CONTENT UNDERSTANDING FOR IMAGING SYSTEMS: PAGE CLASSIFICATION, FADING DETECTION, EMOTION RECOGNITION, AND SALIENCY BASED IMAGE QUALITY ASSESSMENT AND CROPPING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Shaoyuan Xu

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Jan P. Allebach, Co-Chair
School of Electrical and Computer Engineering
Dr. Qian Lin, Co-Chair
HP Labs, Palo Alto, CA
Dr. Amy Reibman
School of Electrical and Computer Engineering
Dr. Mary Comer
School of Electrical and Computer Engineering
Dr. Michael Zoltowski
School of Electrical and Computer Engineering

Approved by:

Dr. Dimitrios Peroulis
Head of Electrical and Computer Engineering

To my beloved parents.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my major advisor, Prof. Jan P. Allebach. I first met him in Spring 2016 right before I joined the lab. But even before that, I heard that he is a very friendly, responsible and knowledgeable person from other people. After working with him for five years in the lab, I strongly agree that his reputation is absolutely veritable. He has also been really helpful to all of the lab members. Whenever we have problems, he is the first person to offer help. And for each project, he gives students his own insight as well as suggestions. Prof. Allebach is also a easy-going person. One will never feel nervous or uncomfortable being together or having a meeting with him. I feel very much privileged to be able to learn from such an honorable and respectful man like him.

I would like to thank my family. I have not spent much time with my parents ever since I came to the US to pursue my Bachelor and Ph.D. degrees but they generously supported me both physically and mentally. Especially, I would like to gratefully and sincerely thank my mom. She has been taking care of my education since I was born and she has been really dedicated. I wish that both my parents can share the joy of my academic achievements.

In addition, I would like thank Dr. Qian Lin. She is my mentor at HP Labs where I worked at from Summer 2017 to Summer 2019. She always provided me with helpful instructions and useful feedbacks during our meetings.

I would also like to thank all my friends and lab mates. They helped me with the difficulties I encountered and inspired me with new ideas. I really appreciate that.

Last but not least, I would like to thank all my committee members for reviewing my work and attending my dissertation.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

SYMBOLS

f_i	Chroma Histogram Flatness for the block i
F	Chroma Histogram Flatness for the entire image which is the maximum of f_i
$c(m, n)$	Chroma strength of a target pixel at position of (m, n)
c_i	Chroma Around Text of the block i
C	Chroma Around Text of an image which is the maximum of c_i
R_c	Color Block Ratio
R_w	While Block Ratio
W_m	Weighted Misclassification Rate
I_d	Feature Impact Factor
cc_r, cc_t	Connected components of raster and testing image
ΔE	*Lab* color difference
D_{LM}	Land mark points distance
D_{eye_outer}	Eye corner land mark points distance
T_L, T_U	Lower and upper bound of facial action unit detection
$S(m, n)$	Saliency Pixel Value
T	Importance coefficient for VSIP
λ	Area ratio of the candidate crop
α_l, α_u	Lower and upper bound of the aspect ratio

ABBREVIATIONS

SVM	Support Vector Machine
DPI	Dot Per Inch
LCH	Lightness Chroma Hue
RGB	Red Green Blue
DAG-SVM	Directed Acyclic Graph-Support Vector Machine
AIO	All In One
PQ	Print Quality
EP	Electrophotographic
JND	Just Noticeable Difference
MLESAC	Maximum Likelihood Estimation Sample Consensus
RANSAC	Random Sample Consensus
CC	Connected Component
SSD	Sum Square Difference
CNN	Convolutional Neural Networks
LBP	Local Binary Patterns
CK+	Extended Cohn-Kanade Dataset
JAFFE	Japanese Female Facial Expression Dataset
MUG	Multimedia Understanding Group
KDEF	Karolinska Directed Emotional Faces Dataset
MTCNN	Multi-task Cascaded Convolutional Networks
LM	Land Marks
LMDB	Lightning Memory-Mapped Database
VGG	Visual Geometry Group
FER	Facial Expression Recognition

FAU	Facial Action Units
FACS	Facial Action Coding System
FRIQA	Full-reference Image Quality Assessment
PSNR	Peak Signal-to-noise Ratio
SSIM	Structural Similarity Index
NRIQA	No-reference Image Quality Assessment
QAC	Quality-aware Clustering
RoD	Region of Discard
RoI	Region of Interest
IoU	Intersection over Union
Disp.	Boundary Displacement
DOF	Depth of Field
DSLR	Digital Single-lens Reflex
PDNet	Prior-model Guided Depth-enhanced Network
HED	Holistically-Nested Edge Detector
FCN	Fully Convolutional Neural Network
IP	Image Patch
SIP	Salient Image Patch
VSIP	Valid Salient Image Patch
JP2K	JPEG2000 Distortion
WN	White Noise
GBLUR	Gaussian Blur
DMOS	Difference Mean Opinion Score
MOS	Mean Opinion Score
BN	Batch Normalization
SGD	Stochastic Gradient Descent
SROCC	Spearman Rank Order Correlation Coefficient
LCC	Linear Correlation Coeffieicnt
MSE	Mean Square Error

FCDB	Flickr Cropping Database
Bbox	Bounding Box
GT	Ground Truth

ABSTRACT

Xu, Shaoyuan Ph.D., Purdue University, August 2020. Content Understanding for Imaging Systems: Page Classification, Fading Detection, Emotion Recognition, and Saliency Based Image Quality Assessment and Cropping. Major Professor: Jan P. Allebach.

This thesis consists of four sections which are related with four research projects.

The first section is about Page Classification. In this section, we extend our previous approach which could classify 3 classes of pages: Text, Picture and Mixed, to 5 classes which are: Text, Picture, Mixed, Receipt and Highlight. We first design new features to define those two new classes and then use DAG-SVM to classify those 5 classes of images. Based on the results, our algorithm performs well and is able to classify 5 types of pages.

The second section is about Fading Detection. In this section, we develop an algorithm that can automatically detect fading for both text and non-text region. For text region, we first do global alignment and then perform local alignment. After that, we create a 3D color node system, assign each connected component to a color node and get the color difference between raster page connected component and scanned page connected. For non-text region, after global alignment, we divide the page into "super pixels" and get the color difference between raster super pixels and testing super pixels. Compared with the traditional method that uses a diagnostic page, our method is more efficient and effective.

The third section is about CNN Based Emotion Recognition. In this section, we build our own emotion recognition classification and regression system from scratch. It includes data set collection, data preprocessing , model training and testing. We extend the model to real-time video application and it performs accurately and smoothly. We also try another approach of solving the emotion recognition prob-

lem using Facial Action Unit detection. By extracting Facial Land Mark features and adopting SVM training framework, the Facial Action Unit approach achieves comparable accuracy to the CNN based approach.

The forth section is about Saliency Based Image Quality Assessment and Cropping. In this section, we propose a method of doing image quality assessment and recomposition with the help of image saliency information. Saliency is the remarkable region of an image that attracts people's attention easily and naturally. By showing everyday examples as well as our experimental results, we demonstrate the fact that, utilizing the saliency information will be beneficial for both tasks.

1. INTRODUCTION

This thesis aims to address four problems: Page Classification, Fading Detection, Emotion Recognition and Image Quality Assessment and Cropping.

1.1 Page Classification for Print Imaging Pipeline

Digital copiers and printers are widely used nowadays. One of the most important things people care about is copying or printing quality. Different defects may have different degree of influence on different types of pages. So we need to classify pages after we print or scan them. In order to improve it, we previously came up with an SVM-based classification method to classify images with only text, only pictures or a mixture of both based on the fact that modern copiers and printers are equipped with processing pipelines designed specifically for different kinds of images [1]. However, in some other applications, we need to distinguish more than three classes. In the next chapter, we develop a more advanced SVM-based classification method using four more new features to classify 5 types of images which are text, picture, mixed, receipt and highlight.

1.2 Color Fading Detection in Customer's Printed Content

People use printers in daily life and one of the most important quality criterion is the fading level of a printed page. If the pages being printed page show fading, people need to change the cartridges as soon as possible. Traditionally, fading detection is done by manually referring to the first several printed pages known as the Raster Page. But this approach requires tremendous labor cost as well as time. In order to be able to do automatic fading detection, in Chapter 3, we come up with an approach to detect fading based on the color difference between the printed pages and the raster page.

1.3 Emotion Recognition Using Convolutional Neural Networks

Emotion plays an important role in daily life, as it helps people better communicate with and understand each other more efficiently. Two mainstream approaches of detecting people's emotions are through facial expressions and voice while automatic facial expression detection algorithm is more straight forward and highly desired. In Chapter 4, we develop an emotion recognition system that can do both still images and real-time (video) emotion recognition using convolutional neural networks. Basically, peoples facial expressions can be classified into 7 categories: Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. For a certain image or a real-time video, our system can show the classification results for all of the 7 emotions. We test the system on 2 data sets and the accuracies are above 80% and the real-time testing performs well so we successfully show the feasibility of implementing convolutional neural networks in real time to detect emotions. In addition, we validate that, Facial Action Unit detection is also a promising option of accomplishing emotion recognition.

1.4 Saliency Based Image Quality Assessment and Cropping

Analyzing images' quality and aesthetics has become an increasingly demanding research topic since the usage of mobile phones and social media has been rapidly increasing. Traditional methods of evaluating an image as a whole is biased. In Chapter 5, we propose an image quality assessment and cropping pipeline based on saliency information. For image quality assessment, we first do saliency detection on the testing image. Then we split the image into patches. The final image quality score is assessed as the average of all salient patches quality scores. For image cropping problem, saliency detection is also applied in the first place. Candidate crops are then generated based on the saliency bounding boxes. Our final resulting crops are guaranteed to contain all saliency information. Trained with deeper network structures and tested with more advanced algorithms, both of our proposed pipelines achieve state-of-the-art results.

2. PAGE CLASSIFICATION FOR PRINT IMAGING PIPELINE

2.1 Introduction

2.1.1 General Introduction of the Project

In everyday life, people have all kinds of images which they want to scan or print. That is why digital copiers or all-in-one printers are needed. And one of the most important things people care about is the printing quality of digital copiers or all-in-one printers.

In order to optimally process different kinds of input images, modern copiers and printers possess multiple processing pipelines. Each one of these pipelines are designed specifically for one type of image based on their unique features. For text images, we require the text area to have clear, sharp edges and also high contrast. But for picture images, we do not want them to have high contrast. Instead, we want them to be more blurred. If an image is copied or printed by the right pipeline, people will get the image of the best quality. However, if an image is copied or printed by the wrong pipeline, the image of significantly bad quality will be produced. For example, if a text image is processed by a picture pipeline, the text area in the output image will have blurred edges and comparably low contrast which is not satisfactory [2]. Figure 2.1 illustrates an example that a text image is processed by both picture and text pipelines. It is obvious that the text image is more clear and has better edge sharpness when it is processed by a text pipeline which is more desirable. So in order to increase the copying or printing quality, we want to come up with a method to classify the images going into copiers or printers and process them in different ways. According to the previous research work done by Cheng Lu from our group, we have already had a SVM-based classification method to classify three types of images: text, picture and mixed and we have three features to train the classifiers [1]. In this paper, we develop four new features to classify two new image classes which are highlight and receipt.

In this chapter[1], we develop four new features to classify two new image classes which are highlight and receipt.

For the rest of Section 2.1, we introduce five types of images and the related work. Section 2.2 describes four new features. Section 2.3 describes the classification structure as well as the feature selection. Experiment results are included in Section 2.4 and finally Section 2.5 is the conclusion of the paper.

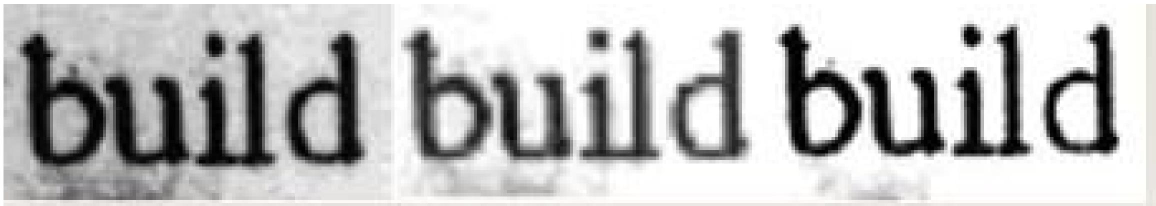


Fig. 2.1.: Image quality of original image/picture mode image/text mode image.

2.1.2 Five Types of Images

We classify all images into five classes which are shown in Figure 2.2: Text, Picture, Mixed, Receipt and Highlight to reach optimal printing quality. And misclassification may cause image quality degradation. Text images contain only text. Picture images contain only pictures. Mixed images contain both text and pictures. Receipt images contain scanned receipts and Highlight images contain highlighted text or pictures. All the input images are scanned at 300 dots per inch (dpi) and the output result is one of the five classes.

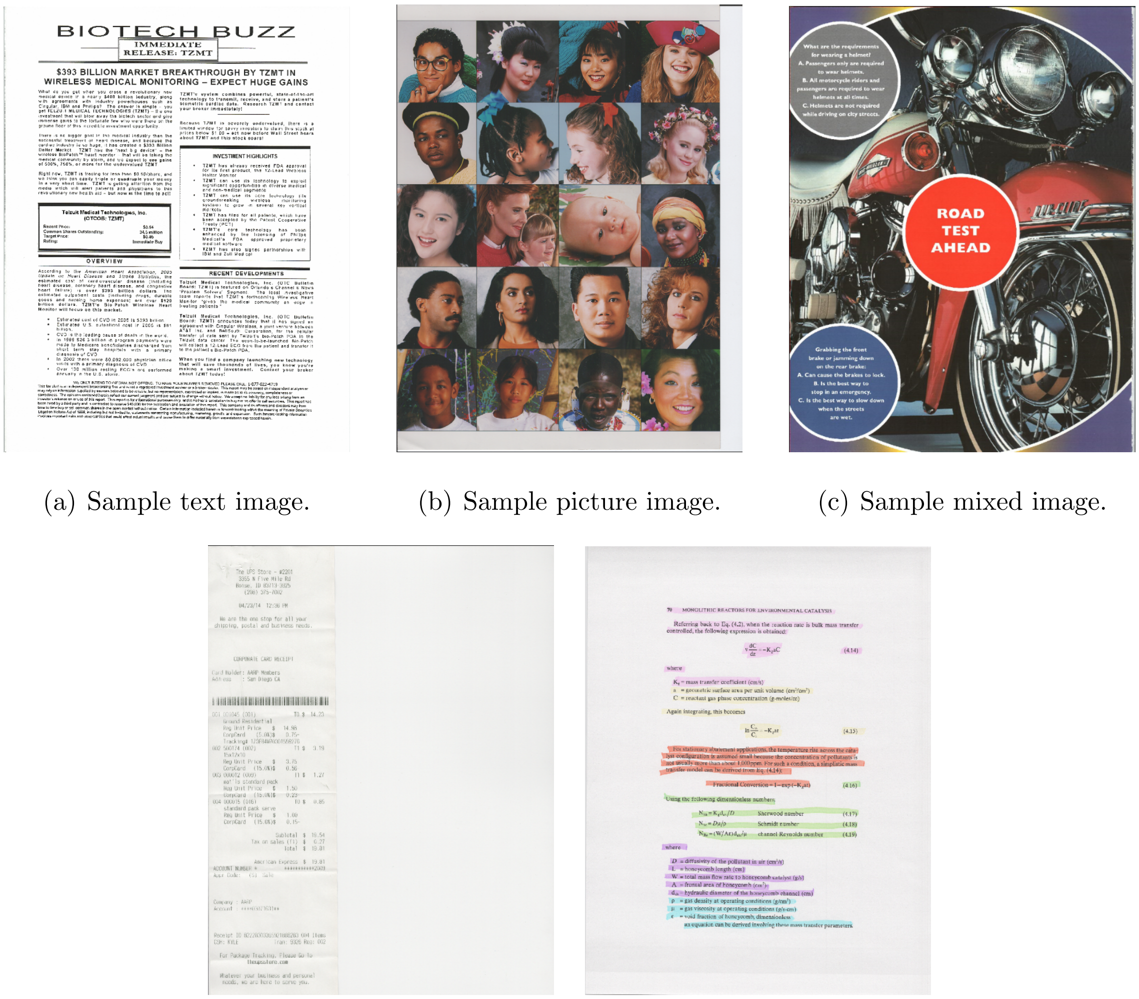
Note that the receipt class and the highlight class are newly added. Receipt class represents a type of document images which have very low contrast and faded text

[1] Research supported by HP Inc., Boise, ID 83714.

which is shown in Figure 2.2(d). Its corresponding processing pipeline needs very strong contrast enhancement and text recovery to guarantee the readability of the copy [3] [4]. Compared with example given in Figure 2.1, receipt type needs even stronger enhancement for better quality. Highlight class typically contains colors that are more saturated compared with color in the natural images. An example is people using fluorescent pen to highlight text line in a document image. Highlighters usually produce very saturated color in order to attract the attention of readers. However, these saturated colors [5] [6] are sometimes difficult to be captured by the scanning devices. Or they are sensed as less saturated colors by the scanner. Figure 2.2(e) shows an captured image with highlight colors, which illustrates the vision of our device when sensing those colors. We can tell the colors in the upper half of the image are very weak from the view of scanners which is not desirable, but colors in the lower half are much more visible. So if we can tell the input image is a highlighted text document, we can produce more saturated color when generating output to match the appearance of actual highlighter ink. This can improve the user experience when copying highlighted document pages for better readability while at the same time preserving important area in the page.

2.1.3 Related Work

There has been a significant amount of research work done by Cheng Lu from our group which is related with an SVM-based classification [7] of three types of images: text, picture and mixed. And he has already developed four features which are Histogram Flatness Score, Color Variability Score, Text Edge Count and Text Color Variance to train the classifiers. For the first feature which is the Histogram Flatness Score, typically, the histogram for a text image tends to have sharper and narrower peaks than the histogram for a non-text image such as a picture image. For the second feature which is the Color Variability Score, it is reasonable to assume that the non-text region of a text document contains only a few gray level values. So we


BIOTECH BUZZ
IMMEDIATE RELEASE: TZMT
$393 BILLION MARKET BREAKTHROUGH BY TZMT IN WIRELESS MEDICAL MONITORING – EXPECT HUGE GAINS
ROAD TEST AHEAD

(a) Sample text image.

(b) Sample picture image.

(c) Sample mixed image.

(d) Sample receipt image.

(e) Sample highlight image.

Fig. 2.2.: Five types of images.

build a block-mean histogram for the image to calculate a Color Variability Score. For the third feature which is the Text Edge Count, we calculate the number of text edges in an image based on the fact that the number of text edges in a text image is more than it is in a picture image. And for the last feature which is the Text Color Variance, we assume that text pixels in one image should have close luminance intensity values. Large variance in a certain area suggests that its a non-text region [1].

The new features are developed based on these four old features.

2.2 Features

2.2.1 Chroma Histogram Flatness

Although both natural images and highlighted text have chroma information, a very significant difference between natural color image and highlight text is that highlighted text tends to include only one or few color while natural images have richer color information. This can be illustrated in Figure 2.3 where we can find only pink in highlighted text image while chroma in natural image is very rich. Then we need a numerical value to represents this difference.

In Section 2.1.3, we introduce the Histogram Flatness Score in the luminance channel and it assumes that a text image has more peaky histogram. Similarly, if we build a histogram for a highlighted text in chroma space, we can expect that there is a single or few peaks, while the histogram for a natural image is more flat. Different from luminance histogram, chroma histogram is two-dimensional. Since our input image can be either RGB or LCH color space, we need to build histograms correspondingly. For RGB input, we transform it into YUV space and build a histogram on UV plane. For LCH input, we can build a histogram on CH space directly. As shown in Figure 2.4(a), UV space is in Cartesian coordinate system so we can uniformly partition it into 8×8 areas and build a histogram based on that. However, CH space is described in polar coordinate system so we need to partition it differently. We uniformly divide the hue and chroma into 8 segments respectively, and thus giving us 64 areas which are not uniform in terms of space. LCH partition for building histogram is shown in Figure 2.4(b). For every input image, we cut it into 32×32-pixel blocks and build a histogram for pixels in the block i which is denoted as h_i. The Chroma Histogram Flatness for the block i is denoted as f_i which is shown below:

$$f_i = \frac{\sqrt[N]{\prod_{n=0}^{N-1} h_i(n)}}{\frac{\sum_{n=0}^{N-1} h_i(n)}{N}} \tag{2.1}$$

which is the geometric mean over the arithmetic where $N = 60$ for YUV and $N = 56$ for LCH in our case. It is worth noticing that we do not use all the bins of histogram to calculate f_i. That is because we should only focus on the chroma flatness and exclude those areas which are close to gray. Highlight-text will also have very low flatness if we consider those gray pixels which correspond to black text and white background. For YUV space, we ignore the central 4 bins and for LCH space we ignore the central 8 bins that are inside the smallest circle.

The Chroma Histogram Flatness F for the entire image is define as maximum f_i of all blocks in the image:

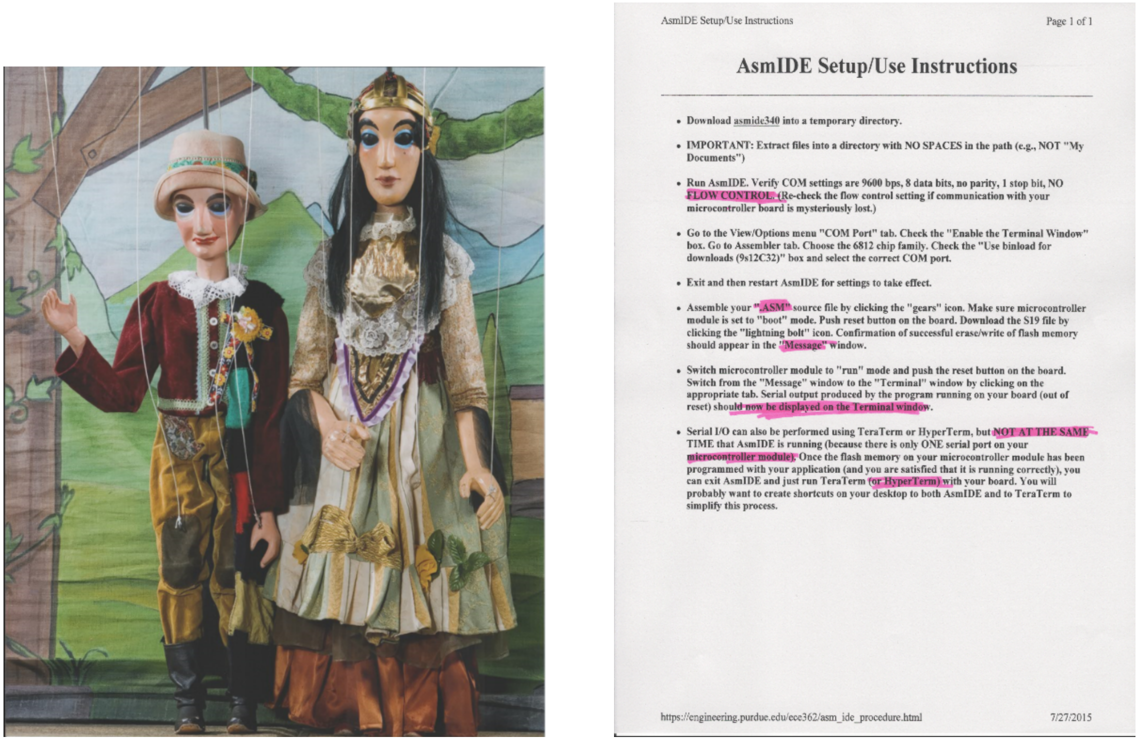
$$F = max(f_i) \tag{2.2}$$

Because we do not expect seeing flat histogram of any block in the highlighted text image.

2.2.2 Chroma Around Text

On top of Chroma Histogram Flatness, we design another feature to detect chroma information of highlighted text image. Typically, we can expect that people use highlighter to emphasize text information, which means text strokes are usually covered and surrounded by highlight colors as shown in Figure 2.5. However for natural images, chroma information does not necessarily exists around edges. Given the fact, we can try to detect if there is chroma existence along the text edges in the image.

To find chroma around text strokes, we first need to find text edges in the image blocks. We follow the method introduced in Sec. 2.1.3 to find text edges. However, we need to note that reverse contrast text should not be considered any more. That is because it is rare that people mark light text with darker highlight which will cause poor readability. After finding a text edge, we search along its luminance increase direction to find if there is any chroma existence. Then we calculate chroma strength



AsmIDE Setup/Use Instructions Page 1 of 1

AsmIDE Setup/Use Instructions

- Download asmide340 into a temporary directory.
- IMPORTANT: Extract files into a directory with NO SPACES in the path (e.g., NOT "My Documents")
- Run AsmIDE. Verify COM settings are 9600 bps, 8 data bits, no parity, 1 stop bit, NO FLOW CONTROL. (Re-check the flow control setting if communication with your microcontroller board is mysteriously lost.)
- Go to the View/Options menu "COM Port" tab. Check the "Enable the Terminal Window" box. Go to Assembler tab. Choose the 6812 chip family. Check the "Use binload for downloads (9s12C32)" box and select the correct COM port.
- Exit and then restart AsmIDE for settings to take effect.
- Assemble your ".ASM" source file by clicking the "gears" icon. Make sure microcontroller module is set to "boot" mode. Push reset button on the board. Download the S19 file by clicking the "lightning bolt" icon. Confirmation of successful erase/write of flash memory should appear in the "Message" window.
- Switch microcontroller module to "run" mode and push the reset button on the board. Switch from the "Message" window to the "Terminal" window by clicking on the appropriate tab. Serial output produced by the program running on your board (out of reset) should now be displayed on the Terminal window.
- Serial I/O can also be performed using TeraTerm or HyperTerm, but NOT AT THE SAME TIME that AsmIDE is running (because there is only ONE serial port on your microcontroller module). Once the flash memory on your microcontroller module has been programmed with your application (and you are satisfied that it is running correctly), you can exit AsmIDE and just run TeraTerm (or HyperTerm) with your board. You will probably want to create shortcuts on your desktop to both AsmIDE and to TeraTerm to simplify this process.

https://engineering.purdue.edu/ece362/asm_ide_procedure.html 7/27/2015

(a) Natural image.

(b) Highlighted text.

Fig. 2.3.: Natural image v.s. Highlighted text.

(c) for two pixels along the luminance increase direction outside of text edge. In YUV space, chroma strength of a target pixel at position of (m, n) is defined as

$$c(m,n) = u(m,n) + v(m,n) \tag{2.3}$$

Note that we use simple summation of u and v here as approximation for faster computation. In case of LCH space, chroma strength is equivalent to $c(m,n)$. The process can be illustrated in Figure 2.6 after finding a text edge.

Similarly, we cut the input image into blocks with 32×32 pixels and then find $c(m,n)$ for all pixels outside text edges. Chroma Around Text (c_i) of a block i is defined as:

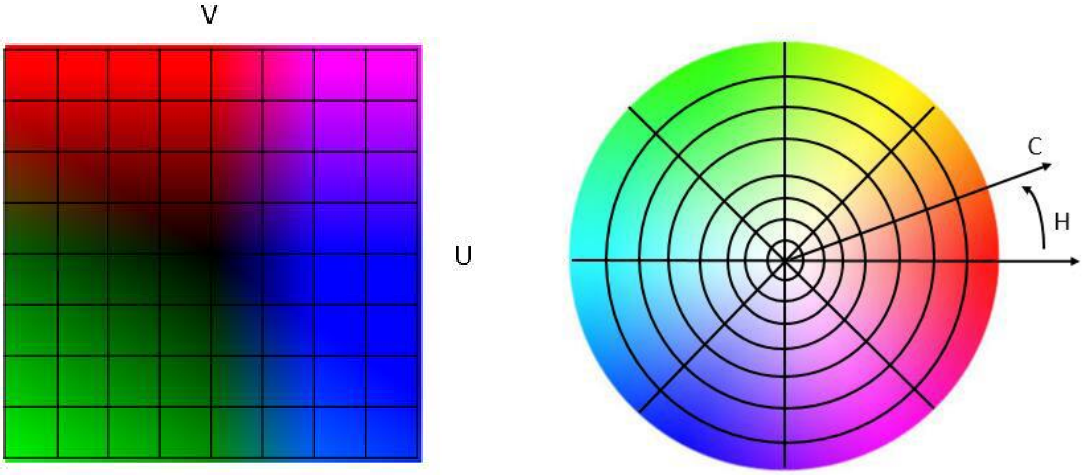
$$c_i = \frac{mean(c(m,n))}{std(c(m,n))} \tag{2.4}$$

We also consider standard deviation of $c(m,n)$ because highlight should be consistent in a single block. Small $std(c(m,n))$ indicates that this block is more likely to contain highlight color. Similarly, chroma around text (C) of an image is defined as the maximum value of c_i of all blocks:

$$C = max(c_i) \tag{2.5}$$

2.2.3 Color Block Ratio

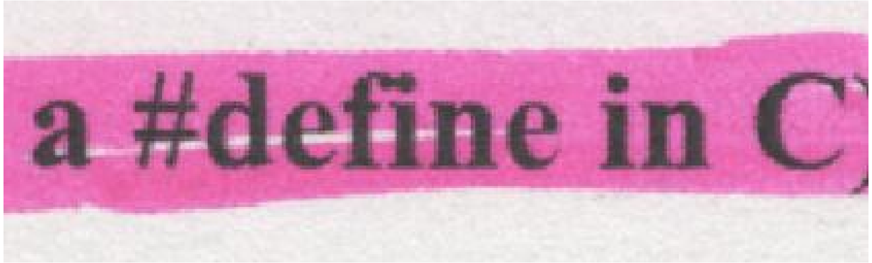
Chroma Flatness and Chroma Around Text together can provide good discriminative power for detecting highlight image. However, they are not capable of handling text image with color background. One such example is given in Fig. 2.7. According to our feature design in Sec. 2.2.1 and Sec. 2.2.2, both features would strongly



(a) YUV space partition.

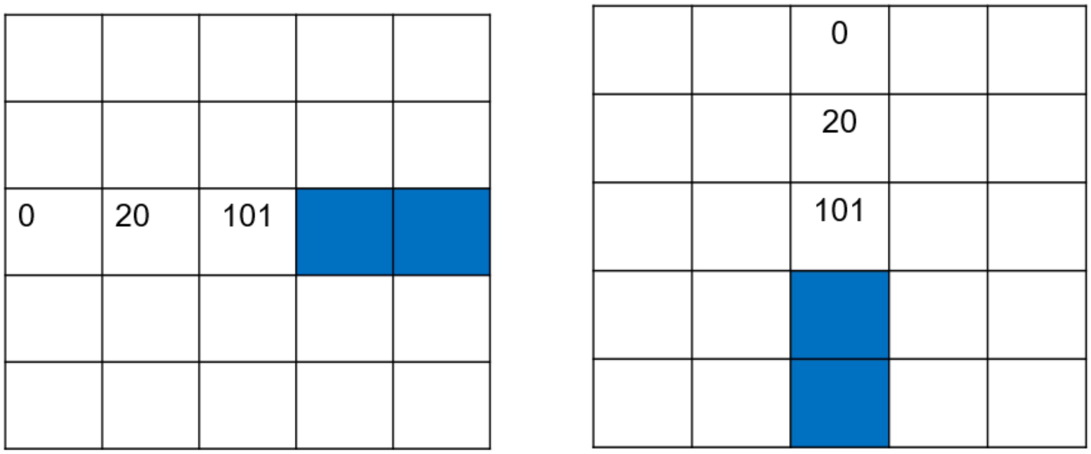
(b) LCH space partition.

Fig. 2.4.: LUV and LCH color space histogram.



a #define in C

Fig. 2.5.: Highlighted colors around text.



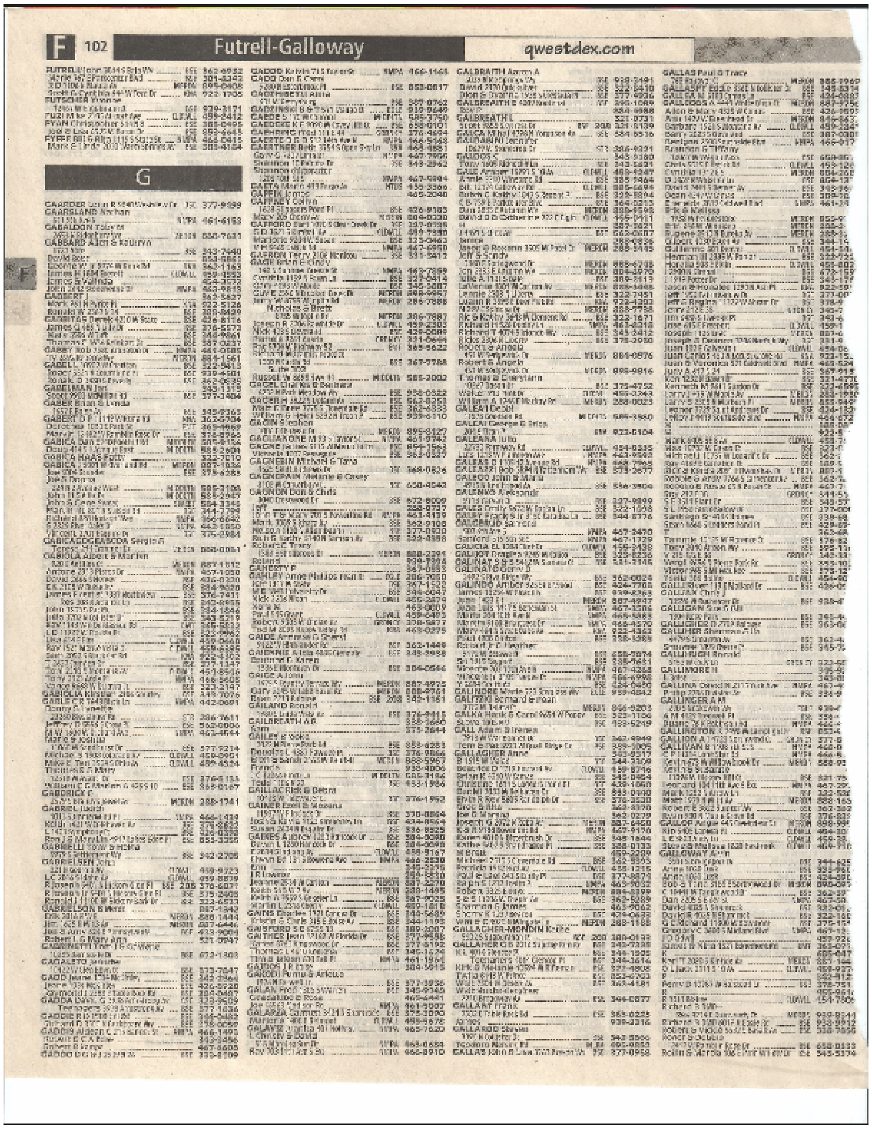
(a) Horizontal direction.

