|_F^2 \\ &= \|\mathbf{PX}-\mathbf{DA}\|_F^2+\|\mathbf{A}\|_1 \\ &\quad +\alpha\|\mathbf{L}+\mathbf{H}\odot\mathbf{M}-\mathbf{PX}\|_F^2+\beta\|\mathbf{E}\|_{2,1}+\gamma\|\mathbf{Z}_1\|_* \\ &\quad +\frac{\mu}{2}\left\|\mathbf{PX}-\mathbf{PXZ}-\mathbf{E}+\frac{\mathbf{Y}_1}{\mu}\right\|_F^2+\frac{\mu}{2}\left\|\mathbf{Z}-\mathbf{Z}_1+\frac{\mathbf{Y}_2}{\mu}\right\|_F^2\end{aligned} \tag{18}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Lagrangian multipliers, and μ is a positive penalty parameter.

For better explanation, we use t to indicate the tth iteration step. $\mathbf{D}_t$, $\mathbf{A}_t$, $\mathbf{M}_t$, $\mathbf{P}_t$, $\mathbf{E}_t$, $\mathbf{Z}_t$, and $\mathbf{Z}_{1,t}$ are defined as the variables updated in the tth iteration. Hence, each variable in the $(t+1)$th iteration can be calculated as follows. All the subproblems have closed-form solutions.

1) For $\mathbf{D}$: Without regard to other unrelated variables, the dictionary $\mathbf{D}$ can be updated by

$$\begin{aligned}\mathbf{D}_{t+1} &= \arg\min_{\mathbf{D}} \mathcal{L} \\ &= \arg\min_{\mathbf{D}} \|\mathbf{PX}-\mathbf{DA}\|_F^2.\end{aligned} \tag{19}$$

To solve this subproblem, we compute the derivative of (18) with respect to $\mathbf{D}$ and obtain

$$\frac{\partial\mathcal{L}}{\partial\mathbf{D}} = -2\mathbf{PXA}^T+2\mathbf{DAA}^T. \tag{20}$$

By setting this partial derivative of (18) to be zero, $\mathbf{D}_{t+1}$ can be updated as

$$\mathbf{D}_{t+1} = \left(\mathbf{PXA}^T\right)\left(\mathbf{AA}^T\right)^{-1}. \tag{21}$$

2) For $\mathbf{A}$: When other variables are fixed, solving $\mathbf{A}_{t+1}$ is equivalent to the following problem:

$$\begin{aligned}\mathbf{A}_{t+1} &= \arg\min_{\mathbf{A}} \mathcal{L} \\ &= \arg\min_{\mathbf{A}} \|\mathbf{PX}-\mathbf{DA}\|_F^2+\|\mathbf{A}\|_1.\end{aligned} \tag{22}$$

This is an l_1-norm minimization problem and can be solved by the Feature-Sign-Search Strategy [49]. By gradient descent, the Feature-Sign-Search Strategy aims to linearly search for

Algorithm 1 Feature-Sign-Search Strategy

Input:
$\mathbf{P}, \mathbf{X}, \mathbf{D}$, The iteration number T_{iter1};
Output:
the learned $\mathbf{A}$;
for $i_{iter1} \leq T_{iter1}$ **do**
  Line search for the optimal $\mathbf{A}$ using gradient descent as $\partial\Psi/\partial\mathbf{A}$;
  if the sign of any element of $\mathbf{A}$ changes **then**
    Set the corresponding element to zero;
  else
    Go back to line search Step 2;
  end if
  Update the iteration variable using $i_{iter1} = i_{iter1} + 1$;
end for
return $\mathbf{A}$.

the optimal $\mathbf{A}$ along the direction of $\partial\Psi/\partial\mathbf{A}$, where Ψ is defined as

$$\Psi = \|\mathbf{PX} - \mathbf{DA}\|_F^2. \tag{23}$$

This process is repeated until any element of $\mathbf{A}$ does not change. Algorithm 1 describes the framework for optimizing $\mathbf{A}$.

3) For M: With other variables fixed, $\mathbf{M}$ is updated by solving the following problem:

$$\begin{aligned}\mathbf{M} &= \arg\min_{\mathbf{M}} \mathcal{L} \\ &= \arg\min_{\mathbf{M}} \|\mathbf{L} + \mathbf{H} \odot \mathbf{M} - \mathbf{PX}\|_F^2 \\ &\quad \text{s.t.} \quad \mathbf{M} \geq 0.\end{aligned} \tag{24}$$

Since the squared Frobenius norm of a matrix can be decomposed element by element, the above optimization problem (24) for $\mathbf{M} \in \mathbb{R}^{c\times n}$ can be splitted into solving $c \times n$ subproblems [46]. Let $\Phi = \mathbf{L} - \mathbf{PX}$, $\mathbf{M}_{ij}$ denotes the (i,j)th entry of $\mathbf{M}$, we have the following formulation:

$$\min_{\mathbf{M}_{ij}} \left(\Phi_{ij} - \mathbf{H}_{ij}\mathbf{M}_{ij}\right)^2, \quad \text{s.t.} \quad \mathbf{M}_{ij} \geq 0 \tag{25}$$

where Φ_{ij} and $\mathbf{H}_{ij}$ are the (i,j)th elements of Φ and $\mathbf{H}$, respectively.

As we defined in (15), the square of the component $\mathbf{H}_{ij}$ is always equal to one, no matter whether the corresponding label is positive or negative. Therefore, we have

$$\left(\Phi_{ij} - \mathbf{H}_{ij}\mathbf{M}_{ij}\right)^2 = \left(\mathbf{H}_{ij}\Phi_{ij} - \mathbf{M}_{ij}\right)^2. \tag{26}$$

