
Memorability Prediction using Deep Learning Techniques

*Thesis submitted to the
Indian Institute of Technology Guwahati
for the award of the degree*

of

Doctor of Philosophy
in
Computer Science and Engineering

Submitted by
Sathisha B

Under the guidance of
Dr. Arijit Sur



Department of Computer Science and Engineering

Indian Institute of Technology Guwahati

12-06-2020

Abstract

In recent times, with huge advancement of artificial intelligence, computational intelligence has been emerging to make more appealing media interfaces like a smart web page, attractive advertisement, cover page of books, etc. To this end, it is required to have some object metrics which essentially describe some subjective media properties. Image or object memorability is such a metric which describes a subjective property of an image. Latest research works show that it is not an incomprehensible concept: variation in remembering images is consistent across viewers. It suggests that independent of a viewers' contexts and biases, some images are intrinsically more memorable than others. This research work proposes few visual factors which play a major role in determining memorability at object and image levels and few deep learning based memorability prediction models are proposed to predict object and image memorability scores individually.

In the first contributory chapter, the relationship between relative spatial characteristics (location and size) of an object and its memorability is explored. It has been experimentally shown that objects of larger size tend to be more memorable than objects of smaller size. Also, the objects present at the center of the image tend to be more memorable whereas objects present at the corners are not. Further, a deep learning based object memorability prediction model is proposed. The proposed model utilizes the object size and location information along with other deep object features to predict the memorability of the given object segment.

The second contributory chapter addresses the relationship between image memorability and two image features: motion and depth. In

this work, it has been experimentally shown that (a) images containing objects in motion tend to be more memorable, (b) images containing objects nearer to the camera at the center tend to be more memorable, and (c) images containing objects farther from the camera at the center tend to be less memorable. Further, deep learning based image memorability prediction models are proposed which utilize motion and depth cues along with the object features to predict memorability scores.

In the third contributory chapter, a *Multiple Instance Learning* (MIL) based deep *Convolutional Neural Network* (CNN) is proposed to utilize visual emotion cues along with other deep object features to predict image memorability scores. Experimental results depicted that incorporation of emotion cues through MIL framework improved the memorability prediction task, and the proposed model performed better than the current state-of-the-art model by achieving a rank correlation close to human consistency.

In this final contributory chapter, it has been addressed how image memorability can be increased. Towards this goal, an end-to-end deep learning model is proposed to enhance the memorability of a generic image. Since the proposed scheme aims to translate an input image to another image having higher memorability, the underlying problem has been considered as memorability based image-to-image translation. The proposed model modifies the given input image to increase its memorability score while retaining its high-level contents. Finally, the dissertation concludes by briefly summarizing the work presented in the dissertation and explaining future research directions.

Declaration

I certify that:

- a. The work contained in this thesis is original and has been done by me under the guidance of my supervisor.
- b. The work has not been submitted to any other Institute for any degree or diploma.
- c. I have followed the guidelines provided by the Institute in preparing the thesis.
- d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
- e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Sathisha B

Copyright

Attention is drawn to the fact that copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author.

This thesis may be made available for consultation within the Indian Institute of Technology Library and may be photocopied or lent to other libraries for the purposes of consultation.

Signature of Author.....

Sathisha B

Certificate

This is to certify that this thesis entitled “**Memorability Prediction using Deep Learning Techniques**” being submitted by **Sathisha B**, to Department of Computer Science and Engineering, **Indian Institute of Technology Guwahati**, for partial fulfillment of the award of the degree of Doctor of Philosophy, is a bonafide work carried out by him under my supervision and guidance. The thesis, in my opinion, is worthy of consideration for award of the degree of Doctor of Philosophy in accordance with the regulation of the institute. To the best of my knowledge, it has not been submitted elsewhere for the award of the degree.

.....

Dr. Arijit Sur

Associate Professor

Department of Computer Science and Engineering

IIT Guwahati

Dedicated to

My Family, Teachers, and Friends

Whose blessing, love, guidance and inspiration paved my path of

success

Acknowledgments

A great many people have contributed to the production of this dissertation. I owe my gratitude to all those people who have made this possible.

I wish to express my deepest gratitude to my supervisor, Dr. Arijit Sur for his valuable guidance, inspiration, and advice. I feel very privileged to have had the opportunity to learn from and work with him. His constant guidance and support not only paved the way for my development as a research scientist also changed my personality, ability, and nature in many ways. I have been fortunate to have such advisor who gave me the freedom to explore on my own and at the same time, the guidance to recover when my steps faltered. Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Prabin Kumar Bora, Prof. G. Sajith, and Dr. Sanasam Ranbir Singh, for their insightful comments and encouragement. Their comments and suggestions helped me to widen my research from various perspectives.

I also like to express my heartfelt gratitude to the director, the deans, and other management of IIT Guwahati whose collective efforts has made this institute a place for world-class studies and education. I am thankful to all faculty and staff members of Dept. of Computer Science and Engineering for extending their co-operation in terms of technical and official support for the successful completion of my research work. Mainly, I like to thank Dr. Arnab Sarkar for supporting and motivating to overcome any problems either in work or otherwise. I am also very much thankful to Dr. Aryabartta Sahu for providing computational resources during the initial stage of my Ph.D. tenure.

I want to extend my special thanks to Mr. Bhriguraj Borah and Mr. Nanu Alan Kachari for their unconditional technical support.

I am thankful to my seniors Sibaji Gaj, Satish Kumar and Shuvendu Rana for mentoring and motivating to overcome any problems either in work and otherwise. I am also grateful to my lab friends Brijesh Singh, Anirban Lekharu, and Prasen Kumar Sharma for being part of this journey. Their encouragement and unconditional support helped me to reach this stage.

I am also grateful to all my well-wishers, friends, seniors, and juniors especially Lipika Rekha Sur, Nilkanta, Rajesh, Shirshendu, Deepak, Nagaraj, Paritosh, Prakriti, Smriti, Sunil, Abhishek Mehta, Dipankar, Natesh, Indranath, Usha, Deepu, Gajendra, Vasumati, Mridumoni, Ravi, and many others for their unconditional help and support.

Most importantly, I would like to thank special people of my life Kavya Basavaraj, Akshay Namdeo, and Vikram CM for their unconditional love and support. You made my life at IIT Guwahati a memorable.

I take this opportunity to submit my revered pranamas to Dr. Sree Sree Shivakumara Mahaswamijigalu and Sree Sree Siddalinga Mahaswamigalu, Sree Siddaganga Math, Tumkur for their continuous blessings for my good health and success. I also like to thank Sree Siddaganga Education Society(R) and community, Sree Siddaganga Math for the constant encouragement and motivation. Especially, I like to thank Dr. HS Jayanna, Gururaj SP, and Renukamba Dinesh for their support and encouragement.

Last but not least, none of this would have been possible without the love and patience of my family. I want to thank my parents, my wife, my sister, and my in-laws, for being a constant source of love, concern, support, and strength all these years.

Contents

1	Introduction	1
1.1	Memorability	2
1.1.1	The Memory Game: Measuring Memorability	2
1.1.2	Human Consistency on Memorability	3
1.1.3	Applications	4
1.2	Literature Survey	5
1.2.1	Understanding Image Memorability	5
1.2.2	Predicting Image Memorability Using Hand-crafted Features	6
1.2.3	Deep Learning Based Image Memorability Prediction . . .	8
1.2.4	Understanding and Prediction of Object Memorability . .	9
1.3	Motivation and Objectives	10
1.4	Contribution of the thesis	12
1.4.1	Object Memorability Prediction: Location and Size Bias .	12
1.4.2	Image Memorability: The Role of Depth and Motion . . .	12
1.4.3	Visual Emotion based Image Memorability Prediction using Multiple Instance Learning	13
1.4.4	Image Memorability Enhancement using Memorability based Image-to-Image Translation	13
1.5	Organization of the Thesis:	14
1.6	Summary	15
2	Research Background	17
2.1	Deep Convolutional Neural Networks	17

CONTENTS

2.1.1	AlexNet	19
2.1.2	VGG-16	21
2.1.3	ResNet	22
2.2	Evaluation Metrics	23
2.2.1	Spearman’s Rank Correlation (ρ)	23
2.2.2	Structural Similarity Index (SSIM)	24
2.3	Experimental Dataset	24
2.4	Summary	25
3	Object Memorability Prediction: Location and Size Bias	27
3.1	Relative Spatial Characteristics	29
3.1.1	Spatial-location	30
3.1.2	Spatial-size	33
3.2	Object Memorability Prediction	36
3.2.1	Input Data Preprocessing	37
3.2.2	Proposed Object Memorability Prediction Models	38
3.2.2.1	SVR-OMP Model	38
3.2.2.2	DCNN-OMP_I	39
3.2.2.3	DCNN-OMP_II	40
3.3	Experimental Results	42
3.3.1	Experimental Set-up	43
3.3.2	Performance Evaluation	43
3.4	Summary	49
4	Image Memorability: The role of Depth and Motion	51
4.1	Role of Motion and Depth in Determining Image Memorability	52
4.1.1	Motion and Memorability	53
4.1.2	Depth and Memorability	55
4.2	Memorability Prediction	57
4.2.1	Proposed OFD-MemNet-I	57
4.2.2	Propose Model OFD-MemNet-II	60
4.3	Experiments and Results	65
4.3.1	Experimental Set-up	66

4.3.2	Performance Evaluation	67
4.4	Summary	69
5	Visual Emotion based Image Memorability Prediction using Multiple Instance Learning	71
5.1	Proposed Model for Image Memorability Prediction	73
5.1.1	Deep CNN for mapping object features to memorability scores	74
5.1.2	MIL based Deep CNN for mapping emotion features to memorability scores	75
5.1.2.1	Multiple Instance Learning (MIL) Framework	75
5.1.2.2	Multi-context patch extraction for MIL based memorability prediction	76
5.1.2.3	<i>MCDR-MemNet</i> : MIL Based Multi-Context Deep Representation Network for emotion-based image memorability prediction	78
5.1.3	<i>Ens-MemNet</i> : Ensemble of Memorability Networks	80
5.2	Experiments and Results	80
5.2.1	Experimental Set-up	80
5.2.2	Results	82
5.2.3	Emotion bias in existing and proposed models	83
5.2.4	Ensembling of VGG-FMemNet, VGG-DMemNet, MCDR-MemNet, and VGG-MemNet	88
5.3	Summary	88
6	Memorability based Image to Image Translation	91
6.1	Proposed Model	94
6.1.1	Translator Network	94
6.1.2	VGG-MemNet	95
6.1.3	Loss Function	96
6.2	Experiments and Results	97
6.2.1	Training of Memorability prediction model VGG-MemNet	97
6.2.2	Training of Translator Network	98