(b) Vertical direction.

Fig. 2.6.: Calculate c(m, n) of two pixels which are cover by blue.

indicate that Figure 2.7 should be classified as highlighted text image. It is because both features only focus on local chroma and ignore global information.

To address this problem, we introduce another feature called Color Block Ratio. We calculate the number of color pixels in every 32×32-pixel block. A pixel is defined as color if its chroma strength is greater than a threshold value $T_c = 10$. The chroma strength is defined in Equation 2.3 for YUV space. In LCH space is simply C channel. A block i is considered a color block if 10% of its pixels are colored. We set $m_i = 1$ if the block is colored otherwise $m_i = 0$. Color block ratio (R_c) of an image is defined as:


F 102 Futrell-Galloway qwestdex.com

G

Fig. 2.7.: Text with yellow background.

$$R_c = \frac{\sum m_i}{\sum i} \tag{2.6}$$

2.2.4 White Block Ratio

Compared with text image, receipt typically only occupies a small part of area on flat-bed scanner. An example is given in Figure 2.2 which shows that a scanned receipt only occupies upper-left corner of the plane. Utilizing this difference, we design the White Block Ratio feature. Again, we cut the input image into 32×32-pixel blocks. For each block i, we check if 95% of the pixels have luminance value which is larger than 230. If this is true, we consider this block to be a white block with $w_i = 1$, otherwise we set $w_i = 0$. White block ratio (R_w) of an image is defined as:

$$R_w = \frac{\sum w_i}{\sum i} \tag{2.7}$$

2.3 Classification

2.3.1 Classification Structure

In order to make a more balanced multi-class classification and for easier tuning, we apply Directed Acyclic Graph-Support Vector Machine (DAG-SVM) [8] to solve the problem. DAG-SVM is a tree-structured classification method that capable of making multi-class classification. Instead of making one vs rest decision, it tentatively make one vs one decision which allows more judgement from lower nodes. DAG-SVM classifiers that we use are shown in Figure 2.8:

In our case, class 1 stands for mix, class 2 stands for text, class 3 stands for picture, class 4 stands for receipt and finally class 5 stands for highlight. It first decides if the input image is non-mix or non-highlight. If the image is classified as non-mix, then it is sent to the right node, otherwise it is sent to the left node. And the whole process goes from the top all the way to the bottom of the tree-structure.

2.3.2 Feature Selection

We introduce four new features to classify receipt and highlight. Together with the features we previously developed, we can represent each image k with a feature vector $f_k \in R^8$. However, we need to evaluate the contribution of each feature to the

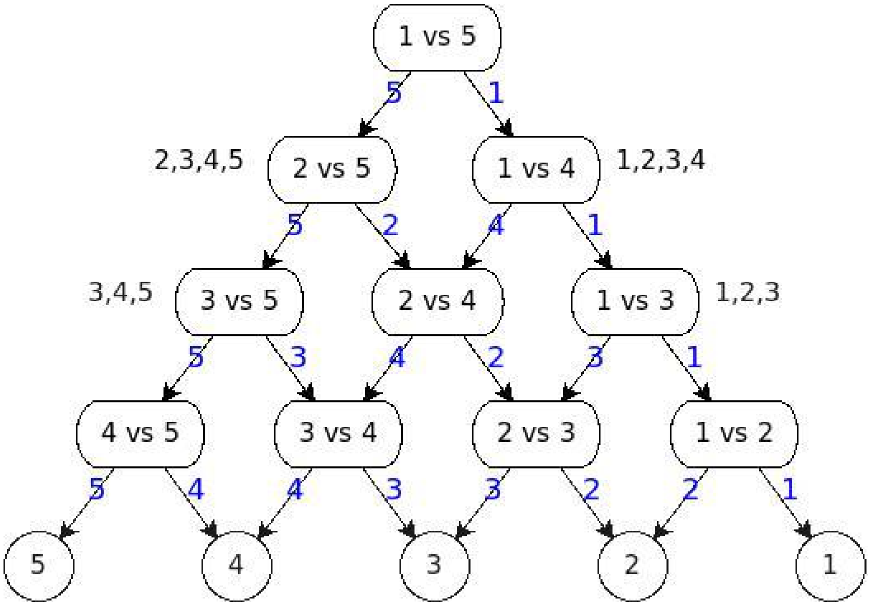


Fig. 2.8.: Tree structure of DAG-SVM.

classification. To test the impact of each feature on the overall classification accuracy and time consumption, we adopt the leave-one-out feature selection.

Because misclassifications are not equally weighted, we need to first consider the metric for feature selection. We define the weighted misclassification rate (W_m) as below:

$$W_m = \frac{\sum_{i,j} w(i,j)n(i,j)}{\sum_j n(i,j)} \tag{2.8}$$

where $w(i,j)$ is the weight of misclassification given in Table 2.2 and $n(i,j)$ is the number of images in this corresponding entry. To avoid biased measure due to different size of image set, the weighted sum needs to be normalized by the size of image set of each type. In our case, we have $\sum_j n(i,j) = 100$ for all the five types. But for completeness, we still keep the normalization term in Equation 2.8.

We first evaluate M_m^8 when we use all 8 features. Then we drop each feature d at a time and evaluate $M_m^{\bar{d}}$. Then the feature impact factor for d is defined as:

$$I_d = \frac{M_m^{\bar{d}} - M_m^8}{M_m^8} \tag{2.9}$$

According to the Equation 2.9, the feature with larger I_d is more important to our classification. Also, we measure the time consumption for each feature and take average over all the images. The results are shown in Table 2.1:

Table 2.1.: Feature impact factor and time consumption.

Feature	Histogram flatness	Color variability	Text edge count	Text color variance	Chroma around text	Chroma histogram flatness	White block ratio	Color block ratio
time(ms)	12.01	12.51	14.97	73.92	36.12	11.45	0.67	0.81
I_d	24.61%	13.84%	11.02%	1.9%	10.2%	3.4%	48.43%	13.65%

According to Table 2.1, we find Text Color Variability has the smallest impact on overall classification accuracy and it consumes most of the time. Subsequently, we exclude this feature from our application and the final feature vector of an image k is a 7-dimensional vector $f_k \in R^7$. This f_k serves as the input to the DAG-SVM classifiers discussed in Section 2.3.1.

2.4 Experimental Results

To test the performance of our designed system, we build an image set which includes 500 images scanned by flat-bed scanner. Each type of images has 100 images.

Different misclassifications are weighted differently due to its impact to image quality. Some misclassifications lead to larger image quality degradation and they should be assigned larger weight. Some other misclassifications cause smaller impact on image quality and should be assigned lower weight. The weight table of misclassification is shown in Table 2.2:

Table 2.2.: Weights of misclassifications $w(i, j)$.

Ground truth	Mix	Text	Picture	Receipt	Highlight
Mix	0	3	5	6	4
Text	3	0	10	6	2
Picture	3	10	0	10	15
Receipt	6	8	3	0	8
Highlight	10	10	10	10	0

To test the performance of our proposed algorithm, we conduct leave-one-out cross validation. We use rbf kernel with σ for every node in Figure 2.8. Then we do exhaustive search for σ and box constraint C to find the best combination that produces the best cross validation result. The goodness of cross-validation result is measured by weight misclassification rate (W_m) which is defined in Equation 8. When $W_m = W_m^*$ reaches minimum, the corresponding confusion matrix $n(i, j)$ in different color spaces are given in Table 2.3 and Table 2.4:

2.5 Conclusion

In this chapter, we introduce four new features to handle multi-class classification for AIO printer with scanning functionality. It extends the scope of the topic about the previous research work on SVM-based image classification of three types of images: text, picture and mix. Our proposed algorithm utilizes the chroma information of input image for better classification. Eventually, we use DAG-SVM with

Table 2.3.: Confusion matrix $n(i, j)$ in YUV space.

Ground truth	Mix	Text	Picture	Receipt	Highlight
Mix	78	5	11	2	4
Text	3	68	0	4	25
Picture	5	0	94	1	0
Receipt	0	3	3	88	6
Highlight	4	8	1	5	82

Table 2.4.: Confusion matrix $n(i, j)$ in LCH space.

Ground truth	Mix	Text	Picture	Receipt	Highlight
Mix	76	6	12	1	5
Text	4	72	0	9	15
Picture	7	0	92	0	1
Receipt	0	6	0	89	5
Highlight	3	14	1	6	76

seven features to classify five classes of images which are text, picture, mixed, receipt and highlight. And based on the classification result, digital copiers or printers will produce images with better quality by choosing corresponding processing pipelines.

3. COLOR FADING DETECTION IN CUSTOMER'S PRINTED CONTENT

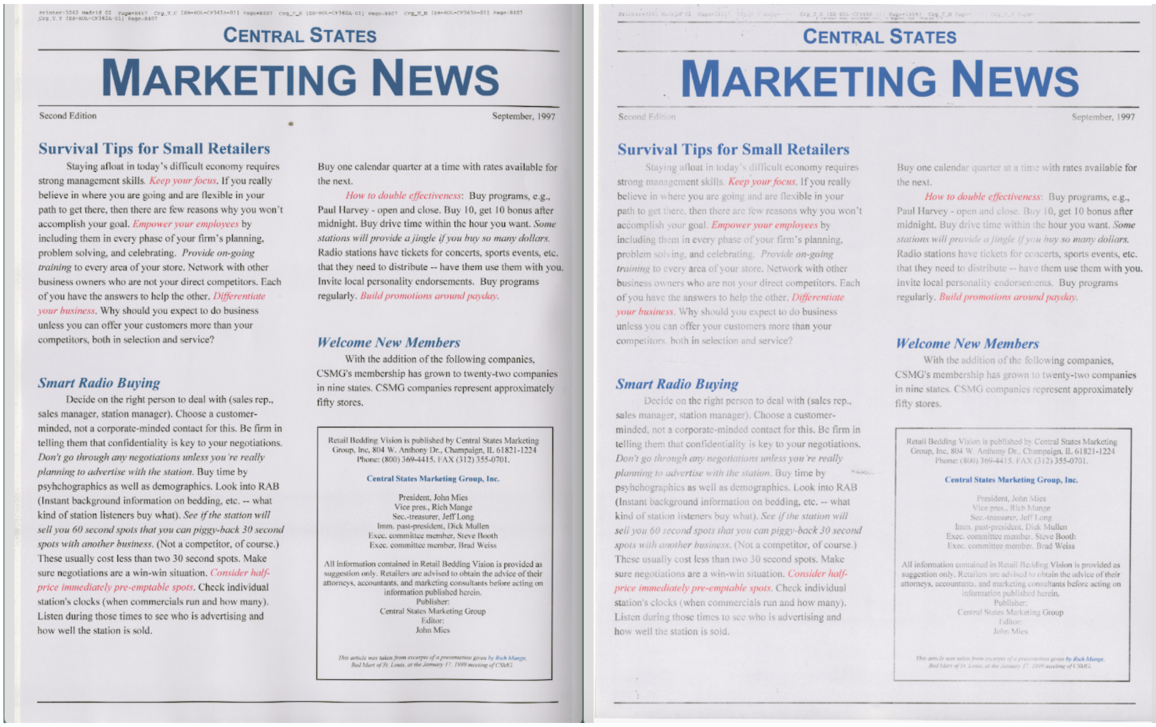
3.1 Introduction

3.1.1 General Introduction of the Project

Print quality (PQ) is one of the most important criterion in printing industry. It profoundly influence the customers' user experience. Page quality is degraded when defects appear on the page which may cause the unsatisfactory of the customers. And the electrophotographic (EP) process which is widely used in modern laser printer is susceptible to various print defects. Among them, color fading is one of the most important defects because it can be detected easily and frequently i.e. whenever one or more of the cartridges' toner is low or depleted. An example can be seen in Figure 3.1, the left page is the raster page which is one of the first printed pages when all of the cartridges are full and the right page is the customer printed page with fading caused by low toner in black cartridge.

3.1.2 Related Work

The traditional method to detect color fading is using a diagnostic page. In order to find out whether the cartridges are low of toner or not, the user need to print the test page and compare it with against the user manual [9]. Although this method may effectively diagnose a page, it is too time consuming for customers. Rather than using this approach, we want to come up with an algorithm that can automatically detect fading while in the mean time, keeps the same effectiveness. Previously, our lab proposed an algorithm [10] to detect text fading. But the previous work was unable to detect fading in non-text region. So the first part of this chapter is the continuation of the previous work which is the fading detection of the text region. Then in the second part, we extend the previous algorithm to be able to detect color fading in non-text region (i.e. image and graphics region) in printed customer content [1].



Printer=3262 Madrid 02 Page=8467 Crg_T_C [SB-BOL-CP361A-01] Page=8407 Crg_T_K [SB-BOL-CP360A-01] Page=8407 Crg_T_M [SB-BOL-CP363A-01] Page=8407
Crg_T_Y [SB-BOL-CP362A-01] Page=8407

CENTRAL STATES

MARKETING NEWS

Second Edition September, 1997

Survival Tips for Small Retailers

Staying afloat in today's difficult economy requires strong management skills. *Keep your focus*. If you really believe in where you are going and are flexible in your path to get there, then there are few reasons why you won't accomplish your goal. *Empower your employees* by including them in every phase of your firm's planning, problem solving, and celebrating. *Provide on-going training* to every area of your store. Network with other business owners who are not your direct competitors. Each of you have the answers to help the other. *Differentiate your business*. Why should you expect to do business unless you can offer your customers more than your competitors, both in selection and service?

Smart Radio Buying

Decide on the right person to deal with (sales rep., sales manager, station manager). Choose a customer-minded, not a corporate-minded contact for this. Be firm in telling them that confidentiality is key to your negotiations. *Don't go through any negotiations unless you're really planning to advertise with the station*. Buy time by psyhchographics as well as demographics. Look into RAB (Instant background information on bedding, etc. -- what kind of station listeners buy what). *See if the station will sell you 60 second spots that you can piggy-back 30 second spots with another business*. (Not a competitor, of course.) These usually cost less than two 30 second spots. Make sure negotiations are a win-win situation. *Consider half-price immediately pre-emptable spots*. Check individual station's clocks (when commercials run and how many). Listen during those times to see who is advertising and how well the station is sold.

Buy one calendar quarter at a time with rates available for the next.

How to double effectiveness: Buy programs, e.g., Paul Harvey - open and close. Buy 10, get 10 bonus after midnight. Buy drive time within the hour you want. *Some stations will provide a jingle if you buy so many dollars.* Radio stations have tickets for concerts, sports events, etc. that they need to distribute -- have them use them with you. Invite local personality endorsements. Buy programs regularly. *Build promotions around payday*.

Welcome New Members

With the addition of the following companies, CSMG's membership has grown to twenty-two companies in nine states. CSMG companies represent approximately fifty stores.

Retail Bedding Vision is published by Central States Marketing Group, Inc, 804 W. Anthony Dr., Champaign, IL 61821-1224 Phone: (800) 369-4415. FAX (312) 355-0701.

Central States Marketing Group, Inc.

President, John Mies
Vice pres., Rich Mange
Sec.-treasurer, Jeff Long
Imm. past-president, Dick Mullen
Exec. committee member, Steve Booth
Exec. committee member, Brad Weiss

All information contained in Retail Bedding Vision is provided as suggestion only. Retailers are advised to obtain the advice of their attorneys, accountants, and marketing consultants before acting on information published herein.
Publisher:
Central States Marketing Group
Editor:
John Mies

This article was taken from excerpts of a presentation given by Rich Mange, Bed Mart of St. Louis, at the January 17, 1990 meeting of CSMG.

CENTRAL STATES

MARKETING NEWS

Second Edition September, 1997

Survival Tips for Small Retailers

Staying afloat in today's difficult economy requires strong management skills. *Keep your focus*. If you really believe in where you are going and are flexible in your path to get there, then there are few reasons why you won't accomplish your goal. *Empower your employees* by including them in every phase of your firm's planning, problem solving, and celebrating. *Provide on-going training* to every area of your store. Network with other business owners who are not your direct competitors. Each of you have the answers to help the other. *Differentiate your business*. Why should you expect to do business unless you can offer your customers more than your competitors, both in selection and service?

Smart Radio Buying

Decide on the right person to deal with (sales rep., sales manager, station manager). Choose a customer-minded, not a corporate-minded contact for this. Be firm in telling them that confidentiality is key to your negotiations. *Don't go through any negotiations unless you're really planning to advertise with the station*. Buy time by psyhchographics as well as demographics. Look into RAB (Instant background information on bedding, etc. -- what kind of station listeners buy what). *See if the station will sell you 60 second spots that you can piggy-back 30 second spots with another business*. (Not a competitor, of course.) These usually cost less than two 30 second spots. Make sure negotiations are a win-win situation. *Consider half-price immediately pre-emptable spots*. Check individual station's clocks (when commercials run and how many). Listen during those times to see who is advertising and how well the station is sold.

Buy one calendar quarter at a time with rates available for the next.

How to double effectiveness: Buy programs, e.g., Paul Harvey - open and close. Buy 10, get 10 bonus after midnight. Buy drive time within the hour you want. *Some stations will provide a jingle if you buy so many dollars.* Radio stations have tickets for concerts, sports events, etc. that they need to distribute -- have them use them with you. Invite local personality endorsements. Buy programs regularly. *Build promotions around payday*.

Welcome New Members

With the addition of the following companies, CSMG's membership has grown to twenty-two companies in nine states. CSMG companies represent approximately fifty stores.

Retail Bedding Vision is published by Central States Marketing Group, Inc, 804 W. Anthony Dr., Champaign, IL 61821-1224 Phone: (800) 369-4415. FAX (312) 355-0701.

Central States Marketing Group, Inc.

President, John Mies
Vice pres., Rich Mange
Sec.-treasurer, Jeff Long
Imm. past-president, Dick Mullen
Exec. committee member, Steve Booth
Exec. committee member, Brad Weiss

All information contained in Retail Bedding Vision is provided as suggestion only. Retailers are advised to obtain the advice of their attorneys, accountants, and marketing consultants before acting on information published herein.
Publisher:
Central States Marketing Group
Editor:
John Mies

This article was taken from excerpts of a presentation given by Rich Mange, Bed Mart of St. Louis, at the January 17, 1990 meeting of CSMG.

Fig. 3.1.: Comparison of the raster page and the sample fading page.

3.1.3 Our Approach

In order to detect fading, our approach is to compare user printed page against the raster page. However, before the two pages are being compared, the printed page needs to be scanned by a scanning device, and the raster page needs to be calibrated into the printer's color space. We also need to make sure that the scanning device is well calibrated. However, even if the two pages are in the correct color space, the scanned page can also be spatially misaligned with respect to the raster

[1] Research supported by HP Inc., Boise, ID 83714.

page. Therefore, the scanned page needs to be spatially aligned with the raster page. The same image registration algorithm used in the previous work is also used in our work which extracts feature points, estimates homography matrix by using MLESAC or RANSAC, and then transforms the scanned page by the homography matrix [11] [12] [13] [14] [15] [16].

For the above-mentioned alignment procedure, we name it global registration. After the global registration, the page is further aligned which we name it local registration. We first do connected component (CC) analysis [17] [18] to find out all the CCs in the page. Then template matching is applied [19] for the local registration.

After the image registration, for the text region, we use Otsu's Thresholding [20] to binarize each CC and calculate the ΔE which is the $L^*a^*b^*$ color difference. And for the non-text region, we propose an algorithm that calculates the "super pixel" $L^*a^*b^*$ value. The value is defined as the average pixel values within a super pixel block of both the raster page, and the customer scanned page. All super pixels are clustered based on their colors. When fading occurs, the ΔE of certain color clusters will increase significantly. This is how fading is detected.

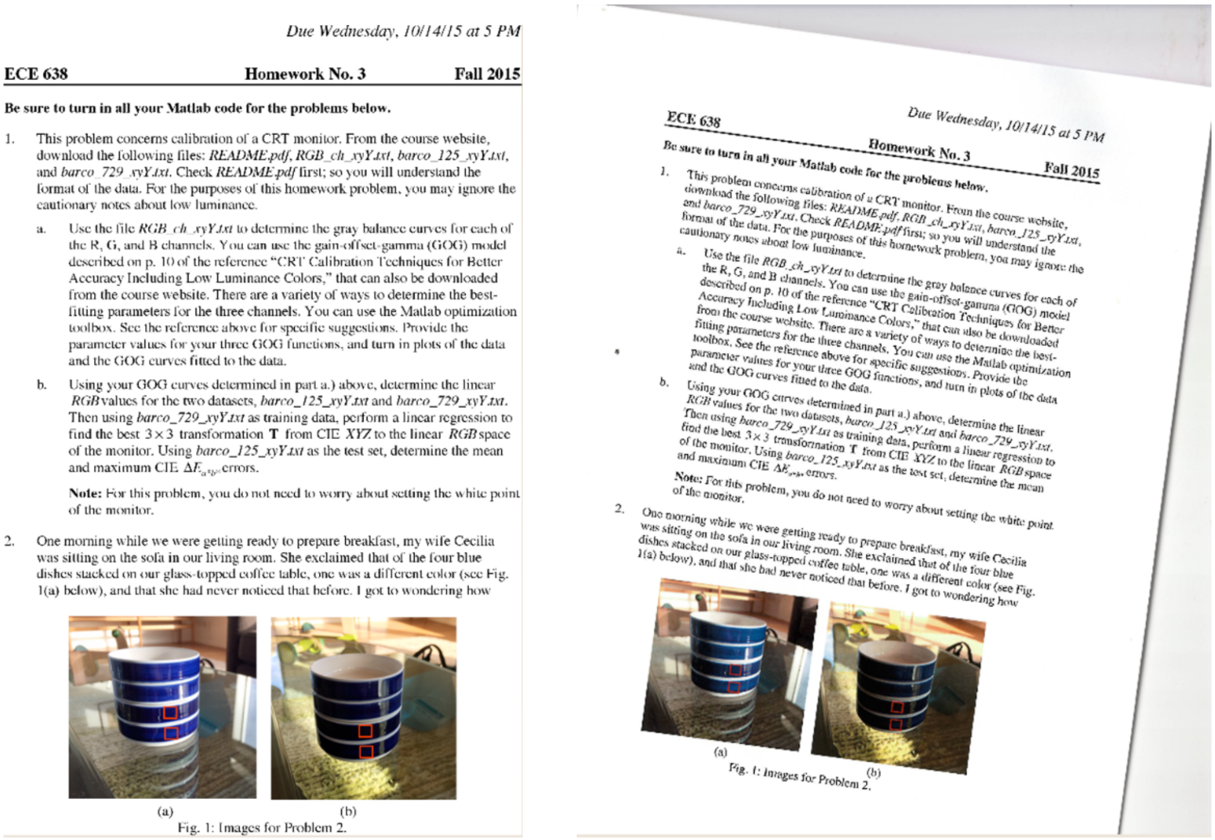
3.2 Text Region Fading Detection

3.2.1 Image Registration

Global Registration

According to the introduction section, it is known that in order to detect fading, the customer scanned image needs to be compared with the raster image. However, during the scanning process, the scanned image may be misaligned, as shown in Figure 3.2:

In our case, there may be three different types of distortion which are: Rotation, Scale and Translation. For Rotation, the image being scanned may not be well aligned with the scanner. For Scale, the scanned image has smaller dimension than the raster



Due Wednesday, 10/14/15 at 5 PM

ECE 638 **Homework No. 3** **Fall 2015**

Be sure to turn in all your Matlab code for the problems below.

1. This problem concerns calibration of a CRT monitor. From the course website, download the following files: *README.pdf*, *RGB_ch_xyY.txt*, *barco_125_xyY.txt*, and *barco_729_xyY.txt*. Check *README.pdf* first; so you will understand the format of the data. For the purposes of this homework problem, you may ignore the cautionary notes about low luminance.

   a. Use the file *RGB_ch_xyY.txt* to determine the gray balance curves for each of the R, G, and B channels. You can use the gain-offset-gamma (GOG) model described on p. 10 of the reference "CRT Calibration Techniques for Better Accuracy Including Low Luminance Colors," that can also be downloaded from the course website. There are a variety of ways to determine the best-fitting parameters for the three channels. You can use the Matlab optimization toolbox. See the reference above for specific suggestions. Provide the parameter values for your three GOG functions, and turn in plots of the data and the GOG curves fitted to the data.

   b. Using your GOG curves determined in part a.) above, determine the linear *RGB* values for the two datasets, *barco_125_xyY.txt* and *barco_729_xyY.txt*. Then using *barco_729_xyY.txt* as training data, perform a linear regression to find the best 3×3 transformation $\mathbf{T}$ from CIE *XYZ* to the linear *RGB* space of the monitor. Using *barco_125_xyY.txt* as the test set, determine the mean and maximum CIE $\Delta E_{a^*b^*}$ errors.

   Note: For this problem, you do not need to worry about setting the white point of the monitor.

2. One morning while we were getting ready to prepare breakfast, my wife Cecilia was sitting on the sofa in our living room. She exclaimed that of the four blue dishes stacked on our glass-topped coffee table, one was a different color (see Fig. 1(a) below), and that she had never noticed that before. I got to wondering how

(a) (b)

Fig. 1: Images for Problem 2.

(a) Raster image.

Due Wednesday, 10/14/15 at 5 PM

ECE 638 **Homework No. 3** **Fall 2015**

Be sure to turn in all your Matlab code for the problems below.

1. This problem concerns calibration of a CRT monitor. From the course website, download the following files: *README.pdf*, *RGB_ch_xyY.txt*, *barco_125_xyY.txt*, and *barco_729_xyY.txt*. Check *README.pdf* first; so you will understand the format of the data. For the purposes of this homework problem, you may ignore the cautionary notes about low luminance.

   a. Use the file *RGB_ch_xyY.txt* to determine the gray balance curves for each of the R, G, and B channels. You can use the gain-offset-gamma (GOG) model described on p. 10 of the reference "CRT Calibration Techniques for Better Accuracy Including Low Luminance Colors," that can also be downloaded from the course website. There are a variety of ways to determine the best-fitting parameters for the three channels. You can use the Matlab optimization toolbox. See the reference above for specific suggestions. Provide the parameter values for your three GOG functions, and turn in plots of the data and the GOG curves fitted to the data.

   b. Using your GOG curves determined in part a.) above, determine the linear *RGB* values for the two datasets, *barco_125_xyY.txt* and *barco_729_xyY.txt*. Then using *barco_729_xyY.txt* as training data, perform a linear regression to find the best 3×3 transformation $\mathbf{T}$ from CIE *XYZ* to the linear *RGB* space of the monitor. Using *barco_125_xyY.txt* as the test set, determine the mean and maximum CIE $\Delta E_{a^*b^*}$ errors.

   Note: For this problem, you do not need to worry about setting the white point of the monitor.

2. One morning while we were getting ready to prepare breakfast, my wife Cecilia was sitting on the sofa in our living room. She exclaimed that of the four blue dishes stacked on our glass-topped coffee table, one was a different color (see Fig. 1(a) below), and that she had never noticed that before. I got to wondering how

(a) (b)

Fig. 1: Images for Problem 2.

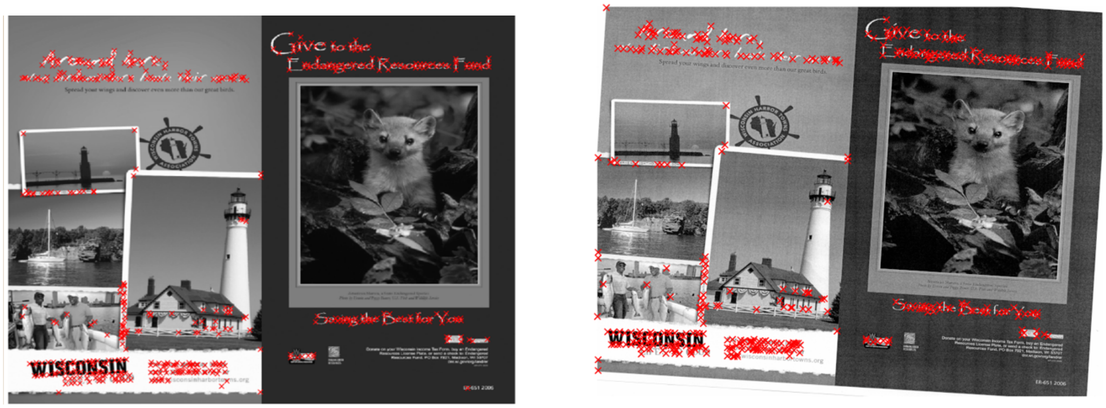
(b) Misaligned scanned image.

Fig. 3.2.: Comparison between the raster and the misaligned scanned image.

page. The scale depends on the user's printing setting. A larger page margin will reduce the dimension of the scanned image. We do not have to consider scale if we can directly read the scale parameter from the printer setting. Last but not least, for Translation, even if the scanned image is deskewed and correctly scaled, its position needs to be aligned with the raster image by translative movement.

Firstly, we use Harris Corner Method [11] to extract feature points from both raster and scanned images, as shown in Figure 3.3.

After the feature points are detected, a certain feature point in the scanned image needs to be matched with its correspondence in the raster image. We use the SSD (Sum Square Difference) as the metric, as shown in Figure 3.4. For each descriptor



(a) Feature points in the raster image.

(b) Feature points in the misaligned scanned image.

Fig. 3.3.: Harris corner feature points in both raster and misaligned scanned image.

in f1, the SSD is calculated relative to all the descriptors in f2. Then we extract the minimum SSD descriptor in f2, which is considered to be a match to the descriptor in f1.

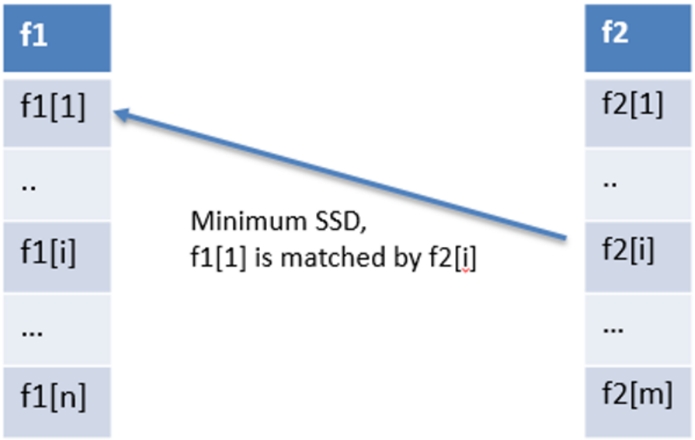


Fig. 3.4.: SSD feature matching.

After all feature points are matched, we need to remove the outliers in order to get the most accurate similarity matrix. RANSAC (Random Sample Consensus) is applied in our experiment, as shown in Figure 3.5.

Algorithm 1 RANSAC

```
1: Select randomly the minimum number of points required to determine the model parameters.
2: Solve for the parameters of the model.
3: Determine how many points from the set of all points fit with a predefined tolerance $\epsilon$.
4: If the fraction of the number of inliers over the total number points in the set exceeds a predefined threshold $\tau$, re-estimate the model parameters using all the identified inliers and terminate.
5: Otherwise, repeat steps 1 through 4 (maximum of $N$ times).
```

Fig. 3.5.: Description of the RANSAC method.

With all the aforementioned steps, the similarity matrix T, is generated. We then apply it to the distorted image to get the final aligned image. Figure 3.6 shows the alignment process.

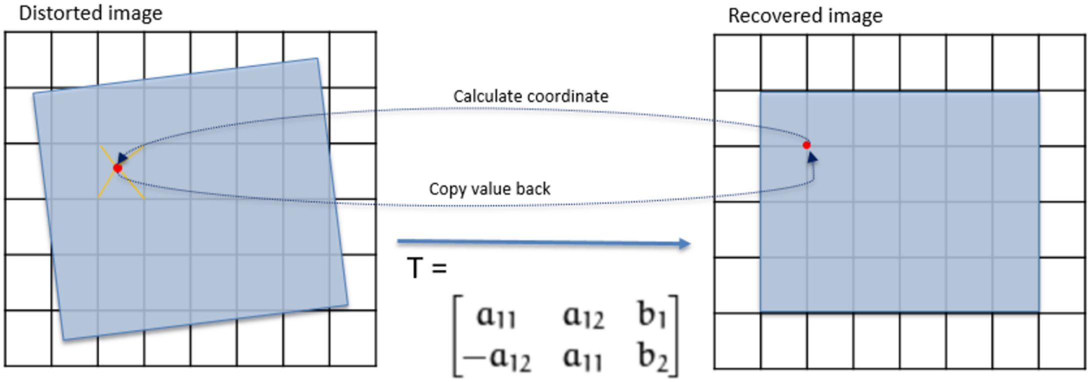
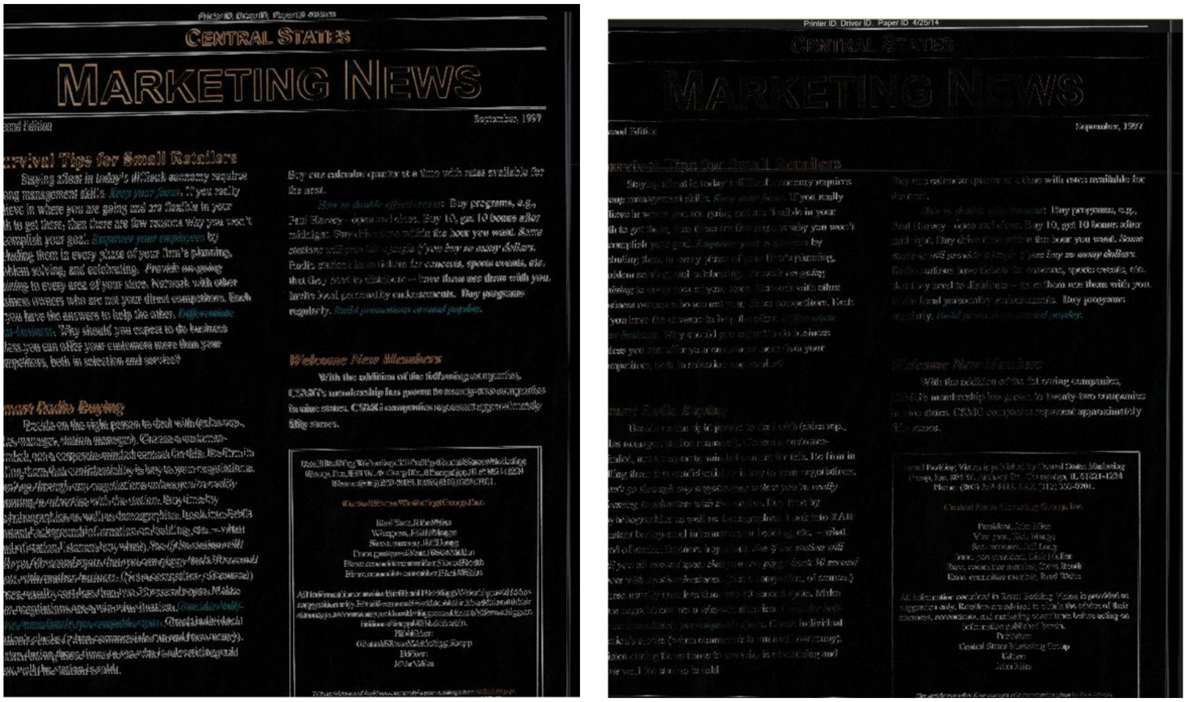

Distorted image
Recovered image
Calculate coordinate
Copy value back
T =
$\begin{bmatrix} a_{11} & a_{12} & b_1 \\ -a_{12} & a_{11} & b_2 \end{bmatrix}$

Fig. 3.6.: Applying the similarity matrix to the distorted image.