Considering the nonnegative constraint about $\mathbf{M}_{ij}$, The optimal solution of $\mathbf{M}_{ij}$ is

$$\mathbf{M}_{ij} = \max\left(\Phi_{ij}\mathbf{H}_{ij}, 0\right). \tag{27}$$

Thus, the optimal solution of $\mathbf{M}$ can be written as

$$\mathbf{M} = \max(\Phi \odot \mathbf{H}, 0). \tag{28}$$

4) For P: By ignoring the terms independent of $\mathbf{P}$ in (18), the optimization of $\mathbf{P}$ is equivalent to minimizing

$$\begin{aligned}\mathbf{P}_{t+1} &= \arg\min_{\mathbf{P}} \mathcal{L} \\ &= \arg\min_{\mathbf{P}} \|\mathbf{PX} - \mathbf{DA}\|_F^2 + \alpha\|\mathbf{L} + \mathbf{H} \odot \mathbf{M} - \mathbf{PX}\|_F^2 \\ &\quad + \frac{\mu}{2}\left\|\mathbf{PX} - \mathbf{PXZ} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu}\right\|_F^2.\end{aligned} \tag{29}$$

The partial derivative of (18) with respect to $\mathbf{P}$ is calculated as

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{P}} &= (2+2\alpha)\mathbf{PXX}^T - 2\mathbf{DAX}^T - 2\alpha\mathbf{G}_1\mathbf{X}^T \\ &\quad + \mu\mathbf{PG}_2\mathbf{G}_2^T - \mu\mathbf{G}_3\mathbf{G}_2^T.\end{aligned} \tag{30}$$

By setting this to zero, we can obtain the optimal $\mathbf{P}$ by

$$\begin{aligned}\mathbf{P}_{t+1} &= \left(2\mathbf{DAX}^T + 2\alpha\mathbf{G}_1\mathbf{X}^T + \mu\mathbf{G}_3\mathbf{G}_2^T\right) \\ &\quad \cdot \left((2+2\alpha)\mathbf{XX}^T + \mu\mathbf{G}_2\mathbf{G}_2^T\right)^{-1}\end{aligned} \tag{31}$$

where $\mathbf{G}_1 = \mathbf{L} + \mathbf{H} \odot \mathbf{M}$, $\mathbf{G}_2 = \mathbf{X} - \mathbf{XZ}$, $\mathbf{G}_3 = \mathbf{E} - \mathbf{Y}_1/\mu$.

5) For E: Then, the variable $\mathbf{E}$ can be updated by

$$\begin{aligned}\mathbf{E}_{t+1} &= \arg\min_{\mathbf{E}} \mathcal{L} \\ &= \arg\min_{\mathbf{E}} \beta\|\mathbf{E}\|_{2,1} + \frac{\mu}{2}\left\|\mathbf{PX} - \mathbf{PXZ} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu}\right\|_F^2.\end{aligned} \tag{32}$$

This is a typical $l_{2,1}$-norm optimization problem. We can derive $\mathbf{E}$ by shrink operation [50]

$$\mathbf{E}_{t+1} = \text{shrink}\left(\mathbf{PX} - \mathbf{PXZ} + \frac{\mathbf{Y}_1}{\mu}, \frac{\beta}{\mu}\right) \tag{33}$$

where $\text{shrink}(\varepsilon, \lambda) = \text{signmax}(|\varepsilon| - \lambda, 0)$.

6) For Z: After dropping other irrelevant variables, solving $\mathbf{Z}$ can be transformed to minimize the following object function:

$$\begin{aligned}\mathbf{Z}_{t+1} &= \arg\min_{\mathbf{Z}} \mathcal{L} \\ &= \arg\min_{\mathbf{Z}} \frac{\mu}{2}\left\|\mathbf{PX} - \mathbf{PXZ} - \mathbf{E} + \frac{\mathbf{Y}_1}{\mu}\right\|_F^2 \\ &\quad + \frac{\mu}{2}\left\|\mathbf{Z} - \mathbf{Z}_1 + \frac{\mathbf{Y}_2}{\mu}\right\|_F^2.\end{aligned} \tag{34}$$

The derivative of (18) with respect to $\mathbf{Z}$ is computed as

$$\frac{\partial\mathcal{L}}{\partial\mathbf{Z}} = \mathbf{X}^T\mathbf{P}^T\mathbf{PXZ} + \mathbf{X}^T\mathbf{P}^T\mathbf{G}_4 + \mathbf{Z} + \mathbf{G}_5. \tag{35}$$

By setting this partial derivative to zero, we can obtain

$$\mathbf{Z}_{t+1} = -\left(\mathbf{X}^T\mathbf{P}^T\mathbf{PX} + \mathbf{I}\right)^{-1} \cdot \left(\mathbf{G}_5 + \mathbf{X}^T\mathbf{P}^T\mathbf{G}_4\right) \tag{36}$$

where $\mathbf{G}_4 = \mathbf{E} - \mathbf{PX} - \mathbf{Y}_1/\mu$, $\mathbf{G}_5 = \mathbf{Y}_2/\mu - \mathbf{Z}_1$.

7) For Z_1: The optimal $\mathbf{Z}_1$ can be derived by minimizing the following function:

Algorithm 2 Proposed Optimization Method

Input:
the input data $\mathbf{X}$, the label $\mathbf{L}$;
Output:
the learned dictionary $\mathbf{D}$, the sparse representation $\mathbf{A}$;
Initialize: $\rho = 1.1$, $\mu_0 = 10^{-6}$, $\mu_{\max} = 10^6$.
while not converged **do**
  Fix others and update $\mathbf{D}_{t+1}$ by Eqn.(21);
  Fix others and update $\mathbf{A}_{t+1}$ by Eqn.(22);
  Fix others and update $\mathbf{M}_{t+1}$ by Eqn.(28);
  Fix others and update $\mathbf{P}_{t+1}$ by Eqn.(31);
  Fix others and update $\mathbf{E}_{t+1}$ by Eqn.(33);
  Fix others and update $\mathbf{Z}_{t+1}$ by Eqn.(36)
  Fix others and update $\mathbf{Z}_{1,t+1}$ by Eqn.(38);
  Fix others and update the multipliers $\mathbf{Y}_{1,t+1}$, $\mathbf{Y}_{2,t+1}$ and penalty parameter μ by Eqn.(39);
  Check the convergence conditions;
  Update the iteration variable using $t = t + 1$;
end while
return the learned dictionary $\mathbf{D}$, the sparse representation $\mathbf{A}$.

$$\begin{aligned}\mathbf{Z}_{1,t+1} &= \arg\min_{\mathbf{Z}} \mathcal{L} \\ &= \arg\min_{\mathbf{Z}} \frac{\gamma}{\mu}\|\mathbf{Z}_1\|_* + +\frac{1}{2}\left\|\mathbf{Z}_1 - \left(\mathbf{Z} + \frac{\mathbf{Y}_2}{\mu}\right)\right\|_F^2. \end{aligned} \tag{37}$$

