LEARNING MULTI-LEVEL DEEP REPRESENTATIONS FOR IMAGE EMOTION CLASSIFICATION

Tianrong Rao[1], *Min Xu*[1], *Dong Xu*[2]

[1] Faculty of Engineering and IT, University of Technology, Sydney
tianrong.rao@student.uts.edu.au; min.xu@uts.edu.au
[2] School of Electrical and Information Engineering, University of Sydney
dong.xu@sydney.edu.au

ABSTRACT

In this paper, we propose a new deep network that learns multi-level deep representations for image emotion classification (MldrNet). Image emotion can be recognized through image semantics, image aesthetics and low-level visual features from both global and local views. Existing image emotion classification works using hand-crafted features or deep features mainly focus on either low-level visual features or semantic-level image representations without taking all factors into consideration. Our proposed MldrNet unifies deep representations of three levels, i.e. image semantics, image aesthetics and low-level visual features through multiple instance learning (MIL) in order to effectively cope with noisy labeled data, such as images collected from the Internet. Extensive experiments on both Internet images and abstract paintings demonstrate the proposed method outperforms the state-of-the-art methods using deep features or hand-crafted features. The proposed approach also outperforms the state-of-the-art methods with at least 6% performance improvement in terms of overall classification accuracy.

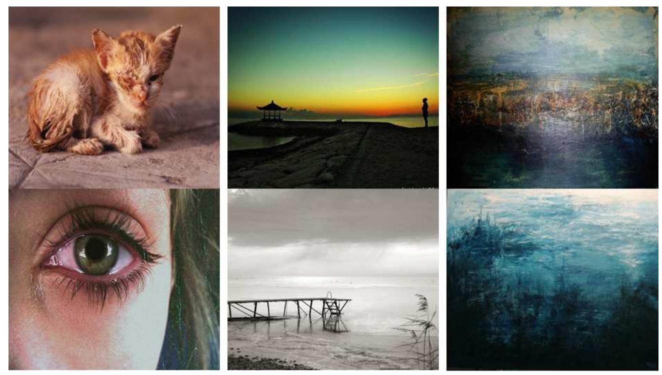

Fig. 1. Sample images from different datasets that evoke the same emotion *sadness*. We can find out that image emotion is related to many factors. Left: web images whose emotions are mainly related to image semantics. Middle: art photos whose emotions are mainly related to image aesthetics, such as compositions and emphasis. Right: abstract paintings whose emotions are mainly related to low-level visual features, such as texture and color.

1. INTRODUCTION

Psychological studies have already demonstrated that humans' emotion reflections vary with different visual stimuli [1, 2]. Inspired by these studies, computer scientists began to predict the emotional reactions of people given a series of visual contents. This creates a new research topic called affective image analysis, which attracts increasing attention in recent years [3, 4, 5, 6]. However, compared to semantic-level image analysis, analyzing images at affective-level is more difficult, due to the two challenges of the complexity and subjectivity of emotions [7].

As shown in Figure 1, image emotion is related to complex visual features from high-level to low-level for both global and local views. Low-level visual features from the local view, such as color, shape, line and texture, were first used to classify image emotions [8, 9, 10, 11]. Joshi *et al.* [2] indicated that image emotion is highly related to image aesthetics and proposed that mid-level features, such as composition, visual balance and emphasis, can be used for image emotion classification. Machajdik and Hanbury suggested that image emotion can be significantly influenced by semantic content of the image [3]. They combined high-level image semantics from the global view with Itten's art theory on relevance of colors [12] to recognize image emotion. However, most of the existing methods rely on hand-crafted features, which are manually designed based on common sense and observation of people. These methods can hardly take all important factors related to image emotion, i.e., image semantics, image aesthetics and low-level visual features, into consideration.

Recently, with the rapid popularity of Convolutional Neural Network (CNN), outstanding breakthroughs have been achieved in many visual recognition tasks, such as image classification [13], image segmentation [14], object