CONTENTS

6.2.3	Performance Evaluation	98
6.3	Summary	99
7	Conclusion and Future Works	101
7.1	Summary of the Contributions	101
7.1.1	Object Memorability Prediction: Location and Size Bias	101
7.1.2	Image Memorability: The Role of Depth and Motion	102
7.1.3	Visual Emotion based Image Memorability Prediction using Multiple Instance Learning	102
7.1.4	Image Memorability Enhancement using Memorability based Image-to-Image Translation	103
7.2	Future Scope	103
	References	105
	Appendix A: Summary of Publications	117

List of Figures

1.1	Mechanical Turk workers played a “Memory Game” in which they watched for repeats in a long stream of images [1].	3
1.2	Sample images along with their ground-truth memorability scores taken from <i>LaMem</i> [2]. The first row shows highly memorable images. The second row shows average memorable images, and the third row shows the least memorable images.	8
2.1	An Example architecture of CNN	19
2.2	The architectural block diagram of the <i>AlexNet</i>	20
2.3	The architectural block diagram of the <i>VGG-16</i>	21
2.4	Residual learning: a building block.	22
3.1	<i>Spatial-location</i> bias on object memorability. Images in each row are showing objects with the same category, but differ in memorability scores and spatial location. The mentioned memorability scores are ground-truth values.	31
3.2	Zone-Map of Image to map the objects into one of the nine possible zones.	32
3.3	Change in memorability score w.r.t. <i>Spatial-location</i> . The <i>Zone-scores</i> are computed for each zone by averaging the memorability scores of all the objects fall into each zone.	33
3.4	<i>Spatial-size</i> bias on object memorability. Each row depicting object segments belonging to the same category, but differ in memorability scores and spatial size. Mem_score represents ground-truth memorability scores.	34

LIST OF FIGURES

3.5	Change in memorability scores w.r.t. <i>Spatial-size</i> . The <i>Size-Score</i> computed for each size range by averaging memorability scores of the object segments belonging to the corresponding size range.	36
3.6	Examples of input object segments prepared according to the existing work [3]. (a) original image. (b), (c), and (d) are extracted object segments. The extracted object segments convey no <i>Spatial-location</i> and <i>Spatial-size</i> information.	37
3.7	Examples of input object segment prepared according to the our proposed method. (a) original image. (b), (c), and (d) are extracted object segments. Images in (b), (c) and (d) represent object segments having <i>Spatial-location</i> and <i>Spatial-size</i> information.	38
3.8	The architecture of the proposed <i>SVR-OMP</i> model.	39
3.9	The architecture of the proposed Deep CNN model <i>DCNN-OMP-I</i> . The architecture is similar to the deep CNN model proposed in [4] except the last layer where the number of outputs is reduced from 1000 to 1 and Softmax layer is replaced by Sigmoid layer to predict object memorability score.	40
3.10	The architecture of the proposed Deep CNN model <i>DCNN-OMP-II</i>	42
3.11	Quality difference between ground-truth and <i>Multiscale Combinatorial Grouping (MCG)</i> [5] object segments. (a) Original Image, (b-k) Ground-truth object segments and (l-u) Top 10 object segments generated using MCG algorithm.	46
4.1	Examples of high-memorable images containing objects with motion.	53
4.2	Examples of images containing objects with motion and no motion along with corresponding optical flow superimposed images.	54
4.3	Examples of high memorable images (first row) with their corresponding predicted depth values (second row).	55
4.4	Examples of low memorable images (first row) with their corresponding predicted depth maps (second row).	56
4.5	The architecture of the proposed Deep CNN model <i>OFD-MemNet-I</i>	58
4.6	The architecture of the proposed model <i>OFD-MemNet-II</i>	61
4.7	Examples of motion cues superimposed images. First row shows original images and the second row shows predicted optical flow superimposed images.	62
4.8	The <i>VGG-16</i> architecture modified and employed to train <i>VGG-FMemNet</i> , <i>VGG-DMemNet</i> , and <i>VGG-MemNet</i>	63

LIST OF FIGURES

4.9	Examples of depth superimposed images. The first row shows original images, and the second row shows the corresponding depth-superimposed images . . .	64
5.1	The architecture of the proposed model <i>Ens-MemNet</i>	73
5.2	The architecture of the proposed <i>VGG-MemNet</i> model.	74
5.3	Example images showing global and local patches. Local patches are extracted from top four salient regions.	77
5.4	The architecture of the proposed <i>MCDR-MemNet</i> model.	78
5.5	Emotion distribution on top ranked images according to the predictions of existing and proposed models (denoted by each sub-figure title). Images are sorted according to the predictions made by existing and proposed models and chosen sets of “Top 10”, “Top 25”, “Top 50”, and “Top 100” images. Emotion distributions are reported for these sets of “Top 10”, “Top 25”, “Top 50”, and “Top 100” images in the first, second, third and fourth rows of the image respectively.	86
5.6	Emotion distribution on least ranked images according to the predictions of existing and proposed models (denoted by each sub-figure title). Images are sorted according to the predictions made by existing and proposed models and chosen sets of “Bottom 10”, “Bottom 25”, “Bottom 50”, and “Bottom 100” images. Emotion distributions are reported for these sets of “Bottom 10”, “Bottom 25”, “Bottom 50”, and “Bottom 100” images in the first, second, third and fourth rows of the image respectively.	87
6.1	Examples of images with their modified versions to increase the memorability score. The Predicted Memorability Score (<i>Predicted Memorability Score (PMS)</i>) is reported for each image.	93
6.2	Framework of the Proposed Method.	94
6.3	Residual Network: Generates residual values which are later used to modify the image to increase its memorability. <i>BN</i> represents Batch Normalization.	95
6.4	<i>VGG-MemNet</i> architecture	96

LIST OF FIGURES

List of Tables

3.1	Comparison of performance between the existing model [3] and the proposed models <i>SVR-OMP</i> , <i>DCNN-OMP-I</i> and <i>DCNN-OMP-II</i> on <i>ground-truth object segments</i>	44
3.2	Comparison of performance between the existing model [3] and the proposed models <i>SVR-OMP</i> , <i>DCNN-OMP-I</i> and <i>DCNN-OMP-II</i> on <i>MCG [5] object segments</i>	45
3.3	Comparison of predicted memorability scores versus <i>Spatial-size</i> of object segments.	47
3.4	Comparison of predicted memorability scores versus the average number of objects located at the center of an image.	48
3.5	Comparison of predicted memorability scores versus the average number of objects located at the corners of an image.	49
4.1	Performance comparison of the existing and proposed models along with Human Consistency (ground-truth).	66
4.2	Comparison of predicted versus ground-truth image memorability scores on LaMem dataset [2]. Images are arranged in descending order of predicted memorability scores. Various ranges of these sorted images are selected. The average ground-truth memorability scores are shown for each set in each row. The reported results are averaged over 5-fold cross-validation models.	69
4.3	Comparison of predicted versus ground-truth image memorability scores on Isola et al. dataset [1]. Uses same measures as detailed in Table 4.2.	69

LIST OF TABLES

4.4	Comparison of predicted versus ground-truth image memorability scores on Dubey et al. dataset [3]. Uses same measures as detailed in Table 4.2. . . .	70
5.1	Performance comparison of the existing (MemNet [2]) and proposed models (VGG-MemNet, VGG-EmoMemNet, MCDR-MemNet and Ens-MemNet). . .	83
5.2	Comparison of predicted versus ground-truth image memorability scores on <i>LaMem</i> dataset [2]. Images are arranged in descending order of predicted memorability scores. Various ranges of these sorted images are selected. The average ground-truth memorability scores are shown for each set in each row. Reported result are averaged over 5-fold cross-validation models	84
5.3	Comparison of predicted versus ground-truth image memorability scores on Isola et al. dataset [1]. Uses same measures as detailed in Table 5.2.	84
5.4	Comparison of predicted versus ground-truth image memorability scores on Dubey et al. dataset [3]. Uses same measures as detailed in Table 5.2. . . .	84
5.5	Performance of the proposed models, VGG-FMemNet, VGG-DMemNet, MCDR-MemNet, VGG-MemNet, OFD-MemNet-II, and Ens-MemNet along with Final-Ens-MemNet.	88
6.1	Comparison of performance between proposed memorability based image-to-image translation method and the style transfer methods [7].	99
6.2	Examples of style trasfer methods along with the proposd method. PMS is reported for each image.	100

List of Acronyms

- AMT** *Amazon Mechanical Turk*
- BEM** *Basic Ensemble Method*
- CNN** *Convolutional Neural Network*
- HOG** *Histogram of Oriented Gradients*
- MCG** *Multiscale Combinatorial Grouping*
- MIL** *Multiple Instance Learning*
- MMSD** *Mean Memorability Score Difference*
- PMS** *Predicted Memorability Score*
- ReLU** *Rectified Linear Unit*
- SIFT** *Scale-Invariant Feature Transform*
- SMS** *Salient Motion Score*
- SSIM** *Structural Similarity Index*
- SVR** *Support Vector Regressor*
- WDS** *Weighted Depth Score*

List of Symbols

ρ	Spearman's rank correlation coefficient
D	Number of channels of an image
M	Number of rows of an image
N	Number of columns of an image
R	Rank variable for the set of memorability scores
S	Set of object segments whose spatial size fall into a particular range
s	Total number of image samples
SS	Size score of the object segment
t	Total number of training samples
W	Trainable neural network weights
W^*	Non-trainable neural network weights
X	Input Image
x_g	Global image patch
x_l	Local image patch
Y_{D-mem}	Memorability score predicted by <i>VGG-DMemNet</i>

