IMAGE SENTIMENT ANALYSIS USING LATENT CORRELATIONS AMONG VISUAL, TEXTUAL, AND SENTIMENT VIEWS

M. Katsurai and S. Satoh

IMAGE SENTIMENT ANALYSIS USING LATENT CORRELATIONS AMONG VISUAL, TEXTUAL, AND SENTIMENT VIEWS

Marie Katsurai

Department of Information Systems Design
Doshisha University
Kyoto, Japan
katsurai@mm.doshisha.ac.jp

Shin'ichi Satoh

Digital Content and Media Sciences Research Division, National Institute of Informatics
Tokyo, Japan
satoh@nii.ac.jp

ABSTRACT

As Internet users increasingly post images to express their daily sentiment and emotions, the analysis of sentiments in user-generated images is of increasing importance for developing several applications. Most conventional methods of image sentiment analysis focus on the design of visual features, and the use of text associated to the images has not been sufficiently investigated. This paper proposes a novel approach that exploits latent correlations among multiple views: visual and textual views, and a sentiment view constructed using SentiWordNet. In the proposed method, we find a latent embedding space in which correlations among the three views are maximized. The projected features in the latent space are used to train a sentiment classifier, which considers the complementary information from different views. Results of experiments conducted on Flickr and Instagram images show that our approach achieves better sentiment classification accuracy than methods that use a single modality only and the state-of-the art method that jointly uses multiple modalities.

Index Terms— image sentiment analysis, multi-view embedding, canonical correlation analysis, SentiWordNet

1. INTRODUCTION

With the popularity of image capturing devices and social media platforms, we have seen a dramatic increase in our ability to collect digital images in various situations and share them on the Web. Two pertinent examples that are currently popular are Flickr, which hosted over 10 billion photos in 2015 [1], and Instagram, which has grown to have more than 400 million monthly active users [2]. These images uploaded by Internet users can be considered to reflect visual aspects of their daily lives. Such ever-growing user-generated images have potential as a new information source to analyze users' opinions and sentiment, which enables several applications including opinion mining about social events, product marketing, and affective human-machine interaction [3]. Thus, automatic inference of the sentiment implied in the images has received increasing research attention in recent years [4–7].

Conventional methods of image sentiment analysis have aimed to design effective visual features for training sentiment polarity classifiers [4–6]. However, due to the *affective gap* between low-level visual features and high-level concepts of human sentiments, it is difficult to directly associate the visual features with sentiment labels. On the other hand, studies about image annotation, not particularly focusing on sentiment analysis, have reported that the collaborative use of textual features around training images (e.g., tags and descriptions) can improve the image content recognition [8, 9]. Inspired from these studies, to bridge images and sentiment, we should investigate how to introduce additional views obtained from textual information to the feature space for training a sentiment classifier.

In this paper, we present a novel image sentiment analysis method that uses latent correlations among visual, textual, and sentiment views of training images. In the proposed method, we first extract features from pairs of images and text to construct visual and textual views. To highlight the sentiment information in the text, we introduce an external sentiment knowledge base, SentiWordNet [10], which forms the sentiment view. Then, using a framework of multi-view canonical correlation analysis (CCA) [11], we calculate a latent embedding space in which correlations among the three views are maximized. Specifically, to capture the nonlinear relationship between features, we introduce explicit feature maps [12, 13] to CCA. Finally, using the features that are projected to the latent embedding space, we train a sentiment classifier. Because the latent space learns the alignments of multiple views, our method corresponds to effectively exploiting the textual information of the training images even if a testing image only has a visual view. Our experiments were conducted on a collection of images from Flickr and Instagram, to which sentiment labels were assigned via crowdsourcing. Results of the experiments show that our three-view approach outperforms the conventional methods.

In summary, the main contributions of this paper are twofold: (i) most conventional methods use only visual features of training images, while we propose a novel image sentiment classification method that can exploit visual, textual, and sentiment views of the training images; and (ii) with experiments designed via crowdsourcing, we show that the complementary use of multiple views of the images can classify image sentiment better than the conventional methods do.