It is also a standard form of the singular value thresholding (SVT) algorithm [51] with a closed-form solution. Under the framework in [51], we first define $\tau = \gamma/\mu$, $\Delta = \mathbf{Z} + \mathbf{Y}_{2,t}/\mu_t$ and perform the singular value decomposition (SVD) as $\Delta = \mathbf{U}\Theta\mathbf{V}^T$, $\Theta = \text{diag}(\{\sigma_i\}_{i=1}^r)$. The matrices $\mathbf{U}$ and $\mathbf{V}$ are left-singular and right-singular matrices, respectively, and r is the rank of Δ. Accordingly, the optimal $\mathbf{Z}_{1,t+1}$ can be updated as

$$\mathbf{Z}_{1,t+1} = \mathcal{D}_\tau(\Delta) \tag{38}$$

where the singular value shrinkage operator is utilized by $\mathcal{D}_\tau(\Delta) = \mathbf{U}\mathcal{D}_\tau(\Theta)\mathbf{V}^T$, $\mathcal{D}_\tau(\Theta) = \text{diag}(\{\sigma_i - \tau\}_+)$ and the subscript "+" denotes the positive part of $\{\sigma_i - \tau\}$.

8) For Y_1, Y_2, μ: Finally, the Lagrange multipliers $\mathbf{Y}_1, \mathbf{Y}_2$ and penalty parameter μ are updated by

$$\begin{aligned}\mathbf{Y}_{1,t+1} &= \mathbf{Y}_{1,t} + \mu_t(\mathbf{P}_{t+1}\mathbf{X} - \mathbf{P}_{t+1}\mathbf{X}\mathbf{Z}_{t+1} - \mathbf{E}_{t+1}) \\ \mathbf{Y}_{2,t+1} &= \mathbf{Y}_{2,t} + \mu_t\big(\mathbf{Z}_{t+1} - \mathbf{Z}_{1,t+1}\big) \\ \mu_{t+1} &= \min(\rho\mu_t, \mu_{\max}) \end{aligned} \tag{39}$$

where $\mu_{\max}$ is the upper bound of μ, and ρ is a positive constant.

Algorithm 2 gives the main steps of the above optimization procedures.

D. Incremental Learning

In the previous sections, we designed a novel objective function for image emotion classification with a fixed quantity of samples. However, the pictures in the database increase dynamically in practical applications. The changes in data amount (training samples) cause repeated dictionary learning, which should be avoided in the case of massive data.

Algorithm 3 Incremental Dictionary Learning

Input:
A new batch of samples $\widehat{X}_{\bar{t}} = [\widehat{x}_1, \ldots, \widehat{x}_i, \ldots, \widehat{x}_{\tilde{n}}] \in \mathbb{R}^{d_{\text{com}} \times \tilde{n}}$; Initial dictionary $\mathbf{D}_0$;
Output:
The updated dictionary $\mathbf{D}_{\bar{t}}$.
for $\bar{t} \le T_{iter3}$ **do**
  Sparse coding with Eqn.(40);
  Update $\Omega_{\bar{t}}$ by:
  $\Omega_{\bar{t}} \leftarrow \Omega_{\bar{t}-1} + \eta_{\bar{t}}\eta_{\bar{t}}^T$;
  Update $\Xi_{\bar{t}}$ by:
  $\Xi_{\bar{t}} \leftarrow \Xi_{\bar{t}-1} + \widehat{x}_{\bar{t}}\eta_{\bar{t}}^T$;
  for $j = 1$ *to* d_{dict} **do**
    Update the j-th column to optimize $\mathbf{D}_{\bar{t}}$ by Eqn.(45) and Eqn.(46):
  end for
  Update $\mathbf{D}_{\bar{t}}$ by each column $\mathbf{d}_j$;
end for
return $\mathbf{D}_{\bar{t}}$.

Therefore, we present an incremental learning approach in the online setting in this section. In this article, we only discuss the setting of supervised incremental learning, since it can be employed in most scenarios [52]. We assume that the training set is independent and identically distributed (i.i.d.) and every incoming sample is labeled. For the unlabeled new samples, we directly use the previous dictionary to predict the labels.

In this section, we denote the time step as $\bar{t}$, to distinguish it from the number of iterations t in previous sections. Inspired by [53], we extend our method to an online version. With the help of least-angle regression (LARS), the sparse representation $\eta_{\bar{t}}$ of sample $\mathbf{x}_{\bar{t}}$ at time $\bar{t}$ is first calculated by dictionary $\mathbf{D}_{\bar{t}-1}$ at time $\bar{t} - 1$

$$\eta_{\bar{t}} = \arg\min_{\eta} \frac{1}{2}\left\|\mathbf{x}_{\bar{t}} - \mathbf{D}_{\bar{t}-1}\eta\right\|_2^2 + \|\eta\|_1. \tag{40}$$

The new dictionary is then computed by the sample $\mathbf{x}_{\bar{t}}$ and the sparse vector $\eta_{\bar{t}}$ at time $\bar{t}$. Assuming $\widehat{X} = \mathbf{P}\mathbf{X} = [\widehat{x}_1, \ldots, \widehat{x}_i, \ldots, \widehat{x}_{\tilde{n}}]$, we define a quadratic function $f_{\bar{t}}(\mathbf{D})$ as

$$f_{\bar{t}}(\mathbf{D}) = \frac{1}{\bar{t}}\sum_{i=1}^{\bar{t}} \|\widehat{x}_i - \mathbf{D}\eta_i\|_2^2 + \|\eta_i\|_1. \tag{41}$$

The new dictionary can be updated by

$$\mathbf{D}_{\bar{t}} = \arg\min_{\mathbf{D}} f_{\bar{t}}(\mathbf{D}). \tag{42}$$