Y_{d-mem}	Memorability score predicted by <i>D-MemNet</i>
Y_{Ens}	Memorability score predicted by <i>Ens-MemNet</i>
Y_{F-mem}	Memorability score predicted by <i>VGG-FMemNet</i>
Y_{f-mem}	Memorability score predicted by <i>F-MemNet</i>
$Y_{MIL-mem}$	Memorability score predicted by MCDR-MemNet
Y_{O-mem}	Memorability score predicted by <i>VGG-MemNet</i>
Y_{o-mem}	Memorability score predicted by <i>MemNet</i>
Y_{OFD-I}	Memorability score predicted by <i>OFD-MemNet-I</i>
Y_{OFD-II}	Memorability score predicted by <i>OFD-MemNet-II</i>
y	Ground truth memorability score
zs	Zone Score

Introduction

Every day people are flooded with data, which is equivalent to 174 newspapers' content [8]. Due to this information overload, it is difficult to reach the target audience with the intended information. One of the possible solutions for this problem may be info-graphics (a visual representation of information) as it can be related to a famous phrase "A picture is worth a thousand words" (or precisely 84.1 words [9]). Also, humans are exceptionally capable of remembering particular images, including those representing daily events and scenes [10] and even the shapes of arbitrary form [11]. However, a person may come across hundreds of images while browsing social media, checking the newspaper, reading a magazine, etc. In spite of enormous storage capacity [12, 13], the human cognitive system may not be able to store all the images (he/she comes across) with the same degree of details. Few images are remembered with more details; few with minor details and remaining are forgotten immediately [1]. For example, photos with natural scenery tend to be less remembered than images with animals, vehicles and faces [2]. In this scenario, it is important to find the answers for the following questions: (a) "Is it possible to create/modify a picture so that it has more chances to be remembered?" and (b) "Is it possible to create/modify an image such that an object within that image has more chances to be remembered?" For these questions to be answered, the existence of a measure to quantify the

1. INTRODUCTION

chances of remembering or forgetting an image or an object is much essential. One such measure is memorability.

1.1 Memorability

Humans are very good at memorizing images. People do not just remember the summary of an image but can identify which picture they have seen and few of its details [11, 12]. There are many general reasons for an image to be remembered including (a) image may contain family members, (b) fun event with friends or (c) a well-known monument. But there are images which do not contain friends, family members, or famous monuments and still highly memorable [11–14]. In this thesis, understanding and predicting the memorability for the latter group of images and objects is considered.

Memorability is defined as an objective measure which determines the degree at which images or objects within images are memorable [1, 3]. While it may seem like forgetting or memorizing an image or an object within images is entirely subjective, recent works [1–3] showed that variation in remembering images is consistent across viewers, indicating that certain images are inherently more memorable than others, independent of a viewers’ contexts and biases.

1.1.1 The Memory Game: Measuring Memorability

Isola et al. [1] made the first attempt to measure the memorability of images. They measured the memorability of an image as the probability that an observer will detect repetition of an image a few minutes after the exposition when shown amidst a stream of images as shown in Figure 1.1. According to cognitive psychological studies, this setting determines which images mark a trace in long-term memory. With the aforementioned setting, Isola et al. [1] presented workers on *Amazon Mechanical Turk* (AMT) [15] with a Visual Memory Game. Participants of the game viewed a stream of images. Every image is displayed with a duration of 1 second and 1.4 seconds of a gap is maintained between every two image

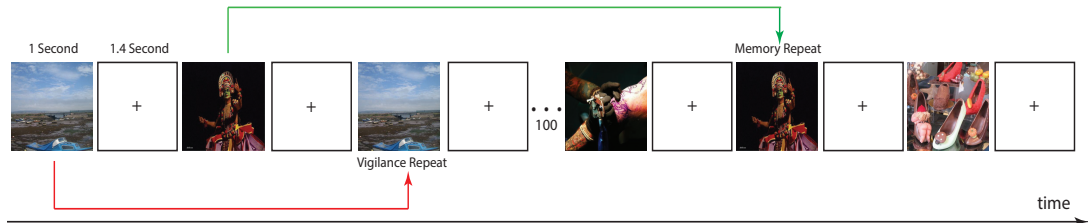


Figure 1.1: Mechanical Turk workers played a “Memory Game” in which they watched for repeats in a long stream of images [1].

presentation, as shown in Figure 1.1. Participants job was to hit the space bar whenever they saw a repeat of an image at any time in the sequence. The image sequence was composed of two categories of images: “target” images (2222) and “filler” images (8220). These images were randomly sampled from the scene categories of the *SUN* dataset [16]. There are two reasons to use filler images. First, to provide the spacing between the first and second presentation of a target image; second, to enable the vigilance task. Vigilance task enabled to check the participants’ attention towards the game by observing the participants responses for the repetition of filler images. For the filler images, repeats occurred with a spacing of 1 to 7 images and for the target images with a spacing of 91 to 109 images. Every target image was repeated precisely once, and except for the vigilance task filler image, every filler image was shown at most once. Once the data collection has been done, the memorability score is assigned to each target image. It was defined as the percentage of correct detections by participants.

1.1.2 Human Consistency on Memorability

Analogous to other image properties like photo quality, saliency, attractiveness, composition, and color harmony, memorability also seems to depend on the viewers’ context, and it appears to be subjective to some inter-subject variability [17]. However, recent works [1-3] showed that there is a sufficiently large degree of agreement between users in remembering images and objects within images, in-

1. INTRODUCTION

dependent of a viewers' contexts and biases.

In order to understand human consistency in remembering images, Isola et al. [1] conducted an experiment. In this experiment, the participants' pool is split into two independent halves and measured how well the memorability scores obtained from the first half of the participants matched with the second half of the participants using Spearman's rank correlation [18]. This process was carried out on 25 random splits and reported the average Spearman's rank correlation. The experiment on Isola et al. [1]'s memorability dataset yielded a rank correlation coefficient values of 0.75, indicating humans are consistent remembering or forgetting images. The same approach is adopted in other existing memorability datasets to evaluate human consistency.

1.1.3 Applications

Image and object memorability is a recent topic in the field of computer vision and has the following promising applications in various domains:

Educational Domain: Memorable academic materials such as flow diagrams or figures representing a particular methodology can be designed to help students to memorize.

User Interface Design Domain: Memorable user interface design can be created for easier navigation for a complex website with hundreds of web pages.

Commercial Domain: Various commercial products, including the logo of a product/company, the cover page of a magazine/book, advertisements can be made memorable.

Computation Photography Domain: By incorporating memorability concept, it is possible to create more memorable pictures of an event or a trip.

1.2 Literature Survey

The idea of memorability and its association with other aspects of the human cognition system have been well studied from a psychological perspective [12, 17, 19–22]. These studies are focussed completely on visual memory related aspects such as the capability of memorizing object-related information [12], visual emotional effect on memorability [20–22] or the brain’s learning technique, e.g., the role of the amygdala in memory [19, 22]. From last few years, many researchers shed light on measuring memorability of images and objects within images. It enabled the research community to understand the cause of visual memorability and its associations with various visual factors, for instance, evoked emotions, saliency, aesthetics, etc. The subsequent subsections details about the existing work on understanding image memorability. It also details about existing methods for predicting image memorability using hand-crafted features and deep learning techniques. Further, the existing literature on understanding and predicting memorability of an object within an image is also presented in the following subsections.

1.2.1 Understanding Image Memorability

Isola et al. [1] defined the image memorability score as the probability that a viewer will detect a repeat of an image within a stream of pictures and measured the memorability scores for 2222 images to understand what makes images memorable. From their study, it is reported that object and scene semantics such as ‘*Labeled Object Counts*,’ ‘*Labeled Object Areas*,’ ‘*Labeled Multiscale Object Areas*,’ ‘*Object Label Presences*,’ and ‘*Scene Category*’ are the primary reasons of memorability. Authors, in [23], attempted to understand the intrinsic memorability of images by discovering the relationship between human-understandable visual attributes and image memorability. To achieve this, they have augmented the image memorability dataset [1] with interpretable spatial, content, and aesthetic image properties using AMT. From their analysis, it is reported that images

1. INTRODUCTION

containing people with visible faces in an enclosed space are more memorable, whereas pictures with a pleasing view and pleasant scenery are not. Mancas and Meur [24] attempted to investigate the relationship between image memorability and saliency. From their investigation, it is reported that image memorability is well correlated with two parameters related to saliency: (1) eye fixation duration and (2) inter-observer congruency. Baveye et al. [25] analyzed the relationship between visual emotions and memorability and reported that images evoking negative emotions tend to be more memorable than positive emotions. Recently, Khosla et al. [2] annotated memorability scores for the large-scale image dataset containing more than 58,000 images to understand the relationship between image memorability and a set of high-level image attributes such as emotions, saliency, popularity, and aesthetics. Authors discovered that emotions and saliency have a positive influence on making images memorable, but popularity and aesthetics are not. To be specific, images evoking negative emotions like *disgust*, *fear*, and *anger* are statistically more memorable than positive emotions like *excitement*, *awe*, and *contentment*. However, the images portraying the *amusement* emotion are exceptional and are equally memorable like images evoking negative emotions. With respect to saliency, Khosla et al. [2] reported that memorability and human eye fixation consistency are positively correlated.

From all of the aforementioned investigations, it is found that object and scene semantics, saliency, and emotions are the primary reasons to make images memorable or forgettable. However, there are few visual cues such as depth and motion information of an image whose relationship with image memorability is uncovered from any of these studies.