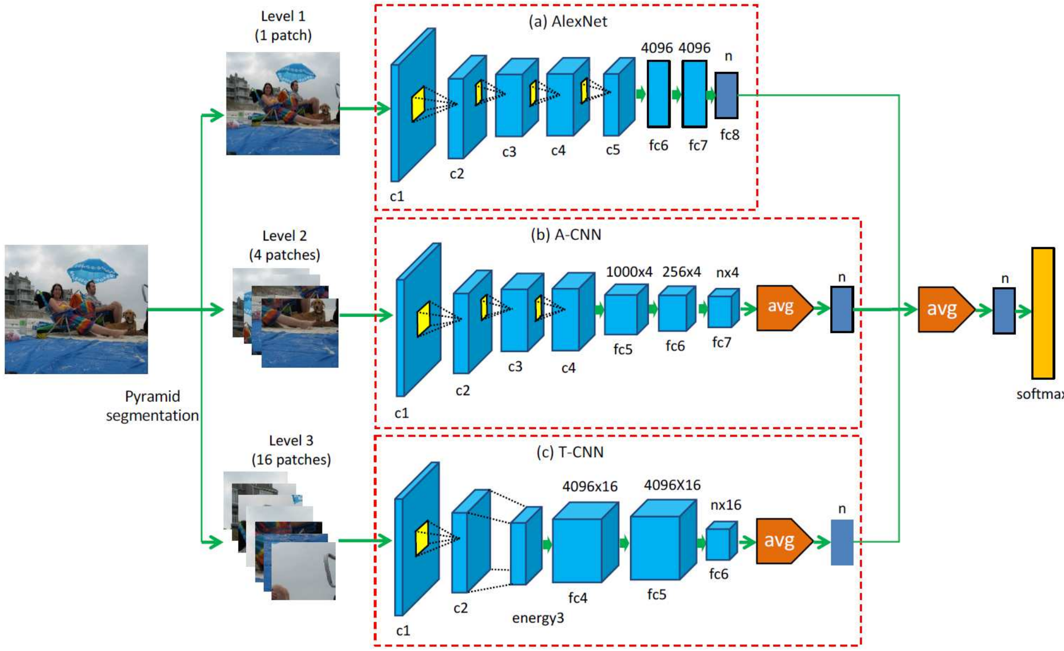

Fig. 2. Overview of our multi-level deep representation network (MldrNet). The input image is first segmented into different levels of patches. Then three different CNN models, including (a) Alexnet, (b) aesthetics CNN (A-CNN) and (c) texture CNN (T-CNN), is applied to extract different levels of deep representations. Finally, MIL is employed to generate the emotion label of an input image.

detection [15] and scene recognition [16]. Instead of designing visual features manually, CNN provides an end-to-end feature learning framework, which can automatically learn deep representations of images. Several researchers have also applied CNN to image emotion classification. However, the currently used CNN methods for visual recognition cannot well deal with mid-level image aesthetics and low-level visual features, such as AlexNet [13]. In [5], the authors indicated that AlexNet is not effective enough to extract emotion information from abstract paintings, whose emotions are mainly conveyed by image aesthetics and low-level visual features.

What's more, the CNN-based methods usually rely on the large scale manually labeled training datasets like the ImageNet dataset [17]. People coming from different culture background may have very different emotion reactions to a same image. Therefore, the emotional textual context associated with Internet images, e.g., titles, tags and descriptions, may not be reliable enough, and result the datasets collected from the Internet for emotion classification may contain noisy and inaccurate emotion labels. The emotion classification accuracy using existing methods, such as AlexNet, could be degraded when using these noisy labeled data as training data [18].

Considering the above mentioned two challenges, in this paper, we propose a new deep network that learns multi-level deep representations (MldrNet) for image emotion classification. Figure 2 shows overview of the proposed MldrNet network. The traditional CNN method is designed for center-position object classification, which cannot effectively extract mid-level image aesthetics and low-level visual features from the local view. To perform end-to-end feature learning for different levels of features from the entire image, we generate image patches from three level by using pyramid segmentation [19]. Three different CNN structures are applied for different scale of patches to extract different levels of deep representations. The multiple patches can naturally fit the assumption of multiple instance learning (MIL) if we consider each patch as an instance and the entire image as a bag. Instead of using emotion label of bag for training directly, generating instances labels into bag label can reduce the need for correct label and effectively deal with the noisy labeled dataset. Therefore, we combine the different deep representations through MIL. Finally, we conduct extensive experiments on several publicly available datasets for emotion recognition from Internet images and abstract paintings to evaluate the effectiveness of our proposed method.

Our main contributions are three-fold.

- We propose a deep multi-level patch learning network, which combines deep representations at three levels, i.e., image semantics, image aesthetics and low-level visual features, for image emotion classification.

- We employ multiple instance learning in our proposed method to unify different deep representations and effectively deal with the noisy labeled dataset.

- We demonstrate that our deep multi-level patch learning network outperforms the state-of-the-art methods using deep features or hand-crafted features through comprehensive experiments on both Internet images and abstract paintings.

2. RELATED WORK

Pervious works of visual emotion classification can be roughly divided into dimensional approaches [20, 21, 22]and categorical approaches [3, 4, 23]. Dimensional approaches represent emotion in a two or three dimensional space, while categorical approaches map emotion directly to one of a few basic emotion categories, which is straightforward for people to understand and label and thus have been widely applied in recent studies. Yanulevskaya *et al.* [24] proposed to categorize emotions for artworks based on Gabor and Wiccest features and Support vector machine (SVM). Solli and Lenz [25] introduced a color-based emotion-related image descriptor, which is derived from psychophysical experiments, to classify images. In [6], different emotional features extracted from both global view and local view were used for emotion prediction. Machajdik *et al.* [3] defined a combination of rich hand-crafted features based on art and psychology theory, including composition, color variance and image semantics. Zhao *et al.* [4] introduced more robust and invariant visual features, which were designed according to art principles to capture information about image emotion. These hand-crafted visual features have only been proven to be effective on several small datasets, whose images are selected from a few specific domains, e.g. abstract paintings and art photos. Considering the recent success from CNN-based approaches in many computer vision tasks, CNN has also been employed in image emotion analysis [5, 18]. However, it fails to extract emotion information related to image aesthetics and low-level visual features.

