Affective Image Content Analysis: A Comprehensive Survey*

Sicheng Zhao[†], **Guiguang Ding**[‡], **Qingming Huang**[♯],
Tat-Seng Chua[§], **Björn W. Schuller**[◊] and **Kurt Keutzer**[†]
[†]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA
[‡]School of Software, Tsinghua University, China
[♯]School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[§]School of Computing, National University of Singapore, Singapore
[◊]Department of Computing, Imperial College London, UK

Abstract

Images can convey rich semantics and induce strong emotions in viewers. Recently, with the explosive growth of visual data, extensive research efforts have been dedicated to affective image content analysis (AICA). In this paper, we review the state-of-the-art methods comprehensively with respect to two main challenges – affective gap and perception subjectivity. We begin with an introduction to the key emotion representation models that have been widely employed in AICA. Available existing datasets for performing evaluation are briefly described. We then summarize and compare the representative approaches on emotion feature extraction, personalized emotion prediction, and emotion distribution learning. Finally, we discuss some future research directions.

1 Introduction

With the rapid development of photography technology and the wide popularity of social networks, people have become used to sharing their experiences and expressing their opinions online using images and videos together with text. This gives rise to a huge volume of multimedia data, which results in urgent demand of processing and understanding the visual content. Compared to low-level visual appearances, humans could better perceive and understand the high cognitive level and affective level of semantics [Zhao *et al.*, 2014a]. Existing works on image content analysis from the computer vision community mainly focus on understanding the cognitive aspects of images, such as object detection. Recently, a lot of research attention has been paid from the psychology, affective computing and multimedia communities to affective level analysis of image content. As what people feel may directly determine their decision making, affective image content analysis (AICA) is of great importance, which can enable wide applications [Chen *et al.*, 2014], ranging from human-computer interaction to image retrieval.

*Corresponding author: Sicheng Zhao (schzhao@gmail.com)

Specifically, the task of AICA is often composed of three steps: human annotation, visual feature extraction and learning of mapping between visual features and perceived emotions [Zhao *et al.*, 2017a]. One main challenge for AICA is the affective gap, which can be defined as "the lack of coincidence between the features and the expected affective state in which the user is brought by perceiving the signal" [Hanjalic, 2006]. Recently, various hand-crafted or learning-based features have been designed to bridge this gap. Current AICA methods mainly assign an image with the dominant (average) emotion category (DEC) with the assumption that different viewers react similarly to the same image. This task can be performed as a traditional single-label learning problem.

However, labeling the emotions in ground-truth generation is in fact highly inconsistent. Different viewers may have totally different emotional reactions to the same image, which is caused by many personal and situational factors, such as the cultural background, personality and social context [Peng *et al.*, 2015; Zhao *et al.*, 2016; Yang *et al.*, 2017b]. This phenomenon causes the so-called subjective perception problem, as shown in Figure 1. In such cases, just predicting the DEC is insufficient for this highly subjective variable.

To tackle the subjectivity issue, we can conduct two kinds of AICA tasks [Zhao *et al.*, 2016]: for each viewer, we can predict personalized emotion perceptions; for each image, we can assign multiple emotion labels. For the latter one, we can employ multi-label learning methods, which associates one instance with multiple emotion labels. However, the importance or extent of different labels is in fact unequal. In such cases, emotion distribution learning would make more sense, which aims to learn the degree to which each emotion describes the instance [Yang *et al.*, 2017b].

In this paper, we concentrate on reviewing the state-of-the-art methods on AICA and outlining research trends. First, we introduce the widely-used emotion representation models and the available datasets for performing AICA evaluation. Second, we summarize and compare the representative approaches on emotion feature extraction, personalized emotion prediction, and emotion distribution learning, corresponding to the affective gap and perception subjectivity challenges. Finally, we discuss potential research directions to pursue.

(a) Original image

Title: Light in Darkness
Tags: london, stormy, dramatic weather, …
Description: …there was a break in the clouds such that a strip of the far bank and bridge were lit up while the sky behind remained somewhat foreboding. I thought it made for a pretty intense scene….

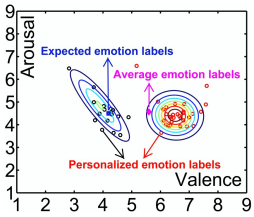
(b) Image metadata

Expected emotion:
Emotion category: fear
Sentiment: negative
Valence: 4.1956
Arousal: 4.4989
Dominance: 4.8378

(c) Expected emotion

Comments from viewers	Personalized emotions
Wow, that is fantastic...it looks so incredible, ….. That sky is amazing.	Emotion: awe Sentiment: positive V: 7.121 A: 4.479 D: 6.635
Yup a fave for me as well. Exciting drama at its best.	Emotion: excitement Sentiment: positive V: 7.950 A: 6.950 D: 7.210
Hey, it really frightened me! My little daughter just looked scared.	Emotion: fear Sentiment: negative V: 2.625 A: 5.805 D: 3.625

(d) Personalized emotion labels



(e) Emotion distribution

Figure 1: Illustration of the image emotion's perception subjectivity phenomenon [Zhao *et al.*, 2017c]. The expected emotions in (c) and the personalized emotions in (d) are assigned using the keywords in red based on the metadata (b) from the uploader and the comments from different viewers. (e) shows the differences of expected, personalized and average emotions, while the contour lines are the estimated continuous emotion distributions by the expectation maximization algorithm based on Gaussian mixture models.