1.2.2 Predicting Image Memorability Using Hand-crafted Features

This subsection details the existing image memorability prediction models using hand-crafted features. As mentioned in the previous subsection, Isola et al. [1] created the first image memorability dataset and showed that image memorabil-

ity prediction task could be addressed using current computer vision techniques with machine learning tools. They have developed the first image memorability prediction model by mapping a combination of global image features with memorability scores using a *Support Vector Regressor* (SVR). The chosen global features are GIST [26], spatial pyramid histograms of *Scale-Invariant Feature Transform* (SIFT) [27], *Histogram of Oriented Gradients* (HOG) [28], *Structural Similarity Index* (SSIM) [29], and pixel color histograms. These are the standard visual features that have been previously found to be effective at scene and object recognition tasks. The same group of researchers [1], annotated images of memorability dataset [1] with various visual properties, including spatial layout, image aesthetics, visual emotions, image dynamics, location, and contain a person. These annotations are used along with global image features (such as GIST, spatial pyramid histograms of SIFT, HOG, SSIM, and pixel color histograms) to predict image memorability scores using SVR. Khosla et al. [30] proposed a probabilistic model to predict image memorability maps. They have approached the problem in reverse direction by introducing a data-driven method that fuses local and global image features to determine how and which local regions of the image may be forgotten instead of which regions of the image may be remembered. The local and global features used in their methods include HOG, color name features, local binary pattern, SSIM, Object Bank, and Saliency. Mancas and Meur [24] used saliency coverage and visibility features [31] along with SIFT, HOG, SSIM, and pixel color histograms to predict image memorability scores using SVR. In [32], a multi-view adaptive regression method is proposed to predict image memorability scores. This method also uses similar hand-crafted features to utilize properties like texture, gradient, and shape to predict memorability scores. From the aforementioned studies, it is visible that all these works have been used standard visual features that have been previously found to be effective at scene and object recognition tasks along with few other image features such as saliency and color histograms.

1. INTRODUCTION

1.2.3 Deep Learning Based Image Memorability Prediction

Image memorability prediction is a complex task. Consider Figure 1.2, where the first row shows highly memorable images, the second row shows average memorable images, and the third row shows the least memorable images. From

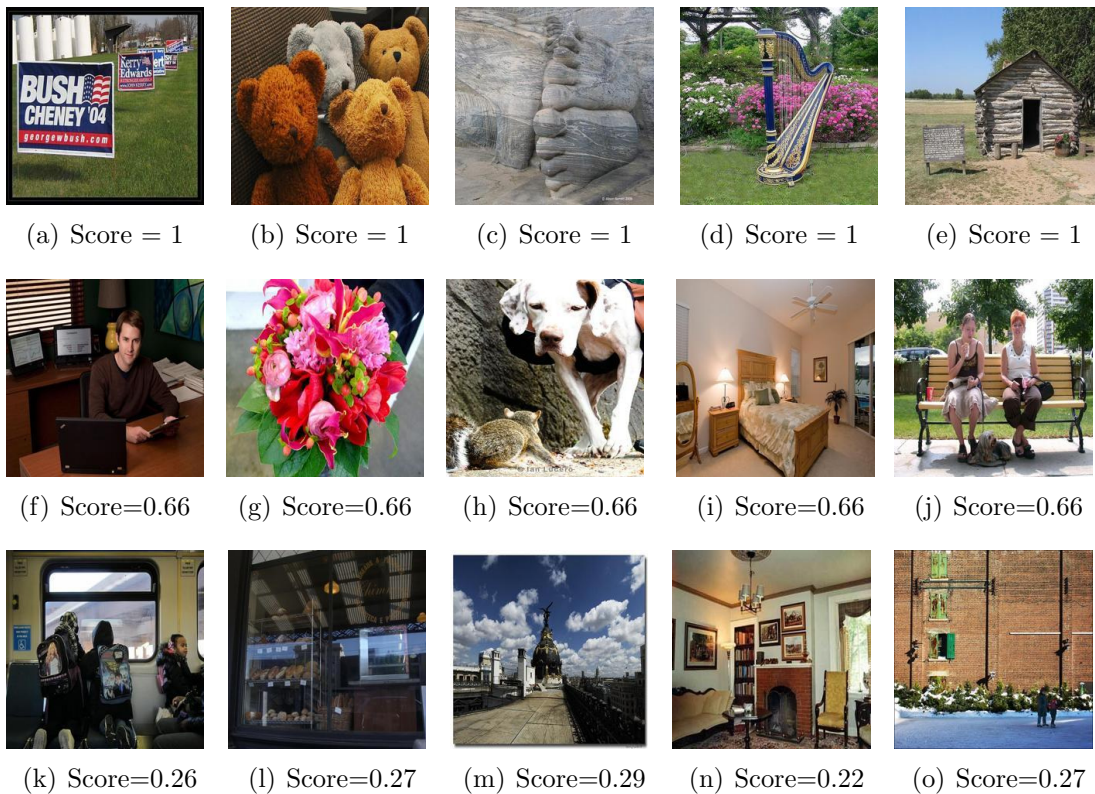


Figure 1.2: Sample images along with their ground-truth memorability scores taken from LaMem [2]. The first row shows highly memorable images. The second row shows average memorable images, and the third row shows the least memorable images.

Figure 1.2, it is clear that images that share the same level of memorability do not look alike; hence, it may require complex features designed specifically for the memorability prediction task to achieve better performance. Recently, deep CNN techniques have shown huge success in various computer vision tasks by automatically learning the task-specific features. Inspired by the success of deep CNN techniques in computer vision applications [33–38], few CNN based image

memorability models have been reported recently. To begin with, Khosla et al. [2] proposed deep CNN based image memorability prediction model, *MemNet* through transfer learning technique. *MemNet* is obtained by fine-tuning the deep CNN model [39] on *LaMem* dataset. *LaMem* is a large scale image memorability dataset with more than 58,000 images created by Khosla et al. [2]. For the fine-tuning purpose, the deep CNN model [39] was loaded with weights which were pre-trained on two datasets: *ILSVRC 2012* [40] and *Places* [39]. In parallel, Baveye et al. [25] fine-tuned the deep CNN model [35] on image memorability dataset [1] to predict memorability scores. This model is similar to [2] but uses *GoogLeNet* [35] and memorability dataset created by Isola et al. [1] in place of *AlexNet* [39] and *LaMem* [2].

1.2.4 Understanding and Prediction of Object Memorability

From the aforementioned literature survey, it can be observed that a series of studies have shed light on what distinguishes the memorability of different images and the intrinsic properties which make images memorable. However, understanding memorability is limited to image level. Though Khosla et al. [30] attempted to determine which local regions within an image are memorable or forgettable, clear comprehension of the memorability of a particular object(s) within an image remained elusive until recently.

Dubey et al. [3] made the first attempt to understand how and which objects inside an image are memorable or forgettable. They augmented both images and object segments from the *PASCAL-S* dataset with ground-truth memorability scores, as explained in Section 1.1.1 and hence, become the first to create object memorability dataset. Their work shed light on various factors and properties that may influence the memorability of an object. From their analysis following observations are reported: (a) simple pixel statics such as mean and variance of HSV color channels do not play a significant role in determining the object memorability in images, (b) saliency is a good predictor of object memorability

1. INTRODUCTION

when image contains a few objects but in complex scenes it is a weaker predictor, (c) object categories play a vital role in determining object memorability, and (d) image memorability is greatly affected by the memorability of its most memorable object. Further, they extracted features from *AlexNet* [33], which is pre-trained on *ImageNet* dataset [40]. Then an object memorability prediction model has been developed by mapping these deep object features to object memorability scores using [SVR](#).

1.3 Motivation and Objectives

From the aforementioned literature survey, it can be observed that a series of studies have been devoted to understand which properties of an image (or an object) are responsible for making an image (or an object) memorable or forgettable. Moreover, a series of research works have developed prediction models to determine memorability score for the given image (or object). Indeed all these studies shed light on many factors and properties which influence visual memorabilities. However, the existing literature suffers from the following limitations:

- With respect to object memorability, the current literature has the following limitations:
 - The state-of-the-art research on object memorability prediction is still in its rudimentary phase with only one prediction model proposed by Dubey et al. [3].
 - Though it is clear from the existing literature that relative spatial characteristics of an object (such as object location and object size) have a strong correlation with image memorability [1], the relationship between object memorability and the relative spatial characteristics of an object is unclear and to the best of our knowledge, no efforts have been made in the existing literature to understand the same.

- With respect to image memorability, the existing literature has the following limitations:
 - Although the existing studies shed light on many factors and properties which influence image memorability, the relationship between image memorability and two visual cues: depth and motion, is unclear.
 - Though it is discovered that visual emotions have a significant role in making an image memorable [2], no existing models used the same in predicting image memorability scores.
 - Though the existence of many memorability prediction models, no end-to-end deep learning model exists to enhance the memorability of a generic image.

Motivated by the aforementioned shortcomings of the existing methods and inspired by the success of deep learning techniques in many computer vision applications, the primary objectives of this dissertation are defined as follows:

- Understand the relationship between memorability (of an image/object) and visual characteristics such as spatial characteristics of an object, depth distribution of an image, motion cues within an image, etc.
- Develop a deep learning based object memorability prediction model which utilizes the object features such as the category of an object, spatial-size, and spatial-location of an object, etc.
- Devise a deep learning based prediction model which uses the image properties like depth, motion, emotion, etc. to determine the memorability score of the given input image.
- Propose an end-to-end deep learning model to enhance the memorability of the given generic image.

1.4 Contribution of the thesis

The major contributions of this dissertation are as follows.

1.4.1 Object Memorability Prediction: Location and Size Bias

In the first contributory chapter, the relationship between relative spatial characteristics (*Spatial-location* and *Spatial-size*) of an object and its memorability is explored. Various experiments are conducted to understand the relationship between object memorability and *spatial-size* of an object. Through experimental results, it is showed that objects present at the center of an image tend to be more memorable than objects present at the corners. Further, a deep learning based object memorability prediction model is proposed. The proposed model can utilize the object size and location information along with other deep object features to predict the memorability of the given object segment. Experimental results highlight that the spatial-location and spatial-size of an object play a significant role in object memorability prediction and the proposed deep learning model outperforms the existing method.

1.4.2 Image Memorability: The Role of Depth and Motion

This chapter explores the relationship between image memorability and two image properties: motion and depth, which, to the best of our knowledge, has not been studied by the existing methods. In this work, motion and depth cues of an image have been represented by predicted optical flow and depth map, respectively. Various experiments have been conducted to understand the association of image memorability with its two image properties: motion and depth. From the experimental analysis, it has been shown that (a) images containing objects with motion tend to be more memorable, (b) images containing objects nearer to the camera at the center tend to be more memorable (c) images containing objects

farther from the camera at the center tend to be less memorable. Further, deep learning based image memorability prediction models are proposed which utilize motion and depth cues along with object features to predict memorability scores.