2. RELATED WORK

The idea of associating low-level visual features with sentiments has been investigated based on psychology and art theory using relatively small and controlled datasets [14, 15], while recent works have started to analyze the sentiments of unconstrained real-world images on social media [4–7]. Typically, the goal is to determine the sentiment polarity of images, i.e., positive or negative. To train a sen-

This research has been partly funded by Harris Science Research Institute of Doshisha University.

timent polarity classifier, color histogram and SIFT-based features of images are used in [4]. In [5], emotion-related adjective-noun pairs were selected for image sentiment analysis, and their classifiers, called SentiBank, were trained based on low-level visual features. The detector response of SentiBank was used to form a mid-level representation of an image. Similarly, attribute features including facial expression were used as mid-level features in [6]. These conventional methods focus on how to design visual representation for sentiment analysis, and other available views of the data (e.g., tag concurrence) are discarded in training classifiers. Recently, Wang et al. [7] exploited both visual content and textual information for sentiment-based image clustering in a nonnegative matrix factorization framework. However, the method in [7] has severe sensitivity to the initialization, and the experiments in this paper demonstrate that our method outperforms the conventional method.

The use of correlations among visual and textual features associated to images has improved several image annotation and cross-modal retrieval tasks [8, 9, 16–20], but its effectiveness has not been fully demonstrated in image sentiment analysis. Thus, this paper aims to use the latent correlations among multiple views for better sentiment analysis. Canonical correlation analysis (CCA) [21] is one of the techniques typically used to learn the alignments of multiple views, but it only models the linear relationship between random variables. Several nonlinear extensions such as kernel CCA [11] and Deep CCA [22] have been proposed to reveal nonlinear relationship between the variables. However, these methods are intractable for large-scale datasets due to their high computational complexity and memory use. In contrast, recent advances of explicit feature maps [12, 13] can convert nonlinear problems to linear problems, which can be solved by linear frameworks with a low computation cost [9, 23]. Following these studies, we introduce the explicit feature maps to CCA in the proposed method.

3. IMAGE SENTIMENT ANALYSIS USING LATENT CORRELATIONS AMONG MULTIPLE VIEWS

This section presents a novel image sentiment analysis method that uses latent correlations among multiple views. An overview of the proposed method is shown in Fig. 1. As shown, we first extract features from each view (See 3.1). Then, after learning the multi-view embedding space (See 3.2), the latent embedding space is used to train an image sentiment polarity classifier (See 3.3).

3.1. Design of views for learning a latent embedding space

Our image sentiment analysis approach exploits three types of features: visual, textual, and sentiment views. This subsection describes the details of feature extraction from each view.

Visual features: Following the feature design used in recent visual classification methods [9, 18, 19], we represent image appearance using a combination of different visual descriptors: a 3×256 dimensional histogram extracted from RGB color channels, a 512 dimensional GIST descriptor, a Bag-of-Words quantized descriptor using a $1,000$ word dictionary with a 2-layer spatial pyramid and max pooling. We also extract the following mid-level features: 2,000-dimensional attribute features [24] and 1,200-dimensional SentiBank outputs [5]. For GIST features, attribute features, and SentiBank features, we use the random Fourier feature mapping [12] to approximate the Gaussian kernel. All other histogram-based features were mapped using the exact Bhattacharyya kernel mapping [13]. Finally, similar to [9], we reduce each kernel-mapped feature to 500 dimensions using PCA and the final concatenated feature results in a 2,500-dimensional vector.

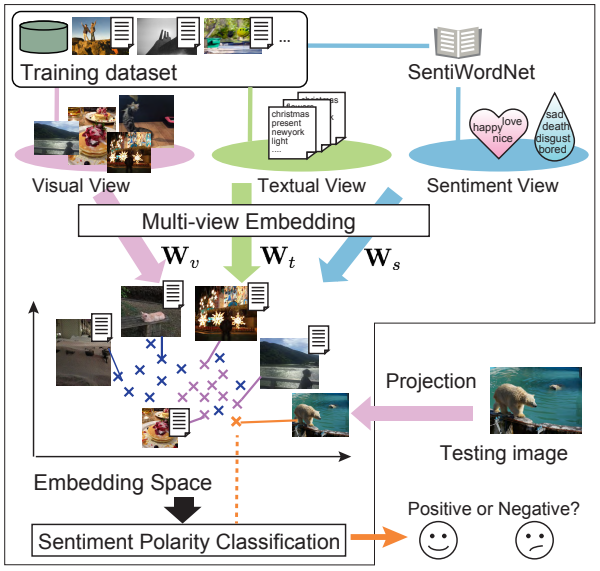

Fig. 1. An overview of the proposed method.