2 Emotion Models from Cognitive Science Community

Psychologists mainly employ two typical models to represent emotions: categorical emotion states (CES), and dimensional emotion space (DES). CES models consider emotions to be one of a few basic categories, such as *fear* and *joy*, *etc*. Some typical models include Ekman's six basic emotions [Ekman, 1992] and Mikels's eight emotions [Mikels *et al.*, 2005]. Specifically, emotions can be classified into *positive* and *negative* [Zhao *et al.*, 2018b], sometimes including *neutral*. In such case, "emotion" is usually called "sentiment". DES models usually employ a 3D or 2D Cartesian space to represent emotions, such as valence-arousal-dominance (VAD) [Schlosberg, 1954] and activity-temperature-weight [Lee and Park, 2011]. VAD is the most widely used DES, where valence represents the pleasantness ranging from happy to unhappy, arousal represents the intensity of emotion ranging from excited to calm, while dominance represents the degree of control ranging from controlled to in control. Dominance is difficult to measure and is often omitted, leading to the commonly used two dimensional VA space [Hanjalic, 2006], where V is easier to recognize in AICA than A.

The relationship between CES and DES and the transformation from one to the other are studied in [Sun *et al.*, 2009]. For example, positive valence relates to a happy state, while negative valence relates to a sad or angry state. CES is easier for users to understand and label, but may not reflect the subtlety and complexity of emotions. DES is more flexible and richer in descriptive power, but absolute continuous values are difficult for users to distinguish and may be unmeaningful because of the lack of agreed-upon standards for subjective emotion rating. CES is mainly used in classification task, while DES is often employed for regression. If discretized into several constant scores, DES can also be used for classification [Lee and Park, 2011]. Ranking based labeling can be applied to ease DEC comprehension difficulties in raters.

Expected emotion is the emotion that the image creator intends to make people feel, while perceived emotion is the actual emotion that is perceived by the image viewers. Discussing the difference or correlation of various emotion models is out of the scope of this paper. We just list some typical emotion models that have been widely used in AICA, as shown in Table 1.

Model	Reference	Type	Emotion states/dimensions
Ekman	[Ekman, 1992]	CES	happiness, sadness, anger, disgust, fear, surprise
Mikels	[Mikels *et al.*, 2005]	CES	amusement, anger, awe, contentment, disgust, excitement, fear, sadness
Plutchik	[Plutchik, 1980]	CES	($\times$ 3 scales) anger, anticipation, disgust, joy, sadness, surprise, fear, trust
Sentiment		CES	positive, negative, (or neutral)
VA(D)	[Schlosberg, 1954]	DES	valence-arousal(-dominance)
ATW	[Lee and Park, 2011]	DES	activity-temperature-weight

Table 1: Representative emotion models employed in AICA.

3 Datasets

The early datasets for AICA mainly come from the psychology community with small-scale images. Recent large-scale datasets are constructed using images from social networks. The released datasets are summarized in Table 2.

The International Affective Picture System (**IAPS**) is an emotion evoking image set in psychology [Lang *et al.*, 1997]. It consists of 1,182 documentary-style natural color images depicting complex scenes, such as *portraits*, *babies*, *animals*, *landscapes*, *etc*. Each image is associated with an empirically derived mean and standard deviation (STD) of VAD ratings in a 9-point rating scale by about 100 college students.

The (**IAPSa**) dataset includes 246 images [Mikels *et al.*, 2005] selected from IAPS, which are labeled by 20 undergraduate participants with a mean age 19.55 years.

The (**Abstract**) includes 279 peer rated abstract paintings without contextual content, while the Artistic database (**ArtPhoto**) consists of 806 artistic photographs from a photo sharing site [Machajdik and Hanbury, 2010], the emotions of which are determined by the artist who uploaded the photo.

The IAPSa, ArtPhoto, and Abstract datasets are categorized into eight discrete categories [Mikels *et al.*, 2005]. Each image is labeled with only one emotion in ArtPhoto, the images in IAPSa may be labeled with more than one emotion, while the detailed votes of all emotions are provided for the images in Abstract, from which one can obtain the DEC and emotion distributions. Totally, 395 emotions are labeled on the 246 images in IAPSa. 228 Abstract images are usually used for emotion classification after DEC selection.

The Geneva affective picture database (**GAPED**) is composed of 520 negative images, 121 positive images and 89 neutral images [Dan-Glauser and Scherer, 2011], labeled by 60 participants with a mean age of 24 years (ranging from 19