Experimental Results

The ”Marketing News” image which is shown in Figure 3.1 is used as an example. In order to visualize the global registration result more straightforwardly, we use the difference image to demonstrate the results. A difference image is generated by subtracting the aligned image directly from the raster image. Noting that, for a difference image, the darker the image is, the better it is aligned.

From Figure 3.7, it is observed that, after the global registration, the raster and the scanned images are mostly aligned.



(a) Before global registration.

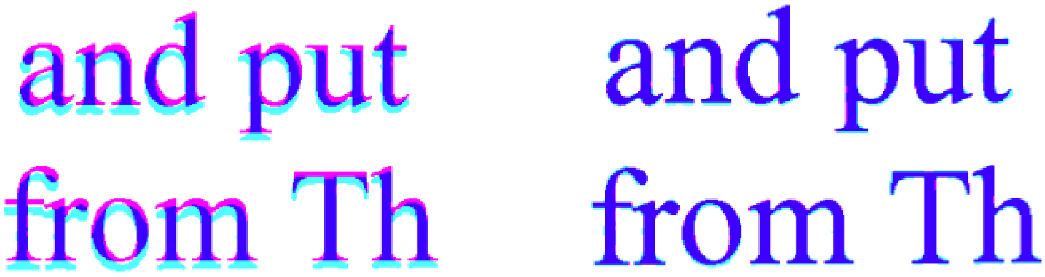
(b) After global registration.

Fig. 3.7.: Difference images before and after global registration.

Local Registration

After global registration, the image needs to be further aligned to reduce the pixel-wise error. We perform it by using the CC algorithm. After a scanned image is globally aligned, we locate all the CCs within both raster and scanned images. The largest value of cross-validation between the master CC and the test CC is calculated and it leads to the translation shift. After applying the shift, all CCs will be perfectly aligned resulting in the alignment of the scanned page.

More specifically, let cc_t be a sample CC in the testing image, cc_r be its correspondence in the raster image. The way we calculate the shift of cc_t with respect to cc_r is to evaluate the normalized cross correlation value at each pixel position (i, j) within cc_t and cc_r. Lets suppose that (i, j) has been shifted by k pixels in the x direction and l pixels in the y direction. After calculating the normalized cross correlation for cc_t and cc_r, the shift (k_{max}, l_{max}) with the maximum normalized cross correlation value ρ_{max} indicates the displacement for which the two images will be best aligned. In this case, (k_{max}, l_{max}) is the desired (k, l). As shown in Figure 3.8, after local registration, the raster image and the scanned image are perfectly aligned.



(a) Before local registration.

(b) After local registration.

Fig. 3.8.: Cyan is the scanned image, Magenta is the raster image and Blue is the overlapping.

3.2.2 Calculating Color Difference

After the global and local registration, we calculate the color difference ΔE in $L^*a^*b^*$ color space as given in Equation 3.1:

$$\Delta E^*_{ab} = \sqrt{(L^*_2 - L^*_1)^2 + (a^*_2 - a^*_1)^2 + (b^*_2 - b^*_1)^2} \tag{3.1}$$

Noting that the previous work does ΔE calculation pixel by pixel, but as Figure 3.9 shows, even if both raster and scanned images are well aligned, because of the toner scattering phenomenon (toner being emitted on the paper by printing needle will spread around its designated position), the pixel-by-pixel based ΔE is still larger than 5 for non-faded pages. But in common cases, ΔE should be no larger than 2.3 for JND (Just Noticeable Difference). So ΔE should not be calculated in a pixel based manner.

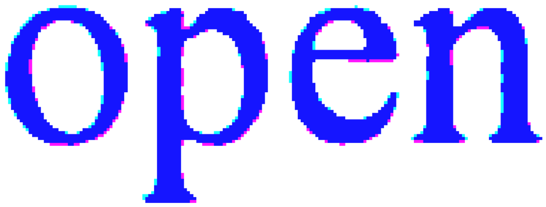
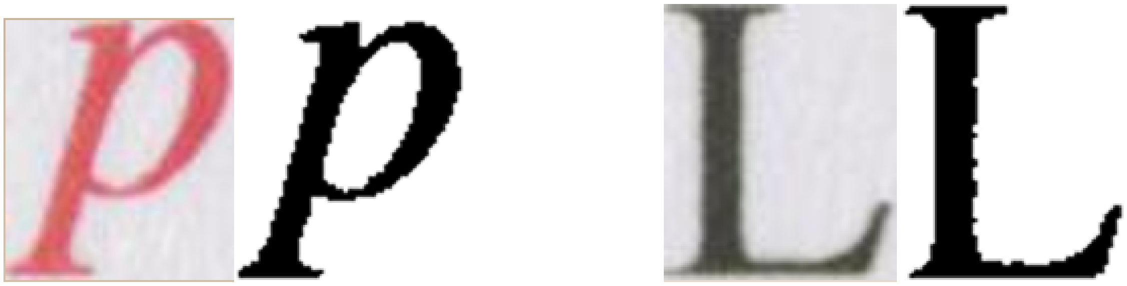


Fig. 3.9.: Even if both raster and scanned pages are well locally aligned, because of the toner scattering phenomenon, the pixel-by-pixel based ΔE will still be very large.

Instead, we calculate the ΔE component by component. We first use Otsu's Method [20] to threshold all the CC bounding boxes of both raster and scanned images. Then we apply bounding box enlargement to make sure all of the CC pixels are in the bounding box. Figure 3.10 shows two examples before and after Otsu's Thresholding. Afterwards, we calculate the mean $L^*a^*b^*$ value of both raster and scanned images' CCs and get the ΔE value for that specific CC pair. The final ΔE value of the image will be the average of all the CCs' ΔE values.



(a) Letter P before and after Otsu Thresholding.

(b) Letter L before and after Otsu Thresholding.

Fig. 3.10.: We will be able to separate foreground and background using Otsu's Thresholding.

Figure 3.11 shows the flowchart of our text fading detection pipeline.

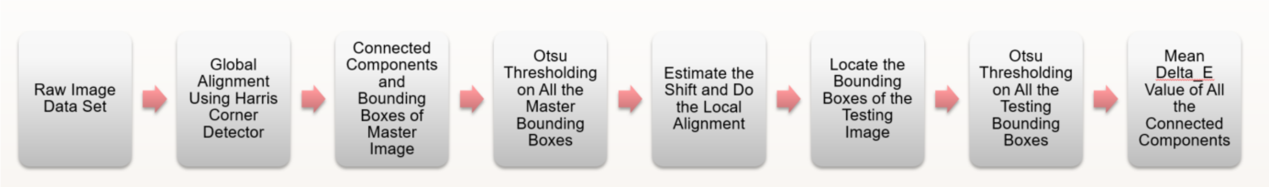

Raw Image Data Set
Global Alignment Using Harris Corner Detector
Connected Components and Bounding Boxes of Master Image
Otsu Thresholding on All the Master Bounding Boxes
Estimate the Shift and Do the Local Alignment
Locate the Bounding Boxes of the Testing Image
Otsu Thresholding on All the Testing Bounding Boxes
Mean Delta_E Value of All the Connected Components

Fig. 3.11.: Text fading detection process.

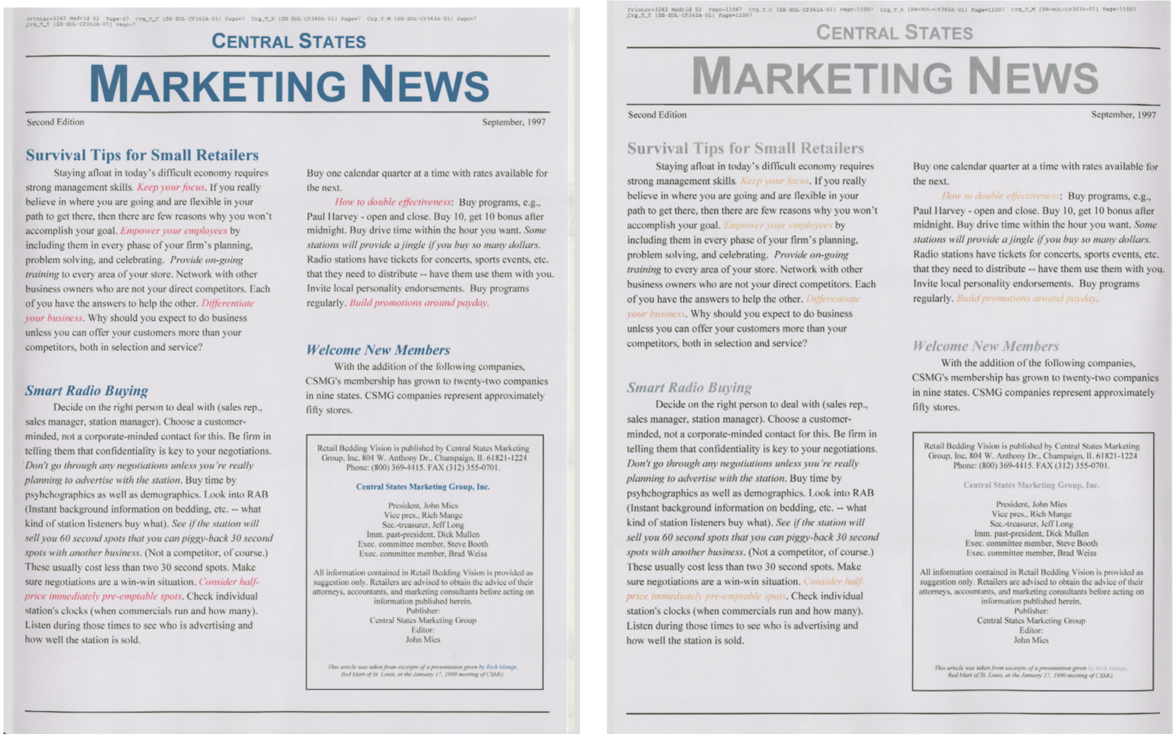
3.2.3 Experimental Results

In this section, two sets of sample images and their testing results are displayed.

The first set of images are the "Marketing News". The ground truth is given that fading appears at Page 80 and disappears at Page 89. Fading then reappears at Page 178 and lasts until the last page which is Page 188. In order to visualize the fading more easily, the comparison between the raster page which is Page 1 with the most faded pages which are Page 88 and Page 186 are displayed. Figure 3.12 and Figure 3.13 show the comparison between the raster page and the most faded pages.

The ground truth is that Magenta and Cyan fades from Page 80 and Black fading from Page 180 so that we expect to see the ΔE value rising at those page numbers. The experimental result is shown in Figure 3.14. Noting that for some cases, fading only appears in part of the image instead of the whole image. So in order to represent the fading level more accurately, we evenly divide each image into four strips and the ΔE value for the image is the maximum ΔE value of the four strips. It can be observed that ΔE value rises at Page 80 and reaches its local peak at Page 88, then it rises again at Page 178 and reaches its local peak at Page 186.

Another set of examples are the "Durango". As shown in Figure 3.15, Figure 3.16 and Figure 3.17, We have the ground truth that Magenta and Cyan fading begins from Page 80, Yellow fading begins from around Page 107 and Black fading begins from Page 179. We expect to observe ΔE value rising at those three page numbers.



Printer=3262 Madrid 02 Page=67 Crg_T_C [SB-EOL-CF361A-01] Page=7 Crg_T_K [SB-EOL-CF360A-01] Page=7 Crg_T_M [SB-EOL-CF363A-01] Page=7
Crg_T_Y [SB-EOL-CF362A-01] Page=7

CENTRAL STATES

MARKETING NEWS

Second Edition September, 1997

Survival Tips for Small Retailers

Staying afloat in today's difficult economy requires strong management skills. *Keep your focus*. If you really believe in where you are going and are flexible in your path to get there, then there are few reasons why you won't accomplish your goal. *Empower your employees* by including them in every phase of your firm's planning, problem solving, and celebrating. *Provide on-going training* to every area of your store. Network with other business owners who are not your direct competitors. Each of you have the answers to help the other. *Differentiate your business*. Why should you expect to do business unless you can offer your customers more than your competitors, both in selection and service?

Smart Radio Buying

Decide on the right person to deal with (sales rep., sales manager, station manager). Choose a customer-minded, not a corporate-minded contact for this. Be firm in telling them that confidentiality is key to your negotiations. *Don't go through any negotiations unless you're really planning to advertise with the station*. Buy time by psyhchographics as well as demographics. Look into RAB (Instant background information on bedding, etc. -- what kind of station listeners buy what). *See if the station will sell you 60 second spots that you can piggy-back 30 second spots with another business*. (Not a competitor, of course.) These usually cost less than two 30 second spots. Make sure negotiations are a win-win situation. *Consider half-price immediately pre-emptable spots*. Check individual station's clocks (when commercials run and how many). Listen during those times to see who is advertising and how well the station is sold. Buy one calendar quarter at a time with rates available for the next.

How to double effectiveness: Buy programs, e.g., Paul Harvey - open and close. Buy 10, get 10 bonus after midnight. Buy drive time within the hour you want. *Some stations will provide a jingle if you buy so many dollars.* Radio stations have tickets for concerts, sports events, etc. that they need to distribute -- have them use them with you. Invite local personality endorsements. Buy programs regularly. *Build promotions around payday*.

Welcome New Members

With the addition of the following companies, CSMG's membership has grown to twenty-two companies in nine states. CSMG companies represent approximately fifty stores.

Retail Bedding Vision is published by Central States Marketing Group, Inc, 804 W. Anthony Dr., Champaign, IL 61821-1224
Phone: (800) 369-4415. FAX (312) 355-0701.

Central States Marketing Group, Inc.

President, John Mies
Vice pres., Rich Mange
Sec.-treasurer, Jeff Long
Imm. past-president, Dick Mullen
Exec. committee member, Steve Booth
Exec. committee member, Brad Weiss

All information contained in Retail Bedding Vision is provided as suggestion only. Retailers are advised to obtain the advice of their attorneys, accountants, and marketing consultants before acting on information published herein.
Publisher:
Central States Marketing Group
Editor:
John Mies

This article was taken from excerpts of a presentation given by Rich Mange, Bed Mart of St. Louis, at the January 17, 1990 meeting of CSMG.

Printer=3262 Madrid 02 Page=11067 Crg_T_C [SB-EOL-CF361A-01] Page=11007 Crg_T_K [SB-EOL-CF360A-01] Page=11007 Crg_T_M [SB-EOL-CF363A-01] Page=11007
Crg_T_Y [SB-EOL-CF362A-01] Page=11007

CENTRAL STATES

MARKETING NEWS

Second Edition September, 1997

Survival Tips for Small Retailers

Staying afloat in today's difficult economy requires strong management skills. Keep your focus. If you really believe in where you are going and are flexible in your path to get there, then there are few reasons why you won't accomplish your goal. Empower your employees by including them in every phase of your firm's planning, problem solving, and celebrating. Provide on-going training to every area of your store. Network with other business owners who are not your direct competitors. Each of you have the answers to help the other. Differentiate your business. Why should you expect to do business unless you can offer your customers more than your competitors, both in selection and service?

Smart Radio Buying

Decide on the right person to deal with (sales rep., sales manager, station manager). Choose a customer-minded, not a corporate-minded contact for this. Be firm in telling them that confidentiality is key to your negotiations. Don't go through any negotiations unless you're really planning to advertise with the station. Buy time by psyhchographics as well as demographics. Look into RAB (Instant background information on bedding, etc. -- what kind of station listeners buy what). See if the station will sell you 60 second spots that you can piggy-back 30 second spots with another business. (Not a competitor, of course.) These usually cost less than two 30 second spots. Make sure negotiations are a win-win situation. Consider half-price immediately pre-emptable spots. Check individual station's clocks (when commercials run and how many). Listen during those times to see who is advertising and how well the station is sold. Buy one calendar quarter at a time with rates available for the next.

How to double effectiveness: Buy programs, e.g., Paul Harvey - open and close. Buy 10, get 10 bonus after midnight. Buy drive time within the hour you want. Some stations will provide a jingle if you buy so many dollars. Radio stations have tickets for concerts, sports events, etc. that they need to distribute -- have them use them with you. Invite local personality endorsements. Buy programs regularly. Build promotions around payday.

Welcome New Members

With the addition of the following companies, CSMG's membership has grown to twenty-two companies in nine states. CSMG companies represent approximately fifty stores.

Retail Bedding Vision is published by Central States Marketing Group, Inc, 804 W. Anthony Dr., Champaign, IL 61821-1224
Phone: (800) 369-4415. FAX (312) 355-0701.

Central States Marketing Group, Inc.

President, John Mies
Vice pres., Rich Mange
Sec.-treasurer, Jeff Long
Imm. past-president, Dick Mullen
Exec. committee member, Steve Booth
Exec. committee member, Brad Weiss

All information contained in Retail Bedding Vision is provided as suggestion only. Retailers are advised to obtain the advice of their attorneys, accountants, and marketing consultants before acting on information published herein.
Publisher:
Central States Marketing Group
Editor:
John Mies

This article was taken from excerpts of a presentation given by Rich Mange, Bed Mart of St. Louis, at the January 17, 1990 meeting of CSMG.

(a) Raster page: Page 1.

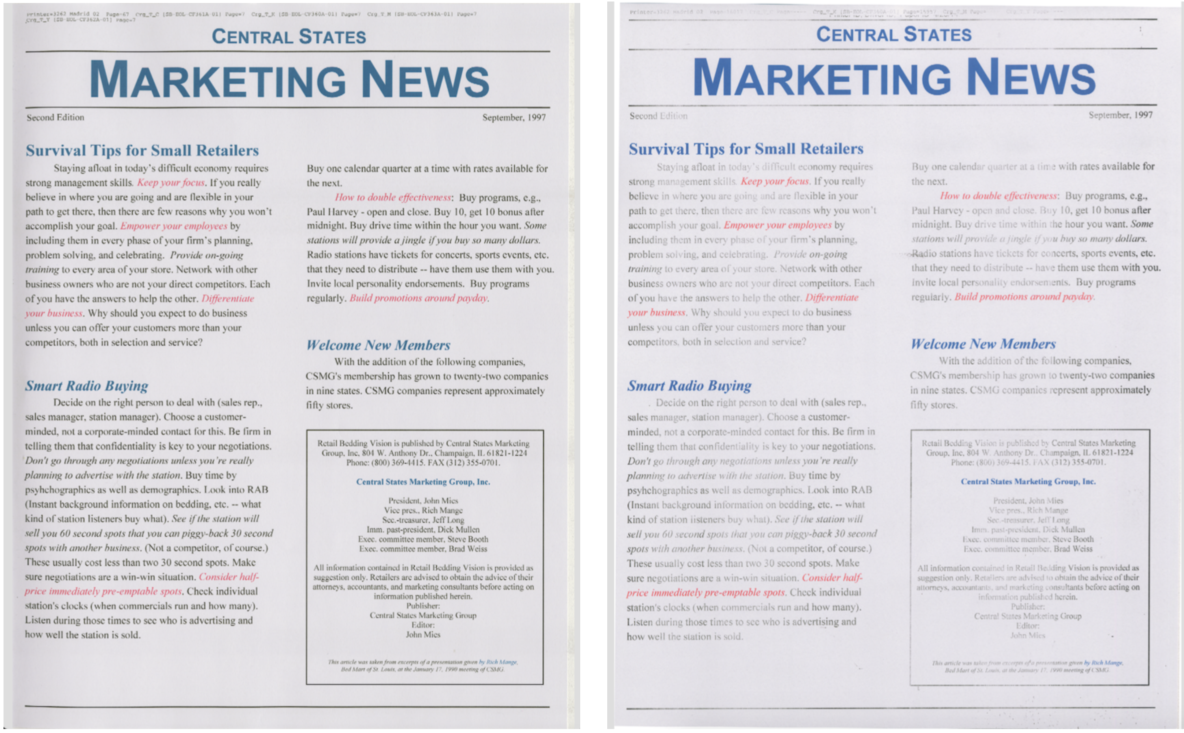
(b) Scanned page: Page 88.

Fig. 3.12.: Comparison between the raster page and one of the most faded pages.

Noting that Yellow fading also happens in "Marketing News" set but since there is no yellow color in the image, no ΔE value rising is detected in the plot. Figure 3.18 shows the experimental result. The consistency between our result and the ground truth validates the fact that our algorithm works very well. We also test our algorithm on several other image sets and the results are promising and accurate. It can be concluded that our method of detecting text fading works effectively and efficiently.

3.3 Non-text Region Fading Detection

Even though our text region fading detection algorithm works well, there are two aspects that need to be improved. One is that we can only observe one or more colors



printer=3262 Madrid 02 Page=67 Crg_T_C [SB-EOL-CF361A-01] Page=7 Crg_T_K [SB-EOL-CF360A-01] Page=7 Crg_T_M [SB-EOL-CF363A-01] Page=7 Crg_T_Y [SB-EOL-CF362A-01] Page=7

CENTRAL STATES

MARKETING NEWS

Second Edition September, 1997

Survival Tips for Small Retailers

Staying afloat in today's difficult economy requires strong management skills. *Keep your focus*. If you really believe in where you are going and are flexible in your path to get there, then there are few reasons why you won't accomplish your goal. *Empower your employees* by including them in every phase of your firm's planning, problem solving, and celebrating. *Provide on-going training* to every area of your store. Network with other business owners who are not your direct competitors. Each of you have the answers to help the other. *Differentiate your business*. Why should you expect to do business unless you can offer your customers more than your competitors, both in selection and service?

Smart Radio Buying

Decide on the right person to deal with (sales rep., sales manager, station manager). Choose a customer-minded, not a corporate-minded contact for this. Be firm in telling them that confidentiality is key to your negotiations. *Don't go through any negotiations unless you're really planning to advertise with the station*. Buy time by psyhchographics as well as demographics. Look into RAB (Instant background information on bedding, etc. -- what kind of station listeners buy what). *See if the station will sell you 60 second spots that you can piggy-back 30 second spots with another business*. (Not a competitor, of course.) These usually cost less than two 30 second spots. Make sure negotiations are a win-win situation. *Consider half-price immediately pre-emptable spots*. Check individual station's clocks (when commercials run and how many). Listen during those times to see who is advertising and how well the station is sold. Buy one calendar quarter at a time with rates available for the next.

How to double effectiveness: Buy programs, e.g., Paul Harvey - open and close. Buy 10, get 10 bonus after midnight. Buy drive time within the hour you want. *Some stations will provide a jingle if you buy so many dollars.* Radio stations have tickets for concerts, sports events, etc. that they need to distribute -- have them use them with you. Invite local personality endorsements. Buy programs regularly. *Build promotions around payday.*

Welcome New Members

With the addition of the following companies, CSMG's membership has grown to twenty-two companies in nine states. CSMG companies represent approximately fifty stores.

Retail Bedding Vision is published by Central States Marketing
Group, Inc, 804 W. Anthony Dr., Champaign, IL 61821-1224
Phone: (800) 369-4415. FAX (312) 355-0701.

Central States Marketing Group, Inc.

President, John Mies
Vice pres., Rich Mange
Sec.-treasurer, Jeff Long
Imm. past-president, Dick Mullen
Exec. committee member, Steve Booth
Exec. committee member, Brad Weiss

All information contained in Retail Bedding Vision is provided as suggestion only. Retailers are advised to obtain the advice of their attorneys, accountants, and marketing consultants before acting on information published herein.
Publisher:
Central States Marketing Group
Editor:
John Mies

This article was taken from excerpts of a presentation given by Rich Mange, Bed Mart of St. Louis, at the January 17, 1990 meeting of CSMG.

(a) Raster page: Page 1.

CENTRAL STATES

MARKETING NEWS

Second Edition September, 1997

Survival Tips for Small Retailers

Staying afloat in today's difficult economy requires strong management skills. *Keep your focus*. If you really believe in where you are going and are flexible in your path to get there, then there are few reasons why you won't accomplish your goal. *Empower your employees* by including them in every phase of your firm's planning, problem solving, and celebrating. *Provide on-going training* to every area of your store. Network with other business owners who are not your direct competitors. Each of you have the answers to help the other. *Differentiate your business*. Why should you expect to do business unless you can offer your customers more than your competitors, both in selection and service?

Smart Radio Buying

Decide on the right person to deal with (sales rep., sales manager, station manager). Choose a customer-minded, not a corporate-minded contact for this. Be firm in telling them that confidentiality is key to your negotiations. *Don't go through any negotiations unless you're really planning to advertise with the station*. Buy time by psyhchographics as well as demographics. Look into RAB (Instant background information on bedding, etc. -- what kind of station listeners buy what). *See if the station will sell you 60 second spots that you can piggy-back 30 second spots with another business*. (Not a competitor, of course.) These usually cost less than two 30 second spots. Make sure negotiations are a win-win situation. *Consider half-price immediately pre-emptable spots*. Check individual station's clocks (when commercials run and how many). Listen during those times to see who is advertising and how well the station is sold. Buy one calendar quarter at a time with rates available for the next.

How to double effectiveness: Buy programs, e.g., Paul Harvey - open and close. Buy 10, get 10 bonus after midnight. Buy drive time within the hour you want. *Some stations will provide a jingle if you buy so many dollars.* Radio stations have tickets for concerts, sports events, etc. that they need to distribute -- have them use them with you. Invite local personality endorsements. Buy programs regularly. *Build promotions around payday.*

Welcome New Members

With the addition of the following companies, CSMG's membership has grown to twenty-two companies in nine states. CSMG companies represent approximately fifty stores.

Retail Bedding Vision is published by Central States Marketing
Group, Inc, 804 W. Anthony Dr., Champaign, IL 61821-1224
Phone: (800) 369-4415. FAX (312) 355-0701.

Central States Marketing Group, Inc.

President, John Mies
Vice pres., Rich Mange
Sec.-treasurer, Jeff Long
Imm. past-president, Dick Mullen
Exec. committee member, Steve Booth
Exec. committee member, Brad Weiss

All information contained in Retail Bedding Vision is provided as suggestion only. Retailers are advised to obtain the advice of their attorneys, accountants, and marketing consultants before acting on information published herein.
Publisher:
Central States Marketing Group
Editor:
John Mies

This article was taken from excerpts of a presentation given by Rich Mange, Bed Mart of St. Louis, at the January 17, 1990 meeting of CSMG.

(b) Scanned page: Page 186.

Fig. 3.13.: Comparison between the raster page and one of the most faded pages.

fading at a certain page number, but the color type is unknown. The other one is that we cannot deal with non-text region fading problem. In this section, we demonstrate how the two improvements are made.

3.3.1 3D Color Node System

In order to represent different colors, we construct a 3D Color Node System. Firstly, the three channels of RGB color space is evenly divided into 4 sections by 5 nodes which are 0, 63, 127, 191 and 255. We will have $5 \times 5 \times 5$ which are 125 nodes in the RGB coordinate system. Each node represents a specific color. All the nodes from RGB color space are then converted to $L^*a^*b^*$ color space as shown in Figure 3.19.

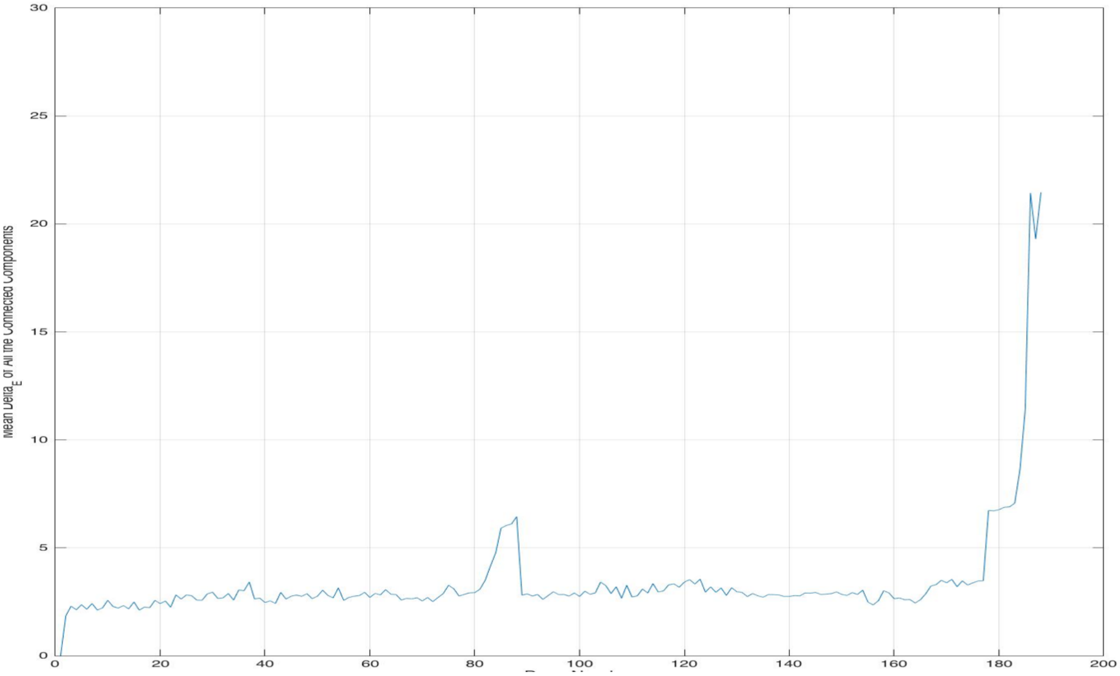
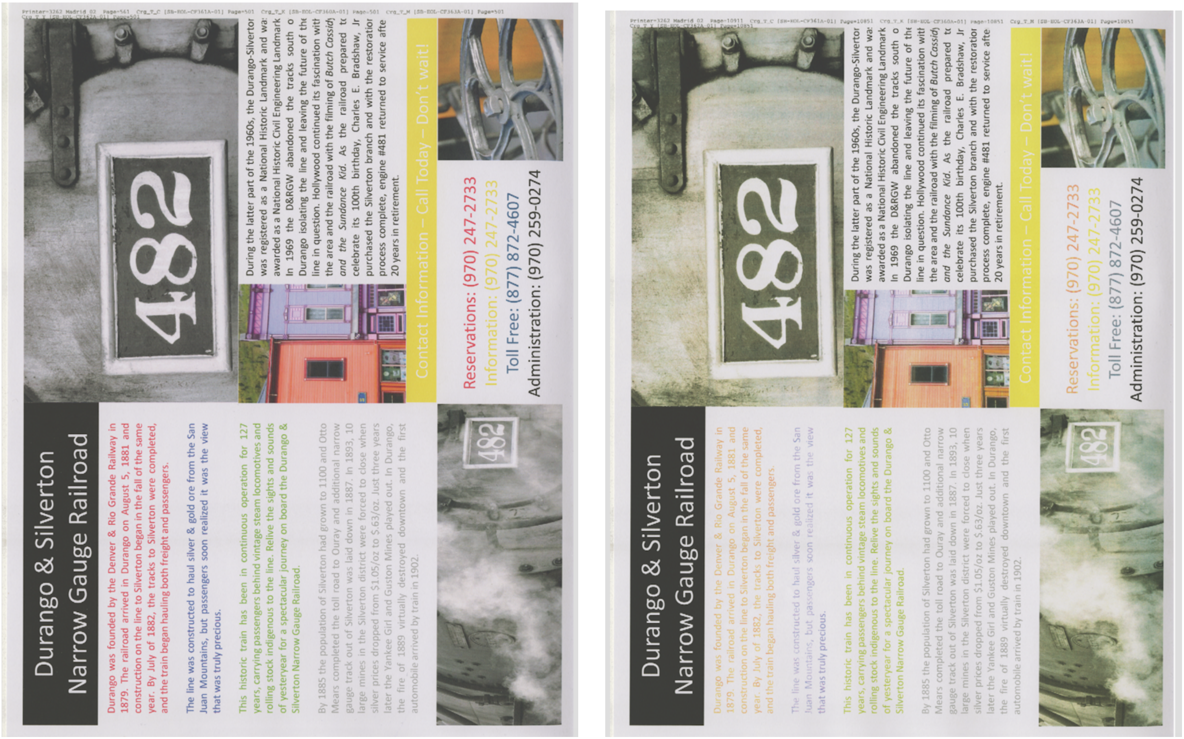


Fig. 3.14.: ΔE value v.s. page numbers plot.

Now for the $L^{*}a^{*}b^{*}$ value of each raster image CC, we assign it to one of the 125 nodes based on the minimum Euclidean Distance between the CC $L^{*}a^{*}b^{*}$ value and the node $L^{*}a^{*}b^{*}$ value. For example, a raster CC has the minimum Euclidean Distance from the 10^{th} node, then this CC is assigned to the color of the 10^{th} node. The color assignment of the scanned page follows the raster page color assignment. The corresponding CC has the same node assignment i.e. the corresponding scanned image CC is also assigned to the 10^{th} node.

The color shift can be easily visualized through the 3D $L^{*}a^{*}b^{*}$ coordinate system. As shown in Figure 3.20, which is the color shift for the ”Marketing News” images. It can be seen that in Page 88, Magenta and Cyan color shift appears and in Page 185, Black color shift takes place.