1.4.3 Visual Emotion based Image Memorability Prediction using Multiple Instance Learning

From the existint literature, it has been observed that visual emotions have a significant role in making an image memorable and hence, emotion cues need to be considered in predicting memorability scores. However, existing methods have not been considered emotion cues in predicting memorability scores. In this chapter, multiple instance learning based deep CNN is proposed to utilize visual emotion cues along with other deep object features to predict image memorability scores. It has been experimentally shown that incorporation of emotion cues through MIL framework improves the memorability prediction task.

1.4.4 Image Memorability Enhancement using Memorability based Image-to-Image Translation

In the final contributory chapter, an end-to-end deep learning model is proposed to enhance the memorability of a generic image. Since the proposed scheme aims to translate an input image to another image having higher memorability, the underlying problem has been considered as memorability based image-to-image translation. The proposed model modifies the given input image to increase its memorability score while retaining its high-level contents. Also, the proposed method learned the mapping between two image domains without using paired (input, label) image dataset. To the best of our knowledge, the proposed model is the first of its kind. Through experimental results, it is showed that the proposed method increases the memorability score of the input image higher than that of the state-of-the-art general image-to-image translation techniques.

1.5 Organization of the Thesis:

This PhD dissertation consists of seven chapters. The first chapter includes an introduction to image and object memorability, followed by a literature survey, research motivations and objectives of the thesis, and contributions of the thesis.

The rest of the thesis is organized as follows:

- Chapter 2 describes the necessary background of the research which includes some preliminary concepts such as deep convolutional neural networks, evaluation metrics, and experimental datasets which are to be used in the later chapters.
- In Chapter 3, the relationship between relative spatial characteristics (location and size) of an object and its memorability is explored. Further, a deep learning based object memorability prediction model is proposed by utilizing these relative spatial characteristics.
- Chapter 4 first explores the relationship between image memorability and two image features: motion and depth. Further, deep learning based image memorability prediction models are proposed which utilize motion and depth cues along with object features to predict memorability scores.
- Chapter 5 presents a novel deep CNN image memorability prediction model based on multiple instance learning framework. The proposed prediction model utilizes visual emotion cues through MIL framework to predict the memorability scores.
- Chapter 6 devises an end-to-end deep learning model to enhance the memorability of a generic image using memorability based image-to-image translation technique. The proposed model modifies the given input image to increase its memorability score while retaining its high-level contents.
- In Chapter 7, this Ph.D. dissertation has been concluded along with the possible future directions.

1.6 Summary

In this introductory chapter, a brief introduction is presented over image and object memorability domain to formulate the scope of research in this field. First, the concept of image and object memorability are explained. Then, brief literature on image and object memorability prediction is described. Based on the limitations of the existing literature, the objectives of the research work are formulated. Finally, a brief description of the contributions and the organization of the thesis have been presented.

1. INTRODUCTION

Chapter 2

Research Background

In this chapter, a brief overview of some fundamental concepts relevant to the topics of interests is presented. It includes a brief introduction to deep convolutional neural networks such as *AlexNet* [33], *VGG-16* [34], and *ResNet* [36]. These networks are used in this dissertation to build various deep learning models to predict image and object memorability scores. Further, the evaluation metrics used to evaluate the proposed memorability prediction models and corresponding datasets used for experimentations are also presented in this chapter.

2.1 Deep Convolutional Neural Networks

Deep CNNs are a class of deep neural networks, which have shown the considerable success on various competitive tasks related to computer vision domain. The capability of learning automatically task-specific features enabled the CNNs more powerful in achieving the state-of-the-art results on various competitive benchmarks. Recent improvements in computational hardware and availability of very large-scale data have made the research in CNNs feasible. In general, the CNN architecture consists of three main types of layers: (1) convolutional layer, (2) pooling layer, and (3) fully-connected or dense layer. Typically, a CNN architecture is comprised of a series of alternate convolutional and pooling layers followed by one or more dense layer at the end. An example architecture of a

2. RESEARCH BACKGROUND

convolutional neural network is shown in Figure 2.1. The detailed descriptions of these layers are as follows:

1. **Convolutional Layers:** A convolutional layer is nothing but a set of neurons where each neuron represents a learnable convolution kernel. Each kernel is small in terms of spatial size and extends for each channel of the input data. When an input image is given to a convolution layer, each kernel is moved across the height and width of the input image and obtains the dot product between the entries in the kernel and the input at any position. As the filter move across the input volume, it produces a two-dimensional activation map. Each value in the activation map is the response of the kernel at every spatial location of the input. These activation maps will be given as input to an activation function. Activation functions get activated when they see the visual features which are extracted from these learned kernels. *Rectified Linear Unit* (ReLU) is one of the most common activation function used immediately after the convolutional layer.
2. **Pooling Layers:** The function of the pooling layer is to gradually minimize the spatial volume of the learned representations to decrease the number of parameters and amount of computation in the network and also to control the overfitting problem. Pooling layer with filter size $N \times N$ reduces $N \times N$ spatial data by a single value. For example, a 2×2 pooling filter replaces 2×2 spatially arranged values by single value. The replacement strategies include Max pooling, Average pooling, and so on. In Max pooling, the entire data volume of spatial size $N \times N$ is replaced by the maximum value computed on the data where the filter is applied. The Average pooling replaces the entire data volume of spatial size $N \times N$ by the arithmetic mean computed on the data where the filter is applied.
3. **Fully Connected Layers:** These layers are nothing but multilayer perceptrons. Each neuron in a layer receives input from all the neurons of the

2.1 Deep Convolutional Neural Networks

previous layer. The output of each neuron in a layer is sent to all the neurons of the next layer. Therefore, the activations of the neurons belong to a particular layer is calculated using matrix multiplication and then, by a bias offset.

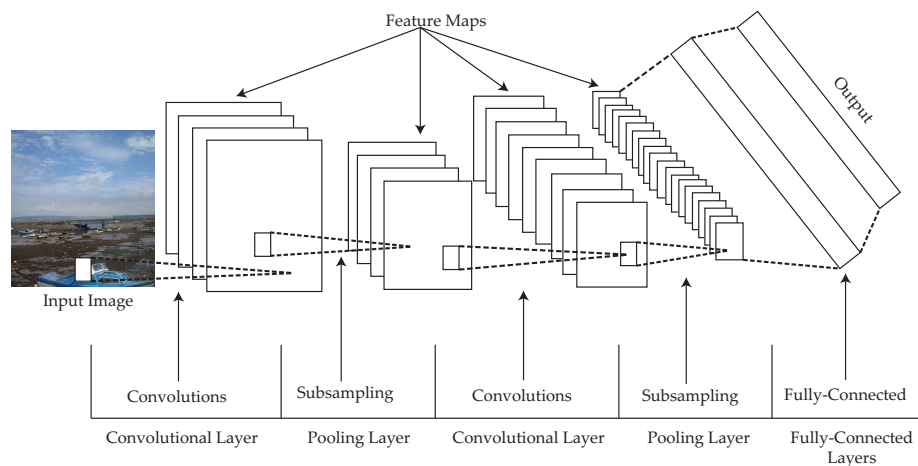


Figure 2.1: An Example architecture of CNN

These three layers are used to build deep CNN, as shown in Figure 2.1. The convolutional layers learn to extract the features required for the intended task. These features are fed to the fully connected layers to learn more abstract level feature representations to perform the intended classification or regression task. In this dissertation, three existing deep CNN models: 1) AlexNet [33], 2) VGG-16 [34], and 3) ResNet [36], are utilized to devise memorability prediction models. Brief description of these three models is explained in the following subsections.

2.1.1 AlexNet

AlexNet proposed by Krizhevsky et al. [33] is probably the first deep CNN model, which showed huge success in large-scale image recognition task [4] by achieving Top-5 error rate of 15.3%. The architectural block diagram of the AlexNet is shown in Figure 2.2. AlexNet is a feed-forward neural network that constitutes of

2. RESEARCH BACKGROUND

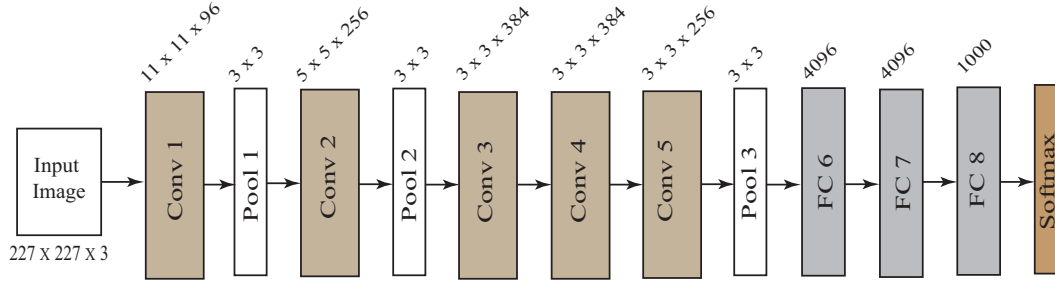


Figure 2.2: *The architectural block diagram of the AlexNet*

five convolutional layers (for feature extraction) and three fully-connected layers (for classification task). As shown in Figure 2.2, the first convolutional layer contains 96 kernels of size 11×11 . Convolutional operation is applied with a stride of 4, which produces feature maps of dimension $55 \times 55 \times 96$ from the input image of dimension $227 \times 227 \times 3$. The first convolutional layer is followed the max-pooling layer with pixel window size 3×3 and stride of 2, which sub-samples $55 \times 55 \times 96$ dimension feature maps to $27 \times 27 \times 96$. The details of the other layers are shown in Figure 2.2. Similar to the first convolutional layer, the second convolutional layer is also followed by a max-pooling layer. However, the third, fourth, and fifth convolutional layers are connected without any sub-sampling layers. Similar to the first and second convolutional layer, the fifth convolutional layer is also followed by a sub-sampling layer. The features extracted from the fifth convolutional layer are passed through three fully-connected layers to learn more abstract level feature representations to perform image classification task using the final layer, which is the soft-max layer.