Multiple instance learning (MIL) is a weakly supervised learning method, which is first introduced for drug design [26]. Since then, a lot of MIL algorithms have been developed, such as the MI-SVM for bag-level classification and mi-SVM for instance-level classification [27]. Since the assumption of MIL can naturally fit into computer vision tasks, MIL has also been widely applied in the field of computer vision, including object detection [28], image retrieval [29] and visual categorization [30]. However, most of these methods applied the hand-crafted features designed for specific tasks.

Recently, researchers started to combine deep representations with MIL framework for various computer vision tasks. Xu *et al.* [31] proposed a MIL framework using features extracted through CNN to analyze medical images. Song *et al.* [32] applied deep MIL for weakly supervised object localization. Wu *et al.* [33] built a MIL framework to learn the correspondences between keywords and image regions for image annotation. In [34], MIL is combined with CNN to estimate pixel-level label of images for image segmentation. Lu *et al.* [35] introduced MIL to combine deep representations from both global and local views for image aesthetics analysis. Different from these methods that learn the same kind deep representations using a single CNN structure, our MldrNet combines deep representations from different levels using different CNN structures.

3. EMOTION ANALYSIS

In this section, we introduce our network that learns multi-level deep representations (MldrNet) for image emotion classification. Specifically, we attempt to combine different levels of learned deep representations within a MIL framework. Previous deep multiple instance learning frameworks usually employ a single CNN structure to extract deep representations, which cannot effectively take all levels of factors related to image emotion into consideration. Consider image emotion is related to different levels of features, i.e., high-level image semantics, mid-level image aesthetics and low level visual features, we attempt to unify different CNN structures within the newly proposed MIL framework. Following the aforementioned discoveries, we divide the images into 8 emotion categories (positive emotion *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotion *Anger*, *Disgust*, *Fear*, *Sadness*) for visual emotion classification.

3.1. Extract Multi-scale Patches

Traditional CNN models are usually applied for semantic-level image analysis in order to extract deep representations about high-level image semantics from a global view. In [6], the authors indicated that both image semantics reflected in the global view and images aesthetics and low-level visual features reflected in the local view are related to image emotion. To extract different levels of deep representations about image emotion, we employ pyramid segmentation to segment the input image into image patches at three different scales and experimentally choose suitable CNN model for each scale. The number of image patches at each scale is 1, 4, 16, separately.

3.2. Convolutional Neural Network

Before introducing our MldrNet, we first review the CNN model that has been widely used for computer vision tasks [13]. Given one training sample $\{(x, y)\}$, where x is the image and y is the associated label, CNN extracts layer-wise representations of input images using convolutional layers and fully-connected layers. Followed by a softmax layer, the output of the last fully-connected layer can be transformed into a probability distribution $\mathbf{p} \in \mathbb{R}^m$ for image emotions of n categories. In this work, $n = 8$ indicates eight emotion categories. The probability that the image belongs to a certain emotion category is defined blow:

$$p_i = \frac{exp(h_i)}{\sum_i exp(h_i)}, i = 1, ..., n, \tag{1}$$

where h_i is the output from the last fully-connected layer. The loss of the predicting probability can also be measured by using cross entropy

$$L = -\sum_i y_i \log(p_i), \tag{2}$$

where $y = \{y_i | y_i \in \{0, 1\}, i = 1, ..., n, \sum_{i=1}^{n} y_i = 1\}$ indicates the true emotion label of the image.

CNN models contain a hierarchy of filters and the level of representations learned from CNN models are higher if one goes "deeper" in the hierarchy [36]. To extract deep representations about mid-level image aesthetics and low-level visual features, we employ different kinds of CNN models inspired by AlexNet, which contain less number of convolutional layers.

As introduced before, AlexNet contains five convolutional layers followed by max-pooling layers, and three fully-connected layers, which contains 4,096, 4,096 and 8 neurons, respectively. The structure of AlexNet is shown in Fig 2(a). AlexNet is mainly trained for semantic-level image classification and tends to extract high-level deep representation about image semantics. It cannot effectively extract emotion information from abstract painting whose emotion is mainly convey by image aesthetics and low-level visual features [5]. To this end, we introduce two other CNN models, aesthetics CNN (A-CNN) [37] and texture CNN (T-CNN) [38], which contain less convolutional layers than AlexNet.

A-CNN is used for image aesthetic analysis, which has a close relationship with image emotion analysis [2]. As shown in Figure 2(b), A-CNN consists of four convolutional layers and three fully-connected layers, which contains 1,000, 256 and 8 neurons, respectively. The first and second convolutional layers are followed by max-pooling layers. We apply A-CNN for the image patches to extract deep representations about image aesthetics.

Texture has been proven as one of the important low-level visual features related to image emotion classification [3]. To extract deep representations about texture of images, we employ T-CNN, which is an efficient CNN model for texture classification [38]. As shown in Figure 2(c), T-CNN removes the last three conventional layers of AlexNet, and adds an energy layer (average-pooling layer with the kernel size as 27) behind the second convolutional layers. Following the energy layer, there are still three fully-connected layers, which contains 4,096, 4,096 and 8 neurons, respectively.

To choose suitable CNN models for different scales of image patches, we conduct comprehensive experiments for the above mentioned three CNN models on image patches of at different scales. The emotion classification accuracies of the three different CNN models are discussed in Section 4.3.