Dataset	Reference	# Images	Type	# Annotators	Emotion model	E/P	Label detail
IAPS	[Lang *et al.*, 1997]	1,182	natural	≈100 (half f)	VAD	P	empirically derived mean and standard deviation
IAPSa	[Mikels *et al.*, 2005]	246	natural	20 (10f,10m)	Mikels	P	at lease one emotion category for each image
Abstract	[Machajdik and Hanbury, 2010]	279	abstract	≈230	Mikels	P	the detailed votes of all emotions for each image
ArtPhoto	[Machajdik and Hanbury, 2010]	806	artistic	–	Mikels	E	one DEC for each image
GAPED	[Dan-Glauser and Scherer, 2011]	730	natural	60	Sentiment, VA	P	one DEC and average VA values for each image
MART	[Alameda-Pineda *et al.*, 2016]	500	abstract	25 (11f,14m)	Sentiment	P	one DEC for each image
devArt	[Alameda-Pineda *et al.*, 2016]	500	abstract	60 (27f,33m)	Sentiment	P	one DEC for each image
Tweet	[Borth *et al.*, 2013]	603	social	9	Sentiment	P	one sentiment category for each image
FlickrCC	[Borth *et al.*, 2013]	≈500,000	social	–	Plutchik	P	one emotion category for each image
Flickr	[Yang *et al.*, 2014]	301,903	social	6,735	Ekman	P	one emotion category for each image
Emotion6	[Peng *et al.*, 2015]	1,980	social	432	Ekman+neutral	P	the discrete probability distribution
FI	[You *et al.*, 2016b]	23,308	social	225	Mikels	P	one DEC for each image
IESN	[Zhao *et al.*, 2016]	1,012,901	social	118,035	Mikels, VAD	B	the emotion of involved users for each image
FlickrLDL	[Yang *et al.*, 2017b]	10,700	social	11	Mikels	P	the discrete probability distribution
TwitterLDL	[Yang *et al.*, 2017b]	10,045	social	8	Mikels	P	the discrete probability distribution

Table 2: Released and freely available datasets for AICA, where '# Images' and '# Annotators' respectively represent the total number of images and annotators (f: female, m: male), 'E, P, B' are short for expected emotion, perceived emotion and both, respectively.

to 43, STD = 5.9). Besides, these images are also rated with valence and arousal values, ranging from 0 to 100 points.

The **MART** dataset is a collection of 500 abstract paintings from the Museum of Modern and Contemporary Art of Trento and Rovereto [Alameda-Pineda *et al.*, 2016], which were realized by professional artists. The **devArt** dataset is a collection of 500 amateur abstract paintings obtained from the "DeviantArt" online social network [Alameda-Pineda *et al.*, 2016], one of the largest online art communities. Using the relative score method in [Sartori *et al.*, 2015], these abstract paintings are labeled as positive or negative sentiment.

Tweet dataset (**Tweet**) includes 470 positive tweets and 133 negative tweets [Borth *et al.*, 2013].

The **FlickrCC** dataset is constructed by retrieving the Flickr creative common (CC) images for the 3,000 adjective noun pairs (ANPs) [Borth *et al.*, 2013]. By excluding the images that do not contain the ANP string in the title, tag or description, about 500k Flickr CC images in total are generated for 1,553 ANPs. These images are then mapped to the Plutchnik's Wheel of Emotions with 8 basic emotions, each with 3 scales, such as ecstasy→joy→serenity.

The **Flickr** dataset consists of 301,903 images based on the Ekman emotion model [Yang *et al.*, 2014]. A word list for each of the six emotion categories is manually defined based on WordNet and HowNet. The emotion category whose word list has the most same words as the adjective words of an image's tags and comments is assigned to the image.

The original **FI** dataset consists of 90,000 noisy images collected from Flicker and Instagram by searching the emotion keywords [You *et al.*, 2016b]. The weakly labeled images are further labeled by 225 Amazon Mechanical Turk (AMT) workers, which are selected through a qualification test. The 23,308 images that receive at least three votes from their assigned 5 AMT workers are kept. The number of images in each Mikels emotion category is larger than 1,000.

The **Emotion6** dataset [Peng *et al.*, 2015] consists of 1,980 images collected from Flickr by using the emotion keywords and synonyms as search terms. There are 330 images for each emotion category. AMT workers were invited to label the images into the Ekman's 6 emotions and neutral to obtain the emotional responses. Each image was scored by 15 subjects. The discrete emotion distribution information is released.

The **IESN** dataset is constructed for personalized emotion prediction [Zhao *et al.*, 2016], with 1,012,901 images from Flickr. Lexicon-based methods are used to segment the text of metadata from uploaders for expected emotions and comments from viewers for actual emotions. Synonym based searching is employed to obtain the Mikels' emotion category by selecting the most frequent synonyms. The average VAD values of the segmentation results are computed as DES ground truth based on the VAD norms of 13,915 English lemmas [Warriner *et al.*, 2013]. 7,723 active users with more than 50 involved images are selected. The DEC and emotion distributions can also be easily obtained.

Two image datasets for discrete emotion distribution learning are released in [Yang *et al.*, 2017b]. One is **FlickrLDL** dataset, a subset of FlickrCC. FlickrLDL contains 10,700 images, which are labeled by 11 viewers using Mikels' emotion model. 30,000 images are collected by searching various sentiment key words from Twitter. After duplication removal, the images are labeled by 8 viewers. In this way, the **TwitterLDL** dataset is generated with 10,045 images. In both datasets, the ground truth emotion distribution for each image is obtained by integrating the votes from the workers.

4 Emotion Feature Extraction

As shown in [Zhao *et al.*, 2014c], there are various types of features that may contribute to the perception of image emotions. In this section, we introduce the hand-crafted features of different levels (Table 3) and the learning based features (Table 5) that have been widely extracted for AICA.

4.1 Low-level Features

Low-level features suffer from being difficult to be interpreted by humans. The widely used features include GIST, HOG2x2, self-similarity and geometric context color histogram features as in [Patterson and Hays, 2012], because they are each individually powerful and can describe distinct visual phenomena in a scene perspective.