Block-coordinate descent with warm restarts is taken for updating each column of $\mathbf{D}_{\bar{t}}$ at time $\bar{t}$ consecutively. The jth column of $\mathbf{D}$ is defined as $\mathbf{d}_j$. For clearness, we introduce two matrices $\Omega_{\bar{t}}$ and $\Xi_{\bar{t}}$, which carry all the information of the past coefficients

$$\Omega_{\bar{t}} = [\omega_1, \ldots, \omega_k] = \sum_{i=1}^{\bar{t}} \eta_i\eta_i^T \tag{43}$$

$$\Xi_{\bar{t}} = [\xi_1, \ldots, \xi_k] = \sum_{i=1}^{\bar{t}} \widehat{x}_i\eta_i^T. \tag{44}$$

By setting the partial derivative of (41) with respect to $\mathbf{d}_j$ to be zero, we obtain the following update rule:

$$\varphi_j \leftarrow \frac{1}{\Omega_{jj}}\left(\xi_j - \mathbf{D}\omega_j\right) + \mathbf{d}_j \tag{45}$$

$$\mathbf{d}_j \leftarrow \frac{\varphi_j}{\max\left(\|\varphi_j\|_2, 1\right)}. \tag{46}$$

The details are listed in Algorithm 3. After learning the updated dictionary $\mathbf{D}_{\tilde{t}}$, we use it to learn the sparse coefficients for final classification. Since our model employs sparse representation $\mathbf{A}$ as feature sets, we omit the update process of other irrelevant variables in the proposed objective function in this section. The main purpose of this article is to design a novel objective function for image sentiment classification and an optimization algorithm for solving it. The comprehensive performance comparison of different types of online learning schemes is beyond the scope of this article.

IV. Discussion

A. Complexity Analysis

Our algorithm consists of the offline part and the online part. In the offline part, the computational complexity is mainly caused by the matrix inversions and multiplication operations for updating $\mathbf{D}, \mathbf{A}, \mathbf{P}, \mathbf{E}$, and $\mathbf{Z}$, and SVD operation for updating $\mathbf{Z}_1$. For convenience, we denote $d_{\text{com}}, d_{\text{fea}}$, and d_{dict} as d_1, d_2, and d_3 in this section. The numbers of iterations in Algorithms 1–3 are defined as $T_{\text{iter1}}, T_{\text{iter2}}$, and T_{iter3}, respectively. In step 1, the computational complexity is $\mathcal{O}(d_1d_2d_3n + d_3^2n + d_3^3)$. In step 2, the computational complexity of the Feature-Sign-Search Strategy is $\mathcal{O}(T_{\text{iter1}}(d_1d_2d_3n + d_1d_3^2n + d_3n))$. In step 4, optimizing $\mathbf{P}$ would cost about $\mathcal{O}(d_1d_2d_3n + cnd_2 + nd_1d_2 + nd_2^2 + nd_2^2 + d_2^3 + d_1d_2^2)$. In step 5, the complexity is $\mathcal{O}(n^2)$ when updating $\mathbf{E}$. In step 6, updating $\mathbf{Z}$ takes $\mathcal{O}(d_1d_2^2n^2 + n^3 + d_1d_2n^2)$ complexity. In step 7, the complexity for performing SVD is $\mathcal{O}(n^3)$. As for the online part, the complexity is $\mathcal{O}(d_1d_3^2)$ for each sample. Assuming the number of new training samples $\tilde{n}$ is larger than the number of offline training samples n, we have $\tilde{n} > n \gg \{d_1, d_2, d_3\}$. Since the number of samples is often larger than the feature dimensions, the speed bottleneck lies in the former. The total complexity is simplified to $\mathcal{O}(T_{\text{iter2}} \cdot n^3 + T_{\text{iter3}} \cdot \tilde{n})$. However, if we directly train the whole data set, the complexity is $\mathcal{O}(T_{\text{iter2}} \cdot (n + \tilde{n})^3)$. Thus, this attempt can save the computational time.

Most recent approaches for image sentiment classification are implemented on GPU platforms, while our method is run on CPU machines. It is difficult to compare the actual running time of our method with other baselines. However, we compared the running time difference between the online version and the offline version of the proposed method. When a new labeled image comes, it will take several hours to retrain a new dictionary. The online version only needs less than 5 min to update the dictionary. Detailed results for online learning is illustrated in Section V-G.

TABLE II
Data Sets We Used in This Article

Dataset	# Images	# Classes	Ref.
FI	23,308	8	[22]
ArtPhoto	806	8	[55]
Abstract	228	8	[55]

B. Convergence

It is difficult to prove the convergence of the whole objective function. However, in each step of the optimization, the subproblem is convex for each variable when the others are fixed. All the subproblems have closed-form solutions. Since the optimization of $\mathbf{D}$, $\mathbf{A}$, $\mathbf{P}$, and $\mathbf{Z}$ are typical optimization problems, we mainly focus on the convergence of $\mathbf{M}$, $\mathbf{E}$, and $\mathbf{Z}_1$. Based on the fact that the squared Frobenius norm of a matrix can be decoupled element by element, (24) can be converted to several subproblems of vector multiplication [46]. Thus, the step of optimizing $\mathbf{M}$ is convex. For $\mathbf{E}$, the iterative shrinkage-thresholding algorithm (ISTA) is used to learn the optimal solution and the convergence has been proven in [50]. For $\mathbf{Z}_1$, the convergence of this step has been proven in [51] and [54]. Additionally, we also provide the experimental study for convergence analysis in Section V-B. The results also show the convergence characteristic of the proposed method.

V. Experiments

In this section, we carry out extensive experiments to test and verify the effectiveness of the proposed method. First, we elaborate on the experimental setup, including data sets and implementation details. Then, we analyze the method from different aspects, e.g., convergence, parameter sensitivity, components, feature dimensions, etc. Finally, we compared our methods with several state-of-the-art methods and discuss the effects of online learning.

A. Experimental Setup

1) Data Sets: We evaluate our method on three publicly available data sets. The Flickr&Instagram (FI) data set [22] contains 23 308 images, which are labeled by hundreds of participants. Each person classifies pictures according to eight emotion categories, e.g., anger, amusement, awe, contentment, disgust, excitement, fear, and sadness. The labels are finally assigned if at least three individuals reach a consensus. Besides this large-scale data set collected from websites, we also conduct experiments on two small-scale data sets. The ArtPhoto (Artistic Photograph) data set [55] includes 806 images from a photo sharing site, whose emotion labels rely on the artists that uploaded the photograph. Abstract (Paintings) data set [55] consists of 228 abstract paintings. Pictures in this data set only contain color and texture, without any recognizable objects. The images are peer rated in a Web survey. For each image, the category with the most votes is selected as the ground truth. The datasets we used in this paper are summarized in Table I. Some examples are illustrated in Fig. 2.

