ArtTalk: Labeling Images with Thematic and Emotional Content

Yifan Wang
Stanford University, VMware
3401 Hillview Ave, Palo Alto, CA, 94304
yifanwang@vmware.com

Mana Lewis
Chez Mana Co.
mana@chezmana.com

Abstract

One picture is worth a thousand words. Artistic work touches us, inspires us and changes us, by creating a strong emotional impact. Artists try to pass two messages via paintings: theme, which is the world from their perspective and emotion that they feel.
In this paper, a CNN (Convolutional Neural Network) model, called ArtTalk, is proposed to uncover the connection between artworks and the two hidden messages. ArtTalk is trained to view an art painting and to tell what is painted in the picture and what the emotion it conveys. In this paper, the design and implementation of ArtTalk are discussed and the model is tested using art paintings from wikiart. It has achieved an accuracy rate of nearly 70% for theme detection and 50% for emotion detection.

1. Introduction

Art paintings are playing a more and more important role in everybody's life, no matter how far the distance from art-related field an individual perceives. Paintings are not quietly sitting in museums and waiting for visitors to admire their beauty, in stead, they can be found everywhere. Election candidates use cartoons to satirize their competitors, advertisement companies use paintings to grab attention and to broadcast their product, and nearly all houses are decorated using art paintings that demonstrate the householders' taste and interest.
To cater for the great need from the society, artists work hard. Hundreds of thousands of art paintings are created on a daily basis in every corner of the world. The ease of communication, a benefit brought by Internet, inspires the creation of artists by allowing them to better understand the work of their counterparts. What is more, Internet exposes artists and their work to the population and this allows artists to grab both commercial and artistic opportunities, which, in turn, leads to the creation of more and more paintings.
Then, with so many artworks, new and old, it has become a problem of how to find the right picture in limited time. It has become infeasible for an individual to view all the artworks before making a selection. This poses a even harsher problem for advertisement companies, where they need to make selections of the right pictures for all the events they are responsible for. Therefore, a tool that can view an image and provides important labels in the content of the drawing and the feel that people might have for it in a timely manner is greatly needed.
Teaching a computer to 'read' an art painting and telling the theme and emotion is the main goal of ArtTalk. The work can be divided into two parts: theme detection and emotion detection. There are researches in both fields but very few effort is made to combine the results together to render well-rounded labels for artworks. In ArtTalk, I create a working model using CNN that can generate tags in both areas and to give a full descriptions of a painting.

First let's look at the work for theme detection. Different from object detections in photography, theme detection in art paintings is more difficult since it needs to cater for strangely shaped objects and abstract concepts. For instance, abstract themes like religion form an important part of artworks, and in fact, that is almost the sole theme for the paintings in medieval period. But this poses a problem for the current object detection CNN models since there is no physical object that can be interpreted as a religion. Moreover, for abstract paintings like the ones by Pablo Picasso, it is very difficult to distinguish his portrait from blocks of color. Thus are the two major problems for theme detection. In order to solve the problem, I used transfer learning by adding additional layers after the mature object detection networks and I have reached an accuracy of more than 67%.
Then for emotion detection, not much progress has been made in this area. Researchers complain that it is not easy to distinguish subtle feelings like happiness and love. In order to conquer the issue, I have used a novel structures with reference to the researches in object detection fields. I have

Mana Lewis is a non-CS231N contributor. She helped label 2000 images using emotion tags.

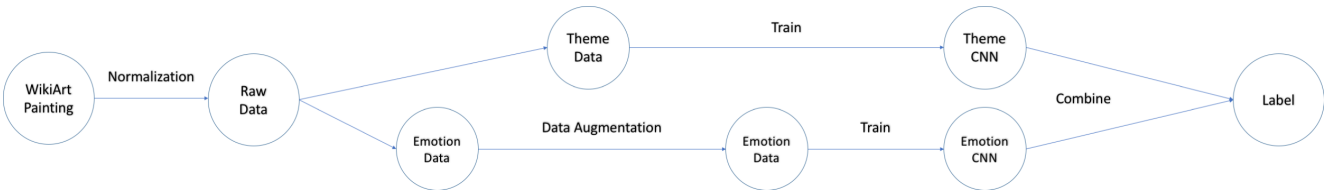

Figure 1. System Design Illustration, this figure depicts the data flow and important operations in ArtTalk.

also compared my result with the results in published papers and mine looks promising.