Features derived from elements of art, including *color* and *texture* are extracted [Machajdik and Hanbury, 2010]. Lee and Park [2011] used the MPEG-7 visual descriptors, including four color-related ideas and two texture-related ideas. Lu *et al.* [2012] investigated how shape features in natural images influence emotions by modeling the concepts of

Feature	Reference	Level	Short description	# Feat
LOW_C	[Patterson and Hays, 2012]	low	GIST, HOG2x2, self-similarity and geometric context color histogram features	17,032
Elements	[Machajdik and Hanbury, 2010]	low	color: mean saturation, brightness and hue, emotional coordinates, colorfulness, color names, Itten contrast, Wang's semantic descriptions of colors, area statistics; texture: Tamura, Wavelet and gray-level co-occurrence matrix	97
MPEG-7	[Lee and Park, 2011]	low	color: layout, structure, scalable color, dominant color; texture: edge histogram, texture browsing	≈200
Shape	[Lu *et al.*, 2012]	low	line segments, continuous lines, angles, curves	219
IttenColor	[Sartori *et al.*, 2015]	low	color co-occurrence features and patch-based color-combination features	16,485
Attributes	[Patterson and Hays, 2012]	mid	scene attributes	102
Sentributes	[Yuan *et al.*, 2013]	mid	scene attributes, eigenfaces	109
Composition	[Machajdik and Hanbury, 2010]	mid	level of detail, low depth of field, dynamics, rule of thirds	45
Aesthetics	[Wang *et al.*, 2013]	mid	figure-ground relationship, color pattern, shape, composition	13
Principles	[Zhao *et al.*, 2014a]	mid	principles-of-art: balance, contrast, harmony, variety, gradation, movement	165
BoVW	[Rao *et al.*, 2016a]	mid	bag-of-visual-words on SIFT, latent topics	330
FS	[Machajdik and Hanbury, 2010]	high	number of faces and skin pixels, size of the biggest face, amount of skin w.r.t. the size of faces	4
ANP	[Borth *et al.*, 2013]	high	semantic concepts based on adjective noun pairs	1,200
Expressions	[Yang *et al.*, 2010]	high	automatically assessed facial expressions (anger, contempt, disgust, fear, happiness, sadness, surprise, neutral)	8

Table 3: Summary of the hand-crafted emotion features at different levels. '# Feat' indicates the dimension of each feature.

Reference	Feature	Fusion	Learning model	Dataset	Target	Result
[Machajdik and Hanbury, 2010]	Elements, Composition, FS	early	Naive Bayes	IAPSa, Abstract, ArtPhoto	cla	0.471, 0.357, 0.495
[Lee and Park, 2011]	MPEG-7	–	K nearest neighbor	unreleased	cla	0.827
[Lu *et al.*, 2012]	Shape, Elements	early	SVM, SVR	IAPSa, IAPS	cla, reg	0.314; V-1.350, A-0.912
[Li *et al.*, 2012]	Segmented objects	–	bilayer sparse learning	IAPS, ArtPhoto	cla	0.612, 0.610
[Yuan *et al.*, 2013]	Sentributes	–	SVM, logistic regression	Tweet	cla	0.824
[Wang *et al.*, 2013]	Aesthetics	–	Naive Bayes	Abstract, ArtPhoto	cla	0.726, 0.631
[Zhao *et al.*, 2014a]	Principles	–	SVM, SVR	IAPSa, Abstract, ArtPhoto, IAPS	cla, reg	0.635, 0.605, 0.669; V-1.270, A-0.820
[Zhao *et al.*, 2014c]	LOW_C, Elements, Attributes, Principles, ANP, Expressions	graph	multi-graph learning	IAPSa, Abstract, ArtPhoto, GAPED, Tweet	ret	0.773, 0.735, 0.658, 0.811, 0.701
[Sartori *et al.*, 2015]	IttenColor	–	sparse group Lasso	MART, devArt	cla	0.751, 0.745
[Rao *et al.*, 2016a]	BoVW	–	multiple instance learning	IAPSa, Abstract, ArtPhoto	cla	0.699, 0.636, 0.707
[Alameda-Pineda *et al.*, 2016]	IttenColor	–	matrix completion	MART, devArt	cla	0.728, 0.761

Table 4: Representative DEC works on AICA using hand-crafted features, where 'Fusion' indicates the fusion strategy of different features, 'cla, reg, ret' in the Target column are short for classification, regression and retrieval (the same below), respectively, 'Result' is the reported best accuracy for classification, mean squared error for regression, and discounted cumulative gain for retrieval on the corresponding datasets.

roundness-angularity and simplicity-complexity. Based on Itten's color wheel, Sartori *et al.* [2015] designed two types of features to represent the color combinations.

4.2 Mid-level Features

Mid-level features are more interpretable, more semantic and more relevant to emotions than low-level features. 102 attributes are detected [Patterson and Hays, 2012], including 5 different types: materials, surface properties, functions or affordances, spatial envelop attributes and object presence. Besides the 102 attributes, Yuan *et al.* [2013] also incorporated eigenfaces corresponding to different emotions, which contribute a lot to the images containing faces. Rao *et al.* [2016a] extracted SIFT features as basic feature and adopted bag-of-visual-words (BoVW) to represent the multi-scale blocks. The latent topic distribution estimated by probabilistic latent semantic analysis is used as another mid-level representation.