3.3. Deep Network Learning Multi-level Deep representations

Using different CNN models, we can estimate the labels of patches at different levels. Since we only have the emotion label of the whole image instead of each image patch, we are unable to train the fully supervised CNN model. Therefore, we employ multiple instance learning to conduct our deep network to learn multi-level deep representations (MldrNet).

Based on the property of MIL, we define the image x as the bag and the nth image patch x_{mn} extracted from mth level as the instance. Through different CNN models, we can extract the output of the last fully connected layer, i.e., the representation, h^i_{mn} of the instance. The representation of the bag can be generated by concatenating the representations of the instances using our MIL framework. However, unlike the traditional MIL frameworks, the instances in our approach represent the image from the global view and the local view respectively. Therefore, we first concatenate the representations of each level

$$\hat{h}_{im} = f(h^i_{m1}, h^i_{m2}, ..., h^i_{mn}). \tag{3}$$

Then, the representations of the entire bag can be aggregated by using the representations of each level

$$\hat{h}_i = f(\hat{h_{i1}}, \hat{h_{i2}}, ..., \hat{h_{im}}). \tag{4}$$

The aggregation function f can be $max_n(h^i_{mn})$ for each level and $max_m(h_{im})$ for each bag or $avg_n(h^i_{mn})$ for each level and $avg_m(h_{im})$ for each bag. We can easily find out that the function $avg(\cdot)$ assigns the same weight to the patches in the same scale, which represent the entire image. While the function $max(\cdot)$ would encourage the model to increase the probability of the patch which is considered to the be most important for image emotion classification. As we mention before, both image semantics from global view and image aesthetics and low-level visual features from local view can influence image emotion. Therefore, we prefer $avg(\cdot)$, which fits the property of image emotion, as aggregate function.

Previous deep multiple instance learning frameworks usually use $max(\cdot)$ as the aggregate function according to the experimental results. In Section 4.3, we show our MldrNet with the above two different aggregation function.

The distribution of emotion categories of the bag can be calculated as follow:

$$p_i = \frac{exp(\hat{h}_i)}{\sum_i exp(\hat{h}_i)}, i = 1, ..., n, \tag{5}$$

which is the final emotion distribution of the image. The emotion category, which is with the highest emotion distribution, is selected as the emotion category of the image.

4. EXPERIMENTS

In this section, we evaluate the performance of our MldrNet on different datasets. The recently published large scale dataset for emotion recognition [18] and three popular used small datasets: IAPS-Subset [39], ArtPhoto and Abstract [3] are used to evaluate the classification result on different kinds of affective images over 8 emotion categories. The MART dataset [40] is used to evaluate the classification result on abstract paintings over 2 emotion categories (positive and negative).

4.1. Datasets

Large Scale Dataset For Emotion classification: This dataset is newly published in [18] to evaluate the classification result over 8 different emotion categories (positive emotions *Amusement*, *Awe*, *Contentment*, *Excitement* and negative emotions *Anger*, *Disgust*, *Fear*, *Sad*). To collect this dataset, 90,000 noisy labeled images are first downloaded from Instagram and Flickr by using the names of emotion categories as the key words for searching. Then, the downloaded images were submitted to Amazon Mechanical Turk (AMT) for further labeling. Finally, 23,308 well labeled images were collected for emotion recognition [1].

Small Scale Datasets For Emotion Classification: Three small datasets that are widely used in previous works for image emotion classification are introduced below.

(1)**IAPS-Subset**: The *International Affective Picture System* (IAPS) is a standard stimulus image set which has been widely used in affective image classification. IAPS consists of 1,182 documentary-style natural color images depicting complex scenes, such as portraits, puppies, babies, animals, landscapes and others [41]. Among all IAPS images, Mikels *et al.* [39] selected 395 images and mapped arousal and valence values of these images to the above mentioned eight discrete emotion categories.

(2)**ArtPhoto**: In the ArtPhoto dataset, 806 photos are selected from some art sharing sites by using the names of emotion categories as the search terms [3]. The artists, who take the photos and upload them to the websites, determine emotion categories of the photos. The artists try to evoke a certain emotion for the viewers of the photo through the conscious manipulation of the emotional objects, lighting, colors, etc. In this dataset, each image is assigned to one of the eight aforementioned emotion categories.

(3)**Abstract**: This dataset consists of 228 abstract paintings. Unlike the images in the IAPS-Subset and ArtPhoto dataset, the images in the Abstract dataset represent the emotions through overall color and texture, instead of some emotional objects [3]. In this dataset, each painting was voted by 14 different people to decide its emotion category. The emotion category with the most votes was selected as the emotion category of that image.

MART: The MART dataset is a collection of 500 abstract paintings from the Museum of Modern and Contemporary Art of Trento and Rovereto. These artworks were realized since the beginning of the 20 century until 2008 by professional artists, who have theoretical studies on art elements, such as colors, lines, shapes and textures, and reflect the results of studies on their paintings. Using the relative score method in [42], the abstract paintings are labeled as positive or negative according to the emotion type evoked by them.