2) Implementation Details: Due to the powerful representation capability of deep learning in recent years, the

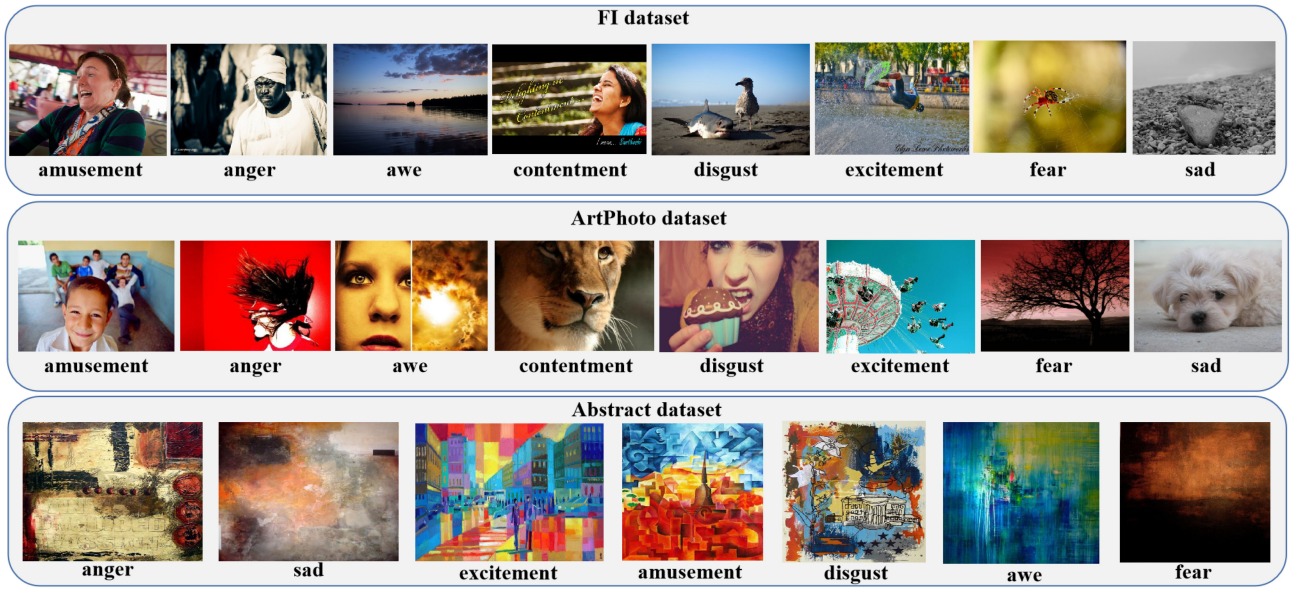

Fig. 2. Some image examples in the data sets used in this article. Since there are only three images in the "anger" category in the Abstract data set, these are not enough for cross-validation. Thus, we neglect this sentiment class in this data set.

performance of many visual tasks has gained improvements. In this article, we extract two types of features as inputs. The well-known "VGGNet-19" [56] and "AlexNet" [57] models are employed to characterize images. The feature extraction is implemented by MATLAB R2019a with the pretrained version of these networks, which are trained on the ImageNet database [58]. The object features are expressed by the outputs of the last fully connected layer {fc7} in VGGNet19, while the scene features are represented by the last fully connected layer {fc8} in AlexNet. Each type of features contains a vector of 4096 dimensions. After concatenating them, we obtain a feature vector of 8192 dimensions to describe an image. We further reduce them to 300 dimensions using principal component analysis (PCA). The Feature-Sign-Search Strategy in the optimization process is implemented with the codes provided by [59].[1] The variable $\mathbf{A}$ outputted from the proposed method is used as a transformed feature matrix. The widely used libSVM [60] is adopted for classification. As other published work about sentiment analysis, the accuracy is taken as a metric for performance comparison. Since there are only three images in the "anger" category in the Abstract data set, these are not enough for cross-validation. Thus, we neglect this sentiment class in this data set. We randomly split each data set 20 times and report the average performance as the final results. In each trial, we randomly choose 80% for training, the remaining for tests. Our experiments were performed on the machine which includes 16-GB of RAM and an Intel Core i7-7770 CPU.

B. Convergence Analysis

In this section, we conduct experiments to evaluate the convergence of the proposed objective function. The convergence condition is defined as the difference of the objective function

[1]http://www.ifp.illinois.edu/ jyang29/codes/CVPR09-ScSPM.rar

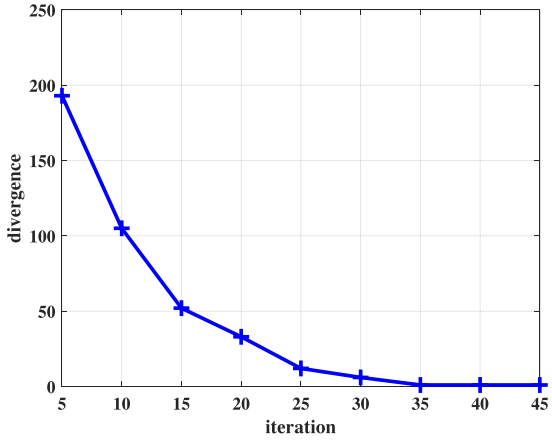

Fig. 3. Convergence curve of the proposed method. The horizontal axis denotes the index of iterations. The vertical axis represents the absolute value of the difference between Obj_t and Obj_{t-1} at the tth iteration.

from two consecutive iterations

$$\text{Dist}_t = |\text{Obj}_t - \text{Obj}_{t-1}|$$

where $|\cdot|$ denotes the absolute value, and Obj_t and Obj_{t-1} represent the value of objective function at the tth and $(t-1)$th iterations, respectively. Fig. 3 presents a typical convergence curve of our method. The horizontal axis denotes the index of iterations. The vertical axis represents the absolute value of the difference between Obj_t and Obj_{t-1} at the tth iteration. From Fig. 3, we can see that our algorithm can reach convergence with a limited number of iterations. Based on previous studies, the stopping criteria are critical to achieve the convergence of the objective function. In this article, we used the relative change between two consecutive iterations falling below a threshold of 1.0e-6 and a maximum of 40 iterations as the stopping criteria for our model.