Durango & Silverton Narrow Gauge Railroad
482
Contact Information – Call Today – Don't wait!
Reservations: (970) 247-2733
Information: (970) 247-2733
Toll Free: (877) 872-4607
Administration: (970) 259-0274

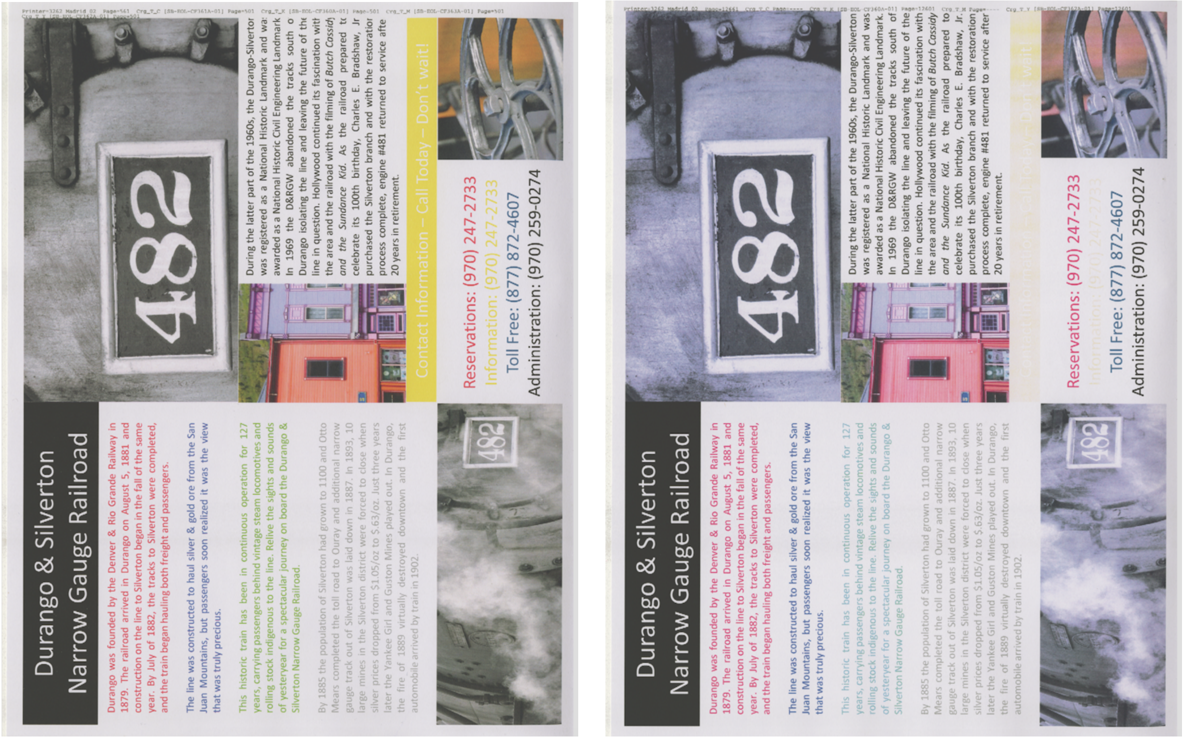
(a) Raster page: Page 1.

(b) Testing page: Page 84.

Fig. 3.15.: Comparison between the raster page and the first appeared most faded page.

3.3.2 Object Map

An object map of a raster page tells the printer how this page should be rendered. An example is provided in Figure 3.21(b), which is the object map of the raster image shown in Figure 3.21(a). Three types of objects are commonly seen in a customer page: Symbol, Raster, and Vector. Symbol mainly represents text characters which are not of our interests since we focus on non-text region in this section. Raster represents those rough regions that contain many details and Vector represents smooth graphic areas. Raster regions are colored in the dark gray and Vector objects are in light gray. Print quality can be improved by using object-oriented halftoning which requires an object map [21].



Durango & Silverton
Narrow Gauge Railroad

Durango was founded by the Denver & Rio Grande Railway in 1879. The railroad arrived in Durango on August 5, 1881 and construction on the line to Silverton began in the fall of the same year. By July of 1882, the tracks to Silverton were completed, and the train began hauling both freight and passengers.

The line was constructed to haul silver & gold ore from the San Juan Mountains, but passengers soon realized it was the view that was truly precious.

This historic train has been in continuous operation for 127 years, carrying passengers behind vintage steam locomotives and rolling stock indigenous to the line. Relive the sights and sounds of yesteryear for a spectacular journey on board the Durango & Silverton Narrow Gauge Railroad.

By 1885 the population of Silverton had grown to 1100 and Otto Mears completed the toll road to Ouray and additional narrow gauge track out of Silverton was laid down in 1887. In 1893, 10 large mines in the Silverton district were forced to close when silver prices dropped from $1.05/oz to $.63/oz. Just three years later the Yankee Girl and Guston Mines played out. In Durango, the fire of 1889 virtually destroyed downtown and the first automobile arrived by train in 1902.

During the latter part of the 1960s, the Durango-Silvertor was registered as a National Historic Landmark and wa: awarded as a National Historic Civil Engineering Landmark In 1969 the D&RGW abandoned the tracks south o Durango isolating the line and leaving the future of the line in question. Hollywood continued its fascination witl the area and the railroad with the filming of *Butch Cassid) and the Sundance Kid*. As the railroad prepared tc celebrate its 100th birthday, Charles E. Bradshaw, Jr purchased the Silverton branch and with the restoratior process complete, engine #481 returned to service afte 20 years in retirement.

Contact Information – Call Today – Don't wait!

Reservations: (970) 247-2733
Information: (970) 247-2733
Toll Free: (877) 872-4607
Administration: (970) 259-0274

(a) Raster page: Page 1.

Durango & Silverton
Narrow Gauge Railroad

Durango was founded by the Denver & Rio Grande Railway in 1879. The railroad arrived in Durango on August 5, 1881 and construction on the line to Silverton began in the fall of the same year. By July of 1882, the tracks to Silverton were completed, and the train began hauling both freight and passengers.

The line was constructed to haul silver & gold ore from the San Juan Mountains, but passengers soon realized it was the view that was truly precious.

This historic train has been in continuous operation for 127 years, carrying passengers behind vintage steam locomotives and rolling stock indigenous to the line. Relive the sights and sounds of yesteryear for a spectacular journey on board the Durango & Silverton Narrow Gauge Railroad.

By 1885 the population of Silverton had grown to 1100 and Otto Mears completed the toll road to Ouray and additional narrow gauge track out of Silverton was laid down in 1887. In 1893, 10 large mines in the Silverton district were forced to close when silver prices dropped from $1.05/oz to $.63/oz. Just three years later the Yankee Girl and Guston Mines played out. In Durango, the fire of 1889 virtually destroyed downtown and the first automobile arrived by train in 1902.

During the latter part of the 1960s, the Durango-Silverton was registered as a National Historic Landmark and was awarded as a National Historic Civil Engineering Landmark. In 1969 the D&RGW abandoned the tracks south of Durango isolating the line and leaving the future of the line in question. Hollywood continued its fascination with the area and the railroad with the filming of *Butch Cassidy and the Sundance Kid*. As the railroad prepared to celebrate its 100th birthday, Charles E. Bradshaw, Jr. purchased the Silverton branch and with the restoration process complete, engine #481 returned to service after 20 years in retirement.

Contact Information – Call Today – Don't wait!

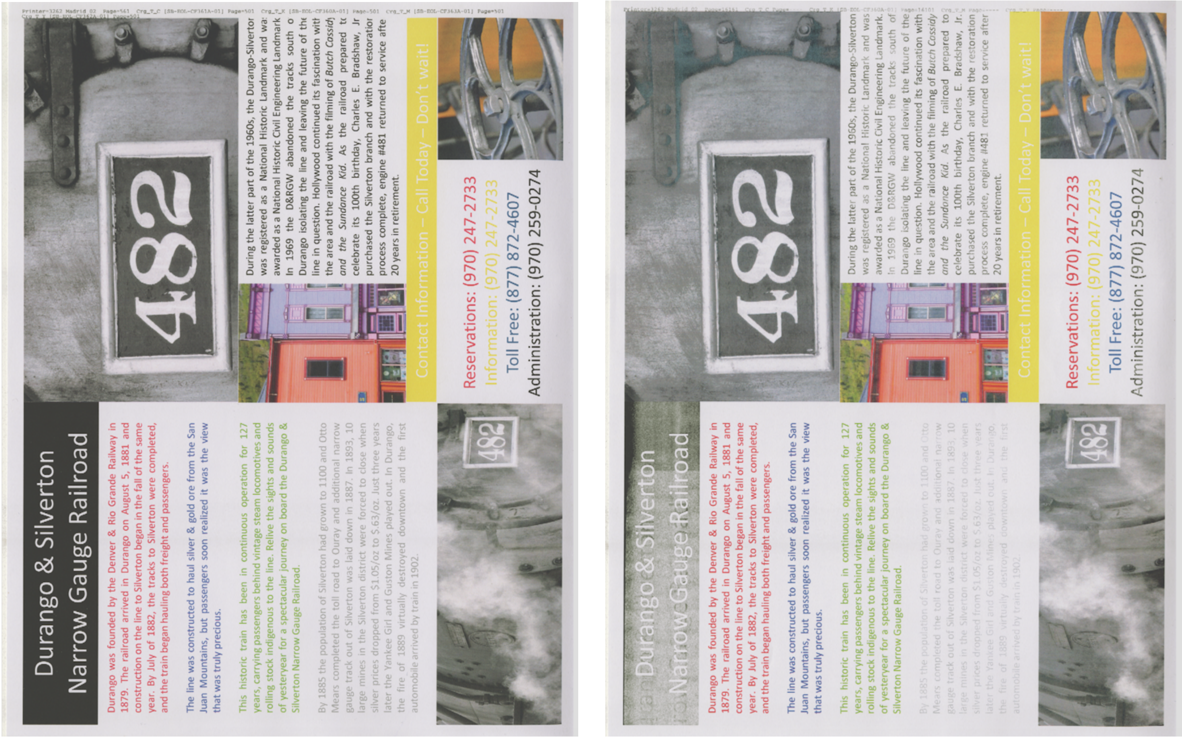
Reservations: (970) 247-2733
Information: (970) 247-2733
Toll Free: (877) 872-4607
Administration: (970) 259-0274

(b) Testing page: Page 119.

Fig. 3.16.: Comparison between the raster page and the second appeared most faded page.

3.3.3 Non-text Region Fading Detection

With the help from the object map to locate the Raster and Vector regions, we are able to do the non-text region fading detection. But a different approach of image registration is applied compared to the text region fading detection in the previous section. Recall that after the global registration, local registration is performed so that the raster page and the scanned page are further aligned to improve the experimental results. Since cross-correlation is applied for local registration but with regard to Raster and Vector regions, they usually occupy large areas in an image which causes the cross-correlation calculation extremely time consuming. For example, if a Raster region has size 600 × 800, the cross-correlation calculation can take up to an hour to



(a) Raster page: Page 1.

(b) Testing page: Page 186.

Fig. 3.17.: Comparison between the raster page and the last appeared most faded page.

finish. Accordingly, local registration is abandoned. But in the meantime, since we have given up local registration, a new way of calculating ΔE is needed.

In order to solve this problem, the object map, the raster image and the scanned image is partitioned into size 60×60 blocks. We name each block a ”super pixel”. Recall that even before local registration, the result from global registration is accurate enough. It usually has only 3-5 pixel-offset. Compared with the size of the super pixel, 3-5 pixel-offset error is not significant which will be ”diluted”.

After the image is partitioned, we assign every super pixel to one of the 125 nodes same as the previous section and do the fading detection. Figure 3.21 is used as our testing sample. Noting that for Figure 3.21(a), after color node assignment, it has 14 color clusters for Raster and 14 color clusters for Vector.

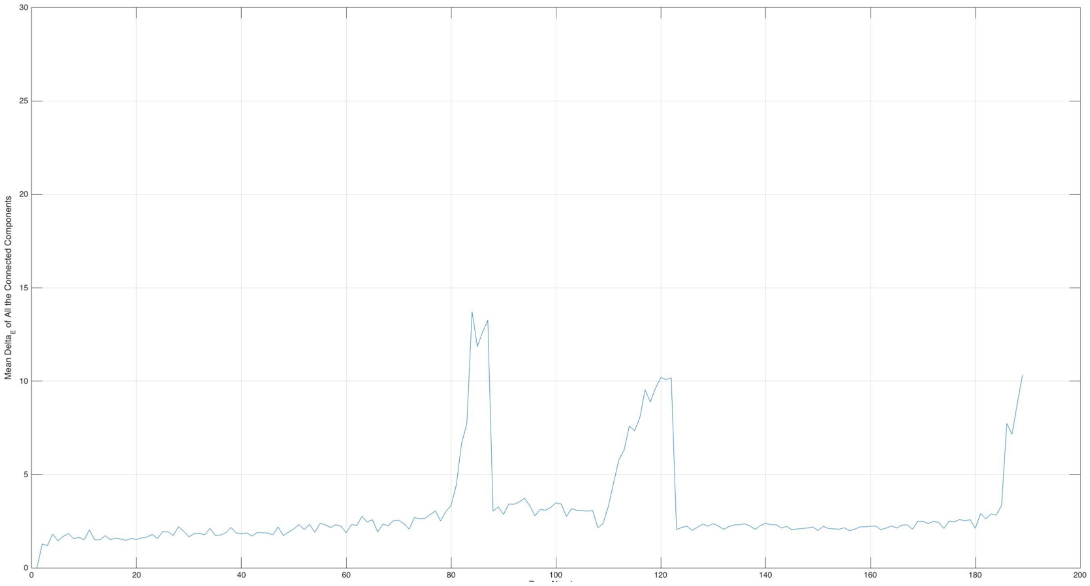
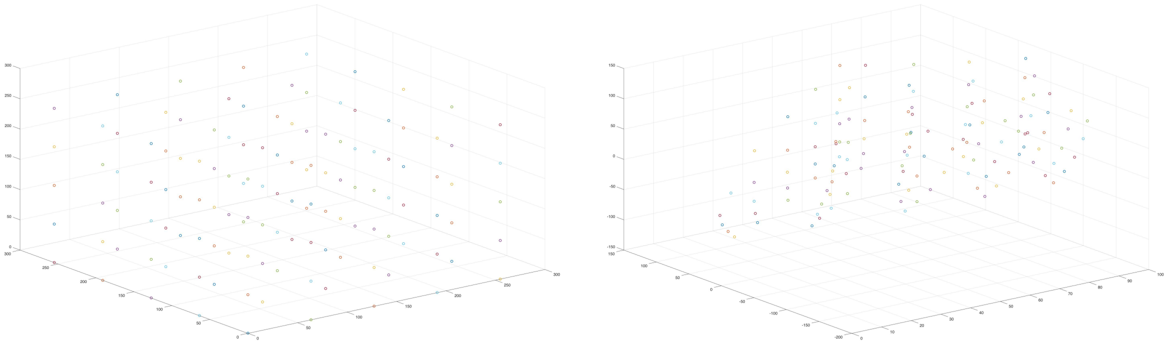


Fig. 3.18.: ΔE value v.s. page numbers plot.

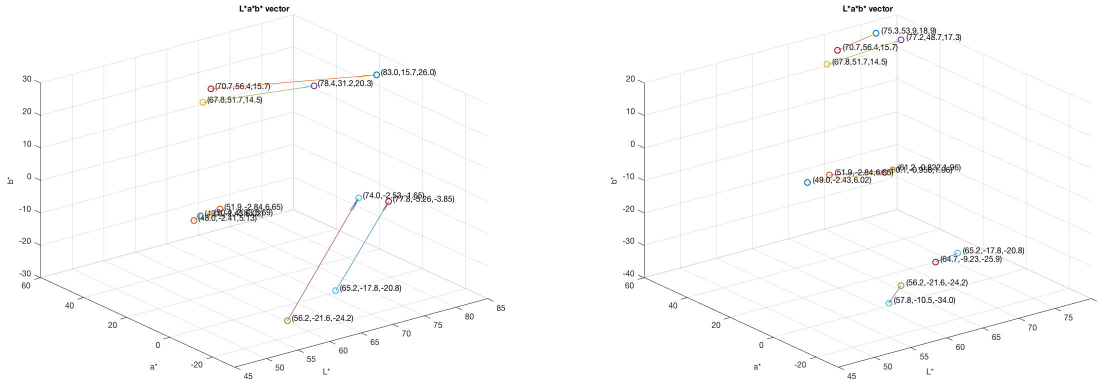


(a) 125 points RGB color nodes.

(b) 125 points $L^*a^*b^*$ color nodes.

Fig. 3.19.: 3D Color Nodes System.

Figure 3.22 shows the ΔE plot for Raster. Recall that Magenta fades at Page 80, Cyan fades at Page 82, Yellow fades at Page 107 and Black fades at Page 180. Since 14 lines in a single plot is extremely crowded, the main plot will be divided into 3 subplots as shown in Figure 3.23, Figure 3.24 and Figure 3.25. The plots also show the color of each cluster. Noting that the colors of the lines correspond to the colors of the clusters.



(a) Color shift between Page 1 and Page 88.

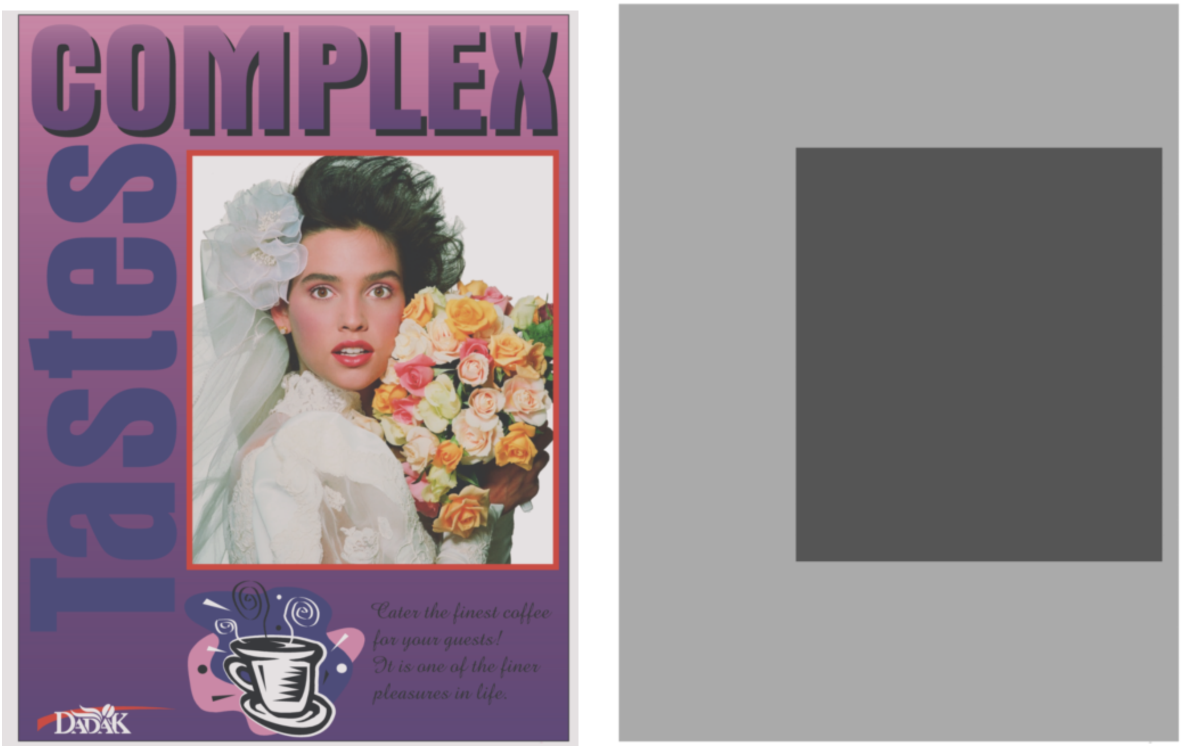
(b) Color shift between Page 1 and Page 185.

Fig. 3.20.: Color shift for "Marketing News" images.

Figure 3.26 shows the ΔE plot for Vector. Figure 3.27, Figure 3.28 and Figure 3.29 are its subplots.

3.4 Conclusion

In this chapter, we propose an approach of detecting color fading for both text region and non-text region from a scanned customer page by comparing it to its raster page automatically. Our algorithm first assigns each CC or super pixel to one of the color nodes in a 3D Color Node System, then locally analyzes each cluster, to predict the depleted cartridge based on its color. From the resulting plots, it can be clearly observed that our proposed algorithm works accurately and efficiently which has obvious advantages compared to the traditional method which depends on a diagnostic page.



(a) Sample raster image.

(b) Object map of the sample raster image.

Fig. 3.21.: Sample raster image with its object map.

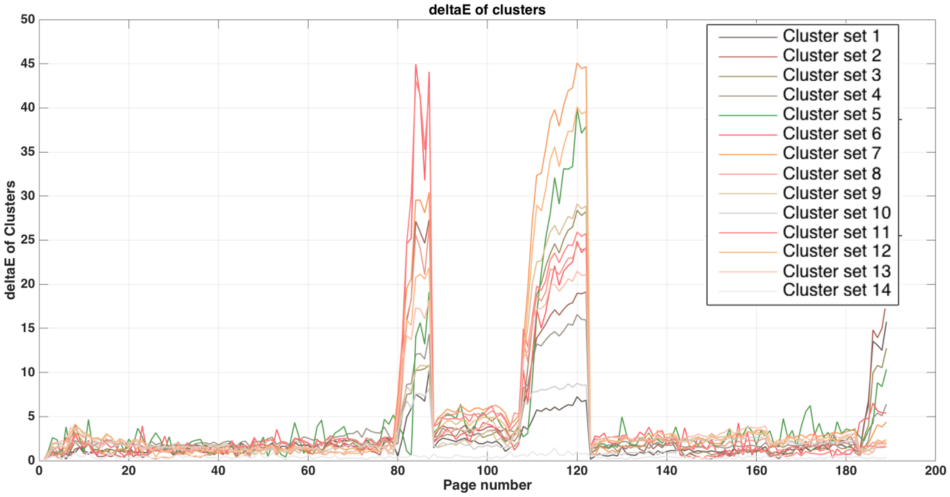


Fig. 3.22.: ΔE value v.s. page number plot for Raster.

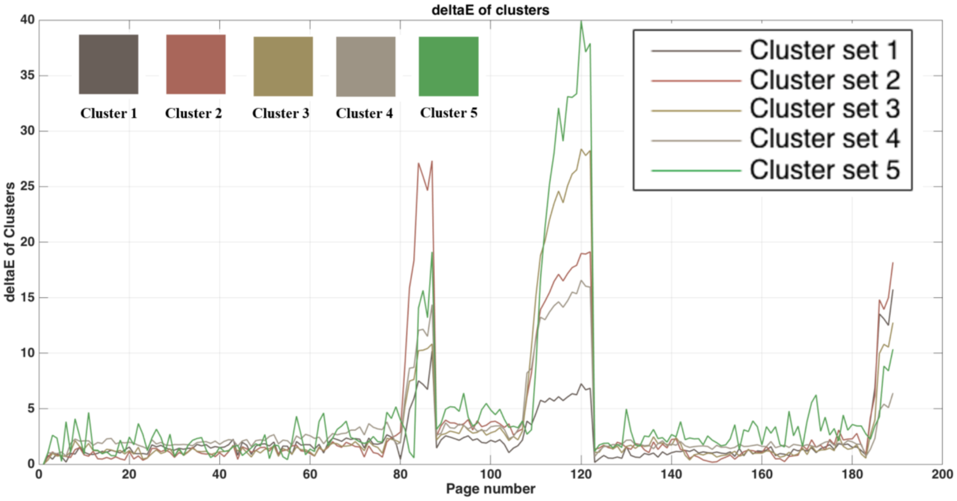


Fig. 3.23.: ΔE value v.s. page numbers plot for Raster Cluster 1-5.

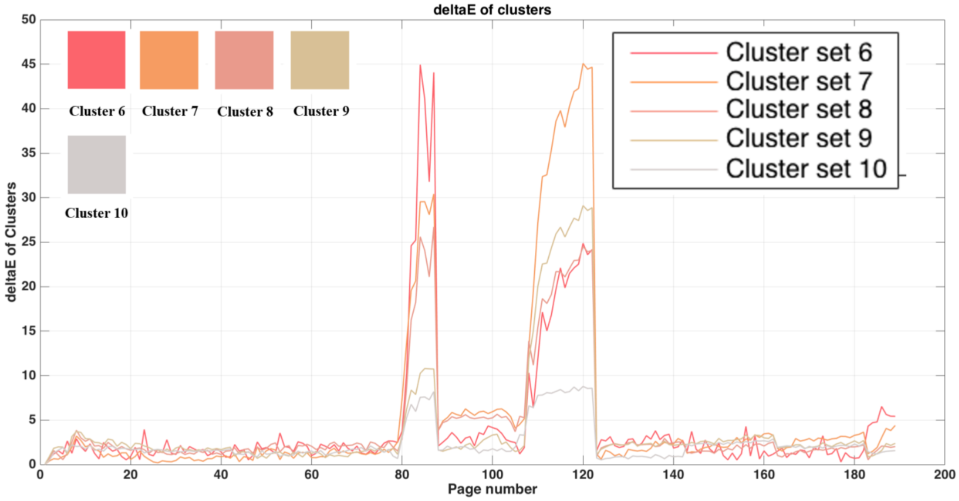


Fig. 3.24.: ΔE value v.s. page numbers plot for Raster Cluster 6-10.

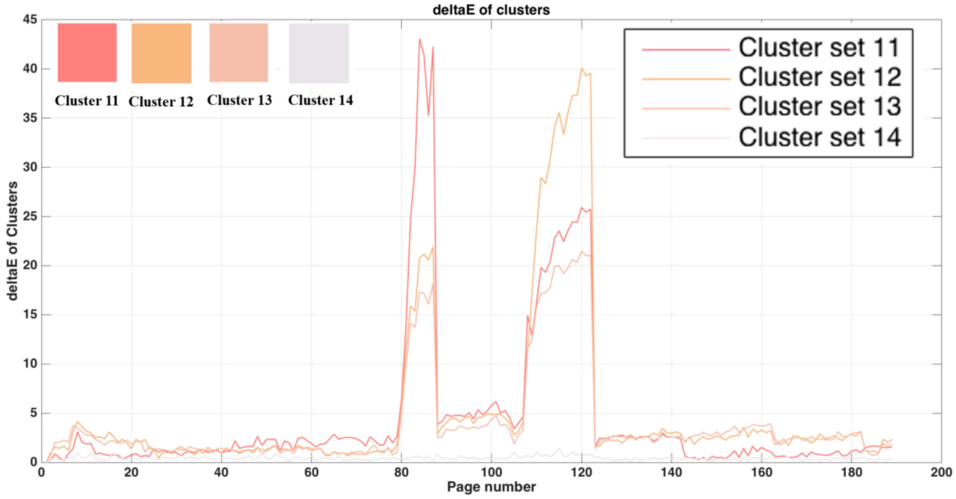


Fig. 3.25.: ΔE value v.s. page numbers plot for Raster Cluster 11-14.

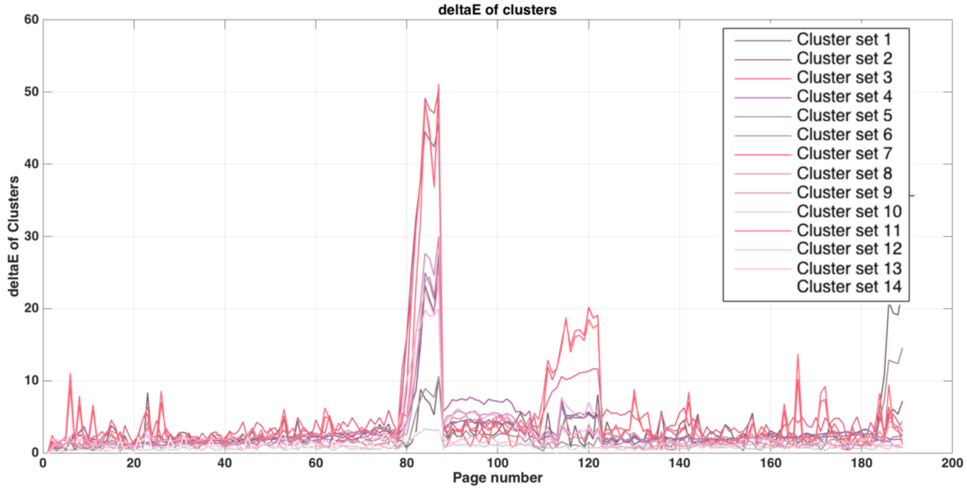


Fig. 3.26.: ΔE value v.s. page number plot for Vector.

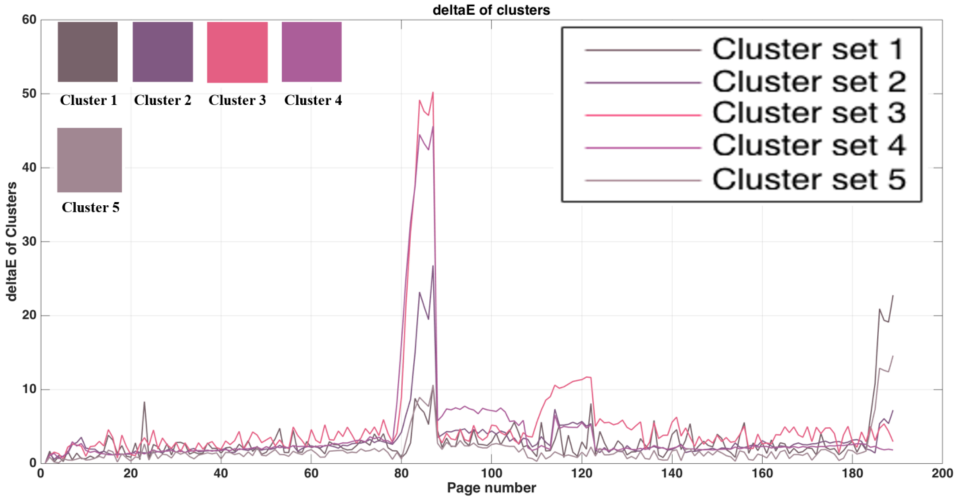


Fig. 3.27.: ΔE value v.s. page numbers plot for Vector Cluster 1-5.

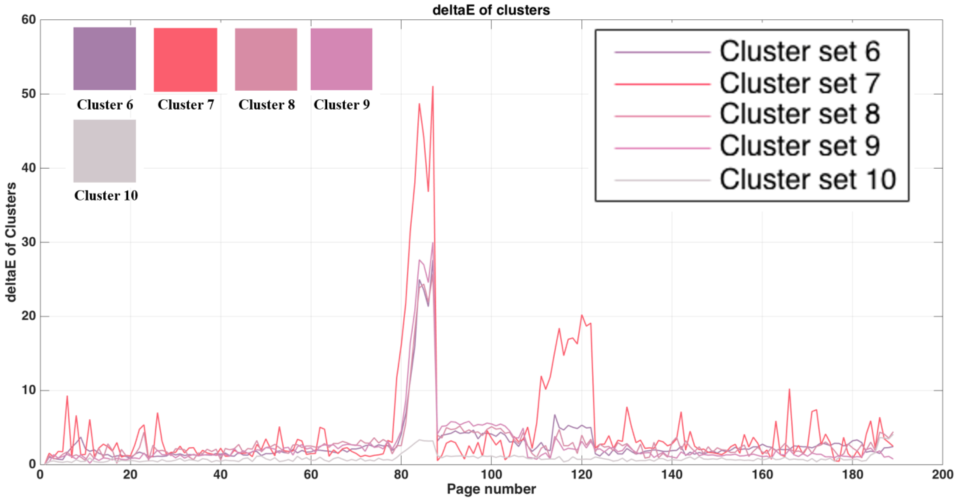


Fig. 3.28.: ΔE value v.s. page numbers plot for Vector Cluster 6-10.

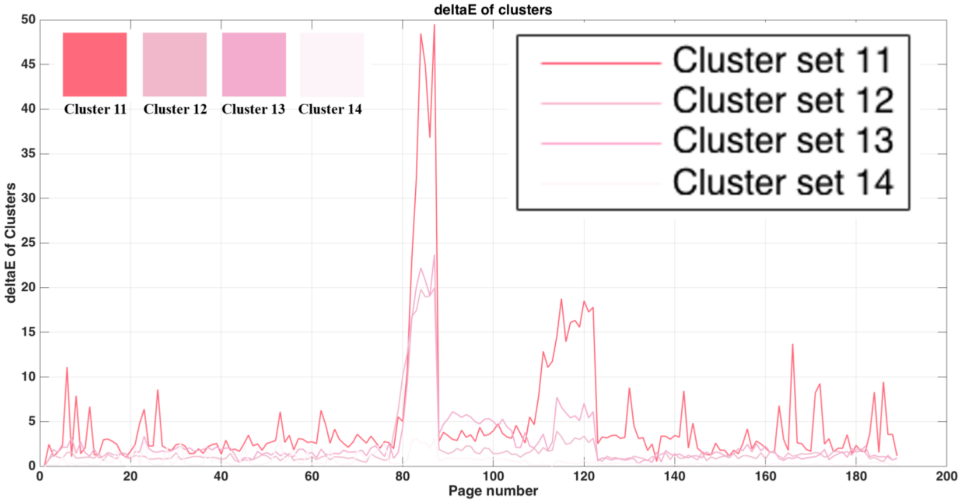


Fig. 3.29.: ΔE value v.s. page numbers plot for Vector Cluster 11-14.

4. EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS

4.1 Introduction

4.1.1 General Introduction of the Project

Emotion has always been extremely significant to us since it is one of the key features of human beings. It helps people to communicate and understand each other. So understanding various emotions have always been a popular topic. There have been a variety of means to detect emotions, such as voice intonation, body language and even electroencephalography [22]. However, the most intuitive and practical way of detecting and recognizing emotions is still through facial expressions. So our approach to detect emotions is examining facial expressions. In this chapter, we propose a system to detect emotions by examining facial expressions. In our system, we follow the research work proposed by Paul Ekman [23], where the emotions are categorized into seven classes: angry, disgust, fear, happy, contempt, sad and surprise, except that the category contempt is replaced with neutral in our work.

4.1.2 Related Work

There has been a lot of research work on emotion recognition, most of which uses traditional computer vision methods, such as LBP [24]; and machine learning classification methods, such as SVM [25]. However, satisfying results could not be achieved due to the limitations of these methods, such as inadaptability to the change of facial muscles. Therefore, we have put much effort in investigating a new approach that take advantages of deep learning [1].

[1] Research supported by HP Labs, Palo Alto, CA 94304.

There is also substantial research work done on emotion recognition using deep learning such as traditional model training methods using a specific network [26], or combining deep learning with machine learning such as LBP [27]. Although they obtain comparably high accuracy, there are two aspects that need to be improved. Firstly, most of them use traditional network structures such as VGG Net, Alex Net, or Google Net (including the improved versions of these network structures). This results in a large model size; so that it is extremely difficult to do real time emotion recognition. Secondly, most of the proposed systems only consider the classification scenario where the intensity information is missing in the results. But in practical usage, intensity information is as important as the classification result, because we want to know not only what emotions people have, but also the level of those emotions. In this chapter, we solve these two problems by selecting an appropriate network structure for an accurate real-time emotion recognition. At the same time, we extend our classification results to the regression scenario so that the intensity information can be concluded from the results. We trained our emotion recognition model for both the classification scenario and the regression scenario.