Despite the fact that deeper networks yield better accuracy, higher depth of the network may also bring the problem of overfitting. Therefore, *AlexNet* incorporated the dropout layers proposed by Srivastava et al. [41]. The dropout layer randomly skips some neurons' output during the training process to enforce the model to learn critical features. To address the overfitting problem further, local response normalization is applied on the feature maps of the first two convolu-

tional layers.

2.1.2 VGG-16

VGG-16 another popular deep convolutional neural network proposed by K. Simonyan and A. Zisserman [34] to perform large-scale image recognition task [4]. The architectural block diagram of *VGG-16* is shown in Figure 2.3 and other variants of VGG can be found in [34]. Input to the *VGG-16* is an RGB image

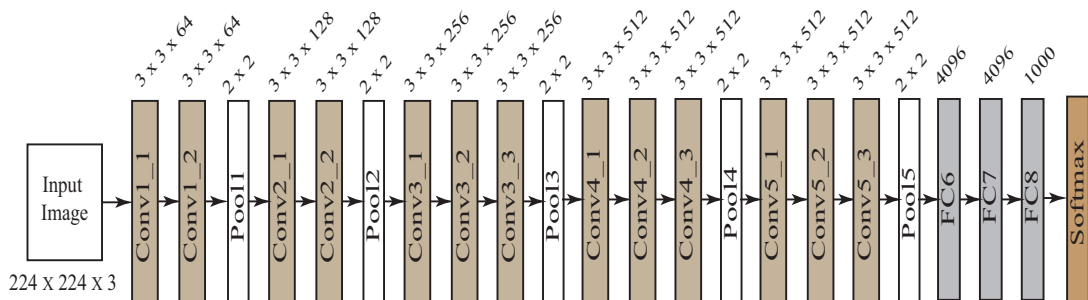


Figure 2.3: The architectural block diagram of the *VGG-16*

with of dimension $224 \times 224 \times 3$. All the convolutional layers use 3×3 kernel size. Except for the last fully-connected layer, all the convolutional and fully-connected layers use ReLU as the activation function. The convolution stride is set to 1 pixel. The spatial padding of convolution layer is set to 1 pixel to retain the input spatial resolution after each convolution. Max pooling is employed after few convolutional layers with a 2×2 pixel window and stride of 2, as shown in Figure 2.3. The features extracted from the last convolutional layer are passed through three fully-connected layers to learn more abstract level feature representations to perform image classification task using the final layer, which is the soft-max layer. *VGG-16* has shown better performance than *AlexNet* on large-scale image recognition task [4] by achieving Top-5 error rate of 7.3%.

2. RESEARCH BACKGROUND

2.1.3 ResNet

Starting from *AlexNet*, the state-of-the-art convolutional neural network architecture has been growing more in-depth with more number of layers. For example, *AlexNet* had eight layers, *VGG-16* introduced 16 layers, and so on. However, merely stacking layers together to increase the network depth may not improve the performance of the model always; instead, the network performance may get saturated and degraded rapidly [36]. He et al. [36] observed that the perfor-

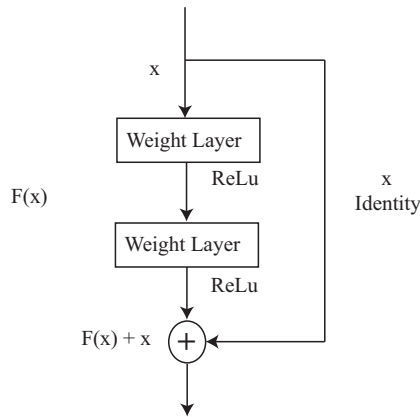


Figure 2.4: *Residual learning: a building block.*

mance degradation of deeper networks is not due to the overfitting problem and adding more layers to the deeper model resulted in higher training error [36]. This degradation problem is addressed in *ResNet* proposed by He et al. [36]. The key idea of *ResNet* is identity mapping, which is achieved by “skip connection”, as shown in Figure 2.4. The skip connection bypasses one or more layers which do not cost any parameters, and it simply adds the output from the previous layer to the layer ahead. The key idea of these skip connections is to expect every few stacked layers to learn the residual mapping in place of desired underlying mapping. The key idea can be formally represented as follows. Let the desired underlying mapping be $H(x)$, and the residual mapping be $F(x) = H(x) - x$. Then the original mapping can be represented as $H(x) = F(x) + x$. The authors hypothesize that optimizing the residual mapping is easier than optimizing the

original mapping. *ResNet* also solves the famous vanishing gradient problem. Vanishing gradient problem is nothing but the gradient value becomes very small when the gradient is backpropagated to earlier layers after several applications of the chain rule. Due to the existence of skip connections in *ResNet*, the gradients can flow back to the earlier layers without vanishing. *ResNet* outperformed all other existing deep CNN models on large-scale image classification task [4] by achieving top-5 error rate of 3.57%.

2.2 Evaluation Metrics

In this dissertation, proposed models rank the images based on their memorability property. In order to verify whether the proposed prediction models rank the images near to humans, the Spearman’s rank correlation coefficient (ρ) is employed. In this dissertation, a deep learning model is proposed to increase the image memorability by modifying the given input image while retaining most of the high-level contents of the given image. To evaluate the retention of the high-level contents, Structural Similarity Index (SSIM) has been employed. The further details of these two evaluation metrics are described in the subsequent subsections.

2.2.1 Spearman’s Rank Correlation (ρ)

Spearman’s Rank Correlation [18] between predicted and ground-truth image memorability scores is computed to measure the consistency between predicted and ground-truth memorability scores. The ρ value ranges from -1 to +1, where +1 represents a complete agreement, and -1 represents the complete disagreement. The ρ between predicted and ground-truth memorability ranks is computed, as shown in Equation 2.1

$$\rho = \frac{cov(R_y, R_Y)}{\sigma_{R_y}\sigma_{R_Y}} \quad (2.1)$$

where R_y and R_Y are the rank variables for the ground-truth and predicted memorability scores, σ_{R_y} and σ_{R_Y} are the standard deviations of the rank vari-

2. RESEARCH BACKGROUND

able R_y and R_Y , and $cov(R_y, R_Y)$ is the covariance of R_y and R_Y .

2.2.2 Structural Similarity Index (SSIM)

The SSIM [29] is a technique to measure the structural similarity between two images. It measures the perception-based quality degradation of an image with respect to a reference image which is of perfect quality. The general form of the SSIM between two images I and J is defined as shown in Equation 2.2:

$$SSIM(I, J) = \frac{(2\mu_I\mu_J + m_1)(2\sigma_{IJ} + m_2)}{(\mu_I^2 + \mu_J^2 + m_1)(\sigma_I^2 + \sigma_J^2 + m_2)} \quad (2.2)$$

where μ_I and σ_I^2 are the mean and variance of I . μ_J and σ_J^2 are the mean and variance of J . σ_{IJ} is covariance of I and J . $m_1 = (a_1r)^2$ and $m_2 = (a_2r)^2$, $a_1 = 0.01$ and where $a_2 = 0.03$ by default and r is the dynamic range of the pixel values.

2.3 Experimental Dataset

There are three standard image memorability datasets and one object memorability dataset which are publically available. These datasets are annotated using the memory game through AMT, as explained in Section 1.1.1 of Chapter 1. The details of these datasets are as follows:

Isola et al. [1] made the first attempt to measure the memorability of images and created the image memorability dataset containing 2222 images. These images were randomly sampled from *SUN* dataset [16] that contains 899 categories and 130,519 images in total. All the images are scaled and cropped about their centers to be 256×256 pixels and images are in RGB color space. Each image is annotated by 78 viewers on an average. The human consistency (refer Section 1.1.2) for this dataset yielded Spearman’s rank correlation (ρ) of 0.75.

Similar to Isola et al. [1], Khosla et al. [2] also developed the memory game to measure the memorability of large-scale image dataset, *LaMem*, consisting of 60,000 images. These images are sampled over a variety of image dataset

including *MIR Flickr* [42], *AVA* dataset [43], affective images dataset [44], image saliency datasets (*MIT1003* [45] and *NUSEF* [46]). Thus, their dataset contains scene-centric images, object-centric images, and other types such as images of art, images evoking certain emotions, and other user-generated images such as ‘selfies.’ All the images are scaled and cropped about their centers to be 256×256 pixels and images are in RGB color space.. Each image is annotated by 80 participants on an average. The human consistency for this dataset yielded Spearman’s rank correlation (ρ) of 0.68.

Dubey et al. [3] created a memorability dataset with 850 annotated images using the memory game. These images were taken from *PASCAL-S* dataset [47]. All the images are scaled and cropped about their centers to be 256×256 pixels and images are in RGB color space. Each image is annotated by 80 participants on an average. Each image is annotated by 80 participants on an average. The human consistency for this dataset yielded Spearman’s rank correlation (ρ) of 0.70.

Dubey et al. [3] made the first attempt to measure the memorability of objects within images. Memorability is measured for a total of 3412 object segments extracted from 850 images. All of these 850 images are sampled from *PASCAL-S* dataset [47]. The *PASCAL-S* dataset is built on the validation set of the *PASCAL VOC 2010* [48] segmentation challenge. Authors used the memory game, which is explained in Section 1.1.1 to measure the object memorability scores. Each object is annotated by 16 participants on an average. The human consistency for this dataset yielded Spearman’s rank correlation (ρ) of 0.76.

2.4 Summary

This chapter presented a brief description of necessary fundamental concepts related to deep convolutional neural networks such as *AlexNet*, *VGG-16*, and *ResNet*. Also, the evaluation metrics which are employed to evaluate the proposed deep learning models are described. Along with these details, image and object

2. RESEARCH BACKGROUND

memorability datasets used for the experimentations are also presented.

With this background, the first contribution of this dissertation will be discussed in the next chapter, where it is shown that the relative size and location of an object play essential roles for understanding the corresponding object memorability within an image. Also, a deep learning based model is devised to utilize the relative size and location details, along with other object features in predicting object memorability scores.

Object Memorability Prediction: Location and Size Bias

Humans selectively process visual information to perform various visual tasks such as object detection, object recognition, scene analysis, etc. Due to this selective nature, the human visual system selects very few visual candidates to carry out afore-mentioned visual tasks. Since most of the computer vision algorithms are designed to help human-visual tasks, such algorithms need to have information about visual candidates or objects. In this chapter, few important visual factors which influence memorability at object level are discovered and also deep learning models are devised to utilize the proposed visual factors in determining memorability scores.