Harmonious composition is essential in an artwork and Machajdik and Hanbury [2010] extracted several features to analyze an image's compositional character. Interpretable aesthetic features [Wang *et al.*, 2013] are designed based on the fact that artists often jointly use figure-ground relationships, color patterns, shapes and their diverse combinations to express emotions in their art creations. Features inspired from principles of art are designed in [Zhao *et al.*, 2014a].

4.3 High-level Features

High-level features are the semantic contents contained in images. People can easily understand the emotions conveyed in images by recognizing the semantics. In the early years, Machajdik and Hanbury [2010] extracted simple semantic content by detecting faces and skins contained in an image. Facial expressions may determine the emotion of the images containing faces. 8 kinds of facial expressions are extracted as high-level features [Yang *et al.*, 2010]. The expressions of images detected without faces are set as *neutral*. An 8 dimensional vector, each element of which indicates the number of related facial expressions in the image, is generated.

Borth *et al.* [2013] proposed to describe the semantic concepts by 1,200 adjective noun pairs (ANPs), which are detected by SentiBank. The advantages of ANP are that it turns a neutral noun into an ANP with strong emotions and makes the concepts more detectable, as compared to nouns and adjectives, respectively. A 1,200 dimensional double vector representing the probability of the ANPs is obtained.

Some representative works based on hand-crafted features of the above 3 levels are summarized in Table 4. Generally, high-level features (such as ANP) perform better for images with rich semantics, mid-level features (such as Principles) perform better for artistic photos, while low-level features (such as Elements) are effective for abstract paintings.

4.4 Learning-based Features

With the advent of deep learning, emphasis has been shifted from designing hand-crafted features to learning features in an end-to-end fashion. To tackle the weakly labeled images, You *et al.* [2015] proposed to progressively select a potentially cleaner subset of the training instances. An initial

Reference	Base net	Pre	# Feat	Cla	Loss	Dataset	Target	Result
[You *et al.*, 2015]	self-defined	no	24	–	–	FlickrCC, (unreleased) Twitter	cla	0.781
[You *et al.*, 2016b]	AlexNet	yes	4096	SVM	–	FI, IAPSa, Abstract, ArtPhoto	cla	0.583, 0.872, 0.776, 0.737
[Rao *et al.*, 2016b]	AlexNet,ACNN,TCNN	yes	4096,256,4096	MIL	-	FI, IAPSa, Abstract, ArtPhoto, MART	cla	0.652, 0.889, 0.825, 0.834, 0.764
[Zhu *et al.*, 2017]	self-defined	no	512	–	contrastive	FI, IAPSa, ArtPhoto	cla	0.730, 0.902, 0.855
[Yang *et al.*, 2018]	GoogleNet-Inception	yes	1024	–	sentiment	FI, IAPSa, Abstract, ArtPhoto	cla, ret	0.676, 0.442, 0.382, 0.400 0.780, 0.819, 0.788, 0.704

Table 5: Representative works on deep learning based AICA methods, where 'Pre' indicates whether the network is pre-trained using ImageNet, '# Feat' indicates the dimension of last feature mapping layer before the emotion output layer, 'Cla' indicates the classifier used after the last feature mapping with default Softmax, 'Loss' indicates the loss objectives besides the common cross-entropy loss, and 'Result' is the reported best accuracy for classification and discounted cumulative gain for retrieval on the corresponding datasets.

convolutional neural network (CNN) model is trained on the training data. According to the prediction score of the trained model on the training data itself, the training samples with distinct sentiment scores between the two classes with a high probability are selected. The trained model is fine-tuned using the newly selected instances. Later, they fine-tuned the pre-trained AlexNet on ImageNet to classify emotions into 8 categories by changing the last layer of the neural network from 1000 to 8 [You *et al.*, 2016b]. An SVM classifier is also trained using features extracted from the second to the last layer of the pre-trained AlexNet model.

Rao *et al.* [2016b] proposed to learn multi-level deep representations for image emotion classification (MldrNet). The input image is segmented into 3 levels of patches, which are input to 3 different CNN models, including Alexnet, aesthetics CNN (ACNN) and texture CNN (TCNN). Multiple instance learning (MIL) is employed to generate the emotion label of an input image. Based on MldrNet, Zhu *et al.* [2017] integrated the different levels of features by a Bidirectional GRU model (BiGRU) to exploit their dependencies. Two features generated from our Bi-GRU model are concatenated as the final features to predict the emotions. Apart from the traditional cross-entropy loss, an additional contrastive loss is jointly optimized to enforce the feature vectors extracted from each pair of images from the same category to be close enough, and those from different categories to be far away.

To explore the correlation of emotional labels with the same polarity, Yang *et al.* [2018] employed deep metric learning and proposed a multi-task deep framework to optimize both retrieval and classification tasks. Besides the cross-entropy loss, a novel sentiment constraint is jointly optimized by considering the relations among emotional categories in the Mikels' wheel, which extends triplet constraints to a hierarchical structure. A sentiment vector based on the texture information from the convolutional layer is proposed to measure the difference between affective images.

The deep representations generally outperform the hand-crafted features, which are designed based on several small-scale datasets for specific domains. However, how the deep features correlate to specific emotions is unclear.