In the following of the paper, Section 2 covers the state-of-art research results, some of which are used as a baseline to demonstrate the effectiveness of ArtTalk. Section 3 discusses the system model of ArtTalk and it dives deep into the structures of two CNNs, one for theme detection and the other for emotion detection. Then in Section 4, we will learn about the source of data and how I change the format of input data to make them fit into the system model. Also we will discuss the data augmentation method used for emotion tagging. The results and evaluations are provided in Section 5, where we can see the effectiveness of ArtTalk. After that, I conclude my research in Section 6 and uncover the future direction in Section 7.

2. Related Work

In this section, we will discuss the state-of-art research results in related fields. First we will look at the work in theme detection, which are mostly transferred from object detection. Then we will look at the progress on emotion detection.

2.1. Theme Detection

Few efforts have been directly targeted at the theme detection for artistic works. But there exist a lot of work on object detection in photography. Though the application area is different, but the results can be effectively transferred into the theme detection.

First, in similar areas, researchers have proposed methods to study painting styles, like abstraction, vintage and so on in [9]. It uses data from Flickr and other art paintings. In addition to the raw pixels as input data, it adds Lab color histogram [12], GIST [11], Graph-based visual saliency [6] and meta-class binary features [1], and it reaches an accuracy of 77% for style analysis.

In object detection, there are many mature CNN structures like VGG-19 [13], GoogleNet [15] and ResNet [7], which have reached an accuracy rate of 92.7%, 93.3% and 96.6% respectively. There are many new optimization techniques which are proved to be useful in training CNNs, and the following three are used in my network.

- Batch normalization [8] layer after each convolutional layer is proven to benefit the performance of object detection in photographs.
- A few dropout layers [14] inserted into the network will help fight overfitting and encourage better training without introducing much additional computation.
- I have also used average max pooling as the last layer to replace the fully connected layers for generating output tags. This idea is borrowed from GoogLeNet[15].

2.2. Emotion Detection

There are two similar research areas for emotion detection: facial expression emotion study and picture semantic study.

For facial expression emotion study, the closest study is presented in [10]. It encodes the picture using LBP extraction. LBP codes render pictures into something like grey-scale pictures, which is represented by a 2-d array. Then different from traditional approaches, it uses Multi-dimensional scaling (MDS) to map LBP codes into 3-d array. Then it uses ensembles of CNNs to detect the emotions depicted in pictures. Finally it provides an accuracy rate of around 51%.

The key idea is that LBP encoding can help us get rid of illumination and other photometric features and hence it eases the training process. However, I perceive that painters use more illuminations to show emotions, compared to the facial expression. Therefore, I believe that illumination and color rendering is much more important in emotion detection for artistic paintings.

The above study is revisited in [4], where LBP model is used as a feature extraction algorithm. The research then tunes mature models directly for emotion detection, and gets an accuracy of less than 40%. It also applies the results in semantic detection for the pictures, and reports an accuracy rate around 65%.

When the problem comes to semantic detection, it has become much easier, since it only needs to group pictures in two categories: positive and negative. Gudi studies facial expression using the semantic detection and he reports an accuracy rate of 67% in [5]. Using CNN, Xu proposes a simple network structure [16] for sentiment analysis with images coming from twitter and tumblr, and he has reached an accuracy of 65%. Similar structures have also been used

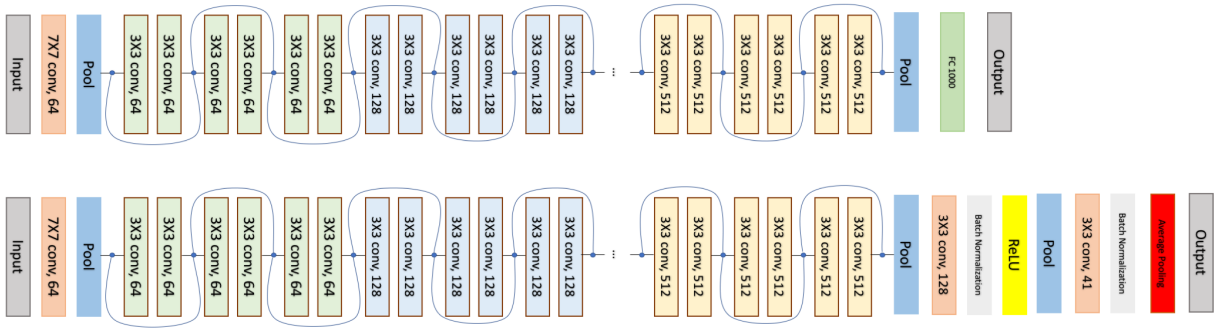

Figure 2. CNN structure for Theme Detection. The structure on the top is vanilla ResNet[7] and the one on the bottom is the novel structure I created using transfer learning.