4.1.3 Seven Classes of Emotions

There are seven universal emotions: Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. Examples of these emotions are shown in Figure 4.1 [28] [29].

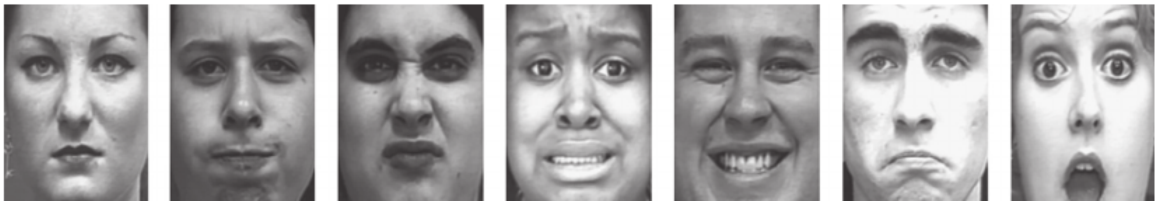


Fig. 4.1.: Seven universal emotions: Neutral, Angry, Disgust, Fear, Happy, Sad and Surprise.

Noting that each emotion is also described by some characteristics [30]:

- Angry: eyebrows pulled down, upper lids pulled up, lower lids pulled up, margins of lips rolled in and lips may be tightened.
- Disgust: eyebrows pulled down, nose wrinkled, upper lip pulled up and lips loose.
- Fear: eyebrows pulled up and together, upper eyelids pulled up, mouth stretch.
- Happy: muscle around the eyes tightened, "Crows Feet" wrinkles around eyes, cheeks raised and lip corners raised diagonally.
- Sad: inner corners of eyebrows raised, eyelids loose and Lip corners pulled down.
- Surprise: entire eyebrow pulled up, eyelids pulled up and mouth hangs open.

4.1.4 Our Approaches

Our emotion recognition project includes two parts: Classification Training and Regression Training. More details will be explained in the later sections. Figure 4.2 shows the flowchart of our emotion recognition project.

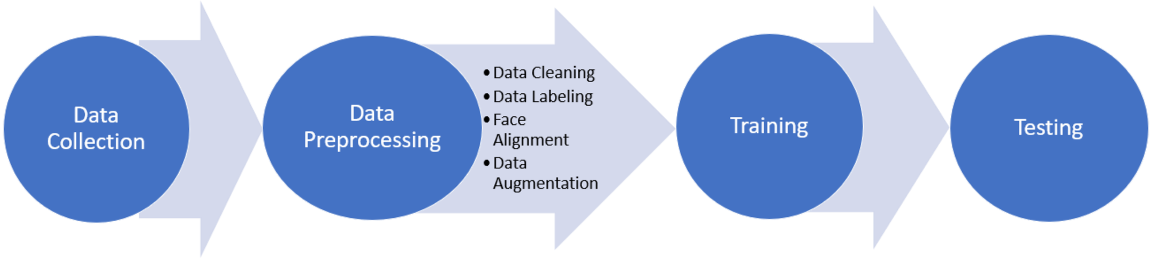


Fig. 4.2.: Flowchart of Emotion Recognition Project.

4.2 Emotion Recognition Classification Training

4.2.1 Data Collection

Due to the lack of public data sets for emotion recognition tasks and the low quality of existing data sets, collecting enough data sets and examining them becomes the first challenging task.

Firstly, there are 4 publicly available data sets: MUG-FED [31], CK+ [28] [29], Jaffe (Japanese Female Facial Expression) [32] and KDEF [33]. Figure 4.3 shows some sample images from these four data sets and Table 4.1 contains the statistics.

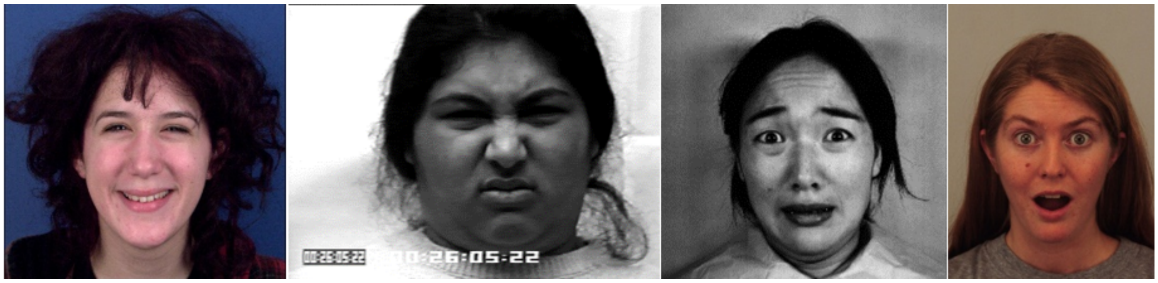


Fig. 4.3.: Sample images of MUG-FED, CK+, Jaffe and KDEF.

4.2.2 Data Preprocessing

Data Set Cleaning

Since most of the public data sets contain raw images, very few of them can be directly used without further examination. Therefore, these data sets need to be cleaned in the first place.

There are 52 subjects in the MUG-FED data Set. For each subject, it has 5-7 emotions and for each emotion, it has 3-7 attempts. And since all of the images are video frames, each emotion starts from neutral to the emotional expression of the strongest intensity and returns to neutral. Therefore, only the images that contain facial expressions of strong intensity should be selected. Table 4.2 shows the statistics of the MUG-FED data set after it is cleaned. It also provides us with 161 manually labeled images, which is used for validation. Besides these 4 data sets obtained from online sources, an additional 490 images were collected by us and are used to validate the model.

Eventually, there are 3 data sets for training and 3 data sets for validation, as shown in Table 4.3.

Data Labeling

After we finish cleaning up the data set, we need to do the data labeling so that we will have the ground truths of the images. Since the original file names of all the collected data sets are different, we need to rename all of the images with consistent names. We have seven emotions so that we have seven labels for all of the images: AN (angry), DI (disgust), FE (fear), HA (happy), NE (neutral), SA (sad) and SU (surprise). So we rename all the images with their corresponding emotions for example: img000.HA.jpg.

Face Alignment

Face alignment is another key step in data set pre-processing. The purpose is to remove potential uncertainties when applying our emotion recognition approach to real-time videos. For example, the position and the angle of the subjects head are changing as the video plays, which could affect the accuracy of the classification results if the face is not aligned in advance. With face alignment, the position of the head is aligned and the scale of the head is adjusted to have the same size, which eliminates the influence of any existing distortions on the recognition results.

We propose a novel face alignment algorithm that shows superior results compared to any existing method. Firstly, an MTCNN [34] face detector is used to detect the face in an image, then the LM detector [35] is used to detect 68 landmark points of the face. A rotation matrix is then obtained based only the eye center coordinates.

The traditional method uses the rotation matrix for face alignment. However, the resultant images can contain comparably useless background of large area and the eyes in different images are not at the same horizontal level, resulting in unsatisfying face classification results.

We improve the traditional face alignment by adding one more step. The aforementioned rotation matrix gets the coordinates of the 68 landmark points in the new coordinate system. Then we use the 1st, 9th and 17th landmark points to get the left, bottom and right boundary of the face. The definition and the location of these facial landmark points will be explained in Section 4.4.1. To get the top boundary, we stipulate that the length from the top boundary to the eye center is one-third of the height of the image. We use the boundary information to crop the original image into the one that contains smaller margins. Finally, the eyes of different images are adjusted to be at the same horizontal level. Figure 4.4 shows some sample images before and after face alignment. Note that all the images after face alignment are re-scaled to 128 × 128.

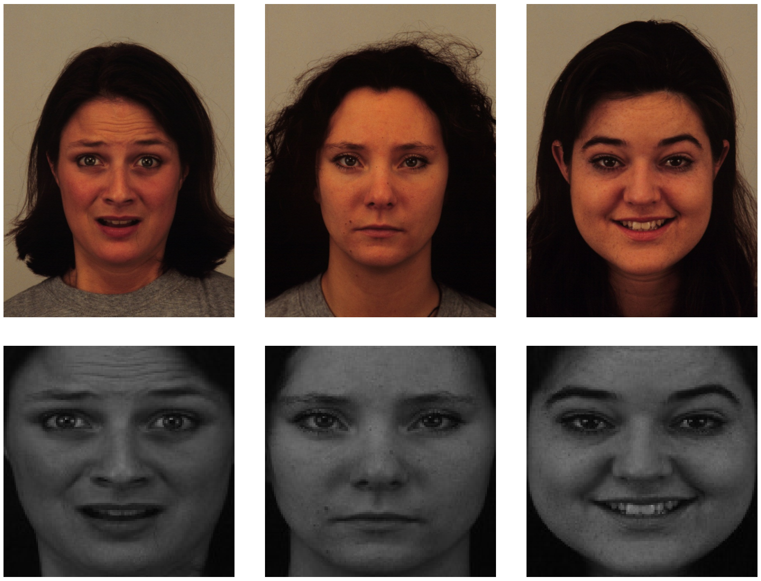


Fig. 4.4.: Comparison of images before and after face alignment.

Data Augmentation

To increase the robustness of our model and to prevent it from being over-fitted, we apply data augmentation on the training data set after face alignment. For each image, 7 images of different brightness and 28 images of different degree of blurring are created, resulting in a final training set with 1,148,812 images. All the training and validation images are then converted to LMDB (Lightning Memory-Mapped Database) format and be ready for training. Table 4.4 shows the statistics of the data augmentation.

4.2.3 Model Training

In order to apply our emotion recognition system to real-time video, the model needs to be comparably small in size and fast in speed. We have tested several pre-trained models, such as the VGG-S [36], on real-time video with multiple rounds of finetuning. The validation accuracies are less than 60% which is far below our requirements. Moreover, the size of the VGG-S model is more than 500 MB which is too large to be implemented efficiently.

To reduce the model size, we modify the original VGG-S [37] model by reducing the kernel size and channel number as shown in Figure 4.5 [35]. Compared to the original VGG-S model, our model has a size of only 12.1 MB. It takes only 4.5 hours to train on more than 1 million images for 50,000 iterations. Besides a smaller size of the model, the validation accuracy obtained by using the new model reaches 85%, which is significantly higher than our previous results. The reason of getting higher accuracy with a model of smaller size is that, if we want to train with the original large-size VGG-S model, it requires millions of raw images and weeks of time to train from scratch which is impossible. Which means, with our training set, smaller model will have higher accuracy. And the modified model can be trained faster and more efficiently with much smaller data set and much less time, in our case, 40k images and 4.5 hours.

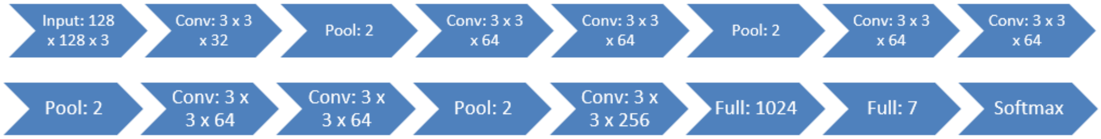


Fig. 4.5.: Framework of the classification model.

4.2.4 Experimental Results

As introduced in Section 4.2.3, the classification validation accuracy of the classification model is 85%. Figure 4.7(a) shows the validation confusion matrix for the validation data set.

In order to test our classification model, we collect our own data set which is called the HP Facial Expression Test Set, which contains 2443 images. The data set is collected with 5 subjects doing 7 emotions while being video recorded and the images were selected from the video frames. Our model achieves an accuracy of 82% and takes only 13.68 seconds to test on the whole testing data set (0.0056 s/image). This test was conducted on a workstation with an Nvidia Titan X GPU. Figure 4.6 shows some sample testing images and Figure 4.7(b) shows the testing confusion matrix.

Real-time Emotion Recognition

Currently, there are not many real-time emotion recognition frameworks, while the existing ones achieve only comparably low accuracy. However, our real-time demo version can detect peoples frontal facial expressions accurately. Figure 4.8 shows some sample results from our real-time emotion recognition demo.

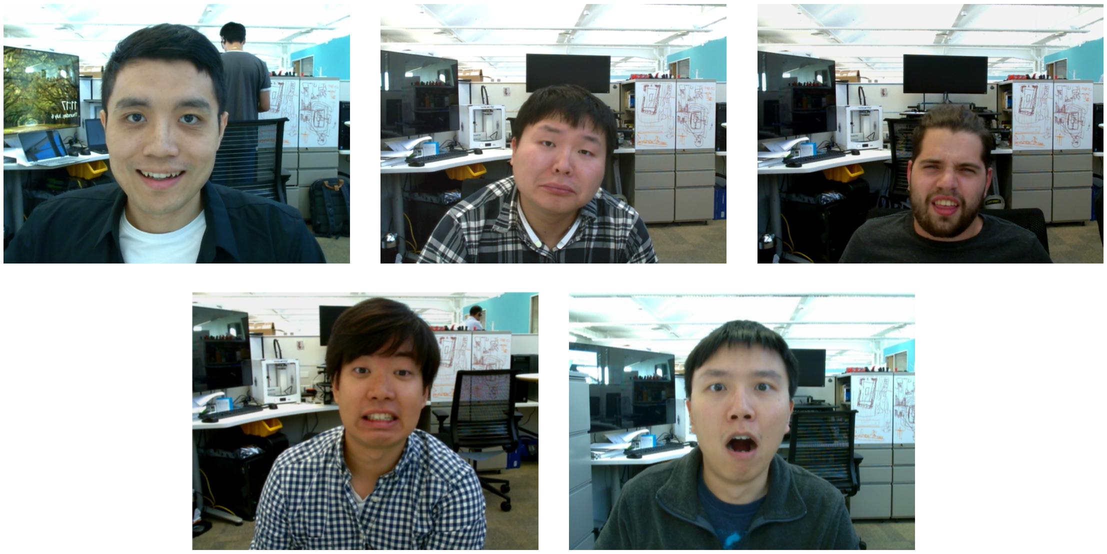
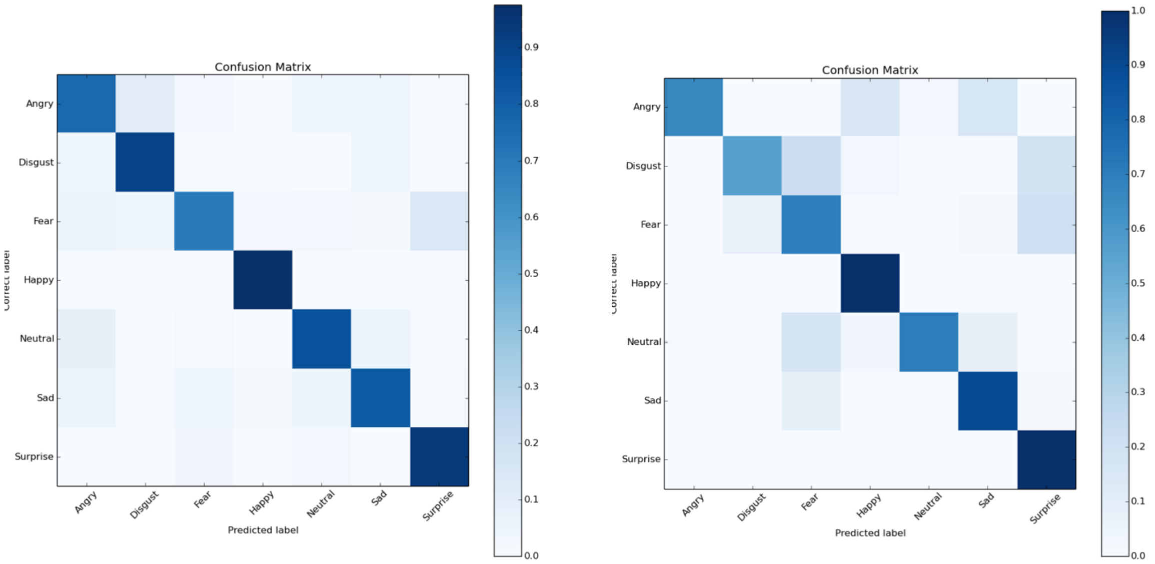


Fig. 4.6.: Sample images from HP Facial Expression Test Set.



(a) Confusion matrix for validation set.

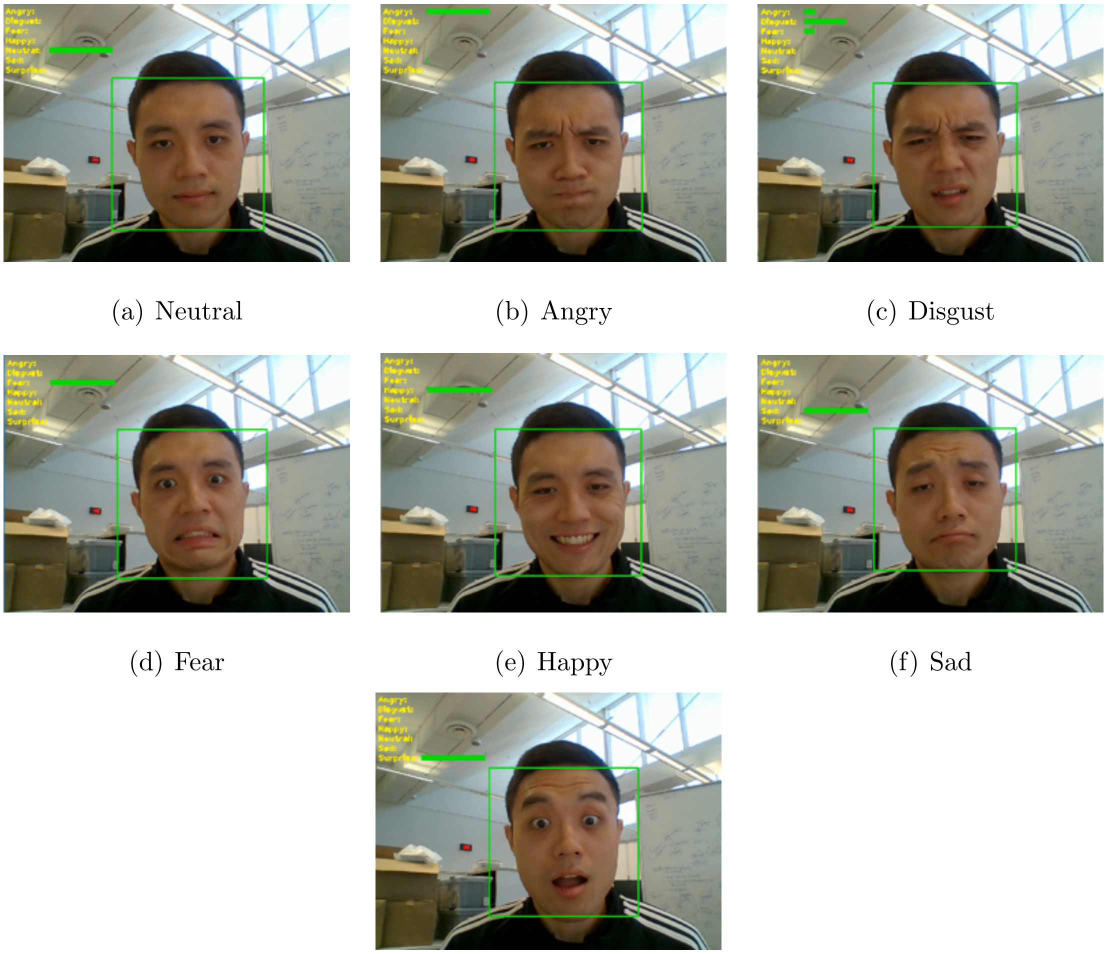
(b) Confusion matrix for self-collected HP data set.

Fig. 4.7.: Our emotion recognition classification experimental results.

4.3 Emotion Recognition Regression Training

4.3.1 Introduction

Although our emotion recognition classification model works well, it has its own drawback. And it is especially obvious when the classification model is applied in a real-time demo.



(a) Neutral

(b) Angry

(c) Disgust

(d) Fear

(e) Happy

(f) Sad

(g) Surprise

Fig. 4.8.: Sample result frames from real-time video demo for all seven emotions. The reader should zoom in to be able to see the labeling of the emotion and intensity provided in the upper left corner of each frame.

Since our classification training data set includes a lot of emotions that are not obvious or are of low intensity, this making them similar to one of the emotion categories in particular: neutral. This causes the prediction on the real-time video to be jittery, since in most cases, for example, a person does not need to express his or her happiness by a drop-jaw smile. And also, given an image or a frame of the video, our classification model can only tell if the facial expression is angry, disgust, fear, happy, neutral, sad or surprise; in other words, it is not able to tell the intensity of the emotion.

To solve these two problems, a regression model is used, where the ground truth labels become the intensities of the emotion, such as 20% happy and 80% neutral or 40% sad and 60% neutral. This additional information about the intensity of the emotion can be useful, especially in real-time videos.

4.3.2 Data Collection

Among the four data sets collected from the online sources, only the MUG-FED data set is used because of the large number of images the data set includes. However, the MUG-FED data set is more like an “in-the-lab” data set, where all of the emotions included are standardized and all the images have the same background and consist of a purely frontal face. Since there are very few public in-the-wild data sets, especially for the task of emotion recognition, we collected our own data set to train the model on a more “in-the-lab” data set.

Until now, we have collected more than 7,000 “in-the-wild” images containing facial expressions and we name this data set as Emotion Intensity in the Wild data Set. Table 4.5 shows the statistics of this data set. And Figure 4.9 shows the comparison between the ”in-the-lab” MUG-FED data set and our collected ”in-the-wild” HP data set. It is worth noting that this data set includes images containing heads at different angles, people with different races and ages, and backgrounds of different lighting conditions.

To train the regression model, we use both the MUG-FED and Emotion Intensity In the Wild data sets.

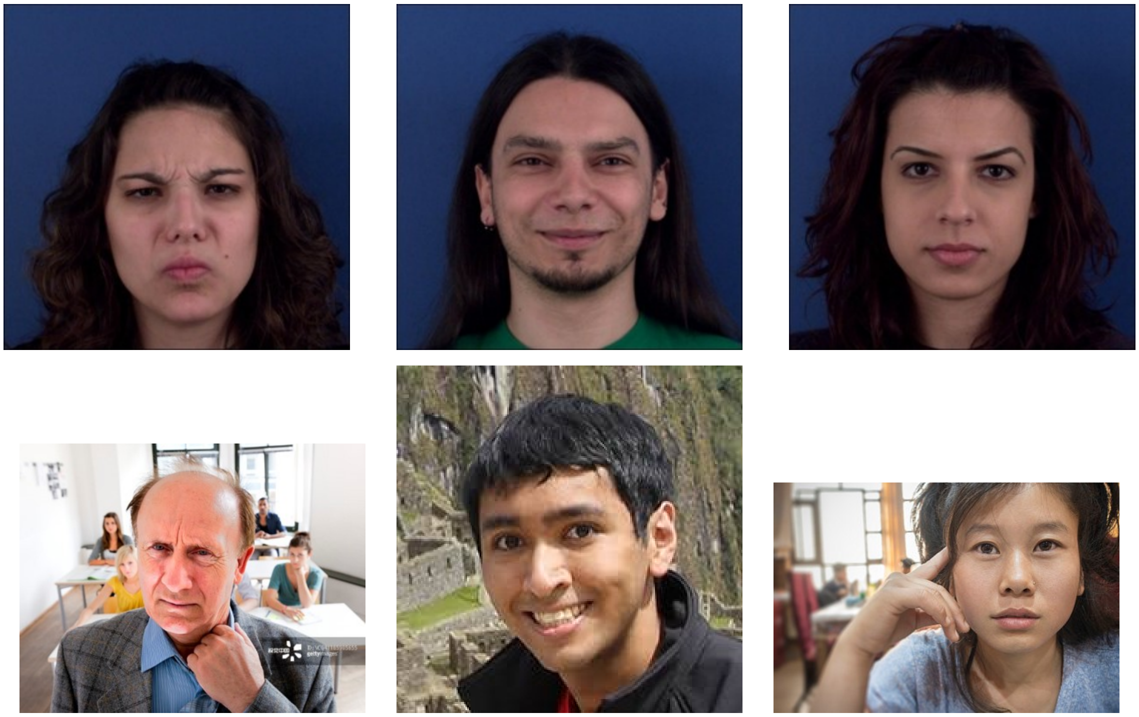


Fig. 4.9.: Comparison between MUG-FED images and HP In the Wild images.

4.3.3 Data Preprocessing

Data Set Cleaning

Before an existing data set obtained from online sources is used, it needs to be examined in the first place. As we introduced in the previous section, the images of the MUG-FED data set are consecutive frames obtained from videos. In a video for a specific emotion, the emotion starts from neutral to 100% facial expression and gradually returns to neutral. Therefore, for each attempt of expressing emotion, we select 9 images that contain facial expression intensities from 20% to 100% and back to 20% with an interval of 20%. An example is shown in Figure 4.10. After data set cleaning, we have collected 7,451 images for training and 981 images for validation for the MUG-FED data set. Each of these 7,451 images is labeled with the intensity of the emotion.

And for the Emotion Intensity In the Wild data set, after excluding some inappropriate images, we have 6,141 images for training and 682 images for validation.

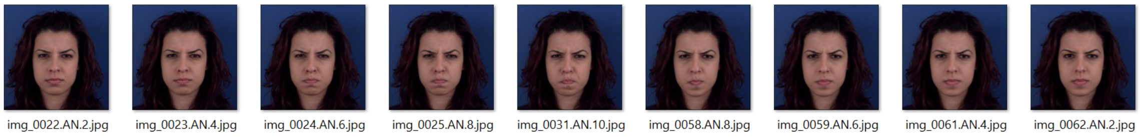


Fig. 4.10.: Sample regression images from MUG-FED data set. These images are from one of the attempts that a subject does which have the intensities from 20% (neutral) to 100%, and back to 20% in steps of 20%. Noting that the numbers: 2, 4, ... after the emotion label AN are the intensity labels, 2 corresponds to 20% etc.

Data Labeling

Similar to our classification data labeling, each image comes with an emotion label. But different from the previous approach, we now need the intensity label for each image. For example, if an image has name img000 with 20% Happy, the final label of the image is: img000.HA.02.jpg. Noting that 02 here means 20% intensity.

Face Alignment

The procedure of face alignment is the same as the one introduced in the previous section. We utilized the 1^{st}, 9^{th} and 17^{th} landmark points to get the boundary of the faces, cropped them, aligned them and rescaled them to 128×128. Figure 4.11 shows some sample face alignment results.

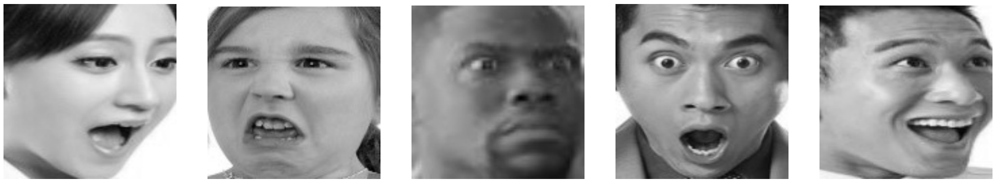


Fig. 4.11.: Sample HP Emotion Recognition In the Wild images after face alignment.

Data Augmentation

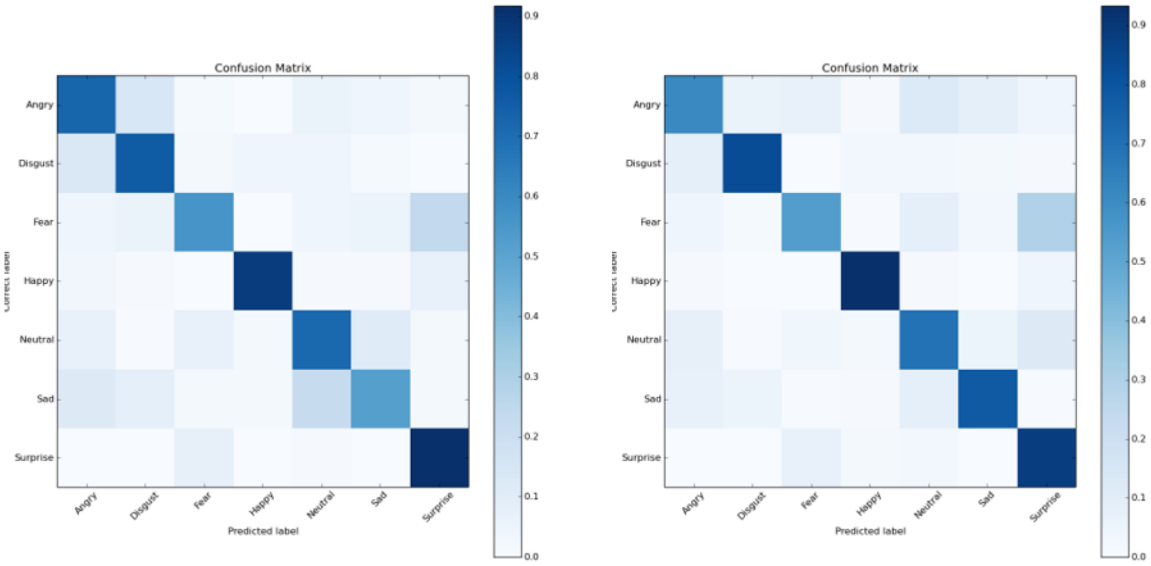
We experimented with two strategies of training, one with only the Emotion Intensity In The Wild data set, another with the combined data set of Emotion Intensity In The Wild data set and the MUG-FED data set. The method of data augmentation remains the same, which includes changing brightness and blurring the images and gives the final training data set, contains 6,141 images before data augmentation and 171,948 images after, and final validation data set, containing 682 images before data augmentation and 19,096 images after for Emotion Intensity In The Wild data set. For the combined data set, the final training data set, containing 13,592 images before data augmentation and 380,576 images after, and the final validation data set, containing 1,663 images before data augmentation and 46,564 images after, as shown in Table 4.6.

4.3.4 Model Training

The model framework used for the regression training is the same as for our classification model, except that the single label input layer is replaced with multi-label input layer and the softmax loss function is replaced with the sigmoid cross entropy loss function.

4.3.5 Experimental Results

The regression training also gets outstanding results. Figure 4.12(a) and Figure 4.12(b) shows the classification confusion matrices; and Table 4.7 shows the regression training results. Noting that, for the training and validation loss values, they are sigmoid cross entropy loss, the smaller the better and for the RMSE values, they represent the standard deviation of the prediction errors and are based on the datum that ranges from 0 to 1.



(a) Confusion matrix for HP In the Wild data set (b) Confusion matrix for the Combined data set

Fig. 4.12.: Our emotion recognition regression experimental results.

As indicated by the high accuracy which is around 77% and the small Root Mean Squared Error (RMSE) value which is below 0.13, our regression model performs well on the emotion recognition task.

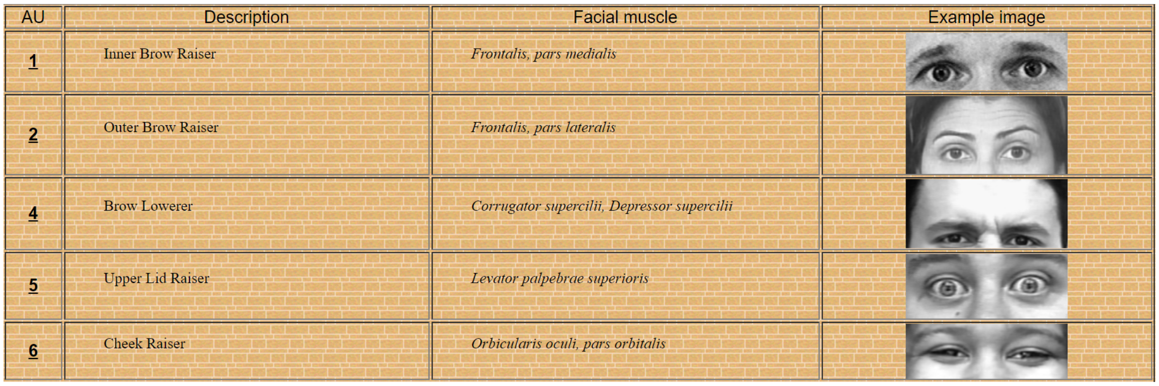
4.4 Facial Action Unit Detection

4.4.1 Introduction

We have shown promising results in our CNN based emotion recognition approach in the previous sections. In this section, we will demonstrate another method of solving emotion recognition problem through Facial Action Unit (FAU) detection and the relationship among Emotion (facial expression), FAU and LM (facial land mark).

Facial Action Coding System

Facial Action Coding System (FACS) is a complex system to describe human facial movements by visually discernible facial movement. It originated from the research work of a Swedish anatomist named Carl-Herman Hjortsj [38]. It was later adopted and popularized by Paul Ekman [39] [40]. Unlike facial expression ratings based on categorization of expressions into prototypical emotions (happiness, sadness, anger, fear, disgust, etc.), FACS can encode ambiguous and subtle expressions, and therefore is potentially more suitable for analyzing the small differences in facial affect [41]. In other words, FAU is a lower level feature compared with facial expression. Figure 4.13 [42] shows some sample FAUs and their descriptions.



AU	Description	Facial muscle	Example image
1	Inner Brow Raiser	*Frontalis, pars medialis*	
2	Outer Brow Raiser	*Frontalis, pars lateralis*	
4	Brow Lowerer	*Corrugator supercilii, Depressor supercilii*	
5	Upper Lid Raiser	*Levator palpebrae superioris*	
6	Cheek Raiser	*Orbicularis oculi, pars orbitalis*	

Fig. 4.13.: Sample Facial Action Units.

Facial Land Marks

Face landmarking, defined as the detection and localization of certain key points on the face, plays arguably the important role as an intermediary step for many subsequent face processing operations that ranges from biometric recognition to the understanding of mental states. [43] It also has an impact on subsequent task focused on the face, like face recognition, gaze detection, face tracking, animation, expression recognition etc. Facial landmarks are the feature points extracted during the face landmarking process. Commonly used LMs are the eyes corners, nose tip, mouth outline, eyebrow arcs, face boundary etc. It is a prominent feature that plays a discriminative role for the above-mentioned tasks or serves as anchor points on a face graph.

There are many different LM systems consisting of different numbers of LM points from 5-points system to 98-points system. For our case, we adopt the 68-points mark up shown in Figure 4.14.