Object memorability is one such important information that may aid in carrying out an intended human-visual task. Object memorability is a task of predicting how well an object within an image can stick on to humans' memory after a single view [3]. It requires the understanding of intrinsic object features which influence the memorability of an object. In recent years, a significant amount of research has been carried out to shed light on inherent characteristics which influence memorability prediction [1, 2, 23–25, 49–52]. However, these studies are limited to image level where memorability prediction is carried out for the entire image.

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

Though Khosla et al. [30] attempted to determine which local regions within an image are memorable or forgettable, clear comprehension of the memorability of particular objects within an image remained elusive until recently. Dubey et al. [3] made the first attempt to understand the object memorability by annotating object segments with memorability scores. This dataset is used to investigate the relationship between object memorability and various visual information such as color, shape, pixel statistics, object category, and saliency. With this investigation, authors stated the following observations: a) shape, color, and other pixel statistics are the poor predictors of object memorability, b) saliency is a good predictor but only when the image complexity is low, and c) object category is a prime visual factor in determining the object memorability score. Further, they have proposed an object memorability prediction model by mapping deep features representing object category information with object memorability scores using SVR. The deep features are extracted using the *AlexNet* [4], which is pre-trained on *ImageNet* dataset [40]. The trained SVR model is tested on ground-truth object segments as well as on automatically generated object segments (to automate the object memorability prediction task). A generic object proposal algorithm, i.e., MCG for image segmentation and object proposal generation [5], is employed to generate the object segments automatically.

Although the existing work on object memorability shed light on various visual factors which influence object memorability, it is suffering from the following limitations:

- Though it is clear from the existing literature that relative spatial characteristics of an object (such as object location and object size) have a strong correlation with image memorability [1], the relationship between object memorability and the relative spatial characteristics of an object is unclear. Also, to the best of our knowledge, no efforts have been made in the existing literature to understand the same.
- The state-of-the-art research on object memorability prediction is still in

its rudimentary phase with only one prediction model proposed by Dubey et al. [3] which uses object category alone as an intrinsic feature to predict object memorability scores.

To overcome these shortcomings, the following contributions are made in this chapter:

- The influence of *Spatial-location* and *Spatial-size* of an object in determining object memorability within an image has been explored. A set of experiments has been carried out to reveal the importance of these characteristics in object memorability prediction.
- Further, a baseline model is proposed to demonstrate the improvement of object memorability prediction due to the influence of the *Spatial-location* and *Spatial-size* of an object.
- Finally, a deep CNN model is devised for automatic feature learning on *Spatial-location*, *Spatial-size*, and category characteristics of an object to accurately predict the object memorability score.

Rest of the chapter is organized as follows: A detailed study on the influence of proposed visual characteristics is shown in Section 3.1. The proposed object memorability prediction models are described in Section 3.2. Performance evaluation of the proposed models is presented through a set of experiments and corresponding results in Section 3.3. Finally, the summary of the chapter is presented in Section 3.4.

3.1 Relative Spatial Characteristics

Spatial-size and *Spatial-location* of an object are two major visual factors to determine the object's importance [53]. The object larger in size and closer to the center is likely to be more important as it has a higher probability of being

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

mentioned by annotators [53]. While the relationship between the object memorability and its relative spatial characteristics (*Spatial-size* and *Spatial-location*) is unclear, it is revealed from the recent study [1] that memorability of an image and relative spatial characteristics of objects within that image do have a strong correlation. In this section, the influence of these two relative spatial characteristics, *Spatial-location* and *Spatial-size*, in determining object memorability prediction has been analyzed.

3.1.1 Spatial-location

Spatial-location of an object can be defined as the position of an object within an image frame. From the object memorability dataset, it can be observed that *Spatial-location* of an object may suppress the effect of object category in predicting its memorability. It is depicted in Figure 3.1 where the objects in each row, are from the same category (the first row contains *Person* category objects, second and third contain *Animal* category objects.) but differ with respect to their memorability scores.

To analyze the influence of object location within the image, an experiment is conducted based on ground-truth object memorability scores. In this experiment, the given image is divided into nine rectangular zones to decide the object position, as shown in Figure 3.2. These zones are named according to their positions. Every object present in the given image is mapped to one of these zones based on the total number of object's pixels located in each of the nine zones. Zone mapping is formally defined as follows:

L : The set of nine zones representing different parts of the image, as shown in Figure 3.2.

O : The object segment whose *Spatial-location* need to be determined.

O_i : The portion of the object segment O belonging to i^{th} zone L_i , where $i \in \{1, 2, \dots, 9\}$.

3.1 Relative Spatial Characteristics

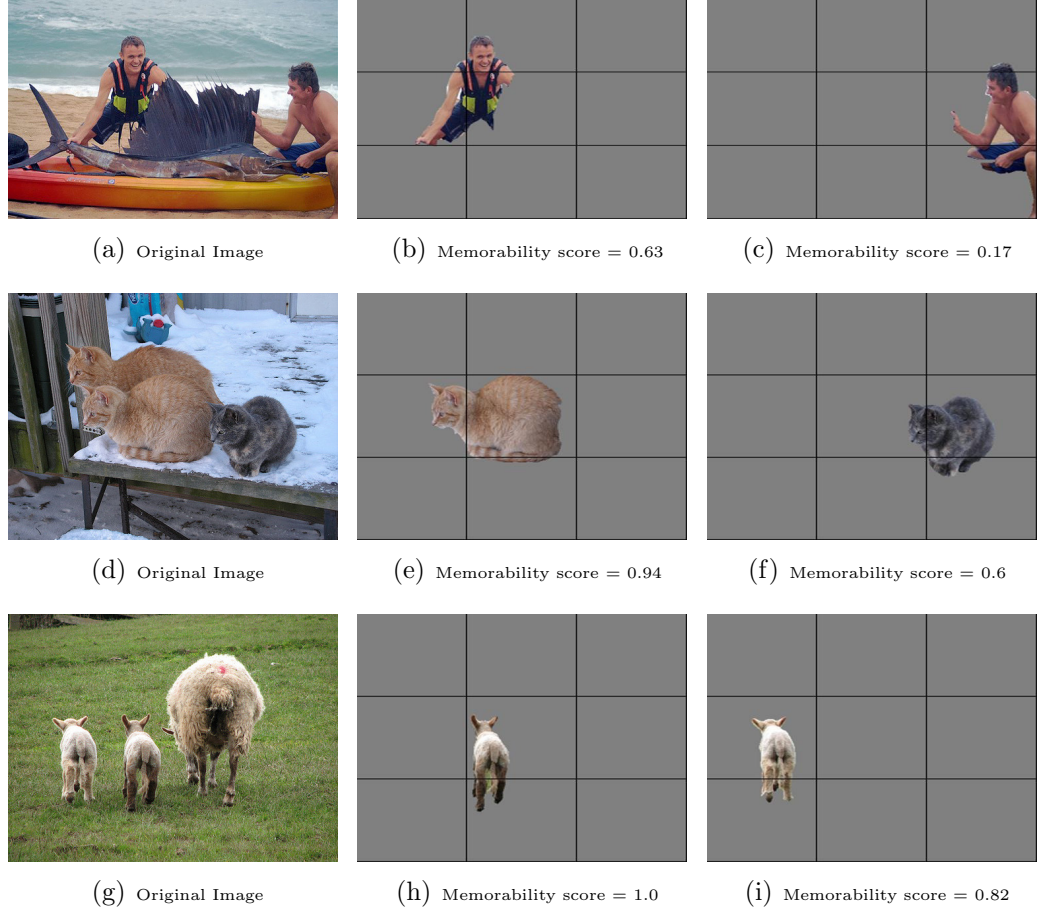


Figure 3.1: *Spatial-location bias on object memorability. Images in each row are showing objects with the same category, but differ in memorability scores and spatial location. The mentioned memorability scores are ground-truth values.*

z : The zone to which object segment O is mapped.

z is computed using following mapping function:

$$z = \arg \max_{\forall L_i \in L} |O_i| \quad (3.1)$$

where $|O_i|$ is the number of pixels belonging in O_i . In this manner, all the object segments present in the dataset are grouped into nine zones. For each zone, a *zone-score* is computed by averaging the memorability scores of the objects belong to the corresponding zone as follows:

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

O_{ik} : The k^{th} object segment belonging to i^{th} zone. Where $i = \{1, 2, \dots, 9\}$, $k = \{1, 2, \dots, n_i\}$ and n_i is the number of object segments from the dataset belonging to i^{th} zone

y_{ik} : The ground-truth memorability score corresponding to the object segment O_{ik}

zs_i : The *Zone-Score* of i^{th} zone L_i .

zs_i is computed as follows:

$$zs_i = \frac{1}{n_i} \sum_{\forall k \in \{1, 2, \dots, n_i\}} y_{ik} \quad (3.2)$$

Computed *zone-scores* zs_i are presented in Figure 3.3. It is evident from the Figure 3.3 that objects that belong to the middle-center zone tend to have higher memorability scores than the objects that belong to the corners of the image.

Top_Left (1)	Top_Center (2)	Top_Right (3)
Middle_Left (4)	Middle_Center (5)	Middle_Right (6)
Bottom_Left (7)	Bottom_Center (8)	Bottom_Right (9)

Figure 3.2: *Zone-Map of Image to map the objects into one of the nine possible zones.*

To investigate the relationship between *Spatial-location* and memorability of an object segment further, another experiment is conducted, which has two steps. In the first step, object memorability dataset is divided into training and testing sets. Then *zone-scores* are computed from the training set using Equation 3.2. In the second step, each object segment of the test set is mapped into one of the nine zones using Equation 3.1. Then memorability score for each object from the

test set is computed by assigning *zone-score* of the corresponding zone (computed in the first step). Finally, Spearman’s rank correlation (ρ) is computed between predicted memorability scores and ground-truth scores. This process is carried out in six-folds for cross-validation purpose. This experiment yielded an average rank correlation of **0.40**. From this result, it is observed that *Spatial-location* and object memorability are positively correlated. Therefore, *Spatial-location* can be considered as one of the important visual characteristics for predicting object memorability scores.