in [17], with a different data source: photos from flickr, and a similar results of 65% is reported.
Based on the research above, we can easily conclude that the smaller emotion categories we have, the better accuracy we are going to get. For semantic detection, where there are only two groups, researchers can reach an accuracy rate near 70%. However, when coming to detect the subtle emotions, the accuracy rate has dropped to less than 40%.

3. Methods

The goal of ArtTalk is to learn what is painted in a picture and what the emotion the picture conveys. In another word, we need to tag a picture with two labels: thematic label and emotional label. In this section, we will discuss the design for the whole system first. And then, we will continue on the network structures for either task.

3.1. System Design

In ArtTalk, the two learning objectives can be achieved via training two CNNs. The first takes care of learning thematic tags and the other provides emotional tags.
The whole work flow can be illustrated in Figure 1. In training stage, two CNNs are fed with two sets of training data. Theme CNN is responsible for thematic tagging and it receives labeled data using theme, while emotion CNN is used for emotional training and it uses our manually labeled data. Though the two CNNs are trained separately, they are used together. In real application, test data are sent to CNN1 for a thematic tagging and CNN2 for an emotional tagging and their results are combined into a description of the graph.

3.2. Theme CNN

We cannot use mature object detection network for two reasons:

- Theme can be an abstract concept. For instance, Christianity is a common theme in artwork, while there is no physical object called Christianity. And also, the number of output theme category is different from the task of ImageNet Challenge.
- Dataset is artwork, which is totally different from photo. The abstract paintings created by Picasso might be a good example to demonstrate the difference.

Thus, mature CNNs for object detection needs to be altered for the new task. The best results come from a transferred ResNet [7] structure where I have maintained the all the layers and their parameters but the final output layers.
In ResNet, the output layer is a fully connected layer to map the results from the output of convolution layers to labels. This is not feasible in the new task since we have different number of categories and we want to save some computation resource. In my structure, as shown in Figure 2, I have added two additional convolution layers to reduce the number of channels to fit into the output categories. Then I use an average pooling layer instead of a fully connected layer to generate the output tags. All the convolutional layers are followed by batch normalization and ReLU activation layer. I have also tried other structures by adding 1, 2 and 3 convolution layers to the end of both ResNet and VGG16, and use average pooling layers to replace fully connected layers for output. The structure depicted in Figure 2 above has provided the best results so far.

3.3. Emotion CNN

As shown in Figure 1, the emotion labels are generated from Emotion CNNs. The difficulty for creating an appropriate neural network structure are twofolds:

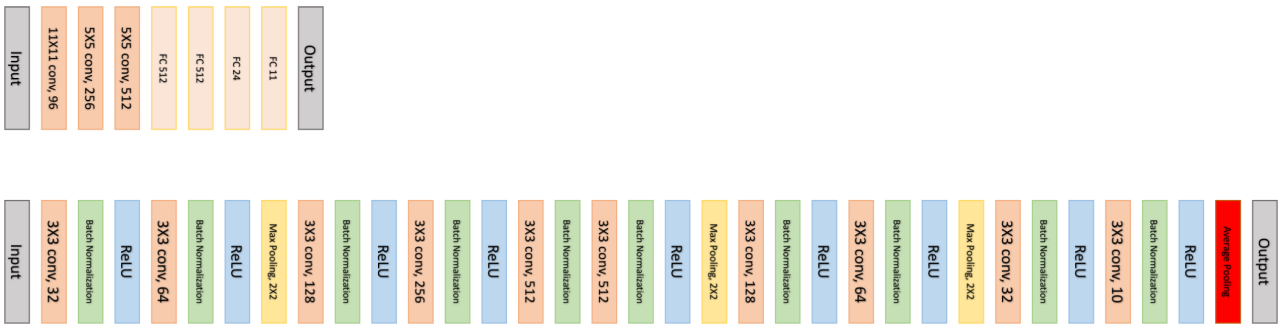

Figure 3. CNN structure for Emotion Detection. The structure on the top is the structure widely used and the one on the bottom is the novel structure I created.