4.2. Experimental Setup

There are two training steps in our MldrNet. We first train three different CNN models repctively. The only change of the three CNN models is to fix the output number of the last fully connected layer to eight. The weights of the pre-trained CNN models are used to initialize the weights in our MldrNet, which would accelerate weight initialization in MldrNet and improve the classification result. Then, during MldrNet training step, only fully connected layers are fine-tuned to fit the emotion labels of the images.

The initial learning rate is empirically set as 0.001 and the gamma is empirically set as 0.1. Each model is fully trained until the loss stops descending. We use the stochastic gradient descent method to train our model with a batch size of 256, momentum and weight decay of 0.00005. We train all models using Caffe with a Linux server with 2 NVIDIA TITAN GPUs.

4.3. Emotion Classification on Large Scale Dataset

The well labeled 23,164 images are randomly split into the training set (80%, 18,532 images), the testing set (15%, 3,474 images) and the validaton set (5%, 1,158 images). Meanwhile, to demonstrate the effectiveness of our approach on noisy labeled dataset, we create a noisy labeled dataset for training by combining the images, which have been submitted to AMT for labeling but labeled from different emotion categories, with the training set of well labeled images. The noisy labeled dataset contains 83,664 images for training. We called the well labeled dataset as *well* dataset and noisy labeled dataset as *noisy* dataset. The *well* dataset and *noisy* dataset are used for training models. The testing dataset is used to test our models.

4.3.1. Results of Different CNN models

To investigate the deep representations of images extracted from different CNN models, we compare them on different

[1]We have 88,298 noisy labeled images and 23,164 manually labeled images as some images no longer exists in the Internet.

scales of patches. The number of patches in each scale is 1, 4, 16, separately. We use the *well* dataset to fine-tune A-CNN, T-CNN and AlexNet. For the second and third scale, the outputs of the last fully connected layers of patches are concatenated as the output of the layer using the function $avg(\cdot)$. Then, the outputs of the three layers are concatenated as the output of the entire image. Through a softmax layer, the emotion distribution of the image can be calculated by using the output.

Table 1 presents the accuracy of 3 different CNN architectures with different scales of patches. As shown in Table 1, AlexnNet achieves the best classification accuracy on patch from scale 1, which is the entire images containing emotion information about the image semantics from the global view. A-CNN and T-CNN perform better on the patches from scale 2 and 3, whose emotions are respectively conveyed by image aesthetics and low-level texture features from local views. Therefore, the CNN models containing less convolutional layers are good at analyzing mid-level and low-level deep representations about image emotion. Based on our experimental results, we apply different CNN models on the patches from different scales.

Model	Scale 1	Scale 2	Scale 3
AlexNet	**58.61%**	48.19%	38.89%
A-CNN	54.12%	**49.88%**	40.31%
T-CNN	49.40%	44.32%	**42.03%**

Table 1. Emotion classification accuracy for different CNN models on different levels.

4.3.2. Result of Our MldrNet

To demonstrate the effectiveness of our MldrNet for image emotion classification on both well labeled dataset and noisy labeled dataset, we compare the average emotion classification accuracy of our MldrNet with four baseline methods by using the *well* dataset and the *noisy* dataset as training dataset respectively. We report the result by using two classical deep learning methods: AlexNet [13] and AlexNet-SVM [18]. Moreover, we also report the result from two state-of-the-art methods for image emotion classification using hand-crafted features: Zhao's method [4] using principle of art features and Rao's method [6] using hand-crafted features extracted from different image patches. Tabel 2 shows the the emotion classification accuracies obtained by the 5 different methods.

First of all, we observe that our MldrNet outperforms the other methods by using both *well* dataset and *noisy* dataset for training. Secondly, deep representations outperforms the hand-crafted features, which are designed based on several small scale datasets for specific domains. Finally, compare to AlexNet, which mainly focuses on extracting emotion information about image semantics, deep representations related to image aesthetics and low-level texture extracted in our MldrNet contribute to the overall performance

To further analyze the performance of different methods, we report the confusion matrix of different methods on the testing dataset. Considering the significant performance improvements by using deep representations compared to hand-crafted features, we only show the confusion matrices of our MldrNet and AlexNet using the *well* dataset as the training dataset (MldrNet-well and AlexNet-well) and the *noisy* dataset as the training dataset (MldrNet-noisy and AlexNet-noisy). As shown in Figure 3, compared to AlexNet, our MldrNet shows a more balanced emotion classification results using both *well* and *noisy* datasets as the training dataset. Meanwhile, using the *noisy* dataset as the training dataset degrades the emotion classification accuracy using AlexNet and it is more likely to confuse the vague emotion, such as *anger* and *contentment*, with other emotion,.

We also visualize a couple of sample images that are correctly classified by our MldrNet but incorrectly classified by AlexNet to qualitatively analyze the influence of image aesthetics and low-level visual features for image emotion classification. As shown in Figure 4, the emotions of the images misclassified by AlexNet are mainly convey by image aesthetics and low-level visual features. Combining the emotion information related to image aesthetics and low-level texture can significantly improve the classification results.

Methods	*well* dataset	*noisy* dataset
Zhao [4]	46.52%	38.43%
Rao [6]	51.67%	43.39%
AlexNet-SVM [18]	57.89%	46.72%
AlexNet [13]	58.61%	47.29%
our MldrNet	**65.23%**	**56.94%**

Table 2. Emotion classification accuracy for different methods by using the *well* dataset and the *noisy* dataset as the training dataset. Different methods are evaluated on the testing dataset.