5 Personalized Emotion Prediction

Zhao *et al.* [2016; 2018d] proposed to predict the personalized emotions (see Figure 1 (d)) of a specified user after viewing an image, associated with online social networks. Different types of factors that may influence the emotion perception are considered: the images' visual content, the social context related to the corresponding users, the emotions' temporal evolution, and the images' location information. Rolling multi-task hypergraph learning is presented to jointly combine these factors. Each hypergraph vertex is a compound triple (u, x, S), where u represents user, x and S are the current image and the recent past images, termed as 'target image' and 'history image set', respectively. Based on the 3 vertex components, different types of hyperedges are constructed, including target image centric, history image set centric, and user centric hyperedges. Visual features (512-dimensional GIST, Elements, Attributes, Principles, ANP, Expressions) in both target image and history image set are extracted to represent visual content. User relationship is exploited from the user component to take social context into account. Past emotion is inferred from history image set to reveal temporal evolution. Location is embedded in both target image and history image set. Semi-supervised learning is conducted on the multi-task hypergraphs to classify personalized emotions for multiple users simultaneously. The average F1 of emotion classification on the IESN dataset is 0.582.

6 Emotion Distribution Learning

According to probability theory, there are typically two types of probability distributions: discrete and continuous. Generally, distribution learning can be formalized as a regression problem. For CES, the task aims to predict the discrete probability of different emotion categories, the sum of which is equal to 1 (see Figure 2). For DES, the task usually transfers to predict the parameters of specified continuous probability distributions, the form of which should be firstly determined, such as Gaussian distribution (see Figure 1 (e)) and exponential distribution.

6.1 Discrete Probability Distribution Learning

Zhao *et al.* [2015] proposed shared sparse learning (SSL) to represent the probability distribution of one test image as the linear combination of the training images' distributions, the coefficients of which are learned from the feature space. Only uni-modal features are considered in SSL, which can simply adopt early or late fusion to handle multi-modal features (indicate multiple features from images unless otherwise specified) without considering the latent correlations. Multi-modal features are fused by weighted multi-modal SSL (WMMSS-L) [Zhao *et al.*, 2017a; 2018a] to explore useful information by the constraint of joint sparsity across different features. The representation abilities of different features are jointly explored with the optimal weight automatically learned.

Reference	Feature	Fusion	Learning model	Dataset	Result
[Zhao *et al.*, 2015]	GIST, Elements, Principles	-	SSL	Abstract	0.134
[Zhao *et al.*, 2017a]	GIST, Elements, Attributes, Principles, ANP, deep features from AlexNet	weighted	WMMSSL	Abstract, Emotion6, IESN	0.482, 0.479, 0.478
[Yang *et al.*, 2017b]	ANP, deep features from VGG16	-	augmented CPNN	Abstract, Emotion6, FlickrLDL, TwitterLDL	0.480, 0,506, 0,469, 0.555
[Zhao *et al.*, 2017b]	GIST, Elements, Attributes, Principles, ANP, deep features from AlexNet	weighted	WMMCPNN	Abstract, Emotion6, IESN	0.461, 0.464, 0.470
Reference	**Base net**	**Pre**	**# Output**	**Loss**	**Dataset**	**Result**
[Peng *et al.*, 2015]	AlexNet	yes	1	Euclidean loss	Emotion6	0.480
[Yang *et al.*, 2017a]	VGG16	yes	6 or 8	KL divergence loss	Emotion6, FlickrLDL, TwitterLDL	0.420, 0,530, 0,530

Table 6: Representative discrete distribution learning works on AICA, where 'Fusion' indicates the fusion strategy of different features, 'Pre' indicates whether the network is pre-trained using ImageNet, '# Output' is the output dimension of the last layer, and 'Result' is the reported best KL divergence on the corresponding datasets except the first line, which is the result on sum of squared difference.

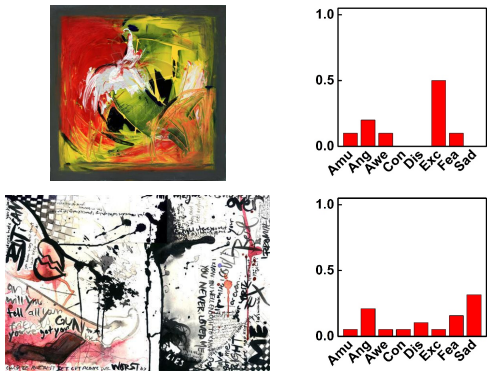


Figure 2: Examples of the image emotion's discrete probability distribution [Zhao *et al.*, 2018a].

Both SSL and WMMSSL can only model one test image each time. For different test images, the shared coefficients have to be learned repeatedly. Geng *et al.* [2013] proposed a conditional probability neural network (CPNN) for distribution learning by modeling the conditional probability of labels given features as a three layer neural network. Yang *et al.* [2017b] replaced the signless integers in CPNN with a binary representation, since adding two emotion labels or subtracting one label from another is meaningless. By adding noises to the ground truth emotion labels, more roughly labeled distributions and samples are generated. Zhao *et al.* [2017b] extended CPNN into multi-modal settings and proposed weighted multi-model CPNN (WMMCPNN) to jointly explore the representation abilities of different features. The linear combination of different CPNN loss functions are minimized with a sparse constraint on the combination coefficients. Once the parameters of the augmented CPNN and WMMCPNN are learned, the emotion distributions of a test image would be easily obtained.