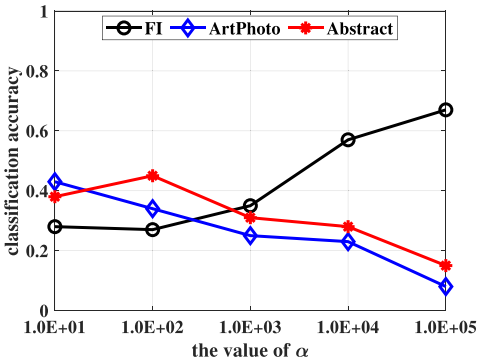

Fig. 4. Impact of the parameter α for the proposed method. The classification accuracy is shown with different values of parameter α. We define a candidate set $\{1.0\mathrm{E}+01, 1.0\mathrm{E}+02, 1.0\mathrm{E}+03, 1.0\mathrm{E}+04, 1.0\mathrm{E}+05\}$ for parameter α.

C. Parameter Sensitivity Analysis

In this part, we conduct experiments to investigate the effects of different parameters. As shown in (16), there are three parameters in our objective function. It is of great importance to select appropriate parameters. Theoretically, the changes of parameters are equivalent to relax the relationship of input deep features and corresponding sentiment classes. It is still an open problem to pick out the optimal parameters for different data sets. Based on preliminary experimental results, we find the selection of parameter α is sensitive to the number of pictures in each data set. Fig. 4 shows the classification performance of different values of α. We define a candidate set $\{1.0\mathrm{E}+01, 1.0\mathrm{E}+02, 1.0\mathrm{E}+03, 1.0\mathrm{E}+04, 1.0\mathrm{E}+05\}$ for parameter α. As we can see in this figure, when the number of samples in the data set becomes larger, increasing the parameter α appropriately can obtain better detection results. For the FI data set, we set balance parameter $\alpha = 1.0\mathrm{E}+05$. For the ArtPhoto data set, the parameter is set as $\alpha = 1.0\mathrm{E}+01$. For the Abstract data set, the balance parameter is defined as $\alpha = 1.0\mathrm{E}+02$. Since the number of samples in different data sets varies a lot, the impact of α on FI and other two data sets may be inconsistent. In addition, we empirically specified the other parameters to be $\beta = 1.0\mathrm{E}+02$, and $\gamma = 1.0e-03$ by default.

D. Component Analysis

In this section, we conduct experiments to verify the necessities of different components in our objective function. We compare the performance by removing some components. The other parameters are retained at their predetermined values, when one of them is set to zero.

1) *noM:* To test the influence of label relaxation item, we remove the label relaxation matrix $\mathbf{M}$ and the luxury matrix $\mathbf{H}$ by setting $\mathbf{M} = \mathbf{0}$.
2) *noSP:* We evaluate the usefulness of subspace learning by setting $\beta = 0$.
3) *noLR:* We consider the effect of low-rank regularization by setting $\gamma = 0$.

From Table III, we can draw the following conclusions.

1) The most influential component is label relaxation, which explicitly embeds label information into our framework. Thus, the distances between different classes can be enlarged. This observation verifies the effectiveness of the label information.
2) The classification results of "noSP" and "noLR" are worse than that of the complete algorithm, which shows the necessity of modeling this problem by subspace learning and low-rank regularization.

TABLE III
CLASSIFICATION ACCURACY COMPARISON OF INVOLVED COMPONENTS IN THE PROPOSED ALGORITHM

Dataset	noM	noSP	noLR
FI	0.1803	0.2826	0.2639
ArtPhoto	0.1325	0.2585	0.2146
Abstract	0.1242	0.2136	0.2483

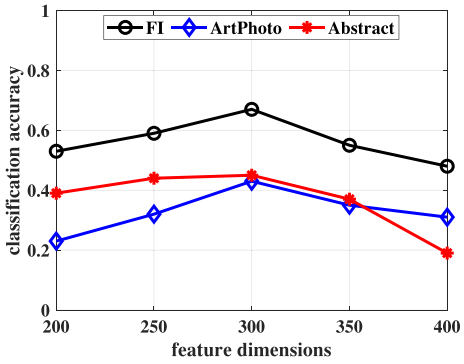

Fig. 5. Impact of feature dimensions for the proposed method. The classification accuracy is shown with different feature dimensions.

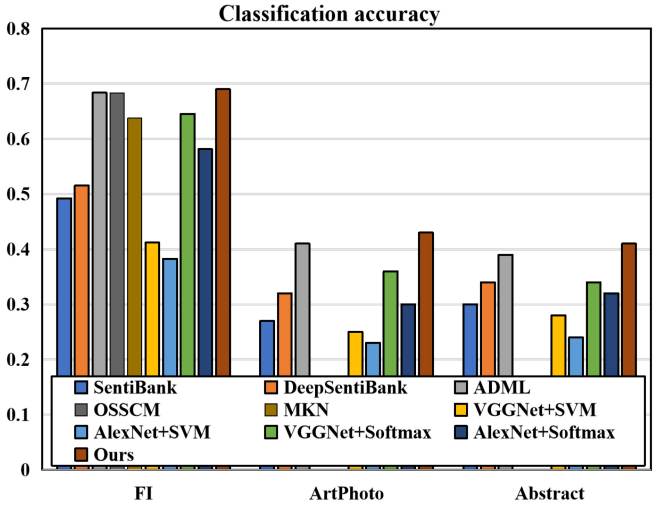

Fig. 6. Classification accuracy on three sentiment classification data sets, including FI, Abstract, and ArtPhoto data sets.