4.3.3. Choice of Aggregation Function

Selecting suitable aggregation function is of critical importance for our MldrNet. Previous deep multiple instance learning methods usually use $max(\cdot)$ as the aggregation function. In Table 3, we present the results of our MldrNet with different aggregation functions. The deep representations we used in our MldrNet are from different levels and $avg(\cdot)$ aggregation function achieves better performances when using both *well* dataset and *noisy* dataset as the training dataset.

4.4. Emotion Classification on small Scale Datasets

We have introduced several image emotion analysis methods using hand-crafted features. To better evaluate the effectiveness of our MldrNet, we compare our method with state-of-the-art methods applied on the three small scale datasets.

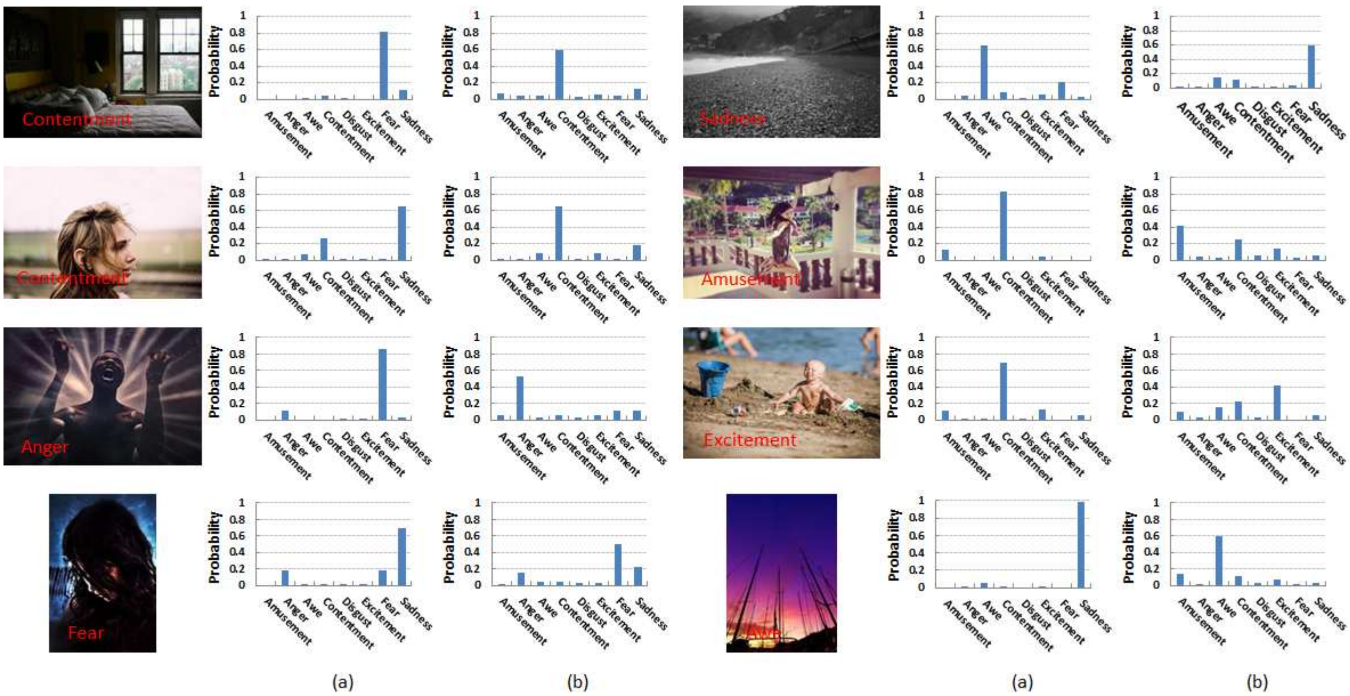

Fig. 4. Sample images correctly classified by our MldrNet but misclassified by AlexNet. The column (a) shows the emotion distribution predicted by AlexNet and the column (b) shows the emotion distribution predicted by our MldrNet. The red label on each image indicates the ground-truth emotion category.

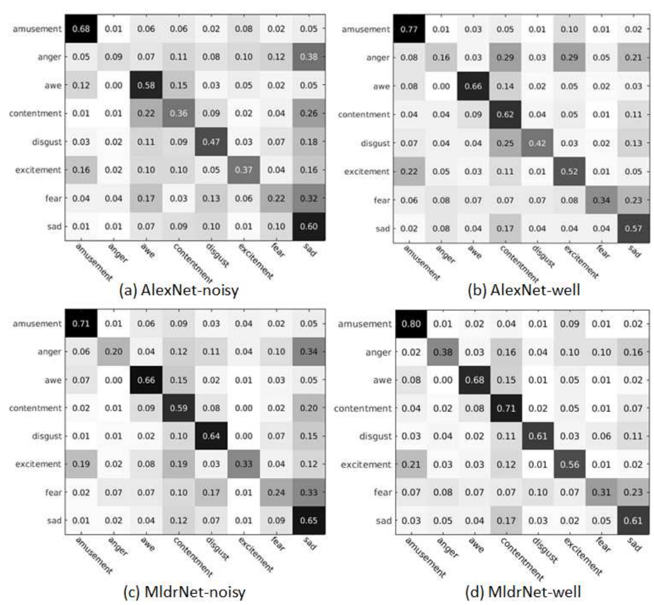

Fig. 3. Confusion matrices for AlexNet and our MldrNet when using the *well* dataset and the *noisy* dataset as training dataset.