Textual features: The second view consists of textual features, which are extracted from text associated to images. We first construct a vocabulary from a training dataset and represent the textual features of an image using a traditional bag-of-words approach, which counts how many times a word appears in text around the image. Following [8, 9], we use the linear kernel for the textual features, which counts the number of words shared between two images. Since this representation is highly sparse, we exploit SVD for large and sparse matrices [25] to reduce the dimensions of the textual feature matrix. In this paper, we experimentally set the dimension of final textual representation to $1,500$.

Sentiment features: The third view aims to characterize the sentiment aspect of the associate text. For this, we use an external knowledge base, called SentiWordNet [10]. It is based on the well-known English lexical dictionary WordNet [26], and has been utilized in text-based opinion mining tasks [27]. In SentiWordNet, three types of sentiment scores, "positivity," "negativity," or "objectivity," are assigned to each WordNet synset. We use these scores to construct a vocabulary of sentiment-related words. Specifically, we select words whose sentiment scores of either positive or negative are larger than a pre-defined threshold. Then, based on the constructed vocabulary, we calculate the sentiment features of an image in the bag-of-words approach. Finally, we apply the SVD to the feature matrix to reduce its dimensionality. The resulting feature is represented as a 20-dimensional vector.

We will use v, t, s to denote the indexes of the visual, textual, and sentiment views, respectively.

3.2. Finding Latent Correlations Among Multiple Views

This subsection describes how to find latent correlations among multiple views using a framework of the generalization of canonical

correlation analysis [11]. Let $\mathbf{X}_i$ ($i \in \{v, t, s\}$) denote the feature matrix of the i-th view, and the similarity between two feature vectors $\boldsymbol{x}, \boldsymbol{x}'$ in the i-th view is defined by a kernel function K_i such that $K_i(\boldsymbol{x}, \boldsymbol{x}') = \varphi_i(\boldsymbol{x})\varphi_i(\boldsymbol{x}')$. We want to find projection matrices $\mathbf{W}_i$ which maps the i-th view into the latent embedding space. The canonical correlation problem can be transformed into a distance problem such that the distances in the resulting space between each pair of views for the same image are minimized [11]. The objective function to learn the latent space is as follows:

$$\begin{aligned} &\min_{\mathbf{W}_v, \mathbf{W}_t, \mathbf{W}_s} \sum_{i,j \in \{v,t,s\}} \|\varphi_i(\mathbf{X}_i)\mathbf{W}_i - \varphi_j(\mathbf{X}_j)\mathbf{W}_j\|_F^2 \\ &\text{subject to } \ \mathbf{W}_i^T \mathbf{\Sigma}_{ii} \mathbf{W}_i = \mathbf{I}, \ \ \boldsymbol{w}_{ik}^T \mathbf{\Sigma}_{ij} \boldsymbol{w}_{jl} = 0, \ \ i, j \in \{v, t, s\}, i \neq j \\ &\qquad k, l = 1, \cdots, d, \quad k \neq l. \end{aligned} \tag{1}$$

where $\mathbf{\Sigma}_{ij}$ is a covariance matrix between $\varphi_i(\mathbf{X}_i)$ and $\varphi_j(\mathbf{X}_j)$, and $\boldsymbol{w}_{ik}$ represents the k-th column of the matrix $\mathbf{W}_i$. In the conventional kernel CCA [11], kernel trick is used in Eq. (1). To reduce the computation complexity, one can use explicit feature maps [12, 13]. Let $\hat{\phi}(\boldsymbol{x})$ denote an explicit feature mapping such that $K_i(\boldsymbol{x}, \boldsymbol{x}') = \hat{\phi}(\boldsymbol{x})\hat{\phi}(\boldsymbol{x})$. Instead of using the kernel trick, the mapping $\hat{\phi}(\boldsymbol{x})$ can be substituted to the objective function [9]. Solving the following generalized eigenvalue problem provides the solution of Eq. (1):