E. Feature Dimension

As a critical issue of inputs, the feature dimension has an impact on the final classification results. In this section, we analyze the influence of different feature dimensions on different data sets. The classification performance of our method with different feature dimensions is shown in Fig. 5. From this figure, we find that too large or too small feature dimensions

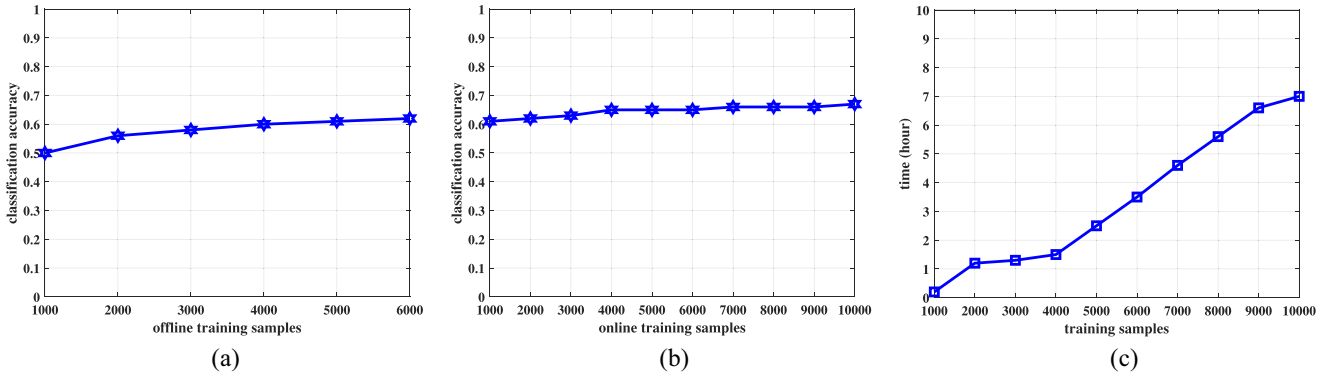

Fig. 7. Impact of online learning for the proposed method. (a) Impact of offline training size. (b) Impact of online training sizes. (c) Running time of offline training sizes.

results in performance degradation. In the following experiments, we report the results of the input features with 300 dimensions.

F. Classification Performance

We evaluate the proposed method against seven baselines, including methods using traditional features and CNN-based methods. Borth *et al.* [61] extracted 1200-D features from the ANP detector of SentiBank. Chen *et al.* [62] employed 2089-D features by the pretrained DeepSentiBank. Yao *et al.* [63] noticed that deep features may be insufficient to interpret the emotions and derived a novel adaptive deep metric learning (ADML) method for emotion classification. Besides, we also add two recent methods to compare the performance of the FI data set. Zhang *et al.* [64] presented an object semantics sentiment correlation model (OSSCM) based on a Bayesian network for image sentiment classification. She *et al.* [9] proposed an MKN to learn the kernel coefficients of sentiment features from the weakly and strongly supervised models. To investigate the effectiveness of the matrix transformation in the proposed method, we also provide the performance comparison with VGGNet-19 [56] and AlexNet [57]. The deep features output from the last fully connected layer of each backbone are fed into the SVM and softmax layer, respectively. For a fair comparison, we directly compare with the performances reported by other methods instead of reimplementing them.

Fig. 6 shows the classification accuracy of our method and other state-of-the-art algorithms. From this figure, we have the following observations.

1) The CNN-based deep features can obtain better performance compared with handcrafted features. This proves the powerful representation ability of deep learning on this visual task.
2) No matter whether we use SVM as a classifier or directly use the softmax layer, the features extracted by VGGNet are more effective than those learned by AlexNet. Moreover, the softmax layer achieves better results than the SVM classifier.
3) The proposed algorithm performs the best among the mentioned methods. Although the data sets vary on sample scales and image contents, our method still outperforms other approaches used for comparison.

G. Online Learning Analysis

In this section, we demonstrate the advantages of online learning on a large-scale data set. The FI data set is divided into three parts: 1) 6800 offline training images; 2) 13 700 online training images; and 3) 2808 test images. To make a fair comparison, we consider three different cases. First, to testify the influence of the offline data, we conduct experiments by gradually increasing the offline training samples from 1 to 6800, when the online learning setting remains unchanged. Second, to analyze the impact of the online part, we test our method by increasing the online training samples from 1 to 13 700, when the offline training data keeps invariable. Third, we calculate the time consumption of different sizes of the offline training data.

Fig. 7 shows the classification results over 2808 test images. From this figure, we can draw the following conclusions.

1) The size of the offline training data affects the final results of the proposed algorithm. However, with the increasing online training images, our method can still obtain competitive performance.
2) The more online training data we involved, the higher classification accuracy it obtains. This observation is consistent with recent work [6]. A large number of training data is helpful to build the intrinsic geometry of the sample space.

Fig. 7 also illustrates the running time corresponding to different sizes of the offline training data. As the number of training samples grows, the amount of running time increases monotonously. If we receive a new labeled image, it may take several hours to retrain a new dictionary. Nevertheless, online learning spends less computing time than traditional methods. The proposed method only needs less than 5 min to update the dictionary, when a new labeled picture is involved. Thus, online learning improves the efficiency of the algorithm.

VI. Conclusion

In this article, we focus on image sentiment classification. In order to solve the previous challenges in this field, we proposed a method based on joint sparse and low-rank feature learning. The deep features extracted by pretrained model are encoded by sparse dictionary learning to select relevant

region features from the whole image. The low-rank subspace learning is adopted to project these features into high-level semantic space. To alleviate the problem caused by label ambiguity, label relaxation is involved in our objective function. We also extend our algorithm to an online learning version, which improves the efficiency for classification. Compared with other state-of-the-art methods, the proposed method achieved better performance on three public data sets. In the future, we will try to train the deep learning model jointly with the proposed method for potential performance improvements.

REFERENCES

[1] D. Joshi *et al.*, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, Sep. 2011.

[2] S. Zhao *et al.*, "Affective image content analysis: Two decades review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 2, 2021, doi: 10.1109/TPAMI.2021.3094362.

[3] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, Sep. 2017.

[4] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Process.*, vol. 14, no. 8, pp. 1440–1456, 2020.

[5] D. She, J. Yang, M. Cheng, Y. Lai, P. L. Rosin, and L. Wang, "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1358–1371, May 2020.

[6] E. Adeli *et al.*, "Semi-supervised discriminative classification robust to sample-outliers and feature-noises," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 515–522, Feb. 2019.

[7] S. Zhao, G. Ding, Q. Huang, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: A comprehensive survey," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 5534–5541.