Method	*well* dataset	*noisy* dataset
MldrNet-max	62.52%	53.40%
MldrNet-avg	65.23%	56.94%

Table 3. Classification accuracies of our MldrNet with different aggregation functions when using the *well* dataset and the *noisy* dataset as the training dataset.

We follow the same experimental settings described in [3]. Due to the inbalanced and limited number of images per emotion category, we employ the "one against all" strategy to train the classifier. The image samples from each category are randomly split into five batches and 5-fold cross validation strategy is used to evaluate the different methods. We just use the images to train the last fully connected layer in our MldrNet and AlexNet. Also, the *true positive rate per class* suggested in [3] is calculated to compare the results. Note that in IAPS-Subset and Abstract dataset, only eight and three images are contained in the emotion category *anger*, so we are unable to perform 5-fold cross validation for this category. Therefore, the *true positive rate per class* of emotion category *anger* in these two datasets is not reported.

The result of four baseline methods are reported, including Machajdik's method (*machajidk* [3]), Zhao's method (*zhao* [4]), Rao's method (*rao* [6]) and AlexNet [13]. Fig-

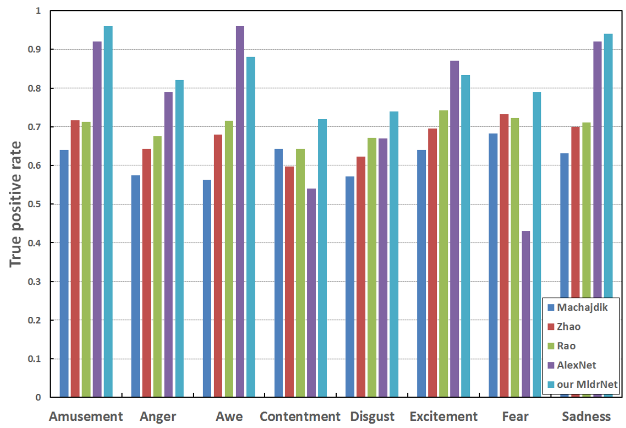

Fig. 5. Classification performance of *machajidk* [3]), *zhao* [4], *rao* [6], AlexNet [13] and our MldrNet on the ArtPhoto dataset.

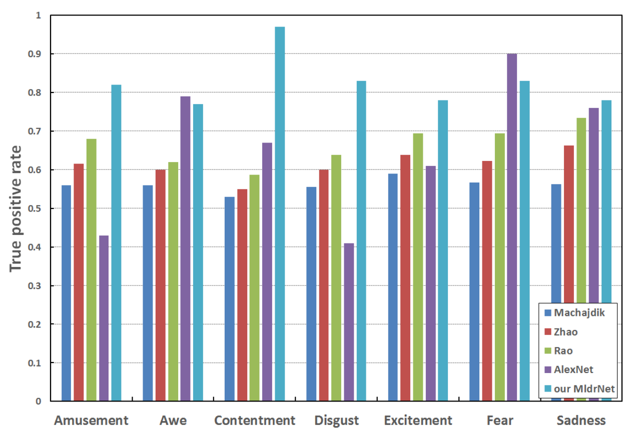

Fig. 6. Classification performance of *machajidk* [3]), *zhao* [4], *rao* [6], AlexNet [13] and our MldrNet on the Abstract dataset.

ure 5,6 and 7 show the performance of different methods on the three small datasets respectively. For most of emotion categories, deep learning methods significantly outperform the state-of-the-art hand-crafted methods. However, the performances of AlexNet in Abstract and ArtPhoto dataset are relatively poor, because emotions of images in these two datasets are mainly conveyed by image aesthetics and low-level visual features. In contrast, our MldrNet achieves the best performance in all three dataset, which shows a robust result.

4.5. Emotion Classification on Abstract Paintings

To further evaluate the benefits of our MldrNet. We also test our MldrNet on the MART dataset, which consists of abstract paintings. Followed the experimental approach in [5], we em-

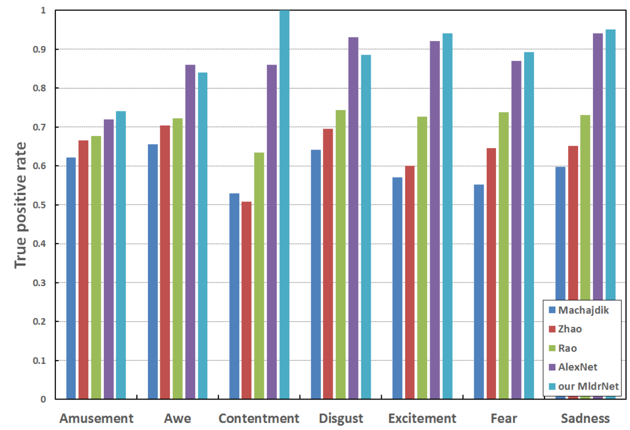

Fig. 7. Classification performance of *machajidk* [3]), *zhao* [4], *rao* [6], AlexNet [13] and our MldrNet on the IAPS-Subset.