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \boldsymbol{w}_3 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{S}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \boldsymbol{w}_3 \end{pmatrix}, \tag{2}$$

where $\mathbf{S}_{ij} = \hat{\varphi}_i(\mathbf{X}_i)\hat{\varphi}_j(\mathbf{X}_j)$ is the covariance matrix between the i-th and j-th views, and $\boldsymbol{w}_i$ is a column of $\mathbf{W}_i$. This multi-view formulation has recently proven to be effective for cross-modal retrieval and image annotation [9, 19]. In the following subsection, we describe how to use the latent space learned from multiple views for image sentiment analysis.

3.3. Sentiment polarity classification using latent correlations among multiple views

Using the projection matrices $\mathbf{W}_i$, the features of the i-th view of the training images can be represented in the latent space as follows:

$$\mathbf{P}_i = \hat{\varphi}_i(\mathbf{X}_i)\mathbf{W}_i\mathbf{D}^p, \tag{3}$$

where $\mathbf{D}$ is a diagonal matrix whose elements are the eigenvalues of each dimension in the embedding space. p is a weighting parameter, which is set to 4 as in [9, 19]. Using Eq. (3) for each view, we represent the final feature matrix of training images as the concatenation of $\mathbf{P}_v$, $\mathbf{P}_t$, and $\mathbf{P}_s$. If we consider the case in which text of testing images is unavailable, we concatenate the projection $\mathbf{P}_v$ to the original feature, following the conventional cross-modal retrieval method [18]. Based on the new feature representation of the training dataset with sentiment labels, we learn a sentiment polarity classifier. In this paper, we exploit a linear SVM, which is also used in the conventional methods [4, 5]. Note that although this paper focuses on binary classification as well as the conventional methods [4, 5], our method can be easily extended to multi-class sentiment classification (e.g., positive, negative, and neural). Given a testing image, we also extract features from available views (either or both of visual and textual views) and classify the features projected to the embedding space.

Table 1. The number of positive and negative images in each dataset.

	Positive	Negative
Flickr dataset	48,139	12,606
Instagram dataset	33,076	9,780

4. EXPERIMENTS

4.1. Dataset construction

To conduct experiments, we collected a set of images from Flickr and Instagram as follows.

- **Flickr dataset**. From Flickr, we first downloaded a set of image IDs provided by [28]. Some images were unavailable, and limiting the number of images for each Flickr user to 70, we obtained 105, 587 images. The most frequent words are "view," "black," "photo," "canon," "nikon," and "film."
- **Instagram dataset**. This dataset was constructed by ourselves from Instagram. Using each of the emotional words listed in SentiWordNet as a query keyword, we crawl a set of images. The total number of images was 120, 000. This dataset contains more images that reflect users' daily lives than Flickr dataset. The most frequent words are "love," "like," "life," "day," and "new."

In this experiment, we extracted textual and sentiment features from tags and descriptions associated to images.

To evaluate the performance of image sentiment classification, we prepared sentiment labels of images via crowdsourcing. Conventional methods exploited pseudo sentiment labels using the automatic annotation algorithm based on image tags [4, 7], but it is unreliable due to the noisy tags or lack of tags. To the best of our knowledge, this paper is the first to provide sentiment polarity labels to large-scale image datasets by crowdsourcing-based human annotations. Specifically, we chose CrowdFlower[1] as a platform, and presented each image for subjective evaluation. For each image, three workers were asked to provide a sentiment score. They could choose on a discrete five-point scale labeled with "highly positive," "positive," "neutral," "negative," and "highly negative." The final construction of the ground truth exploited the majority votes of polarity for each image. Table 1 shows the details of the number of positive and negative images in each dataset. Since this experiment targets on the binary classification problem following the previous works [4, 5], we discarded the images labeled by "neutral" and the images resulting in disagreement among workers. Note that our method can be extended to the multi-class classification problem, which will be performed in our future work. The datasets with sentiment labels is available on the Web[2].