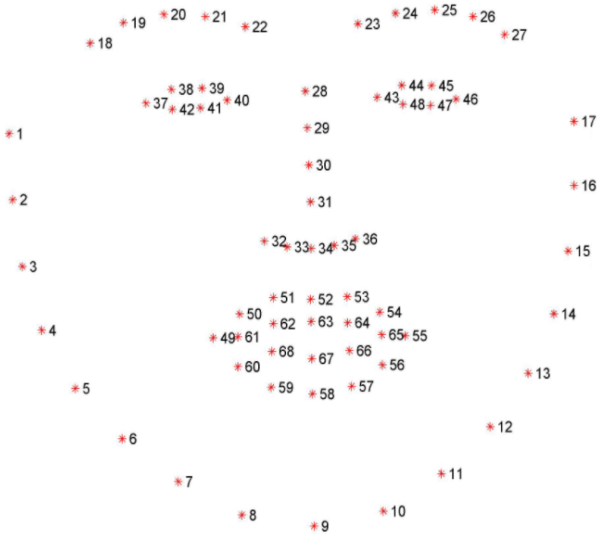
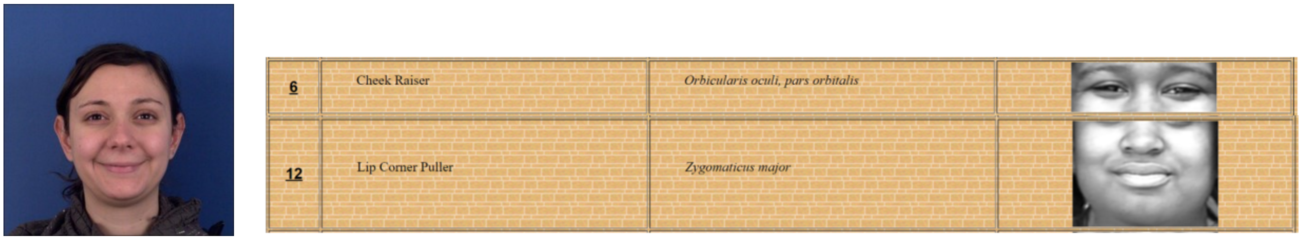


Fig. 4.14.: 68P Facial Land Marks System.

The relationships among Emotions, Facial Land Marks and Facial Action Units

Firstly, Emotions and Facial Action Units are related as shown in Table 4.8. Each emotion is a combination of certain FAUs. For example, Happy, it requires FAU 6^{th} and 12^{th} to be ON as shown in Figure 4.15.

Based on the observation of Figure 4.15, it can be concluded that FAU is related with the movement of the facial muscles such as eyebrows or mouth and they are marked by LMs. Accordingly, when people are making a facial expression which results in the movement the facial muscles, the LMs also have geometrical displacements. So if we are able to accurately track the movement of LMs, we can do FAU detection which leads to Emotion Recognition.



6	Cheek Raiser	*Orbicularis oculi, pars orbitalis*	
12	Lip Corner Puller	*Zygomaticus major*	

Fig. 4.15.: Happy consists of FAU 6 and 12.

There are a total of 13 FAUs related with emotions which are FAU 1, 2, 4, 5, 6, 7, 9, 12, 15, 16, 20, 23 and 26 as shown in Table 4.8.

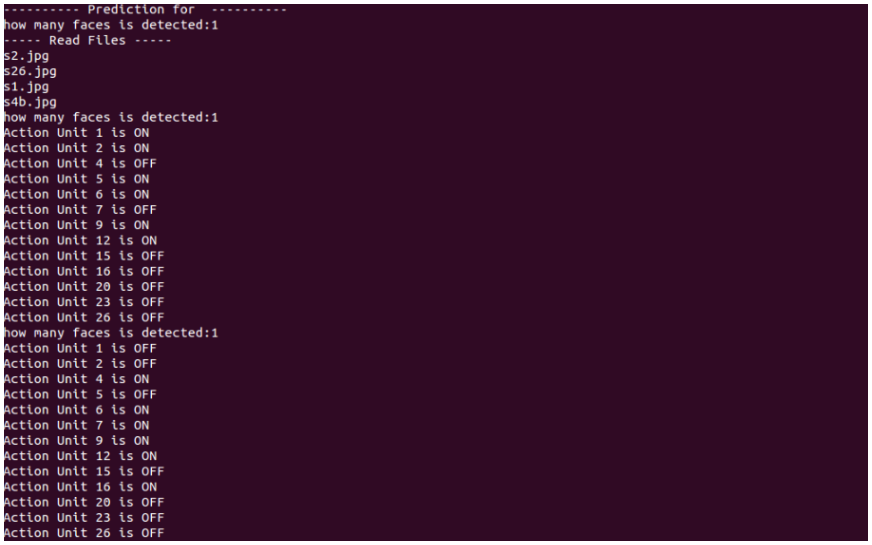
In order to unravel the relationship between FAUs and LMs, we have done careful observations as well as detailed experiments. Table 4.9 elaborates the explanation of their relationship.

4.4.2 Preliminary Results

According to the aforementioned relationships, we first develop a trial version of FAU detector based on the geometrical movements of the LM points. The system takes an image with Neutral facial expression as input, all FAUs are detected based on the position differences of the LMs against the Neutral frame. Figure 4.16 shows some sample demo results.

4.4.3 Improvements on LM Detection Strategy

The preliminary results are based on the absolute displacement of the LM points which are not accurate nor stable. Some contradictories will arise when we adopt the absolute displacement strategy. For example, for FAU 6 and 12, when they are



```
---------- Prediction for  ----------
how many faces is detected:1
----- Read Files -----
s2.jpg
s26.jpg
s1.jpg
s4b.jpg
how many faces is detected:1
Action Unit 1 is ON
Action Unit 2 is ON
Action Unit 4 is OFF
Action Unit 5 is ON
Action Unit 6 is ON
Action Unit 7 is OFF
Action Unit 9 is ON
Action Unit 12 is ON
Action Unit 15 is OFF
Action Unit 16 is OFF
Action Unit 20 is OFF
Action Unit 23 is OFF
Action Unit 26 is OFF
how many faces is detected:1
Action Unit 1 is OFF
Action Unit 2 is OFF
Action Unit 4 is ON
Action Unit 5 is OFF
Action Unit 6 is ON
Action Unit 7 is ON
Action Unit 9 is ON
Action Unit 12 is ON
Action Unit 15 is OFF
Action Unit 16 is ON
Action Unit 20 is OFF
Action Unit 23 is OFF
Action Unit 26 is OFF
```

Fig. 4.16.: Sample FAUs detection demo.

”ON”, they both result in LM 49, 61, 65, 55 to move upwards. But if only FAU 6 is activated, we cannot guarantee that FAU 12 is also activated. Accordingly, we have to exclude some conditions in FAU 6 in order to detect FAU 12 independently.

A more robust and accurate strategy of detecting the FAUs is needed. After some research and observation, instead of measuring the absolute displacement of the LMs, the new approach measures the relative displacement between the LM points. Instead of simply comparing the two LM positions, we now adopt two thresholds for each LM's displacement change. They are both based on the ratio of the distance between the corresponding LM points before and after the activation of the FAU (D_{LM}) and the horizontal distance between LM 37 and 46 (D_{eye_outer}) as shown in Equation 4.1. Each threshold is carefully tested and selected to ensure both accurate classification and regression accuracy.

$$T = D_{LM}/D_{eye\_outer}. \tag{4.1}$$

For a certain FAU, when the intensity is below the lower bound threshold T_L, that FAU is “OFF”. When the intensity of that FAU is above the upper bound threshold

T_U, it is “ON” with 100% intensity. If the intensity of the FAU is between the lower and upper bound threshold, if follows the uniform distribution.

In addition, most of the time when people are doing facial expressions, the FAUs associated with their left and right eyes, left and right eyebrows, left and right mouth corner may have different intensities. In order to describe this phenomenon, the intensities of the left and right sub-FAUs are calculated individually and the intensity of the FAU is the average of the two sub-FAU intensities.

Detailed description of how the activation of each FAU is defined in Table 4.10.

4.4.4 SVM Training Based FAU Detection For Emotion Recognition

In the previous sections, we demonstrate the CNN based emotion recognition approach. In order to extend our algorithm to much faster yet comparably accurate application, we adopt another method of using SVM to solve the emotion recognition problem. Based on our analysis in Section 4.4.3, each FAU has an intensity which corresponds to a feature in the SVM framework. So totally, we have 13 feature numbers for each emotion class which will be fed into our SVM.

Training and Validation Data set

Because of the nature of SVM, we do not need as many training and validation images as the CNN based approach. And for each subject, a Neutral emotion image is selected as the anchor. Based on this anchor image, each training and validation image will generate a 13-dimensional feature vector and all these feature vectors will be written into .txt files for training and validation. An example of how the images are preprocessed is shown in Figure 4.17. We first do LM detection on all of the images in the data set. Then, image preprocessing is done using the same algorithm described in Section 4.2.2. For all of the LM points which has already been detected, the same geometrical transformation in the data preprocessing section is implemented on all of the LMs so that in the new coordinate system, they will appear at the same location on the face of the subject.

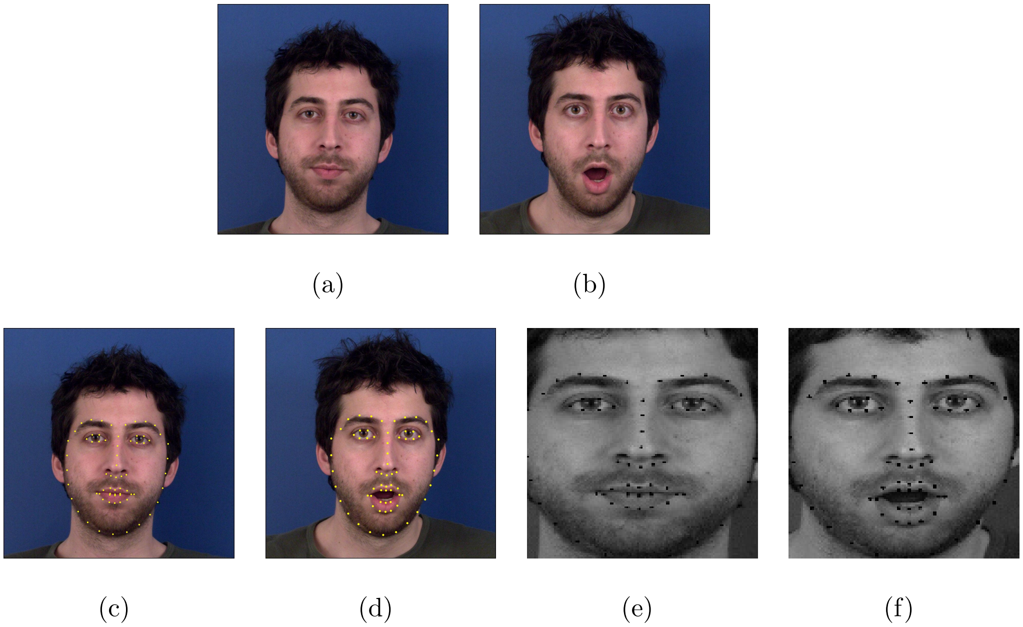


Fig. 4.17.: (a), (b): A Neutral sample image and an sample image with emotion; (c), (d): Both images with marked LMs; (e), (f): Both images preprocessed and aligned with marked LMs.

Detailed statistics of our training and validation data sets can be found in Table 4.11. They are subsets of our previous CNN classification data sets. The images are also from the MUG-FED [31], CK+ [28] [29] and Jaffe [32] databases.

SVM Training Kernels and Parameters

Since our task is a multi-dimensional problem, Radial Basis Function (RBF) Kernel has the best performance classifying our 13-dimensional feature vectors. So RBF kernel is selected over Linear Kernel and Polynomial Kernel.

Except for the selection of the kernel type, two parameters are also related with SVM training which are Gamma and C_{SVM} also known as the Box Parameter.

The Gamma Parameter determines how profoundly a single data sample exerts influence. That is to say, the gamma parameter can be said to adjust the curvature of the decision boundary. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. More intuitively, if gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself which means that all of the decision boundary will be surrounding each support vector and the final result will be extremely biased. When gamma is very small, the model is too constrained and cannot capture the complexity or shape of the data. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model.

C_{SVM} is the cost of misclassification which means that how much a single misclassification penalize the overall performance of our SVM. A large C_{SVM} gives you low bias and high variance and a small C_{SVM} gives you higher bias and lower variance. More intuitively, for larger values of C_{SVM}, the decision boundary will perform better at classifying all training points correctly which will result in overfitting. A lower C_{SVM} will encourage a larger margin, therefore a simpler decision function, at the cost of training accuracy. In other words, C_{SVM} behaves as a regularization parameter in the SVM.

So picking the appropriate Gamma and C_{SVM} values is critical to our SVM training and validation accuracy. For our case, both parameters have 3 candidate values which are $[10^{-1}, 10^{0}, 10^{1}]$ and it leads to 9 trials.

Validation and Testing Results

Totally, 9 trials have been conducted in order to optimize the Gamma and C_{SVM} values. The validation accuracies are shown in Table 4.12. The validation accuracy reaches the maximum value when Gamma is 10^{-1} and C_{SVM} is 10^{0}. It can be concluded that based on our testing, the optimal values for Gamma and C_{SVM} are 10^{-1}, 10^{0}. The final validation results with 87.7% accuracy is shown in Figure 4.18.

In the last phase of this project, we use the same HP Facial Expression Test Set previously used in Section 4.2.4 for the CNN approach to test our SVM model. With the same Gamma and C_{SVM} values we have validated in the previous section, 85% accuracy is achieved.

Compared to our previous CNN approach, the validation accuracy is increased by 2.7% from 85% to 87.7% and the testing accuracy is increased by 3% from 82% to 85%. In addition, because of the light weight nature of SVM, compared against our CNN approach, the model size is reduced from 12 MB to 100 KB.

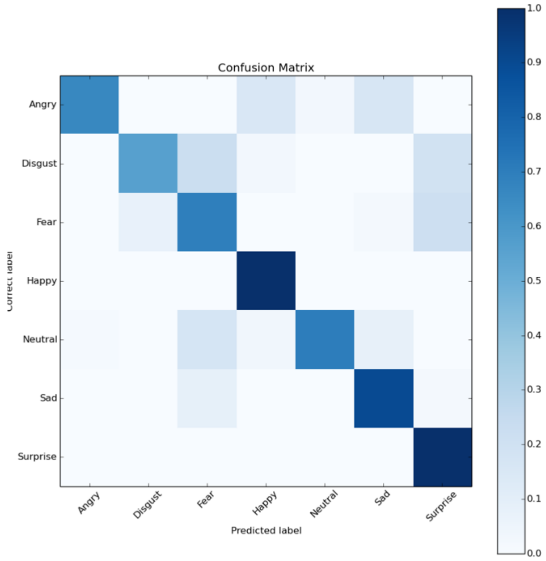


Fig. 4.18.: Confusion matrix for SVM validation result.

4.5 Limitations and Future Work

The limitations of our proposed emotion recognition method come from two aspects: the shallow network structure and the lack of temporal information. Extensive research work has shown that a deeper network structure tends to result in better training performance. Even though for our case, a real-time application is required so that the network may not be too complicated. Still, some most recent network structures, which have more complex network architecture yet faster speed can be considered. Also, for the real-time application, since our proposed method is single-frame-based without utilizing the temporal information, the results jitter among different emotion classes. So the future work may include: (1) Adopting a deeper yet faster network structure. (2) Considering temporal information, e.g. the information from adjacent frames to overcome the result fluctuation.

4.6 Conclusion

In this chapter, we first demonstrate our effort in CNN based Emotion Recognition task. This end-to-end problem includes data collection, data preprocessing, model training and testing. In order to achieve our goal of being able to do real-time facial expression recognition, VGG-S, a light version of VGG is selected to ensure the light weight as well as high accuracy. This model structure greatly reduce the model size from 500 MB to 12 MB and at the same time, achieve a high classification accuracy of 85%. The small model size also makes the real-time application possible. With the frame rate of over 20 FPS using an Intel i5 CPU, our application is able to do accurate real-time facial expression recognition on most of the devices. Later, we extend our algorithm from classification scenario to regression scenario in order to obtain the intensity of each emotion. With the HP Emotion Recognition In the Wild data set we have collected, we are able to achieve as low as 0.12 RMSE for our regression model. In the final stage of this project, we approach the Emotion Recognition task using the FAU detection. Totally, there are 13 FAUs related with the 7 Emotions we want to detect. With our pretrained facial land mark detector, we are able to get very accurate LM coordinates. And together with the relationships we discovered between the FAUs and the LMs, the 13 FAUs can be detected by the relative displacement of LMs with the help of a Neutral frame. We then transform the 13 FAUs into feature vectors and feed them into the SVM framework to train the classification model.

Finally, we achieved the validation accuracy of 87.7% and testing accuracy of 85% along with much smaller model size.

Table 4.1.: Statistics for four collected data sets.

Database	Facial Expression	# of Subjects	# of Images	Gray / Color	Size	Ground Truth	Type
Extended Cohn-Kanade Data set (CK+)	Angry, Contempt, Disgust, Fear, Happy, Neutral, Sad and Surprise	123	593 image sequences (327 sequences having discrete emotion labels)	Mostly gray	640 × 490	Facial expression labels and FACS labels	Posed; spontaneous smiles
Japanese Female Facial Expression (Jaffe)	Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise	10	213 static images	Gray	256 × 256	Facial expression labels	Posed
Multimedia Understanding Group (MUG)	Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise	86	1,462 sequences with more than 100K images	Color	896 × 896	Facial expression and land mark (LM) labels	Posed
The Karolinska Directed Emotional Faces (KDEF)	Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise	70	4,900 images	Color	562 × 762	Facial expression labels	Posed

Table 4.2.: Statistics for MUG-FED after data set cleaning up.

Database	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
MUG-FED	6,220	4,856	4,605	9,329	3,719	5,562	5,623	39,914

Table 4.3.: Training and Validation data sets for classification training.

Data Set	Training	Validation
Name	MUG-FED, CK+ and Jaffe	MUG-FED (Manually Labeled by author), KDEF and Images of myself
Number	41,029	1,867

Table 4.4.: Data augmentation.

Brightness Change	Blur (Gaussian, Average, Median)	Total Multiples	Number of Images
7 ×	4 ×	28 ×	1,148,812

Table 4.5.: Statistics for HP Emotion Recognition In the Wild Data Set.

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total (Without Neutral)
20%	194	147	93	236	1093	210	91	971
40%	323	218	103	230		164	218	1,256
60%	221	243	134	320		137	279	1,334
80%	207	198	151	295		81	316	1,248
100%	227	178	157	286		84	420	1,352
Total	1,172	984	638	1,367	1,093	676	1,324	7,254 (With Neutral)

Table 4.6.: Data set statistics for regression training.

Data set	Training	Validation
Emotion Intensity In the Wild (After data augmentation)	6141 (171948)	682 (19096)
Combined data set (After data augmentation)	13592 (380576)	1663 (46564)

Table 4.7.: Regression training results.

Data Set	Training Loss	Validation Loss	Regression RMSE	Classification Accuracy
HP In the Wild	0.13	0.3	0.129	77.2%
Combined Data set	0.122	0.239	0.123	76%

Table 4.8.: Relationship between emotion and FAUs.

Emotion	Action Units
Angry	4 + 5 + 7 + 23
Disgust	9 + 15 + 16
Fear	1 + 2 + 4 + 5 + 7 + 20 + 26
Happy	6 + 12
Sad	1 + 4 + 15
Surprise	1 + 2 + 5 + 26

Table 4.9.: Relationship between FAUs and LMs.

FAU	Movements of LMs
1	LM 21, 22, 23 and 24 moving upwards
2	LM 19, 20, 25 and 26 moving upwards
4	LM 21, 22 moving rightwards, LM 23, 24 moving leftwards
5	LM 38, 39, 44 and 45 moving upwards
6	LM 41, 42, 47, 48, 49 and 55 moving upwards
7	LM 38, 39, 44 and 45 moving downwards, LM 41, 42, 47 and 48 moving upwards
9	LM 32 and 36 moving upwards
12	LM 49, 55, 61 and 65 moving upwards
15	LM 49, 55, 61 and 65 moving downwards
16	LM 56, 57, 58, 59, 60, 66, 67 and 68 moving downwards
20	LM 49, 60 and 61 moving downwards and leftwards, LM 55, 56 and 65 moving downwards and rightwards
23	LM 50, 51, 52, 53 and 54 moving downwards, LM 56, 57, 58, 59 and 60 moving upwards
26	LM 56, 57, 58, 59, 60, 66, 67 and 68 moving downwards, LM 49 and 61 moving rightwards, LM 55 and 65 moving leftwards

Table 4.10.: Revised relationship between FAUs and LMs. Noting that for each FAU, T_L and T_U are different and carefully tested. A stands for "Activated" and N stands for "Neutral". $|X|$ is the absolute value of X and $\|X\|$ is the norm of X.

FAU	Movements of LMs
1	$T_L <= \|y_{22} - y_{18}\|_A - \|y_{22} - y_{18}\|_N <= T_U$, $T_L <= \|y_{23} - y_{27}\|_A - \|y_{23} - y_{27}\|_N <= T_U$
2	$T_L <= y_{19A} - y_{19N} <= T_U$, $T_L <= y_{20A} - y_{20N} <= T_U$, $T_L <= y_{25A} - y_{25N} <= T_U$, $T_L <= y_{26A} - y_{26N} <= T_U$
4	$T_L <= \big\|(x,y)_{22} - (x,y)_{28}\big\|_N - \big\|(x,y)_{22} - (x,y)_{28}\big\|_A <= T_U$, $T_L <= \big\|(x,y)_{23} - (x,y)_{28}\big\|_N - \big\|(x,y)_{23} - (x,y)_{28}\big\|_A <= T_U$
5	$T_L <= \|y_{38} - y_{42}\|_A - \|y_{38} - y_{42}\|_N <= T_U$, $T_L <= \|y_{39} - y_{41}\|_A - \|y_{39} - y_{41}\|_N <= T_U$, $T_L <= \|y_{44} - y_{48}\|_A - \|y_{44} - y_{48}\|_N <= T_U$, $T_L <= \|y_{35} - y_{47}\|_A - \|y_{45} - y_{47}\|_N <= T_U$
6	$T_L <= \|y_{38} - y_{42}\|_N - \|y_{38} - y_{42}\|_A <= T_U$, $T_L <= \|y_{39} - y_{41}\|_N - \|y_{39} - y_{41}\|_A <= T_U$, $T_L <= \|y_{44} - y_{48}\|_N - \|y_{44} - y_{48}\|_A <= T_U$, $T_L <= \|y_{45} - y_{47}\|_N - \|y_{45} - y_{47}\|_A <= T_U$, $T_L <= \|y_{51} - y_{49}\|_N - \|y_{51} - y_{49}\|_A <= T_U$, $T_L <= \|y_{53} - y_{55}\|_N - \|y_{53} - y_{55}\|_A <= T_U$
7	$T_L <= \|y_{38} - y_{42}\|_N - \|y_{38} - y_{42}\|_A <= T_U$, $T_L <= \|y_{39} - y_{41}\|_N - \|y_{39} - y_{41}\|_A <= T_U$, $T_L <= \|y_{44} - y_{48}\|_N - \|y_{44} - y_{48}\|_A <= T_U$, $T_L <= \|y_{45} - y_{47}\|_N - \|y_{45} - y_{47}\|_A <= T_U$
9	$T_L <= \big\|(x,y)_{22} - (x,y)_{28}\big\|_N - \big\|(x,y)_{22} - (x,y)_{28}\big\|_A <= T_U$, $T_L <= \big\|(x,y)_{23} - (x,y)_{28}\big\|_N - \big\|(x,y)_{23} - (x,y)_{28}\big\|_A <= T_U$, $T_L <= y_{49A} - y_{49N} <= T_U$, $T_L <= y_{51A} - y_{51N} <= T_U$, $T_L <= y_{53A} - y_{53N} <= T_U$, $T_L <= y_{55A} - y_{55N} <= T_U$
12	$T_L <= \|y_{51} - y_{49}\|_N - \|y_{51} - y_{49}\|_A <= T_U$, $T_L <= \|y_{53} - y_{55}\|_N - \|y_{53} - y_{55}\|_A <= T_U$
15	$T_L <= \|y_{51} - y_{49}\|_A - \|y_{51} - y_{49}\|_N <= T_U$, $T_L <= \|y_{53} - y_{55}\|_A - \|y_{53} - y_{55}\|_N <= T_U$
16	$T_L <= \|y_{62} - y_{68}\|_A - \|y_{62} - y_{68}\|_N <= T_U$, $T_L <= \|y_{63} - y_{67}\|_A - \|y_{63} - y_{67}\|_N <= T_U$, $T_L <= \|y_{64} - y_{66}\|_A - \|y_{64} - y_{66}\|_N <= T_U$, $T_L <= y_{9N} - y_{9A} <= T_U$
20	$T_L <= \|x_{65} - x_{61}\|_A - \|x_{65} - x_{61}\|_N <= T_U$, $T_L <= \|x_{55} - x_{49}\|_A - \|x_{55} - x_{49}\|_N <= T_U$
23	$T_L <= \|y_{51} - y_{62}\|_N - \|y_{51} - y_{62}\|_A <= T_U$, $T_L <= \|y_{53} - y_{64}\|_N - \|y_{53} - y_{64}\|_A <= T_U$
26	$T_L <= \|y_{62} - y_{68}\|_A - \|y_{62} - y_{68}\|_N <= T_U$, $T_L <= \|y_{63} - y_{67}\|_A - \|y_{63} - y_{67}\|_N <= T_U$, $T_L <= \|y_{64} - y_{66}\|_A - \|y_{64} - y_{66}\|_N <= T_U$, $T_L <= y_{9N} - y_{9A} <= T_U$

Table 4.11.: Data set statistics for SVM training and validation.

	Training	Validation
Number of Images	2568	350

Table 4.12.: SVM Validation Accuracy and Parameter Tuning. The highest accuracy is highlighted

Accuracy (%)		Gamma		
		10^{-1}	10^{0}	10^{1}
C_{SVM}	10^{-1}	76	82.0	72.9
	10^{0}	87.7	83.7	71.2
	10^{1}	81.7	78.3	73.0

5. SALIENCY BASED IMAGE QUALITY ASSESSMENT AND CROPPING

5.1 Introduction

5.1.1 General Introduction of the Project

Image Quality and Aesthetics Assessment has become a hot research topic in the past decade due to its usefulness in a wide variety of applications such as image capture pipelines (e.g. traditional cameras and mobile phone cameras), photobook generation [44], image thumbnailing [45], storage techniques, online video streaming and sharing media. It is one of the most critical features since it directly determines the user experience. In fact, Image Quality and Aesthetics Assessment consists of two separate yet related topics: Image Quality Assessment and Image Aesthetics Assessment. The aesthetics of an image is considered from an artistic perspective such as composition and colorfulness, while the quality of an image is considered from a technical perspective such as noise and distortion. In this chapter, we propose an alternative approach of analysing image quality and image aesthetics based on image saliency information. More specifically, for image quality, we focus on image noise and for image aesthetics, we focus on image recomposition.

5.1.2 Related Work

A lot of research work has been done on these two topics. For image quality assessment, people start with Full-reference Image Quality Assessment (FRIQA) approach. A reference page with no artifacts nor noise is needed. All testing images are required to be compared with the reference page either manually or automatically on the basis of FRIQA metrics such as PSNR, SSIM [46], etc. FRIQA requires the existence of the reference image as well as substantial human labor which is later replaced by No-reference Image Quality Assessment (NRIQA) approach. NRIQA analyses an image based on both local and global features in the form of a statistical model of distortions in an automatic manner with no need for a reference image. Traditionally, machine learning has shown some promising results in NRIQA. Tang et al. [47] propose a pipeline which first classify all photos into seven categories such as Animal, Human, Landscape, etc. Then, they evaluate the quality of an image based on their proposed regional and global features. Xue et al. [48] approach NRIQA task by partitioning the distorted images into overlapped patches, and use a percentile pooling strategy to estimate the local quality of each patch. Then, a quality-aware clustering (QAC) method is proposed to learn a set of centroids on each quality level. These centroids are then used as a codebook to infer the quality of each patch in a given image, and subsequently a perceptual quality score of the whole image can be obtained. More recently, CNN based NRIQA algorithms have shown predominant advantage and much better results. Kang et al. [49] demonstrates that extracting high level features using CNNs can lead to state-of-the-art blind quality assessment results. They show that by replacing hand-crafted features, the end-to-end feature learning system from CNN framework has become a much easier and reliable approach of solving NRIQA problems.

For image cropping and recomposition problems, earlier methods use heuristic metrics such as Rule of Thirds, Diagonal Dominance, Visual Balance, etc. to supervise the cropping process [50] [51]. With the superiority of CNN based algorithms in Computer Vision domain, recent methods such as [52] and [53] propose to train the neural networks in a pair-wise ranking manner for which the data sets they have collected contain ranked image pairs. Most recently, Zeng et al. [54] propose a more advanced algorithm of utilizing the Region of Discard (RoD). RoD is the opposite of the Region of Interest (RoI) which is the discarded information of an image. The proposed method concatenate the RoD with RoI during training and get state-of-the-art cropping results.

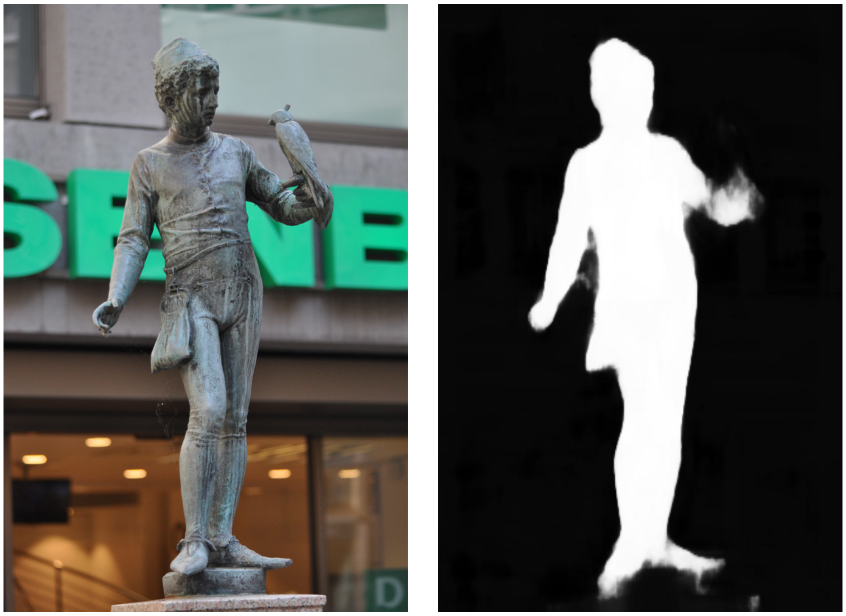
5.1.3 Our Approach

Even though some recent papers achieve good results regarding to some numerical metrics such as PSNR, SSIM for image quality assessment task and IoU, Disp. for image cropping task, they fail to get satisfactory visualized results because they analyze the image as a whole.

Figure 5.1(a) gives us an example for the image quality assessment scenario. Most of the portraits, group photos and close range images have shallow depth of field (DOF) in order to achieve pleasing aesthetic effect, thus creating a blurred background. This statue picture has very high score for both quality and aesthetics when being judged by a human photographer. But when we evaluate it using some current IQA pipelines, the actual quality scores are very low and they all classify this image as a noisy image with blur artifact. The reason is that even though the statue itself is sharp and in-focus, the image contains a large area of blurred background. So when it is being evaluated as a whole, it will be classified as a blurred image with low quality score.

For the image cropping scenario, a bad crop may have high IoU which is undesirable as shown in Figure 5.2. Even though for this image, most of the skier's body is included in the final crop and the IoU is high enough to become a good result, it is not since the man's feet and the snowboard is excluded from the final crop. This example implies that even a slight adjustment or displacement of the view can profoundly influence the composition quality.

Both examples demonstrate the fact that sometimes analysing the quality and aesthetics of an image as whole may not lead to optimal results. Instead, understanding the content of the image as the first step may help improve the final result. In this chapter, we propose both image quality assessment and image cropping algorithms based on saliency detection as shown in 5.1(b). For the statue picture, if we can get accurate saliency map shown in Figure 5.1(b) and do IQA only on the salient area, the image will be classified as a good quality image which meets the anticipation. And



(a) Original image

(b) Saliency map

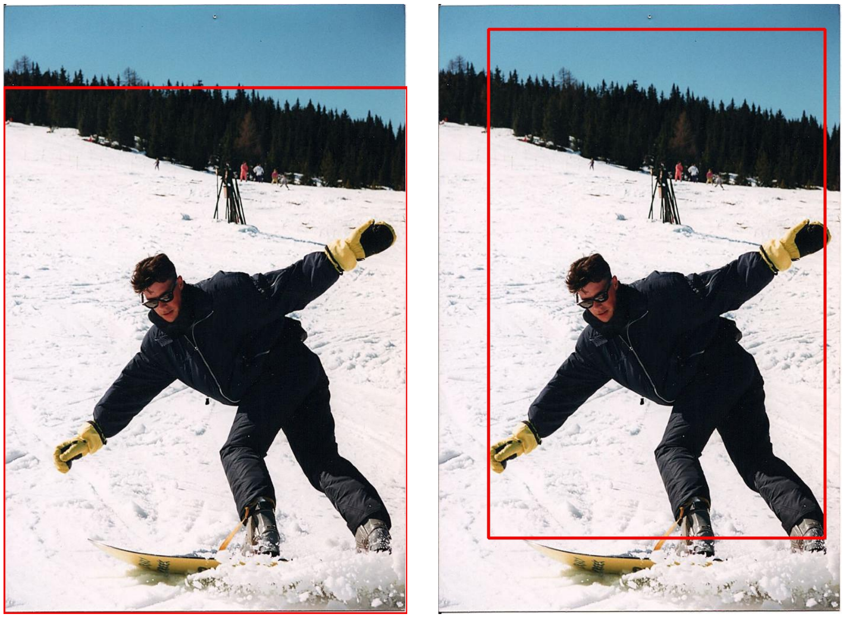
Fig. 5.1.: Shallow DOF image with low measured quality score but high ground truth quality score.

also for the picture of the skier, if we do the cropping based on the bounding box that covers all of the salient region, which in this case, the skier and the snowboard, the final cropping result will be closer to the groundtruth with better composition.

5.2 Saliency Based Image Quality Assessment

5.2.1 Image Saliency Detection

Saliency is the remarkable region of an image that capture people's attention easily, e.g. the statue and the skier in Figure 5.1 and Figure 5.2. Generally speaking, photographs containing visually dominant subjects induce stronger aesthetic interests. Therefore, photographers deliberately avoid an uniform sharpness and illumination in the whole image to achieve higher image aesthetics [55]. Also, in most cases, the saliency of an image represents the true object, that the photographer wants people to notice at the first sight.



(a) Image cropping groundtruth (b) A bad view with high IoU

Fig. 5.2.: An unsatisfactory cropping with high IoU.