[8] A. Liu, Y. Shi, P. Jing, J. Liu, and Y. Su, "Low-rank regularized multi-view inverse-covariance estimation for visual sentiment distribution prediction," *J. Vis. Commun. Image Represent.*, vol. 57, pp. 243–252, Nov. 2018.

[9] D. She, M. Sun, and J. Yang, "Learning discriminative sentiment representation from strongly- and weakly supervised CNNs," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 3s, p. 96, 2020.

[10] M. Meng, J. Wei, and J. Wu, "Learning multi-part attention neural network for zero-shot classification," *IEEE Trans. Cogn. Devel. Syst.*, early access, Dec. 14, 2020, doi: 10.1109/TCDS.2020.3044313.

[11] P. Gong, X. Wang, Y. Cheng, Z. J. Wang, and Q. Yu, "Zero-shot classification based on multitask mixed attribute relations and attribute-specific features," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 1, pp. 73–83, Mar. 2020.

[12] X. Wang, C. Chen, Y. Cheng, and Z. J. Wang, "Zero-shot image classification based on deep feature extraction," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 2, pp. 432–444, Jun. 2018.

[13] Y. Lu, Y. Chen, D. Zhao, B. Liu, Z. Lai, and J. Chen, "CNN-G: Convolutional neural network combined with graph for image segmentation with theoretical analysis," *IEEE Trans. Cogn. Devel. Syst.*, vol. 13, no. 3, pp. 631–644, Sep. 2021.

[14] A. Mahdi, J. Qin, and G. Crosby, "DeepFeat: A bottom-up and top-down saliency model based on deep features of convolutional neural networks," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 1, pp. 54–63, Mar. 2020.

[15] L. Tang, Z.-X. Yang, and K. Jia, "Canonical correlation analysis regularization: An effective deep multiview learning baseline for RGB-D object recognition," *IEEE Trans. Cogn. Devel. Syst.*, vol. 11, no. 1, pp. 107–118, Mar. 2019.

[16] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 311–320.

[17] S. Zhao *et al.*, "Predicting personalized emotion perceptions of social images," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1385–1394.

[18] S. Zhao, H. Yao, Y. Gao, G. Ding, and T. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 526–540, Oct.–Dec. 2018.

[19] Z. Wu, M. Meng, and J. Wu, "Visual sentiment prediction with attribute augmentation and multi-attention mechanism," *Neural Process. Lett.*, vol. 51, pp. 2403–2416, Feb. 2020.

[20] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.

[21] X. Zhu *et al.*, "Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3595–3601.

[22] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 308–314.

[23] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, Mar. 2019.

[24] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2661–2668.

[25] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[26] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[27] V. Papyan, Y. Romano, J. Sulam, and M. Elad, "Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 72–89, Jul. 2018.

[28] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.

[29] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Inf. Fusion*, vol. 40, pp. 57–75, Mar. 2018.

[30] L. Shao, R. Yan, X. Li, and Y. Liu, "From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1001–1013, Jul. 2014.

[31] Z. Li, Z. Zhang, J. Qin, Z. Zhang, and L. Shao, "Discriminative fisher embedding dictionary learning algorithm for object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 786–800, Mar. 2020.

[32] S. Zhang, H. Yao, X. Sun, and X. Lu, "Sparse coding based visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 46, no. 7, pp. 1772–1788, 2013.

[33] N. Akhtar, F. Shafait, and A. Mian, "Discriminative Bayesian dictionary learning for classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2374–2388, Dec. 2016.

[34] Y. Rong, S. Xiong, and Y. Gao, "Double graph regularized double dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 7707–7721, 2020.

[35] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, 2010.

[36] C. Lu, J. Shi, and J. Jia, "Online robust dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 415–422.

[37] Y. Naderahmadian, S. Beheshti, and M. A. Tinati, "Correlation based online dictionary learning algorithm," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 592–602, Feb. 2016.

[38] Z. Zhao, H. Lu, D. Cai, X. He, and Y. Zhuang, "User preference learning for online social recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2522–2534, Sep. 2016.

[39] F. Wu, X.-Y. Jing, X. You, D. Yue, R. Hu, and J.-Y. Yang, "Multi-view low-rank dictionary learning for image classification," *Pattern Recognit.*, vol. 50, pp. 143–154, Feb. 2016.

[40] M. Meng and X. Zhan, "Zero-shot learning via low-rank-representation based manifold regularization," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1379–1383, 2018.

[41] M. Meng, M. Lan, J. Yu, J. Wu, and D. Tao, "Constrained discriminative projection learning for image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 186–198, 2020.

[42] G. Lerman and T. Maunu, "An overview of robust subspace recovery," *Proc. IEEE*, vol. 106, no. 8, pp. 1380–1410, Aug. 2018.

[43] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, Mar. 2011.

[44] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[45] M. Brbić and I. Kopriva, "ℓ_0 -motivated low-rank sparse subspace clustering," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1711–1725, Apr. 2020.
[46] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.
[47] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," *IEEE Trans. Image Process.*, vol. 25, pp. 850–863, 2016.
[48] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," 2010. *arXiv:1009.5055*.
[49] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2007, pp. 801–808.
[50] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithms for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[51] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
[52] V. Losing, B. Hammer, and H. Wersing, "Incremental on-line learning: A review and comparison of state of the art algorithms," *Neurocomputing*, vol. 275, pp. 1261–1274, Jan. 2018.
[53] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 689–696.
[54] K. C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pac. J. Optim.*, vol. 6, no. 3, pp. 615–640, 2010.
[55] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 83–92.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Red Hook, NY, USA: Curran Assoc., 2012, pp. 1097–1105.
[58] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[59] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1794–1801.
[60] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[61] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
[62] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," 2014, *arXiv:1410.8586*.
[63] X. Yao, D. She, H. Zhang, J. Yang, M.-M. Cheng, and L. Wang, "Adaptive deep metric learning for affective image retrieval and classification," *IEEE Trans. Multimedia*, vol. 23, pp. 1640–1653, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9113756
[64] J. Zhang, M. Chen, H. Sun, D. Li, and Z. Wang, "Object semantics sentiment correlation analysis enhanced image sentiment classification," *Knowl. Based Syst.*, vol. 191, Mar. 2020, Art. no. 105245.