The CPNN series have only 3 layers and the input is extracted visual features instead of original images. A deeper CNN regressor (CNNR) for each emotion category is trained in Emotion6 [Peng *et al.*, 2015] with the exact AlexNet architecture. The number of output nodes is changed to 1 to predict a real value and the Softmax loss layer is replaced with the Euclidean loss layer. The predicted probabilities of all emotion categories are normalized to sum to 1. However, the CNNR cannot guarantee that the predicted probability is non-negative. Further, the probability correlations among different emotions are actually ignored, since the regressor for each emotion category is trained independently. Yang *et al.* [2017a] designed a multi-task deep framework based on VGG16 by jointly optimizing the cross-entropy loss for emotion classification and Kullback-Leibler (KL) divergence loss for emotion distribution learning, which achieves the state-of-the-art performances. For the single emotion dataset, the single label is transformed to emotion distribution with emotion distances computed on Mikels' wheel. The representative methods are summarize in Table 6.

6.2 Continuous Probability Distribution Learning

Zhao *et al.* [2017c] proposed to learn continuous emotion distributions in VA space. Based on the assumption that the VA emotion labels can be well modeled by a mixture of 2 bidimensional Gaussian mixture models (GMMs, see Figure 1 (e)), the task turns to predict the parameters of GMMs, including the mean vector and covariance matrix of the 2 Gaussian components as well as the mixing coefficients.

Shared sparse regression (SSR) is proposed as the learning model by assuming that the test feature and test parameters can be linearly represented by the training features and training parameters but with shared coefficients. SSR can only model one test image each time. To explore the task relatedness, multi-task SSR is further presented to simultaneously predict the parameters of different test images by utilizing appropriate shared information across tasks. GIST, Elements, Attributes, Principles, ANP, and 4,096-dimensional deep features from AlexNet are extracted as visual features. Experiments are conducted on a subset of IESN, which consists of 18,700 images each with more than 20 VA labels. The average KL divergence of multi-task SSR using ANP is 0.436.

7 Conclusion and Future Directions

This paper attempted to provide an overview of recent developments on affective image content analysis (AICA). Obviously, it cannot cover all the literature on AICA, and we focused on a representative subset of the latest methods. We summarized the widely employed emotion representation models, released datasets, and compared the existing methods on emotion feature extraction, personalized emotion prediction and emotion distribution learning. We believe that AICA will continue to be an active and promising research area with broad potential applications, such as an emotion-aware personalized music slide show [Chen *et al.*, 2014], emotion based image musicalization [Zhao *et al.*, 2014b], and image

captioning with sentiment [Mathews *et al.*, 2016]. Many issues in AICA, however, are still open.

Understanding image content and context. Accurately analyzing what is contained in an image can significantly improve the performance of AICA. Sometimes we even need subtle analysis of visual contents. For example, if an image is about the laugh of a lovely child, it is more likely that we feel "amused"; but if it is about the laugh of a known evil ruler or criminal, we may feel "angry". The context of image content is also important in AICA. Similar visual content under different contexts may evoke different emotions. For example, we may feel "happy" about beautiful flowers. But if the flowers are placed in a funeral, we possibly feel "sad".

Recognizing group emotions. Recognizing the dominant emotion is too generic, while predicting personalized emotion is too specific. Modeling emotions for groups or cliques of users, who share similar tastes or interests, may be a good choice – e. g., by cultural or societal background. Analyzing the user profiles provided by each individual to classify users into different types of backgrounds, tastes and interests may help to tackle this problem.

Understanding emotions of 3D data and videos. Compared with traditional intensity images, 3D data contain more spatial information, being useful in low light levels and being color and texture invariant, while videos (such as animated GIF) contain rich temporal correlation information. Combining the spatial and temporal correlation together with the visual content would make more sense.

Further, jointly exploring the complementarity of multimodal data (such as image and text) [You *et al.*, 2016a; Zhao *et al.*, 2018c; Chen *et al.*, 2018] would improve the emotion recognition performance. How to adapt the emotions from one labeled domain to another unlabeled domain [Patel *et al.*, 2015] is another interesting research topic.

Acknowledgments

This work is supported by the Berkeley Deep Drive, the National Natural Science Foundation of China (Nos. 61701273, 61571269, 61332016, 61620106009, U1636214), the China Postdoctoral Science Foundation Project (No. 2017M610897), and the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA). This research is part of NExT++ project, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative.

References

[Alameda-Pineda *et al.*, 2016] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5240–5248, 2016.

[Borth *et al.*, 2013] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM International Conference on Multimedia*, pages 223–232, 2013.

[Chen *et al.*, 2014] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia*, pages 367–376, 2014.

[Chen *et al.*, 2018] Fuhai Chen, Rongrong Ji, Jinsong Su, Donglin Cao, and Yue Gao. Predicting microblog sentiments via weakly supervised multimodal deep learning. *IEEE Transactions on Multimedia*, 20(4):997–1007, 2018.

[Dan-Glauser and Scherer, 2011] Elise S Dan-Glauser and Klaus R Scherer. The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2):468–477, 2011.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[Geng *et al.*, 2013] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[Hanjalic, 2006] Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006.

[Lang *et al.*, 1997] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, pages 39–58, 1997.

[Lee and Park, 2011] Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039, 2011.

[Li *et al.*, 2012] Bing Li, Weihua Xiong, Weiming Hu, and Xinmiao Ding. Context-aware affective images classification based on bilayer sparse representation. In *ACM International Conference on Multimedia*, pages 721–724, 2012.