- There is no precursor. ArtTalk is the first trial to categorize art paintings by emotions.
- There is not sufficient training data, we will discuss this problem later.

As discussed in Section 2, I borrow the ideas from facial expression study and semantic study on both paintings and photographs. The simple but promising structure is illustrated in the upper part of Figure 3, where 67% percent correct rate in semantic detection research is reported. I have tested the model but it can only generate a little better than 10% in accuracy. In another word, it is merely better than to randomly select a category.
After that, I have created a new structure as shown in the lower half of Figure 3. The structure is similar to VGG16 with an introduction of batch normalization layers after every convolution layer and average pooling layer for output. Its efficiency will be examined in Section 4.

4. Datasets

In this section, we will discuss the source of raw data and how I transfer the pictures into the training and testing data for the the model presented in Section 3. First I will talk about the data flow to present a large picture of the dataset. Then I will dive into to two detailed operations in the model: data normalization and augmentation.

4.1. Data Flow

The raw images are scraped from WikiArt, where artistic paintings are collected and theme labels are provided. WikiArt provides a dataset of 35750 high-resolution pictures under 60 themes. The pictures are in varied sizes and cannot be used directly. I normalize all the pictures to the same size which will be discusses in Section 4.2 and the normalized data are sent out to train Theme CNN and Emotion CNN.

Table 1. Training Data Summary

	Testing Data	Training Data	Augmented
Thematic	1473	32000	No
Emotional	2000	32000	Yes

This table table summarizes the training datasets for both emotional and thematic tagging.

- Data for Theme CNN:
  I use the theme tags provided by WikiArt directly, but I found that huge imbalance exists in the amount of paintings under each category. The top five popular themes, which are female-portraits, male-portraits, Christianity, forests-and-trees and houses-and-buildings, contain 15070 paintings together, which is nearly 50% of total images I have. This implies there might exist less popular themes that have only tens of pictures, which is not enough for a valid training.
  In order to avoid the problem caused by the imbalance, I have chosen only the themes with more than 150 paintings within the category. After the initial filtering, 41 themes and 33473 paintings are left in my raw data for thematic tagging.

- Data for Emotion CNN:
  Different from thematic tagging, where there exists labeled data, emotional tagging requires human labeling, and the labeling process is very subjective. My project collaborator is responsible for providing emotion-tagged data.
  We have divided the emotions into eleven categories: **neutral**, happy, love, joy, sad, disgust, fear, surprise, lust, anger and envy. This differs from traditional emo-

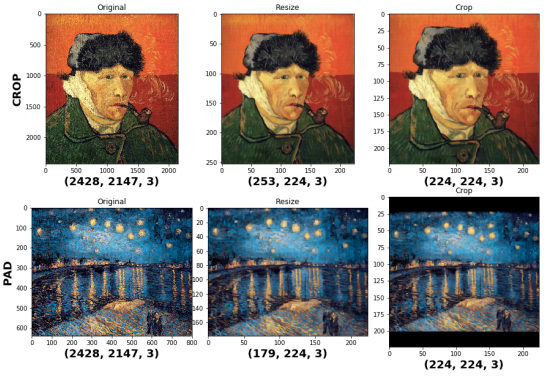

Figure 4. Data Normalization.

Figure 5. Data Augmentation.

tion categories with an additional feeling of neutral. This is because during the labeling process, my collaborator and I have found that some pictures under certain categories do not convey feelings or they only convey a feeling of calmness. For instance, most landscape pictures do not have clear emotional direction.
All together, we have labeled 2000 images for emotional tagging. This might not be insufficient to train complex neural network structures but this is the best human labeler can do within a limited time. I have used data augmentation method, which will be discussed in Section 4.3.

The numbers of training and testing data are summarized in Table 1. For Theme CNN, there are 41 categories, and I use 32,000 images for training and 1473 images for testing. For Emotion CNN, there are 11 categories and I use 32,000 images for training and validation, and 2,000 as testing data. The total 34,000 images used for emotion tagging are generated from data augmentation.