Model	Accuracy
TSVM [43]	69.2%
LMC [44]	71.8%
Lasso [42]	68.2%
Group Lasso [42]	70.5%
NLMC [5]	72.8%
AlexNet [13]	69.8%
our MldrNet	**76.4%**

Table 4. Emotion classification accuracy of different methods on the MART dataset.

ploy 10-fold cross validation to compare our MldrNet with other six baseline methods on the MART dataset. The basline methods are: kernel transductive SVM (TSVM [43]), linear matrix completion (LMC [44]), Lasso and Group Lasso both proposed in [42], non-linear matrix completion (NLMC [5]) and AlexNet [13]. The results shown in Table 4 demonstrate that our MldrNet can effectively extract emotion information from abstract paintings when compared with all other methods. Compared to AlexNet, applying A-CNN and T-CNN clearly improves the emotion classification tperformance.

5. CONCLUSION

In this paper, we have proposed a new network that learns multi-level deep representations for image emotion classification. We have demonstrated that image emotion is not only related to high-level image semantics, but also related to mid-level image aesthetics and low-level visual features. Our MldrNet successfully combine the deep representations from different scales of image patches by employing a new multiple instances learning framework. Our MldrNet outperforms the state-of-the-art methods on both Internet images and abstract paintings.

6. REFERENCES

[1] Peter J Lang, "A bio-informational theory of emotional imagery," *Psychophysiology*, vol. 16, no. 6, pp. 495–512, 1979.

[2] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[3] Jana Machajdik and Allan Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM MM*, 2010, pp. 83–92.

[4] 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 MM*, 2014.

[5] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," in *CVPR*, 2016.

[6] Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, and Ian Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *ICIP*, 2016.

[7] Wang Wei-ning, Yu Ying-lin, and Zhang Jian-chao, "Image emotional classification: static vs. dynamic," in *SMC*, 2004.

[8] Hang-Bong Kang, "Affective content detection using hmms," in *ACM MM*, 2003.

[9] Weining Wang and Qianhua He, "A survey on emotional semantic image retrieval.," in *ICIP*, 2008.

[10] Joel Aronoff, "How we recognize angry and happy emotion in people, places, and things," *Cross-cultural research*, vol. 40, no. 1, pp. 83–105, 2006.

[11] Alan Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[12] Johannes Itten and Ernst Van Haagen, *The Art of Color; the Subjective Experience and Objective Rationale of Colour*, Reinhold, 1962.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[16] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[18] 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*, 2016.

[19] Hendrik James Antonisse, "Image segmentation in pyramids," *Computer Graphics and Image Processing*, vol. 19, no. 4, pp. 367–383, 1982.

[20] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *ACM MM*, 2011.

[21] 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 MM*, 2012.

[22] Sergio Benini, Luca Canini, and Riccardo Leonardi, "A connotative space for supporting movie affective recommendation," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1356–1370, 2011.

[23] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *CVPR*, 2015.

[24] Victoria Yanulevskaya, JC Van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek, "Emotional valence categorization using holistic image features," in *ICIP*, 2008.

[25] Martin Solli and Reiner Lenz, "Color based bags-of-emotions," in *CAIP*, 2009.

[26] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[27] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2002.

[28] Cha Zhang, John C Platt, and Paul A Viola, "Multiple instance boosting for object detection," in *NIPS*, 2005.

[29] Wen Li, Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang, "Text-based image retrieval using progressive multi-instance learning," in *ICCV*, 2011.

[30] Sudheendra Vijayanarasimhan and Kristen Grauman, "Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization," in *CVPR*, 2008.

[31] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, I Eric, and Chao Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *(ICASSP*, 2014, pp. 1626–1630.

[32] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell, "Weakly-supervised discovery of visual pattern configurations," in *NIPS*, 2014.

[33] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu, "Deep multiple instance learning for image classification and auto-annotation," in *CVPR*, 2015.

[34] Pedro O Pinheiro and Ronan Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015.

[35] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *ICCV*, 2015.

[36] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[37] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rapid: rating pictorial aesthetics using deep learning," in *ACM MM*, 2014.

[38] Vincent Andrearczyk and Paul F Whelan, "Using filter banks in convolutional neural networks for texture classification," *arXiv preprint arXiv:1601.02919*, 2016.

[39] 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*, vol. 37, no. 4, pp. 626–630, 2005.

[40] Victoria Yanulevskaya, Jasper Uijlings, Elia Bruni, Andreza Sartori, Elisa Zamboni, Francesca Bacci, David Melcher, and Nicu Sebe, "In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings," in *ACM MM*, 2012.

[41] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert, "International affective picture system (iaps): Affective ratings of pictures and instruction manual," *Technical report A-8*, 2008.

[42] 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 MM*, 2015.

[43] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, 1999.

[44] Ching-Hui Chen, Vishal M Patel, and Rama Chellappa, "Matrix completion for resolving label ambiguity," in *CVPR*, 2015.