[Lu *et al.*, 2012] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. On shape and the computability of emotions. In *ACM International Conference on Multimedia*, pages 229–238, 2012.

[Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, pages 83–92, 2010.

[Mathews *et al.*, 2016] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI Conference on Artificial Intelligence*, pages 3574–3580, 2016.

[Mikels *et al.*, 2005] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005.

[Patel *et al.*, 2015] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2012.

[Peng *et al.*, 2015] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015.

[Plutchik, 1980] Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[Rao *et al.*, 2016a] Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, and Ian Burnett. Multi-scale blocks based image emotion classification using multiple instance learning. In *IEEE International Conference on Image Processing*, pages 634–638, 2016.

[Rao *et al.*, 2016b] Tianrong Rao, Min Xu, and Dong Xu. Learning multi-level deep representations for image emotion classification. *arXiv preprint arXiv:1611.07145*, 2016.

[Sartori *et al.*, 2015] Andreza Sartori, Dubravko Culibrk, Yan Yan, and Nicu Sebe. Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *ACM International Conference on Multimedia*, pages 311–320, 2015.

[Schlosberg, 1954] Harold Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81, 1954.

[Sun *et al.*, 2009] Kai Sun, Junqing Yu, Yue Huang, and Xiaoqiang Hu. An improved valence-arousal emotion space for video affective content representation and recognition. In *IEEE International Conference on Multimedia and Expo*, pages 566–569, 2009.

[Wang *et al.*, 2013] Xiaohui Wang, Jia Jia, Jiaming Yin, and Lianhong Cai. Interpretable aesthetic features for affective image classification. In *IEEE International Conference on Image Processing*, pages 3230–3234, 2013.

[Warriner *et al.*, 2013] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

[Yang *et al.*, 2010] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. Exploring facial expressions with compositional features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2638–2644, 2010.

[Yang *et al.*, 2014] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. How do your friends on social media disclose your emotions? In *AAAI Conference on Artificial Intelligence*, pages 306–312, 2014.

[Yang *et al.*, 2017a] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *International Joint Conference on Artificial Intelligence*, pages 3266–3272, 2017.

[Yang *et al.*, 2017b] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence*, pages 224–230, 2017.

[Yang *et al.*, 2018] Jufeng Yang, Dongyu She, Yukun Lai, and Ming-Hsuan Yang. Retrieving and classifying affective images via deep metric learning. In *AAAI Conference on Artificial Intelligence*, 2018.

[You *et al.*, 2015] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI Conference on Artificial Intelligence*, pages 381–388, 2015.

[You *et al.*, 2016a] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *ACM International Conference on Multimedia*, pages 1008–1017, 2016.

[You *et al.*, 2016b] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI Conference on Artificial Intelligence*, pages 308–314, 2016.

[Yuan *et al.*, 2013] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *ACM International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10, 2013.

[Zhao *et al.*, 2014a] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia*, pages 47–56, 2014.

[Zhao *et al.*, 2014b] Sicheng Zhao, Hongxun Yao, Fanglin Wang, Xiaolei Jiang, and Wei Zhang. Emotion based image musicalization. In *IEEE International Conference on Multimedia and Expo Workshops*, pages 1–6, 2014.

[Zhao *et al.*, 2014c] Sicheng Zhao, Hongxun Yao, You Yang, and Yanhao Zhang. Affective image retrieval via multi-graph learning. In *ACM International Conference on Multimedia*, pages 1025–1028, 2014.

[Zhao *et al.*, 2015] Sicheng Zhao, Hongxun Yao, Xiaolei Jiang, and Xiaoshuai Sun. Predicting discrete probability distribution of image emotions. In *IEEE International Conference on Image Processing*, pages 2459–2463, 2015.

[Zhao *et al.*, 2016] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. Predicting personalized emotion perceptions of social images. In *ACM International Conference on Multimedia*, pages 1385–1394, 2016.

[Zhao *et al.*, 2017a] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. Approximating discrete probability distribution of image emotions by multi-modal features fusion. In *International Joint Conference on Artificial Intelligence*, pages 4669–4675, 2017.

[Zhao *et al.*, 2017b] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. Learning visual emotion distributions via multi-modal features fusion. In *ACM International Conference on Multimedia*, pages 369–377, 2017.

[Zhao *et al.*, 2017c] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. *IEEE Transactions on Multimedia*, 19(3):632–645, 2017.

[Zhao *et al.*, 2018a] Sicheng Zhao, Guiguang Ding, Yue Gao, Xin Zhao, Youbao Tang, Jungong Han, Hongxun Yao, and Qingming Huang. Discrete probability distribution prediction of image emotions with shared sparse learning. *IEEE Transactions on Affective Computing*, 2018.

[Zhao *et al.*, 2018b] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. Personality-aware personalized emotion recognition from physiological signals. In *International Joint Conference on Artificial Intelligence*, 2018.

[Zhao *et al.*, 2018c] Sicheng Zhao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. Real-time multimedia social event detection in microblog. *IEEE Transactions on Cybernetics*, 2018.

[Zhao *et al.*, 2018d] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing*, 2018.

[Zhu *et al.*, 2017] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition. In *International Joint Conference on Artificial Intelligence*, pages 3595–3601, 2017.