The goal of the first step is to come up with a computationally inexpensive saliency detection algorithm, which is able to detect the saliency of an image accurately, so that the points of interest can be located and used for further assessment. Traditionally, previous methods detect saliency by extracting global and local information in both time and frequency domain using self-defined features and mathematical algorithms [56]. But most of the traditional methods can only capture the contours of high-contrast objects along with a lot of noise. More recently, end-to-end CNN training based methods show much stronger ability in saliency map generation task.

Since for both of our problems, Saliency Detection Algorithm is not our research focus, it acts as a tool which serves our further assessment. So a state-of-the-art algorithm that has the highest accuracy and robustness is desirable. In the next section, we will validate our selection by comparing the results from several saliency detection algorithms.

5.2.2 Saliency Detection Algorithm Comparison

Totally, we have selected three candidates. One of them is the most easily implemented OpenCV saliency detection algorithm which is based on [57]. This traditional, mathematical method extracts the spectral residual of an image in spectral domain by analyzing the log-spectrum of an input image, and proposes a fast method to construct the corresponding saliency map in spatial domain. The other two are the most recent state-of-the-art saliency detection methods: PDNet [58] and DSS [59].

In the PDNet paper, Zhu et al. [58] propose a novel prior-model guided depth-enhanced network (PDNet). The PDNet is composed of a master network and sub-network. The master network is a convolution-deconvolution pipeline. The convolution stage serves as a feature extractor that transforms the input image into hierarchical rich feature representation, while the deconvolution stage serves as a shape restorer to recover the resolution and segment the salient object in fine detail from background. The sub-network can be treated like an encoder convolution architecture and it process depth map as input and enhance the robustness of the master network. To address the problem of insufficient RGB-D data for training, they also employ a large dataset to pre-train the master network. This pre-train setup before training our network using RGB-D data has proved to contribute dramatically to accuracy improvement.

In the DSS paper, Hou et al. [59] propose a new salient object detection method by introducing short connections to the skip-layer structures within the HED architecture. Holistically-Nested Edge Detector (HED) provides a skip-layer structure with deep supervision for edge and boundary detection. Their framework takes full advantage of multi-level and multi-scale features extracted from FCNs, providing more advanced representations at each layer, a property that is critically needed to perform segment detection.

Both PDNet and DSS claim state-of-the-art results and they both compare their results with different methods using different data sets which makes it difficult to conclude which one is better directly from their papers. In order to draw a conclusion, we conducted a trial with both algorithms tested on a self-collected data set containing 300 images. We evaluated them by the visualized quality of the generated saliency map. Figure 5.3 provides a visual comparison of the above-mentioned methods. Based on the results of all 300 images, DSS has the highest accuracy as well as the best visualized quality. So for both topics, we use DSS as our saliency map generation algorithm. For more detailed interpretation of how DSS works, please refer to [59].

5.2.3 Saliency Based Image Quality Assessment

Methodology

Inspired by [60], as shown in Figure 5.4, the source image is first split into image patches (IPs) with size 64×64 to train the CNN model. Each IP has the same quality label as the source image. The saliency map is also split into salient image patches (SIPs) same as the source image. The SIP is used to determine whether a certain IP is a valid salient image patch (VSIP) for the model to predict on. The pixel values of the saliency map generated by the DSS [59] method range from 0 to 255, and, the higher saliency value of the pixel, the more salient it is. So we define that for an IP, if it satisfies the condition:

$$\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} S(m,n) >= 255 \times T \times M \times N \tag{5.1}$$

we define the IP as an VSIP. M, N are the height and width of the IP, $S(m,n)$ is the saliency value at pixel location (m,n) in the SIP and T is the importance coefficient which serves as a threshold to select SIPs. For our case, T is a fixed value of 0.5. When an image is being assessed, only the VSIPs are predicted on and the final score for the whole image is the average of those predicted patches.

Data Sets for Image Quality Assessment

Three public data sets are used in our experiment, and they are: LIVE Image Quality Assessment Database [61], CSIQ Image Quality Database [62] and TID2008 Database [63].

LIVE Image Quality Assessment Database

The LIVE IQA Database contains 29 reference images that are distorted using five distortion types: JPEG2000 (JP2K), JPEG, white noise in the RGB components (WN), Gaussian blur (GBLUR), and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model. In this way a total of 982 images, out of which 203 were the reference images, are generated and then evaluated by human subjects with Difference Mean Opinion Scores (DMOS). A DMOS is the difference between reference and tested Mean Opinion Score and a higher DMOS denotes greater visual distortion and lower quality. The DMOS range for LIVE is [0, 99].

CSIQ Image Quality Database

The CSIQ database contains 30 reference images, each image is distorted using one of six types of distortions: JPEG, JP2K, global contrast decrements, additive pink Gaussian noise, additive white Gaussian noise (same as WN), and GBLUR, each at four to five different levels of distortion. Each type of distortion consists of 150 images which makes it a 900 images data set. The DMOS range for CSIQ is [0, 1].

TID2008 Database

The TID2008 Database contains 25 reference images and 1700 distorted images. Each reference image has 17 types of distortions at four levels. The distortion type includes: JPEG, JP2K, WN, GBLUR, JPEG Transmission Errors, Contrast Change, High Frequency Noise, etc. Different from LIVE and CSIQ, TID2008 adopts Mean Opinion Score (MOS) as its evaluation metric. Same as its literal meaning, MOS is the average value of the scores rated by human observers. For MOS, a higher value means less visual distortion and better quality. The MOS range for TID2008 is [0, 9]. TID2008 is used for cross-database evaluation.

Among the above-mentioned three public databases, the only four distortion types in common are: JPEG, JP2K, WN and GBLUR. And since JP2K is not a frequently used image compression technique compared with JPEG, for our experiments, we only train and test on JPEG, WN and GBLUR distortion types as shown in Figure 5.9. Different from [60] which only uses 60% of images of LIVE for training, we use the images from LIVE and CSIQ for training, validation and testing in order to make deeper network structure possible and prevent over-fitting. More specifically, for the three distortion types, LIVE has 587 images and CSIQ has 450 images. The combined data set contains 1037 images and are split into 60% for training, 20% for validation and 20% for testing. In addition, the DMOS range of CSIQ is linearly mapped to it of LIVE in order to make them consistent.

Local Contrast Normalization

It has been proven in [49] and [60] that applying a local contrast normalization will improve the final training performance. Similar to [49] and [60], all images in the combined data set are locally normalized before training. Suppose the pixel value at location (x, y) is $I(x, y)$, then its normalized value $\hat{I}(x, y)$ is as follows:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y) + C} \tag{5.2}$$

$$\mu(x,y) = \frac{\sum_{m=-M}^{m=M} \sum_{n=-N}^{n=N} I(x+m, y+n)}{(2 \times M + 1)(2 \times N + 1)} \tag{5.3}$$

$$\sigma(x, y) = \frac{\sqrt{\sum_{m=-M}^{M} \sum_{n=-N}^{N} (I(x+m, y+n) - \mu(x, y))^2}}{(2 \times M + 1)(2 \times N + 1)} \tag{5.4}$$

where C is a constant to prevent zero divisor. $2 \times M + 1$ and $2 \times N + 1$ are the normalization window sizes. We set $C = 1$ and $M = N = 3$ in our experiments.

Data Augmentation

In order to increase the size of the training set for the purpose of accommodating deeper network structure and more training epochs, we use some simple data augmentation techniques.

During the cropping process, instead of splitting each image into non-overlapping patches, we crop the image into 64 × 64 patches with stride 32 which means that each image patch has overlapping areas. Then, for each image patch, five types of flips and rotations are performed which are: horizontal flip, vertical flip, 90°, 180° and 270° clockwise rotation.

Network Structure

Kang et al. [49] propose a network structure consisting of only one convolutional layer with max and min pooling, two fully connected layers and one output layer. Jia et al. [60] proposes a deeper network structure consisting of ten convolutional layers, four max pooling layers, two fully connected layers and one softmax layer. Together with the evolution from AlexNet [64] to ResNet [65], extensive research works have shown that a CNN architecture with more layers leads to better performance. Taking this into consideration, we propose a network structure modified on top of VGG-19 [37] shown in Figure 5.6.

Different from the original VGG-19 structure, the input size is 64 × 64 × 3. For each convolutional layer and fully-connected layer, except for the ReLu activation layer, we add a Batch Normalization (BN) layer [66] and a Dropout layer [67] with the ratio of 0.5. The function of the activation layer is to bring non-linearity to the network. The function of the BN layer is to reduce internal covariate shift and the function of the Dropout layer is to prevent the model from over-fitting. For the output layer, instead using Softmax, we adopt Euclidean layer to solve this regression problem.

Training Specs

Stochastic Gradient Descent (SGD) with Adam [68] optimization algorithm is used for training. Learning rate is initially set to 0.01, batch size is set to 64 and momentum is set to 0.9. A total of 20 epochs is run for training and the model with the smallest validation error is selected for testing.

Evaluation Metrics

Two universal evaluation metrics are used in our experiments: Spearman Rank Order Correlation Coefficient (SROCC) and Linear Correlation Coeffieicnt (LCC). SROCC ρ measures how well the relationship between two variables can be described using a monotonic function as shown in Equation 5.5, where n is the number of data points of the two variables and d_i is the difference in the ranks of the i^{th} element of each random variable considered. But LCC r measures the linear relationships between two variables as shown in Equation 5.6, where $\overline{x}$ and $\overline{y}$ are the mean values of all samples. Specifically, higher SROCC and LCC values indicates closer relationships and better results.

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)} \tag{5.5}$$

$$r = \frac{\sum_{i=1}^{n}((x_i-\overline{x})(y_i-\overline{y}))}{\sqrt{(\sum_{i=1}^{n}(x_i-\overline{x})^2)(\sum_{i=1}^{n}(y_i-\overline{y})^2)}} \tag{5.6}$$

For the cross-database evaluation using the TID2008 Database, since the MOS ranges from 0 to 9, we perform a non-linear logistic function mapping to convert the LIVE DMOS labels to the TID2008 MOS labels as shown in Figure 5.7. Using the same strategy in [69], 80% images of TID2008 are used for training and estimating the parameters of the logistic function while 20% images are used for cross-database evaluation.

Experimental Results

We compare our results with the approaches proposed in two papers: [49] and [60]. Kang et al. [49] are the first to propose CNN based IQA approach and [60] achieves state-of-the-art result.

Table 5.1 and Table 5.2 shows the SROCC and LCC results of the combined test set and the cross-database evaluation on TID2008 Database.

Table 5.1.: SROCC results on the combined test set and TID2008. The best results are highlighted.

SROCC	Combined Test Set	TID2008
CNN [49]	0.941	0.905
SDCNN [60]	0.962	0.871
Ours	0.966	0.892

Table 5.2.: LCC results on the combined test set and TID2008. The best results are highlighted.

LCC	Combined Test Set	TID2008
CNN [49]	0.936	0.885
SDCNN [60]	0.965	0.870
Ours	0.968	0.881

The results show that our proposed method outperforms both CNN [49] and SDCNN [60] on the combined test set. And we also achieve comparable results on the cross-database evaluation on the TID2008 compared with [49].

Limitations and Future Work

The main limitation of our proposed IQA method comes from our training data sets. Even though LIVE [61], CSIQ [62], and TID2008 [63] have been used as standard databases in the IQA research tasks for more than 10 years, they have their own limitations. The primary one is that the images in all the three databases are of comparably low resolution, e.g. 640×512. Compared to the images taken by modern cell phones which have much larger resolution, the above-mentioned three databases have lost their representativeness and generalizability. And also, since our proposed method is a fixed size patch-based IQA approach, a larger image size means longer processing time. So, the future works may include: (1) Collecting large-scale databases consisting of high-quality images taken by modern cell phones or other devices which have larger resolution. (2) Taking viewing distance into account when collecting ground truth. (3) Adopting more advanced approaches of utilizing a saliency map, e.g. an adaptive patch size or treating the saliency map as an input channel of the CNN network.

Conclusion

In this section, we have proposed a saliency-based approach of solving the IQA task. Compared with [60], we use a better saliency detection algorithm, a deeper network structure and more training data and achieve new state-of-the-art results. Our proposed method not only demonstrates the fact that a CNN-based method offers a promising way of solving the NRIQA problem, but also provides an alternative means of evaluating the image quality by focusing on the salient region.

5.3 Saliency Based Image Cropping

5.3.1 Introduction

Image composition has always been one of the most important ingredients in photography. It directly influences people's judgement on the aesthetics of an image. The purpose of image cropping is to improve the composition by removing redundant information and keeping the salient region. In this section, we propose a saliency-based image cropping pipeline which adopts the saliency detection algorithm introduced in Section 5.2.2 to solve the image recomposition problem. For more details about the saliency detection algorithm, please refer to [59].

5.3.2 Methodology

There are two mainstream approaches of training an image cropping network: (1) Train the network using the database in which the images are crops with a single score labeled by human annotators. The loss layer is traditional Cross-Entropy Loss (measures the performance of a classification model whose output is a probability value between 0 and 1) as shown in Equation 5.7 or Mean-Square Error (MSE, the most commonly used loss function for regression) Loss as shown in Equation 5.8. (2) Train the network using the database in which the crops are in pairs with similarity scores. This type of network is usually called Siamese Neural Network [70] [71] with Ranking Loss [52].

$$L_{Cross-Entropy} = -(y - \log(p) + (1-y)\log(1-p)) \tag{5.7}$$

$$L_{MSE} = \frac{1}{N}\sum_{i=0}^{N}(y - \hat{y}_i)^2 \tag{5.8}$$

In the work of Chen et al. [52], they analyze the most basic behavior of photography: when a photographer is taking a picture, he constantly moves the camera and judges if the current view is more aesthetically pleasing than the previous one until the desired view is obtained. The above phenomenon reveals the essential nature of photo composition to successively rank a pair of views with gradually changing contents. They also build a pairwise ranking data set with the images downloaded from the Flickr website [1].

Inspired by the work of [52], our proposed pipeline adopts a similar but much deeper Siamese network structure with Ranking Loss and use the same training and validation data sets with the assistance of saliency information.

5.3.3 Data Set for Image Cropping

Traditional image cropping data sets always require heavy human labor work involved which means that for every crop of a reference image, the score label is acquired by taking average of the opinion scores from multiple annotators. The whole process is both time and labor consuming. It always results in the limitation of the size of the data set. Chen et al. [52] propose an inexpensive way of collecting pairwise data sets. Their approach is based on the assumption that if an image is a professional photograph with perfect composition, then, any deviations away from the current view will cause aesthetic degradation. More specifically, if every reference image in the data set is a professional photo, then all the crops originated from it will have lower scores. The (reference image - crop) will become a negative pair for the ranking loss without needing extra labelling.

In this way, they have crawled 31,860 images from the Flickr website. After manually filtering out inappropriate images, the resulting image pool contains 21,045 high-quality images. They randomly select 17,000 images for training and the rest are for validation. For every image, 14 crops are generated and the each crop is paired with its reference image. The final data set consists of 294,630 image pairs.

[1] https://www.flickr.com/

Data Augmentation

We do horizontal flipping and five levels of brightness change for data augmentation. After data augmentation, our data set consists of 2,062,410 image pairs. They are then divided into 80% for training and 20% for validation.

5.3.4 Network Structure

Different from conventional CNN network structure whose input is a single image and uses Cross-Entropy or MSE as its loss function, our proposed network uses Siamese structure which takes in two images successively and updates the parameters using the Ranking Loss during the backpropagation process.

The Siamese Neural Network [72] is also called a Twin Neural Network. The input comes in pairs. The reference image is first fed into the network and its feature vector is precomputed at the output of the fully connected layer. It then forms a baseline against which the cropped image is compared. Both inputs share the same weights and all the other parameters. This architecture is widely used in face recognition.

The difference between Ranking Loss and Cross-Entropy or MSE Loss is that the objective of Ranking Loss is to predict the relative distance between the input pairs instead of predicting a label or a score directly. So the label for the input pair can be as simple as a binary similarity score, e.g. negative in our case. This also explains why the data collection in [52] is inexpensive and efficient. Suppose we have a reference sample I_r and a negative sample I_n, the Ranking Loss can be written as:

$$L(I_r, I_n) = max(0, C - d(I_p, I_n)) \tag{5.9}$$

where d is the distance between the reference sample and the negative sample and C is the margin that regularizes the minimal distance between the ranking scores of the image pair.

The backbone of the network proposed in [52] is AlexNet [64], while we propose a much deeper architecture based on ResNet-34 [65]. It has a max pooling layer, 32 convolutional layers, an average pooling layer and a fully connected layer as shown in Figure 5.8.

5.3.5 Training Specs

Stochastic Gradient Descent (SGD) with Adam [68] optimization algorithm is used for training. The learning rate is initially set to 0.01, the batch size is set to 64, and the momentum is set to 0.9. A total of 20 epochs is run for training and the model with the smallest validation error is selected for testing.

5.3.6 Evaluation Metrics

We adopt the same evaluation metrics as [52], i.e., Intersection-Over-Union (IoU), Boundary Displacement ($Disp.$) and $\alpha - recall$. IoU is described as the Area of Overlap over the Area of Union:

$$IoU = \frac{I_g \cap I_c}{I_g \cup I_c} \tag{5.10}$$

where I_g is the ground truth crop and I_c is the crop generated by the network. $Disp.$ is described as the average of the boundary displacement summation of the image edges:

$$Disp. = \frac{\sum_{i=1}^{4} \left\| B_i^g - B_i^c \right\|}{4} \tag{5.11}$$

where B_i^g and B_i^c represent the four edges of I_g and I_c. For the two vertical edges, $\left\| B_i^g - B_I^c \right\|$ denotes the x coordinate difference. For the two horizontal edges, $\left\| B_i^g - B_i^c \right\|$ denotes the y coordinate difference. And $\alpha - recall$ denotes the fraction of best crops that have an overlapping ratio $((I_g \cap I_c)/I_g)$ greater than α with the ground truth. Same as [52], we set α to 0.75 in all our experiments.

5.3.7 Testing Procedure

Based on the observation described in Section 5.1.3 that a crop which has high IoU value may still be a bad one because it does not include the whole saliency region, for the testing stage, four additional steps are introduced including: saliency detection, connected-component and pruning, bounding box merging, and saliency-anchor-based image cropping.

Saliency Detection

For every testing image, we adopt the same saliency detection algorithm [59] used for the IQA project in Section 5.2.1.

Connected-Component And Pruning

After the gray scale saliency map is generated, a connected-component algorithm proposed in [73] is applied to the saliency map in order to further locate the saliency region. In our experiment, we adopt the 8-connectivity connected-component option. The saliency region whose area size is smaller than 1% of the total saliency area size is then removed from the candidate region list in order to reduce noise and keep the most remarkable content.

Bounding Box Merging

After the above-mentioned procedures, an image may contain multiple connected-components which are all our desired salient regions. And they need to be included in the final crop. Each connected-component corresponds to a bounding box. We merge all of the bounding boxes by keeping their topmost, leftmost, bottommost and rightmost edges. In this way, a single bounding box that contains all of these regions is generated and will meet requirements.

Saliency-Anchor-Based Image Cropping

The last thing needs to be done before an image is fed into the network, is the cropping strategy, one that covers majority of the content and different aspect ratios. Inspired by [54], we propose a saliency-anchor-based image cropping scheme. It consists of two cropping sets.

The first set requires us to construct an image grid with $M \times N$ bins on the reference image and select $m \times n$ bins on the top-left and bottom-right corners. Then, we put the top-left and bottom-right vertices of the crop at the center of one of the $m \times n$ bins to form a candidate crop as shown in Figure 5.9(a). In our experiment, M and N are set to be 16, m and n are set to be 4. But the crops are required to satisfy the following conditions:

(1) The area of the crop should be no smaller than a certain proportion of the whole image size:

$$Area_c \geq \lambda Area_r \tag{5.12}$$

where $Area_c$ and $Area_r$ are the size of the reference image and the crop, λ is the ratio threshold. In our experiment, we set λ to be 0.5.

(2) The aspect ratio of the crop should be within a certain range:

$$\alpha_l \leq \frac{W_c}{H_c} \leq \alpha_u \tag{5.13}$$

where α_l and α_u are the lower and upper boundary of the aspect ratio. W_c and H_c are the width and height of the crop. α_l and α_u are set to be 0.5 and 2 in our experiment.

(3) The candidate crop needs to cover the full saliency bounding box.

We build the second set by growing the candidate crops around the saliency bounding box with five popular aspect ratios: 16 : 9, 4 : 3, 3 : 4, 9 : 16 and 1 : 1, as shown in Figure 5.9(b). For example, suppose we want to generate candidate crops for the 16 : 9 aspect ratio. We first adjust the aspect ratio of the original saliency bounding box to 16 : 9. This adjusted bounding box serves as the anchor of all the candidate crops with the same aspect ratio. Then, we enlarge the bounding box by adding 32 to the left and right boundaries in the horizontal direction and 18 to the top and bottom boundaries in the vertical direction in order to maintain the same aspect ratio. Each addition corresponds to one candidate crop. We keep on adding multiples of 32 and 18 until one edge of the bounding box reaches the boundary of the reference image. The crops in this set are guaranteed to cover the saliency bounding box and satisfy the aspect ratio condition. But the crop size also needs be greater than or equal to $\lambda Area_r$.

With our proposed cropping scheme, each reference image will generate approximately $10 - 2000$ candidate crops. More importantly, all these candidate crops will cover the saliency regions.

5.3.8 Experimental Results

FCDB Test Set

In order to validate the performance of our trained model, we first evaluate the aforementioned three metrics: IoU, Disp. and $\alpha{-}recall$ on one of the public cropping databases, Flickr Cropping Database (FCDB) [74], and compare our results against several baselines and state-of-the-art methods both statistically and visually. The FCDB test set contains 348 images.

Table 5.3 summarizes our results compared against other methods. Figure 5.10 illustrates the visual comparison of some sample cropping result from FCDB. Noting that since [54] uses different evaluation metrics, we only include it in the visual comparison.

Statistically, our method beats the other methods and reaches state-of-the-art results. From the visual examples, it can also be concluded that crops that are generated by our method are more visually pleasing than others. Not only because our pipeline is trained with public databases that takes aesthetic factors into consideration, but also, the resulting crops contain full salient regions.

Table 5.3.: Performance comparison on FCDB [74]. The best results are highlighted. Noting that all of the results except for "Ours" are from [52].

Method	IoU (the higher the better)	Disp. (the lower the better)	$\alpha-recall$ (the higher the better)
eDN [75]	0.4929	0.1356	12.68
AlexNet-finetune [52]	0.5543	0.1209	16.092
MNA-CNN [76]	0.5042	0.1361	0.0747
RankSVM+AVA [74]	0.5270	0.1277	12.6437
RankSVM+FCDB [74]	0.602	0.1057	18.1034
AesRankNet [77]	0.4843	0.1401	0.0804
VFN [52]	0.6842	0.0843	35.0575
Ours	0.6931	0.0798	36.9945

Self-collected Test Set

Another very significant observation is that for images of people or animals, both GAIC [54] and VFN [52] tend to crop out the legs which greatly degrades the cropping performance. To validate this observation, we collect a test set containing 140 images of people and animals. Since we do not have the ground truths, the performance is evaluated based on visual preference. We first apply our method, GAIC [54], and VFN [52] on the test set. Then, we manually vote on each resulting crop to select the best one and collect the votes for each method. The same process is done three times and the final votes are the averages of the three trials.

Table 5.4 shows the voting statistics and Figure 5.11 illustrates the visual comparison.

It clearly shows that our crops always include people's legs.

Table 5.4.: Voting statistics of our self-collected test set.

Method	Ours	GAIC	VFN
Votes	72	50	18

5.3.9 Limitations and Future Work

The main limitation of our proposed image cropping method is from the evaluation metrics. Even though IoU, Disp., and $\alpha-recall$ have been widely used as standard metrics for image cropping tasks, they are too objective, and only consider the geometrical information of the crops. However, the true nature of the image cropping problem is a subjective and flexible task without a unique solution. So, new evaluation metrics considering the aesthetics aspects of photography are desired. Future works may include: (1) Designing new evaluation metrics combining both objective and subjective information. (2) Collecting data sets of a wider variety of contents for the benchmarking of the new metrics.

5.3.10 Conclusion

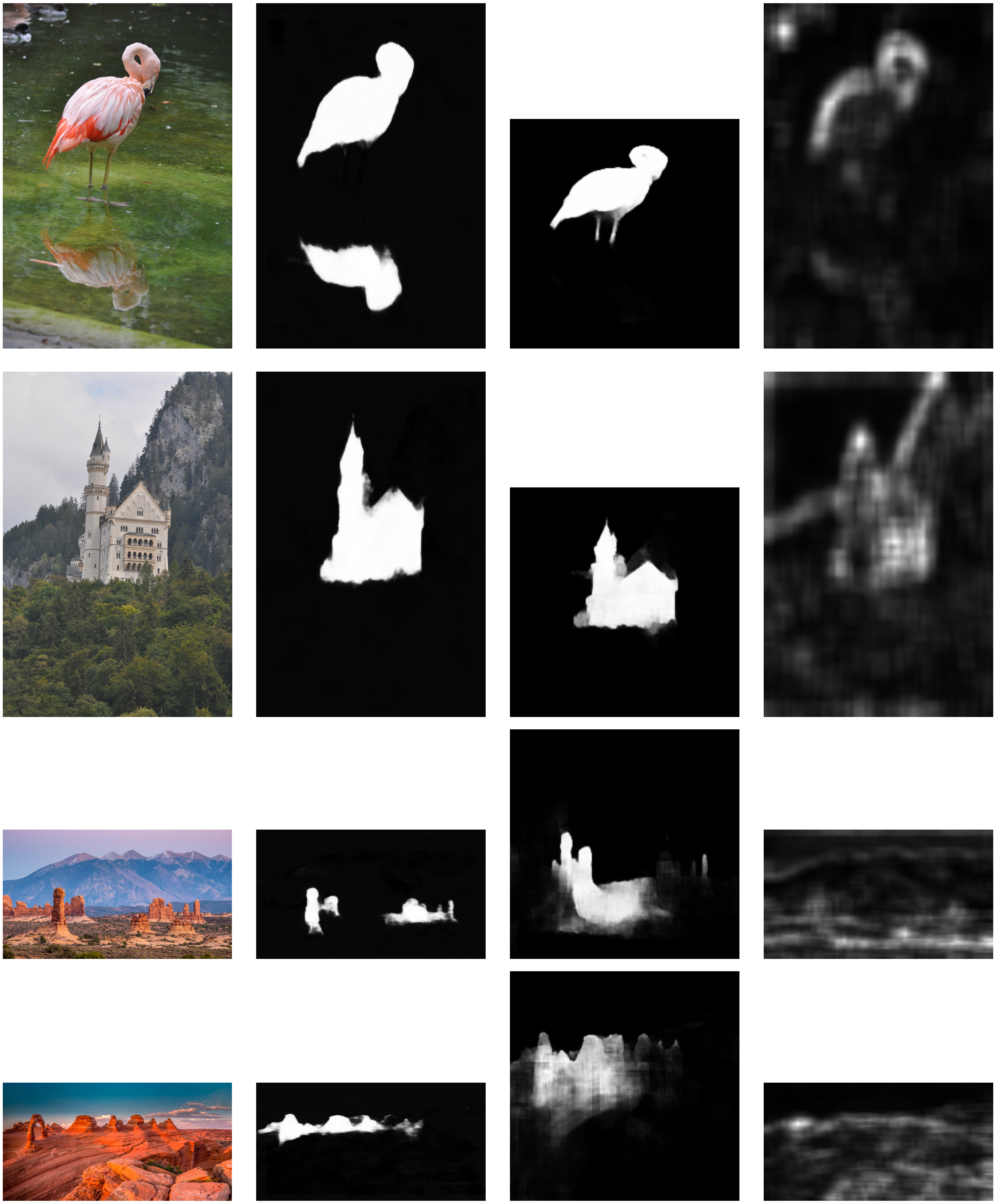
Due to the deeper network structure, robust and accurate saliency detection algorithm, and more advanced cropping scheme, our proposed saliency-based image cropping algorithm outperforms other methods both statistically and visually, and achieve state-of-the-art results.

5.4 Conclusion

In this chapter, we show our efforts in image quality assessment and image cropping tasks. We first demonstrate why saliency is important for both tasks by showing real world examples. Then, we select the best saliency detection algorithm by comparing several state-of-the-art methods.

We start by solving the IQA problem. Three public data sets are used for training, validation, and cross-database evaluation. All training and validation images are split into size 64×64 patches. Each patch has the same quality label as the original image. We adopt the VGG-19 network structure for training. Two universal evaluation metrics are used in our experiments: SROCC and LCC. During the testing phase, saliency detection is applied on each testing image and only the salient patches are evaluated. Our proposed IQA pipeline achieves state-of-the-art results for the testing set and comparable results for the cross-database evaluation.

The solution for the image cropping task is also demonstrated. We train a Siamese network architecture on top of the Resnet-34 backbone with Ranking Loss. The data set is comprised of pair-wise images. Three evaluation metrics are used which are: IoU, Disp., and $\alpha-recall$. During the testing stage, the same saliency detection algorithm is applied on every testing image. We then propose additional steps to generate saliency-anchor-based crops. The best crop ranked by the trained model will be the final result. We compare our method against others using both a public data set and a self-collected data set statistically and visually. Our proposed method outperforms other methods and achieves state-of-the-art performance.



(a) Source (b) DSS (c) PDNet (d) OpenCV

Fig. 5.3.: The visual comparison of the three methods.

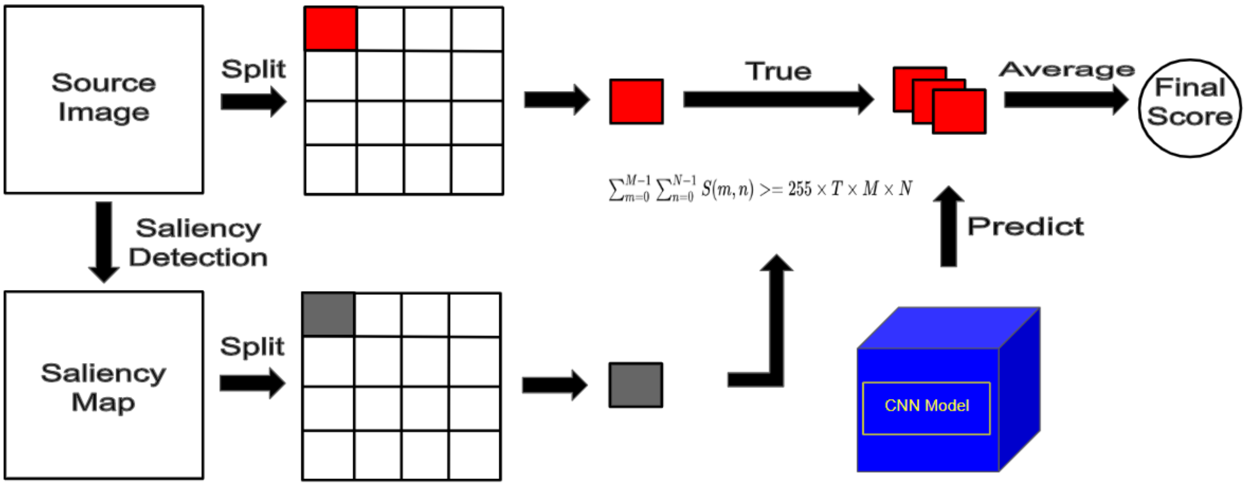
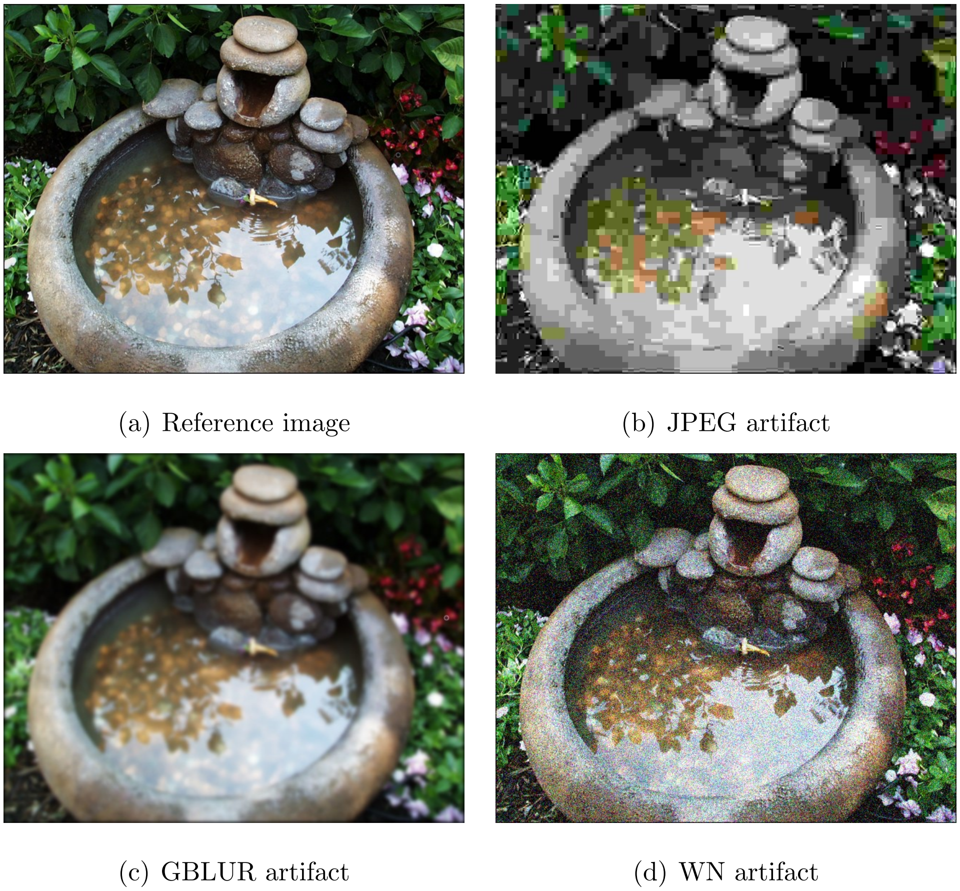


Fig. 5.4.: Our saliency-based IQA flowchart.


(a) Reference image

(b) JPEG artifact

(c) GBLUR artifact

(d) WN artifact

Fig. 5.5.: Sample reference and distorted images.