4.2. Baselines

We compare the performance of our multi-view embedding-based approach with the following conventional methods, which exploit either visual or textual view: a low-level visual feature-based method [4] (denoted as **Low**), a mid-level visual feature-based method [5] (denoted as **SentiBank**), a method that concatenates low-level visual features with the mid-level features (denoted as **Low&SentiBank**), and a textual feature-based method [10] (denoted as **SentiStrength**[3]). Note that for Low [4], we use the same

[1] http://www.crowdflower.com/

[2] http://mm.doshisha.ac.jp/senti/CrossSentiment.html

[3] http://sentistrength.wlv.ac.uk/

Table 2. Average and standard deviation of the classification accuracy of image sentiment polarity for 10 runs in each dataset. Note that for Low [4], we use the same visual feature set as those described in Sec. 3.1, except for SentiBank outputs.

Method	Flickr dataset	Instagram dataset
Random	49.78 ± 1.05%	50.06 ± 1.09%
Low [4]	69.44 ± 0.85%	67.16 ± 1.28%
SentiBank [5]	70.01 ± 0.63%	67.26 ± 1.12%
Low&SentiBank	70.54 ± 1.00%	68.03 ± 1.36%
SentiStrength [29]	59.30 ± 0.87%	62.78 ± 0.91%
USEA [7]	51.87 ± 1.76%	52.61 ± 2.00%
LC(V+T)+P(V)	70.94 ± 0.67%	68.29 ± 1.42%
LC(V+S)+P(V)	70.67 ± 0.78%	65.44 ± 1.16%
LC(V+T+S)+P(V)	72.36 ± 0.41%	68.54 ± 1.14%
LC(V+T)+P(V+T)	74.42 ± 0.67%	72.43 ± 1.54%
LC(V+S)+P(V+S)	68.98 ± 1.01%	69.35 ± 1.08%
LC(T+S)+P(T+S)	64.63 ± 0.91%	66.50 ± 0.49%
LC(V+T+S)+P(V+T+S)	**74.77** ± **0.82**%	**73.60** ± **0.88**%

visual feature set as those described in Sec. 3.1, except for SentiBank outputs. By comparing these methods in terms of using a single view of the testing data, we investigate the effectiveness of our multi-view embedding approach. For each method, we used Liblinear[4] to train a linear SVM, and the soft margin parameter C of the linear SVM was determined by cross validation. We also compare our method with the state-of-the-art method that exploits visual and textual features of the testing data [7] (denoted as **USEA**). For reference, the random classification results are shown as **Random**.

To validate the contribution of the latent correlations among multiple views, we split up different views with and without embedding. The views used for calculating latent correlations are denoted by LC, and the features projected from images for classification are shown by P. **LC(V+T)** will refer to the two-view embedding based on visual and tag features; **LC(V+T+S)** to the three-view embedding based on visual, textual, and sentiment features; **P(V)** to the projection of only visual features of the images; and **P(V+T)** to the projection of visual and textual features of the images.

4.3. Performance evaluation and discussion

Each dataset was randomly separated into a training set and a test set for 10 runs. In the Flickr dataset, for each sentiment polarity at each run, we randomly sampled 11,346 images and 1,260 images for training and testing, respectively. In the Instagram dataset, for each sentiment polarity at each run, we randomly sampled 8,802 images and 978 images for training and testing, respectively. As a performance evaluation metric, we calculated the average and standard deviation of classification accuracy over all runs. The results are shown in Table 2. As shown, our method using the three views of the training and testing images obtained the best average classification accuracy. This result validates the effectiveness of the complementary use of the multiple views for image sentiment analysis. Even in the case in which the textual and sentiment views of a testing image are unavailable due to the lack of associated text (i.e., LC(V+T+S)+P(V)), our approach presents a better representation of visual features because the latent space learns the alignments of multiple views.