4.2. Data Normalization

Via data normalization, I mainly solve the problem of how to render all the images into the same size. The resizing module can be divided into two steps. The first is to resize along one axis, which is the width axis in practice. Then I either crop or pad along the other axis to make all the pictures share the same size.
In practice, I choose to render all the pictures to the size of 224X224X3. The selection of 224 is to make sure that the input size for my model will be same to that of ImageNet, and this will ease the training process since I have utilized mature models which are originally used for ImageNet.
The procedure is best illustrated in Figure-4. In the first row, we first resize Van Gogh's portrait along width, and renders the picture to 253X224X3 from 2428X2147X3. Then, we can see the height is larger than 224 pixel and hence we crop from the center to render the picture to 224X224X3. The parts that I crop out are the backgrounds and the main body of the painting is intact. In the second row of Figure-4, I demonstrate how I pad Van Gogh's starry night to the desired shape.
I have used OpenCV [2] to facilitate the computation needed for data normalization.

4.3. Data Augmentation

I have practiced four data augmentation method [3] against our training data: crop, blur, sharpen and rotate. There are other possible alternatives, and I will discuss why I do not use them.

- I crop the data into four parts, the upper left, lower left, upper right and lower right corner of the original picture and then I resize the cropped data back to the same size as the original input. Using starry night as an example, I have demonstrated how I crop a picture in Figure-5. One of my concern is that I might cut useful information for some of the cropped pieces and this will hurt training results.

- I blur the picture to different 'blur strength', as shown in the second row in Figure-5. From left to right, I have tried the strength from 0.5 to 2. The concern is that picture becomes darker when it gets more blurry, and this might change people's feeling when they view the picture.

- As an opposite to blur operation, I can also sharpen the picture, whose results are shown in the third row of Figure-5. As can be seen, we run the risk of increasing noise when we sharpen the image too much. But it is difficult to determine how much sharpening is too much, which is depending on the picture characteristics. I have demoed a sharpen strength from 0.5 to 2.

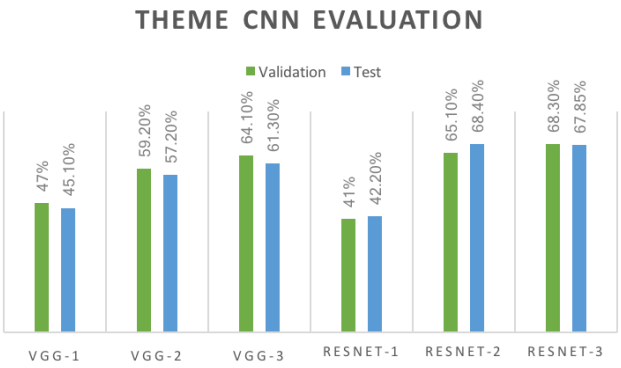

Figure 6. Theme CNN Evaluation. The compared results for different proposed models are illustrated.

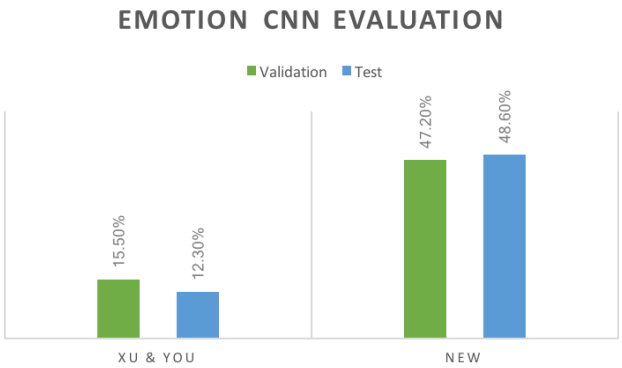

Figure 7. Emotion CNN Evaluation. The compared results for different proposed models are illustrated.

- As a final practice, I have tried to rotate the picture by 45, 90, 180, and 270 degrees. In order not to change the picture size, I cut the area out of 224X224 square and pad the new areas with black blocks.

Other data augmentation method includes changing the illumination, changing the viewing angle and making the picture black-and-white. For fear that this might change the emotional feeling of people who view the altered picture, I do not include them in my data preprocessing module.

5. Experiments

The analysis for the accuracy of Theme CNN and Emotion CNN are carried out separately. To evaluate the the effectiveness of both networks, we use the simple accuracy rate. Firstly, we use training and validation data to train the network. Then we select the best network with highest validation accuracy rate and use the test data to check its effectiveness. Both results of validation and testing accuracy rates are reported.

5.1. Theme CNN

I have tried six different structures for Theme CNN and the results are summarized in Figure 6.