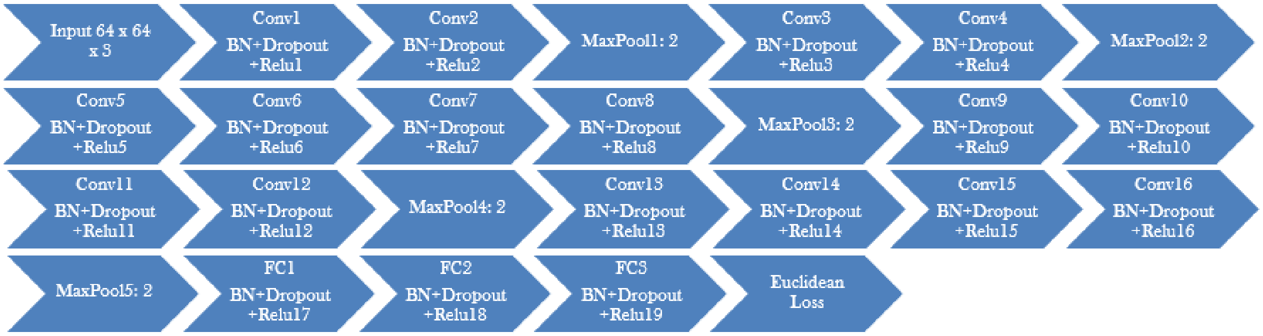


Fig. 5.6.: Our IQA CNN network structure.

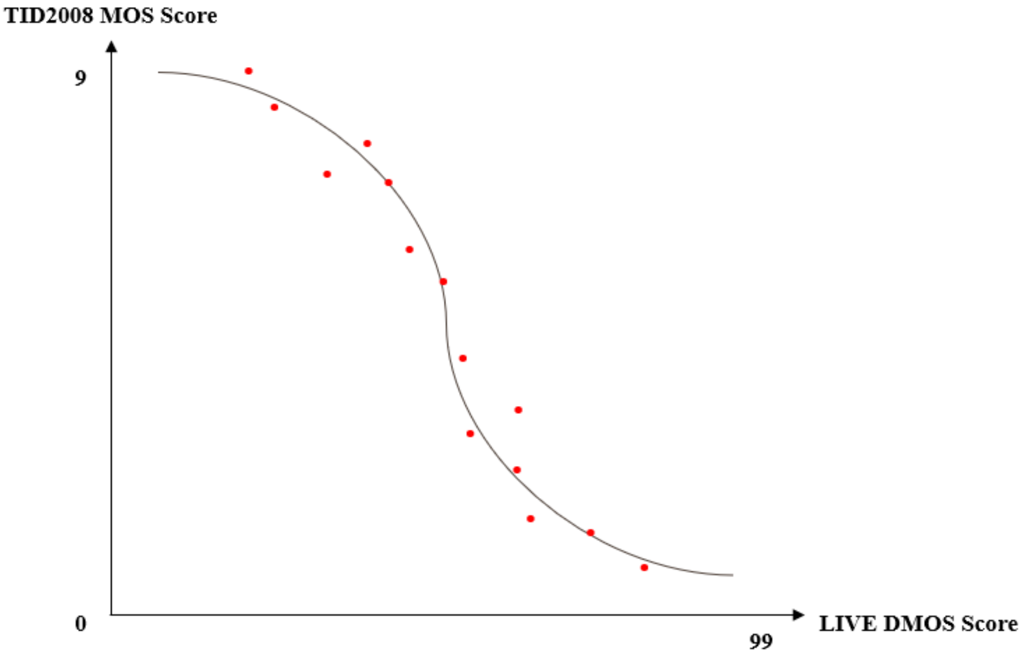


Fig. 5.7.: Non-linear logistic mapping to convert range [0, 99] LIVE DMOS Score to range [0, 9] TID2008 MOS Score. Noting that the red points are the training data from TID2008 Database which have both DMOS and MOS.

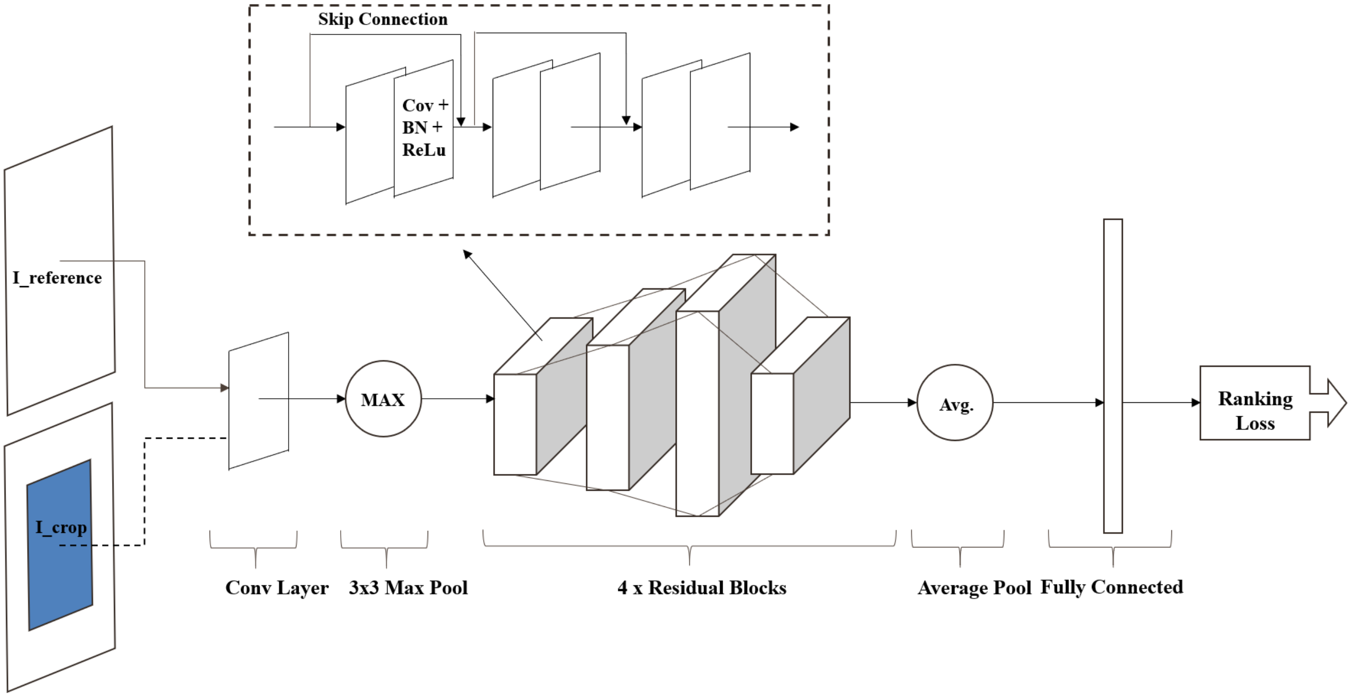
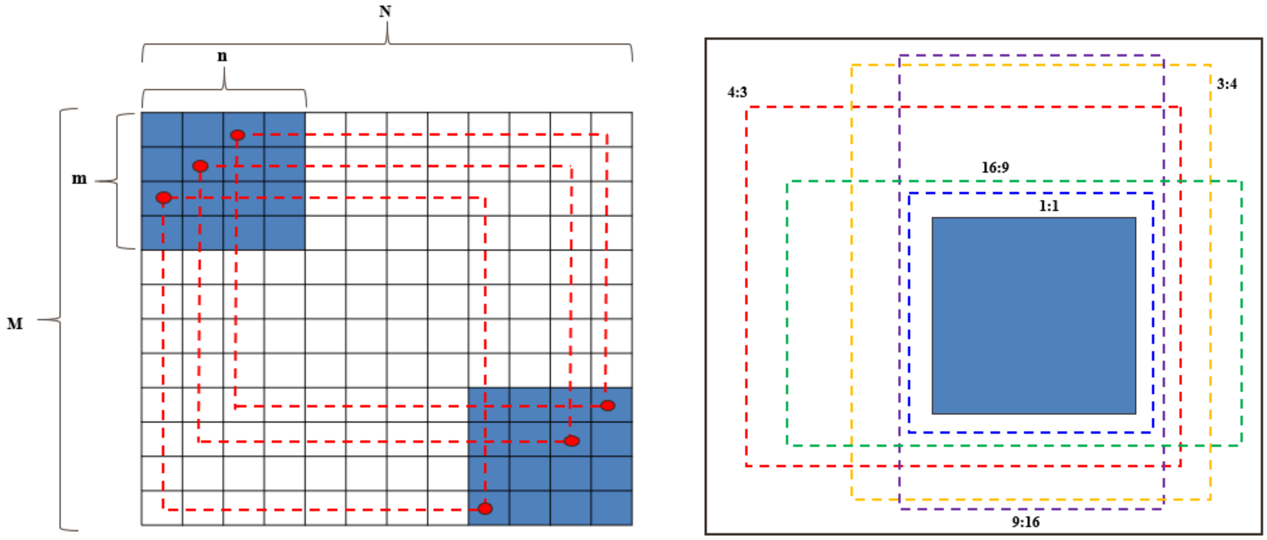


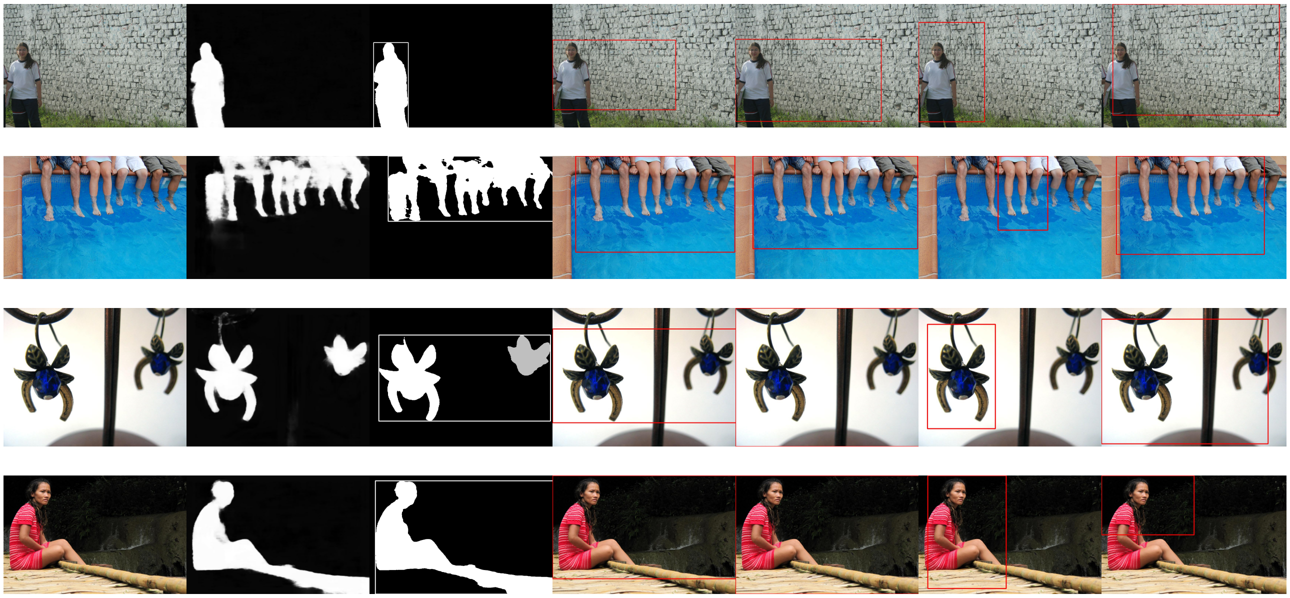
Fig. 5.8.: Our image cropping network structure. The backbone of this network is ResNet-34 [65]. The network takes in the reference image and the crop successively. It starts with a Conv layer and a 3 × 3 Max Pooling layer followed by 4 Residual Blocks. Each Residual Block consists of multiple Conv layer pairs with Skip Connections. Each Conv layer is also connected with a Batch Normalization (BN) layer and a ReLu layer. In the end, we have an Average Pooling layer, a Fully Connected layer, and a Ranking Loss layer.

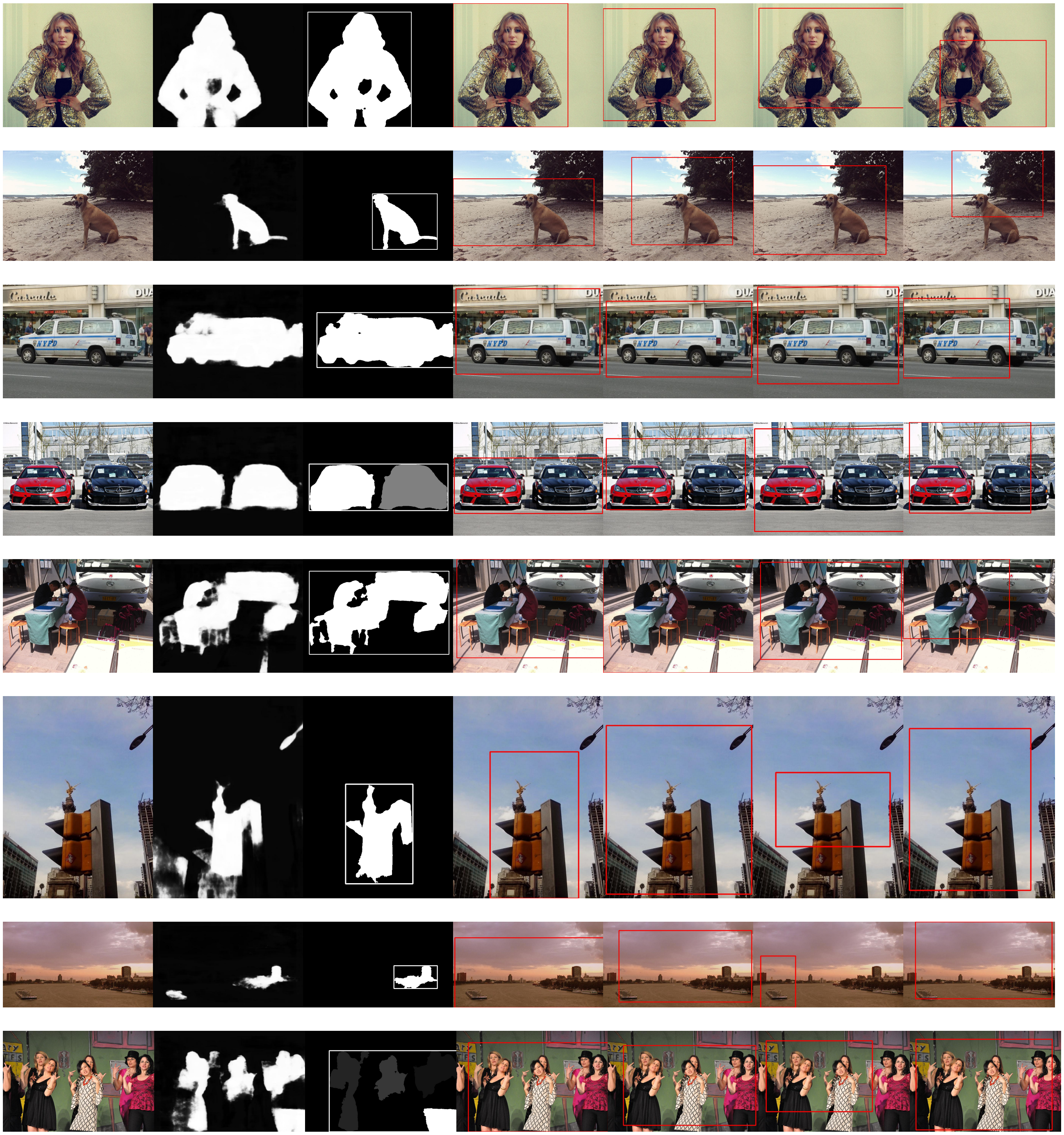


(a) Scheme 1

(b) Scheme 2

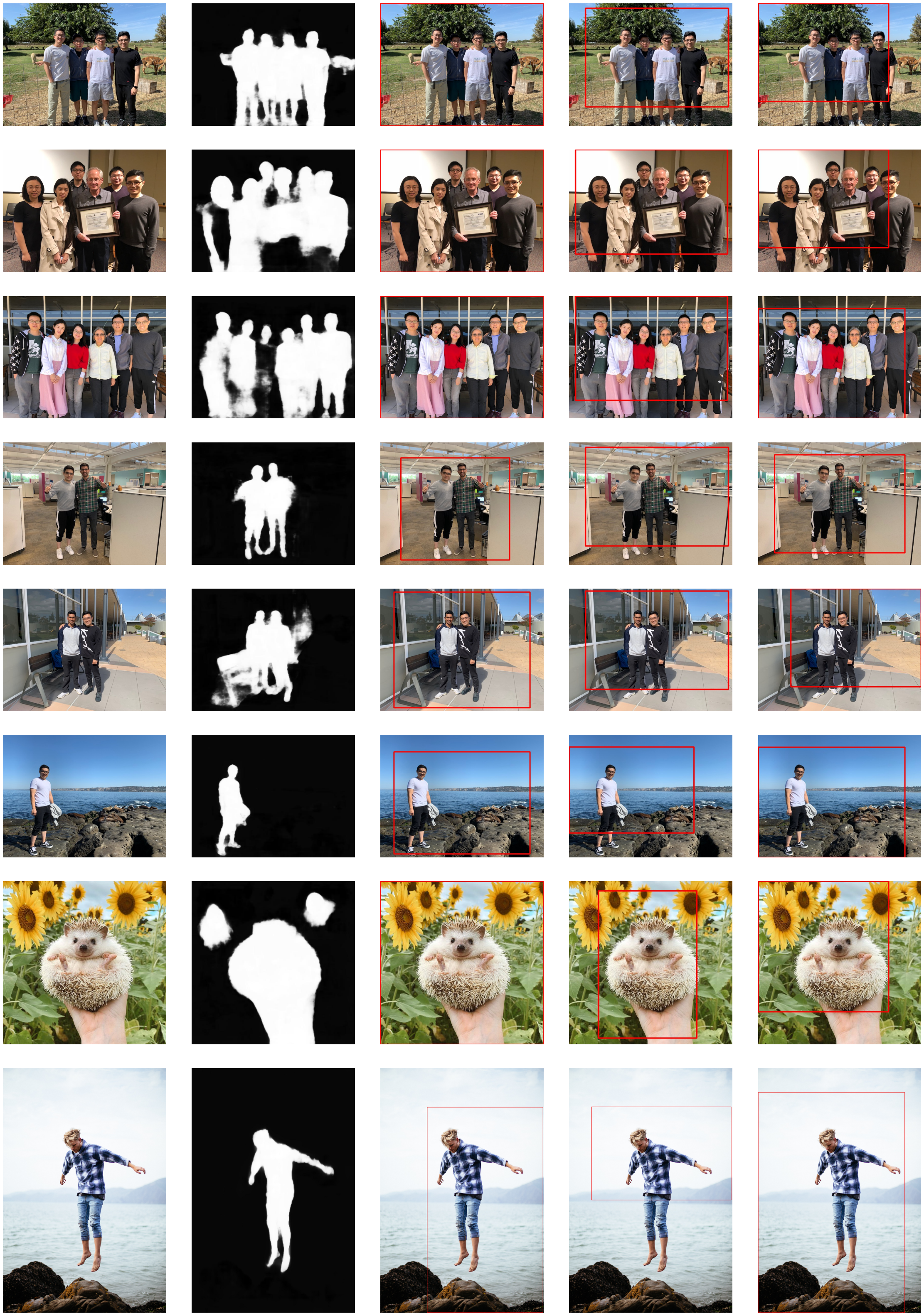
Fig. 5.9.: Saliency-anchor-based cropping scheme. For Scheme 1, we split the reference image into $M \times N$ bins and select $m \times n$ bins on the top-left and bottom-right corners. The vertices of the candidate crop is placed at the center of one of these bins. For Scheme 2, we generate the candidate crops by adding specific values (e.g. 32 and 18) on both horizontal and vertical directions in order to keep the aspect ratio.





(a) Reference (b) Saliency (c) CC with Bbox (d) GT [74] (e) Ours (f) GAIC [54] (g) VFN [52]

Fig. 5.10.: Cropping results comparison of our proposed pipeline against two state-of-the-art methods.



(a) Reference (b) Saliency (c) Ours (d) GAIC [54] (e) VFN [52]

Fig. 5.11.: Cropping results comparison of our proposed pipeline against two state-of-the-art methods.

6. SUMMARY

In this dissertation, we mainly discuss four projects which are: Page Classification for Print Imaging Pipeline, Fading Detection, CNN Based Emotion Recognition, and Saliency Based Image Quality Assessment and Cropping.

In Chapter 2, we develop new features to extend our previous page classification algorithm to detect two new classes of pages which are: Receipt and Highlight. We also show how to use a DAG-SVM structure to classify five classes of pages: Text, Picture, Mixed, Receipt, and Highlight. Our contributions to this part of the work include:

- We developed four new features to extend our previous page classification algorithm to classify two new classes.
- We made a Page Classification User Manual to help users better understand the project.

In Chapter 3, we propose an algorithm for detecting fading in both text region and non-text region. For the text region, we first do the global alignment and then the local alignment using template matching. Then, we create the $L^*a^*b^*$ 3D color node system. We assign each CC to a color node and calculate the ΔE value between the raster page CCs and the scanned page CCs. For the non-text region, especially for the raster and vector region, after the global alignment, we skip the local alignment. Instead, we partition the image into size 60×60 ”super pixels” and assign each super pixel to a color node among 125 nodes and calculate the color difference ΔE. Our contributions to this part of the work include:

- We proposed a novel algorithm for detecting fading in both text and non-text regions.

- For text region fading detection, instead of calculating the ΔE pixel-by-pixel, we used Otsu's method to separate foreground from the background and got the average ΔE of the foreground.

- We created a 3D Color Node System. We divided the $L^*a^*b^*$ color space into 125 nodes where each node corresponds to a certain color. With this system, we can detect what kind of color is fading.

- For non-text region, we divided the object map, raster image, and the scanned image into 60×60 super pixels; so that the inaccuracy of global alignment was reduced. In this way, we are able to detect the fading level for raster and vector regions.

In Chapter 4, we show our ability to do emotion recognition through Convolutional Neural Network training. The whole process includes: Data Collection, Data Preprocessing, Model Training and Model Testing. We also investigate the relations among Emotion, LM, and FAU and accomplish the emotion recognition task through the SVM training of FAUs. Our contributions to this part of the work include:

- We built an emotion recognition framework from scratch. We first collected four public data sets and manually clean them. After that, we implemented data preprocessing, including labeling data, aligning faces and augmenting data. In the training process, we designed our own model based on a VGG-S model but with a much smaller size, better accuracy, and improved efficiency.

- We developed an emotion recognition regression training framework which takes the intensity information of emotions into consideration. We collected our own Emotion Intensity In the Wild data set and defined a 5-level regression labeling scenario. Our experiment results show that the proposed system can recognize the emotion intensities with promising accuracies.

- We showed that our model achieved accurate and smooth real-time recognition of both emotion type and intensity by applying it to real-time scenarios.

- We analyzed the relationship between LM and FAU to solve the FAU detection problem with the geometrical information of the LMs. We then trained an emotion recognition model with an SVM framework using FAUs as the input. Our work successfully validates the possibility that facial expression can be detected using other approaches.

In Chapter 5, we propose a pipeline of solving IQA and image cropping problems with saliency information. We validate the advantage of utilizing saliency information to solve image quality and aesthetic assessment problem. The proposed pipelines for both tasks achieve state-of-the-art results. Our contributions to this part of the work include:

- We demonstrated and validate the advantage of saliency based image quality and aesthetic assessment method.

- We proposed a salient patch based image quality assessment pipeline. Instead of analyzing the quality of an image as a whole, we split the image into salient patches. The quality score of the image will be the average score of all salient patches. With more training and validation data, a deeper network structure and a more robust saliency detection algorithm, our proposed pipeline reaches state-of-the-art result.

- We proposed a saliency-anchor-based image cropping pipeline. The candidate crops are generated with more advanced cropping schemes. All candidate crops are guaranteed to cover the salient regions. Trained with much a deeper network and tested with a more advanced cropping strategy, our propose pipeline achieves the best performance both statistically and visually.

- Our proposed image cropping pipeline also solves an universal problem of other current methods which easily crop out the legs of people's or animals'.

REFERENCES

REFERENCES

[1] C. Lu, J. Wagner, P. Brandi, D. Larson, and J. P. Allebach, "SVM-based automatic scanned image classification with quick decision capability," in *Proc. SPIE 9015, Color Imaging XIX: Displaying, Processing, Hardcopy, and Applications*, vol. 9015, 2014. [Online]. Available: https://doi.org/10.1117/12.2047335

[2] X. Dong, K.-L. Hua, P. Majewicz, G. McNutt, C. Bouman, J. P. Allebach, and I. Pollak, "Document page classification algorithms in low-end copy pipeline," *Journal of Electronic Imaging*, vol. 17, p. 043011, Oct. 2008.

[3] D. Chen, K. Shearer, and H. Bourlard, "Text enhancement with asymmetric filter for video OCR," in *11th International Conference on Image Analysis and Processing. Palermo: IEEE*, vol. 17, Sep. 2001, pp. 192–197.

[4] H. Li and D. Doermann, "Text enhancement in digital video using multiple frame integration," in *Proceedings of the Seventh ACM international conference on Multimedia (Part 1)*, Oct. 1999, pp. 19–22. [Online]. Available: https://doi.org/10.1145/319463.319466

[5] X.-C. Huang and B. D. Nystrom, "Multilevel ink mixing device and method using diluted and saturated color inks for inkjet printers," US Patent 6 172 692B1, Jan., 2001.

[6] V. Bulović, A. Shoustikov, M. Baldo, E. Bose, V. Kozlov, M. Thompson, and S. Forrest, "Bright, saturated, red-to-yellow organic light-emitting devices based on polarization-induced spectral shifts," *Chemical Physics Letters*, vol. 287, no. 3, pp. 455–460, 1998.

[7] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[8] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[9] W. Jang, M.-C. Chen, J. P. Allebach, and G. Chiu, "Print quality test page," *Journal of Imaging Science and Technology*, vol. 48, no. 15, pp. 432–446, Sep. 2009.

[10] Z. Xiao, M. Nguyen, E. Maggard, M. Shaw, J. Allebach, and A. Reibman, "Real-time print quality diagnostics," *Electronic Imaging, Image Quality and System Performance XIV*, vol. 2017, pp. 174–179, 01 2017.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151. [Online]. Available: https://doi.org/10.5244%2Fc.2.23

[12] R. Hartley and A. Zisserman, “Chapter 4. Estimation-2D Projective Transformation,” *Multiple View Geometry in Computer Vision*, pp. 87–131, 2004.

[13] J. Li and N. M. Allinson, “A comprehensive review of current local features for computer vision,” *Journal of Neurocomputing*, vol. 71, no. 10-12, pp. 1771–1787, Jun. 2008.

[14] Y. Lv, Q. Feng, L. Qi, and Q. Chen, “Sub-pixel surface fitting algorithm in digital speckle correlation method,” in *9th International Conference on Electronic Measurement and Instruments*, Beijing, 08 2009, pp. 4–959–4–962.

[15] P. Torr and A. Zisserman, “MLESAC: A new robust estimator with application to estimating image geometry,” *Computer Vision and Image Understanding*, vol. 78, pp. 138–156, 2000.

[16] M. Fischler and R. Bolles, “Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun 1981.

[17] R. Fisher, S. Perkins, A. Walker, and E. Wolfart, “Connected Component Labeling,” 2003. [Online]. Available: https://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm

[18] R. Walczyk, A. Armitage, and D. T. Binnie, “Comparative study on connected component labeling algorithms for embedded video processing systems,” in *International Conference on Image Processing, Computer Vision and Pattern Recognition*, vol. 2, Dec. 2010, pp. 853–859.

[19] J. P. Lewis, “Fast template matching,” *Vision Interface*, vol. 95, pp. 120–123, May 1995.

[20] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on System, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[21] S. J. Park, M. Q. Shaw, G. Kerby, T. Nelson, D.-Y. Tzeng, K. R. Bengtson, and J. P. Allebach, “Halftone blending between smooth and detail screens to improve print quality with electrophotographic printers,” *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 601–614, Nov. 2015.

[22] P. Abhang, S. Rao, B. W. Gawali, and P. Rokade, “Article: Emotion recognition using speech and EEG signal a review,” *International Journal of Computer Applications*, vol. 15, pp. 37–40, Feb. 2011.

[23] P. Ekman, W. Friesen, M. O’Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Pitcairn, and P. Ricci Bitti, “Universals and cultural differences in the judgments of facial expressions of emotion,” *Journal of personality and social psychology*, vol. 53, pp. 712–717, 11 1987.

[24] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[25] S. Xu, C. Lu, M. Shaw, P. Bauer, and J. Allebach, “Page classification for print imaging pipeline,” *Electronic Imaging, Color Imaging XXII: Displaying, Processing, Hardcopy, and Applications*, no. 18, pp. 137-142, 2017.

[26] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 03 2016.

[27] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pp. 503–510, 2015. [Online]. Available: https://doi.org/10.1145/2818346.2830587

[28] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46–53.

[29] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression," in *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis*, 07 2010, pp. 94–101.

[30] Humintell, *The Seven Basic Emotions: Do you know them?*, Jun. 2016. [Online]. Available: https://www.humintell.com/2010/06/the-seven-basic-emotions-do-you-know-them/

[31] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG Facial Expression Database," in *Proc. 11th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 04 2010, pp. 1–4.

[32] M. J. Lyons, S. A. ad Miyuki Kamachi, and J. Gyoba, "Coding facial expressions with Gabor Wavelets," in *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

[33] D. Lundqvist, A. Flykt, and A. hman, "The Karolinska Directed Emotional Faces - KDEF," in *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 1998.

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *CoRR*, vol. 23, no. 10, pp. 1499–1503, 2016. [Online]. Available: http://arxiv.org/abs/1604.02878

[35] R. Mao, Q. Lin, and J. P. Allebach, "Robust convolutional neural network cascade for facial landmark localization exploiting training data augmentation," in *Electronic Imaging, Imaging and Multimedia Analytics in a Web and Mobile World*, 2018, pp. 374–1–374–5.

[36] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM International Conference on Multimodal Interaction(ICMI)*, Nov. 2015.

[37] A. Z. K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 09 2014.

[38] C.-H. Hjortsj, "Man's face and mimic language," Lund : Studentlitteratur, 1970.

[39] P. Ekman and W. Friesen, "Facial Action Coding System: A technique for the measurement of facial movement. consulting psychologists press." Palo Alto: Consulting Psychologists Press, 1978.

[40] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial Action Coding System: The manual on cd rom." Salt Lake City: A Human Face, 2002.

[41] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of Neuroscience Methods*, vol. 200, no. 2, 2011.

[42] P. Ekman and W. Friesen, "FACS - Facial Action Coding System," 1978. [Online]. Available: https://www.cs.cmu.edu/~face/facs.htm

[43] H. Ouanan, M. Ouanan, and B. Aksasse, "Facial landmark localization: Past, present and future," *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pp. 487–493, 10 2016.

[44] S.-F. Xue, H. Tang, D. Tretter, Q. Lin, and J. Allebach, "Automatic photobook: focusing on image selection and image layout based on content and composition," *Proceedings of SPIE, Imaging and Printing in a Web 2.0 World IV*, vol. 8664, 03 2013.

[45] J. Huang, H. Chen, B. Wang, and S. Lin, "Automatic thumbnail generation based on visual representativeness and foreground recognizability," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 253–261, 2015.

[46] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[47] X. Tang, W. Luo, and X. Wang, "Content-Based Photo Quality Assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930-1943, 2013.

[48] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for Blind Image Quality Assessment," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995–1002, 2013.

[49] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for No-Reference Image Quality Assessment," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740, 2014.

[50] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Computer Graphics Forum*, vol. 29, no. 2, pp. 469-478, 2010.

[51] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Learning the change for automatic image cropping," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 971–978, 06 2013.

[52] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," *Proceedings of the 2017 ACM on Multimedia Conference - MM 17*, pp. 37–45, 10 2017. [Online]. Available: http://arxiv.org/abs/1702.00503

[53] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good View Hunting: Learning photo composition from dense view pairs," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5437–5446, 2018.

[54] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5942–5950, 04 2019. [Online]. Available: http://arxiv.org/abs/1904.04441

[55] L.-K. Wong and K.-L. Low, "Saliency retargeting: An approach to enhance image aesthetics," *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 73–80, 2011.

[56] M. Runxin, Y. Yang, and Y. Xiaomin, "Survey on image saliency detection methods," *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 329–338, 2015.

[57] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 06 2007.

[58] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 199–204, 07 2019.

[59] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3203–3212, 2017.

[60] S. Jia and Y. Zhang, "Saliency-based deep convolutional neural network for no-reference image quality assessment," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14 859-14 872, 2017.

[61] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, 2006.

[62] D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, Jan 2010.

[63] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - A database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30–45, 01 2009.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 06 2016. [Online]. Available: http://arxiv.org/abs/1512.03385

[66] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML15. JMLR.org, 2015, pp. 448-456.

[67] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929-1958, 01 2014.

[68] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 12 2014.

[69] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1098–1105, 2012.

[70] S. Chopra, R. Hadsell, and Y. Lecun, “Learning a similarity metric discriminatively, with application to face verification,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*, pp. 539–546, 07 2005.

[71] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, 2014.

[72] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sckinger, and R. Shah, “Signature verification using a siamese time delay neural network,” *Series in Machine Perception and Artificial Intelligence Advances in Pattern Recognition Systems Using Neural Network Technologies*, pp. 25-44, 1994.

[73] C. Grana, D. Borghesani, and R. Cucchiara, “Optimized block-based connected components labeling with decision trees,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1596-1609, 2010.

[74] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, “Quantitative Analysis of Automatic Image Cropping Algorithms: A dataset and comparative study,” *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 226–234, 2017.

[75] E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798–2805, 06 2014.

[76] L. Mai, H. Jin, and F. Liu, “Composition-preserving deep photo aesthetics assessment,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 497–506, 06 2016.

[77] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, “Photo aesthetics ranking network with attributes and content adaptation,” *Computer Vision ECCV 2016 Lecture Notes in Computer Science*, pp. 662-679, 2016.

VITA

VITA

Shaoyuan Xu received his Bachelor of Science degree in Electrical Engineering from University of Illinois at Urbana-Champaign in 2015. He was admitted as a direct Ph.D. and joined Professor Allebach's team in Spring 2016. He then started to work on projects in the area of image processing and image quality. In Summer 2016, he went to HP Boise for a summer internship during which he worked on page classification project. From Summer 2017 to Spring 2019, he worked at HP Labs in Palo Alto as a joint research intern on emotion recognition and image quality assessment projects. During Summer 2019, he had a three-month internship at Google to develop All-in-focus imaging pipeline and he had another internship at Samsung during Fall 2019 to develop image post-processing algorithms. Besides image processing, image quality and deep learning, his research interests also include computer vision and machine learning.