Examples of classification results by our three-view embedding-based method are shown in Fig. 2, in which the red border indicates a misclassified image. We found some difficult cases that cannot be accurately classified by the proposed method. For example, current visual features do not characterize facial expression, letters and drawings in the images. Thus, the design of better features of each view will be performed in our future work.

(a) Images classified as positive.

(b) Images classified as negative.

Fig. 2. Examples of images classified as positive or negative in Flickr dataset. Image captions correspond to Flickr user IDs. The red border indicates a misclassified image.

5. CONCLUSION AND FUTURE WORK

In this paper, we present a novel image sentiment analysis method that uses the latent correlations among multiple views of training images. In the proposed method, we first extract features from visual, textual, and sentiment views. Then, to project the features from these views, we follow the framework of multi-view CCA using explicit feature mappings. Finally, in the embedding space, a sentiment polarity classifier is trained based on the projected features. To validate the effectiveness of the proposed method, we constructed image datasets via crowdsourcing. Experiments conducted on the datasets show that our multi-view embedding space is more effective for classifying image sentiment polarity than methods that use a single modality only and the state-of-the art method that jointly uses multiple modalities.

The features used in our framework should be investigated for further performance improvement. We will introduce additional views or features such as facial expressions [6]. In addition, we will introduce the deep learning-based features [30, 31], which have significantly improved many computer vision tasks, into the proposed framework. Furthermore, we will tackle the multi-class sentiment classification such as positive, negative, and neutral.

[4]http://www.csie.ntu.edu.tw/˜cjlin/liblinear/

6. REFERENCES

[1] flickrBLOG, "Find every photo with flickrs new unified search experience," http://blog.flickr.net/en/2015/05/07/flickr-unified-search/, May 2015, Last accessed: 09/24/2015.

[2] Instagram Blog, "Celebrating a community of 400 million," http://blog.instagram.com/post/129662501137/150922-400million, Sep 2015, Last accessed: 09/24/2015.

[3] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[4] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 715–718.

[5] 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. Int. Conf. Multimedia (MM)*, 2013, pp. 223–232.

[6] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: Image sentiment analysis from a mid-level perspective," in *Proc. Int. Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, 2013, pp. 10:1–10:8.

[7] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li, "Unsupervised sentiment analysis for social media images," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2015.

[8] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 902–909.

[9] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[10] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, 2006, pp. 417–422.

[11] D Hardoon, S Szedmak, and J Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, Dec 2004.

[12] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Neural Information Processing Systems (NIPS)*, 2007.

[13] F. Perronnin, J. Sánchez, and Y. Liu, "Large-scale image categorization with explicit data embedding," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 2297–2304.

[14] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek, "Emotional valence categorization using holistic image features," in *Proc. Int. Conf. Image Processing (ICIP)*, Oct 2008, pp. 101–104.

[15] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 83–92.

[16] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.

[17] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2085–2098, Oct 2015.

[18] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Computer Vision ECCV 2014*, vol. 8692 of *Lecture Notes in Computer Science*, pp. 529–545. Springer International Publishing, 2014.

[19] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Computer Vision ECCV 2014*, vol. 8690 of *Lecture Notes in Computer Science*, pp. 584–599. Springer International Publishing, 2014.

[20] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, June 2014.

[21] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936.

[22] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Machine Learning (ICML)*, 2013, pp. 1247–1255.

[23] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," in *Proc. Int. Conf. Machine Learning (ICML)*, 2014.

[24] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 771–778.

[25] R. M. Larsen, "Lanczos bidiagonalization with partial reorthogonalization," Tech. Rep. 537, Department of Computer Science, Aarhus University, 1998.

[26] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[27] C. Hung and H.-K. Lin, "Using objective words in SentiWordNet to improve word-of-mouth sentiment classification," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 47–54, March 2013.

[28] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, and J. Tang, "How do your friends on social media disclose your emotions?," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2014, pp. 306–312.

[29] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment in short strength detection informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.

[30] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. Int. Conf. Multimedia (MM)*, 2014, pp. 675–678.

[31] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 2121–2129.