- VGG-1: Replace the last fully connected layer with one convolution layer followed by an average pooling layer for output. It reports an accuracy of 45.1% for testing.
- VGG-2: Add a convolution layer, batch normalization layer and ReLU activation layer between last convolution layer and average pooling in VGG-1. It reports an accuracy of 57.2% for testing.
- VGG-3: Add another module made by one convolution layer, one batch normalization layer and one ReLU activation layer before average pooling layer in VGG-2. It reports an accuracy of 61.3% for testing.
- ResNet-1: Replace the last fully connected layer with one convolution layer followed by an average pooling layer for output. It reports an accuracy of 42.2% for testing, which is the worst among all models.
- ResNet-2: Add a convolution layer, batch normalization layer and ReLU activation layer between last convolution layer and average pooling in ResNet-1. It reports an accuracy of 68.4% for testing. This is the best results and its structure are depicted in Figure 2.
- ResNet-3: Add another module made by one convolution layer, one batch normalization layer and one ReLU activation layer before average pooling layer in ResNet-2. It reports an accuracy of 67.7% for testing.

I have checked the wrongly labeled images and found that there are two obvious issues. First, there exists many overlapping themes. For instance, the following three themes are frequently mislabeled: 'Rome', 'streets-and-squares' and 'houses-and-buildings'.
Besides, many images are wrongly labeled because of the strange shape drawn in the paintings. The pictures within abstract genre are nearly all labeled into a wrong theme.

5.2. Emotion CNN

I have tried two models for Emotion CNN and the results are summarized in Figure 7.

- Xu & You's model, which are described in [16] and [17]. They are using the same structure for semantic study on pictures from twitter, tumblr and flickr. For semantic tagging, nearly 70% accuracy rate is reported. However, it has no luck in the field of emotion tagging. It accuracy rate on testing data is 12.3% which is only slightly better than random selection (about 9% in accuracy rate). Xu's network structure is shown in the upper half of Figure 3.

- My model, which is depicted in the bottom half of Figure 3. The structure is similar to VGG16 with additional layers to improve the training efficiency. It has shown an accuracy rate of 48.6%. It can be further tuned for a higher accuracy.

I have looked at the results for wrongly labeled emotions and I found that most of them belong to the same semantic group. For instance, it is difficult to distinguish happiness and love but there is no mistake in grouping them into positive feelings. Therefore, if we change the problem to semantic detection in artworks, the accuracy will be significantly improved.
Moreover, I also noticed that the label for lust are nearly all wrong. This might be because we do not have enough training images that depict an emotion of lust. We only 17 images labeled as lust in the total 2000 original images. Lacking training data might also contribute to the low correctness rate.

6. Conclusion

In this paper, I have proposed a new model using CNN to allow computers to read and analyze artistic paintings. It is novel in both application area, which is artwork, and in the grouping characteristics, which include emotion.
I have demonstrated that the model can reach an accuracy rate of nearly 70% for theme detection and nearly 50% for emotion detection, which is effective enough to facilitate human selections.
I have also tested and compared different network structures and this will lay foundation for next step of research.

7. Future Work

There are three directions to further improve the proposed models:

1. Better pre-processing on the training data for theme detection. If we can have better training data that have less overlapping, we will see an improvement on the overall accuracy rate for theme detection.

2. More emotion labeled data. The training procedure for Emotion CNN suffers from lacking training data. Though data augmentation eases the problem, we can still expect a better accuracy rate if we have more labeled data. And this would also facilitate and finer tuned results.

3. We can also work on data visualization for the training results. I am always curious about what effect mood most from a viewer's perspective. Is it the object that painted or the atmosphere? This can be disclosed with a saliency mapping.

8. Acknowledgement

I have received help and advices from Dr. Vinay Chaudhri, and the teaching assistants from the course CS231N.

References

[1] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. pages 3085–3092, 2012.

[2] G. Bradski. *Dr. Dobb's Journal of Software Tools*.

[3] P. Forsyth. *Computer Vision - A Modern Approach*. Prentice Hall, USA, 2002.

[4] V. Gajarla and A. Gupta. Emotion detection and sentiment analysis of images.

[5] A. Gudi. Recognizing semantic features in faces using deep learning. *CoRR*, abs/1512.00743, 2015.

[6] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. pages 545–552, 2006.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[9] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. *CoRR*, abs/1311.3715, 2013.

[10] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. pages 503–510, 2015.

[11] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

[12] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. pages 499–512, 2012.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[16] C. Xu, S. Cetintas, K. Lee, and L. Li. Visual sentiment prediction with deep convolutional neural networks. *CoRR*, abs/1411.5731, 2014.

[17] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. *CoRR*, abs/1509.06041, 2015.