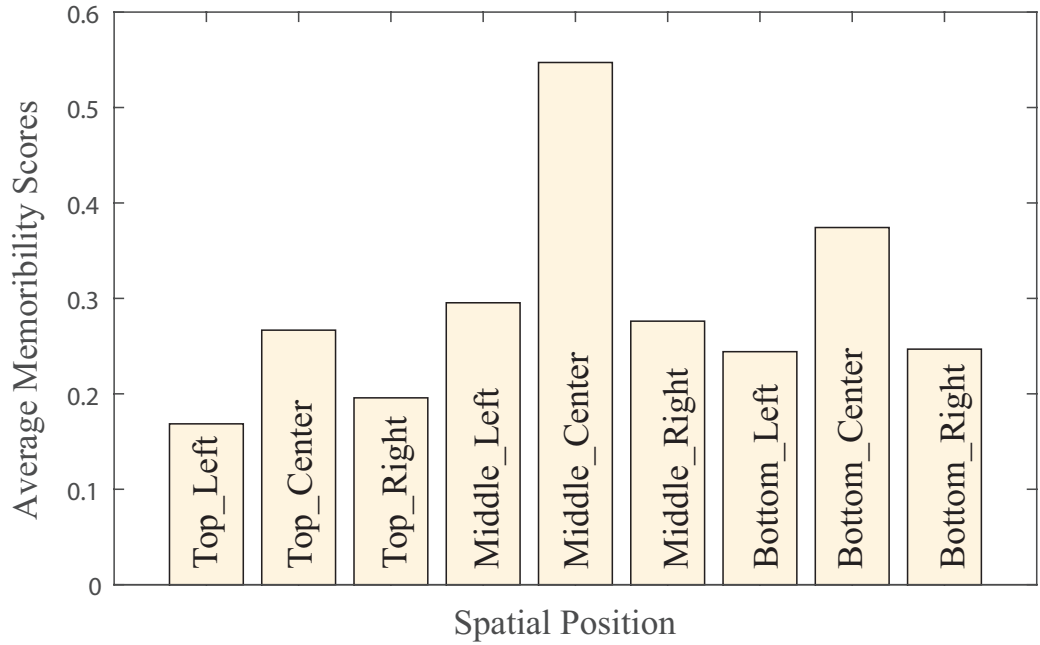


Figure 3.3: Change in memorability score w.r.t. *Spatial-location*. The Zone-scores are computed for each zone by averaging the memorability scores of all the objects fall into each zone.

3.1.2 Spatial-size

Spatial-size of an object can be defined as the relative size of an object with respect to image size. Similar to the *Spatial-location*, *Spatial-size* may also suppress the effect of object category in determining object memorability scores. It is shown in Figure 3.4, where the objects in each row are from the same category

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

(first and the third rows contain *person* category objects, and the second row contains *Animal* category objects) but differ much in their memorability scores. For example, though the object segments shown in Figure 3.4(f) & 3.4(h) belong to the same object category *Animal*, the memorability score (0.29) of object segment depicted in Figure 3.4(f) is relatively lesser than the memorability score (0.66) of the object segment depicted in Figure 3.4(h). Hence, the *Spatial-size* of an object is an intuitive factor to influence its memorability score.

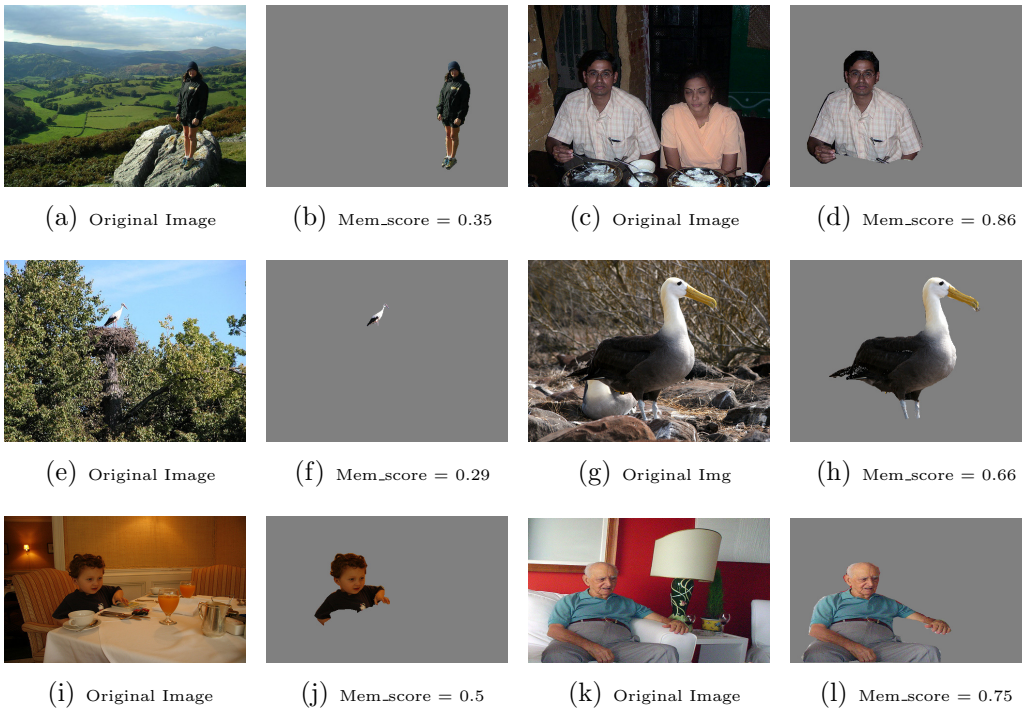


Figure 3.4: *Spatial-size bias on object memorability.* Each row depicting object segments belonging to the same category, but differ in memorability scores and spatial size. *Mem_score* represents ground-truth memorability scores.

To understand the relationship between an object’s *Spatial-size* and its memorability, an experiment is conducted based on the ground-truth object memorability scores. In this experiment, all the object segments present in the object memorability dataset is divided into six non-overlapping sets based on *Spatial-size* as shown in Equation 3.3

$$\begin{aligned}
 S_1 &= \{O_k \mid 0 < OS_k < 2\} \\
 S_2 &= \{O_k \mid 2 \leq OS_k < 5\} \\
 S_3 &= \{O_k \mid 5 \leq OS_k < 10\} \\
 S_4 &= \{O_k \mid 10 \leq OS_k < 20\} \\
 S_5 &= \{O_k \mid 20 \leq OS_k < 30\} \\
 S_6 &= \{O_k \mid 30 \leq OS_k < 100\}
 \end{aligned} \tag{3.3}$$

where

O_k : k^{th} object segment in the object memorability dataset.

OS_k : *Spatial-size* of O_k . It is defined as follows:

$$OS_k = \frac{OP_k}{IP_k} * 100 \tag{3.4}$$

where

OP_k : Total number of pixels belong to O_k

IP_k : Total number of pixels belong to the image which contains O_k

The ranges of values in Equation 3.3 are chosen to ensure the balanced number of objects in each set. For each of these groups, the *size-score*(SS_i) is computed as follows:

$$SS_i = \frac{1}{S_i} \sum_{\forall O_k \in S_i} y_k \tag{3.5}$$

where y_k is the memorability score of O_k and $i = \{1, 2, \dots, 6\}$.

The computed SS_i for $i = \{1, 2, \dots, 6\}$ is given in Figure 3.5, where it is evident that an increase in the size of an object tends to increase in its memorability score.

The relationship between *Spatial-size* and memorability of an object segment is investigated further using another experiment. This experiment is carried out in two steps. In the first step, object memorability dataset is divided into training and testing sets. Then SS_i for $i = \{1, 2, \dots, 6\}$ is computed using Equation 3.5.

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

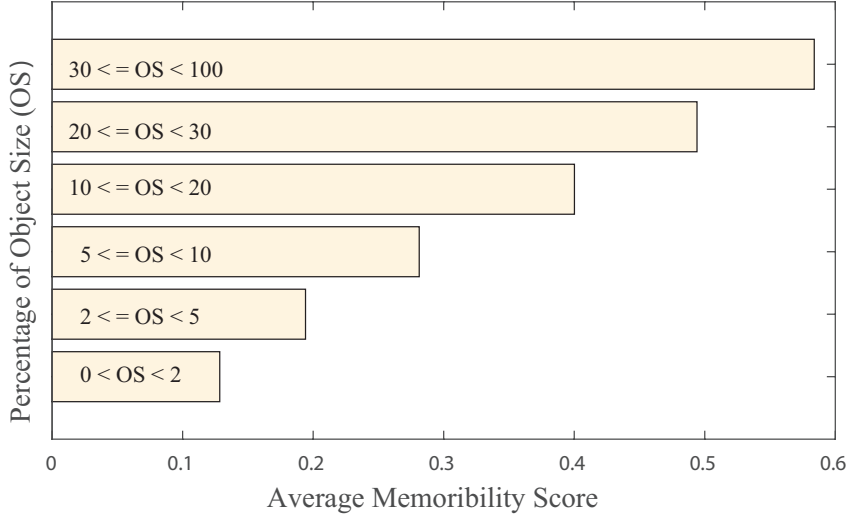


Figure 3.5: Change in memorability scores w.r.t. *Spatial-size*. The *Size-Score* computed for each size range by averaging memorability scores of the object segments belonging to the corresponding size range.

In the second step, each object segment of the test set is mapped into one of the six sets S_i for $i = \{1, 2, \dots, 6\}$ based on the object *Spatial-size*. Then memorability score for each object segment from the test set is computed by assigning corresponding SS_i . Finally, Spearman’s rank correlation is computed between predicted and ground-truth object memorability scores. This process is carried out in six-folds for cross-validation purpose. This experiment yielded an average rank correlation of **0.50**. From this result, it is observed that *Spatial-size* and memorability of an object are positively correlated. Therefore, *Spatial-size* can be considered as one of the important visual factors for determining object memorability scores.

3.2 Object Memorability Prediction

Based on the analysis carried out in Section 3.1, it is observed that *Spatial-size* and *Spatial-location* play a crucial role in determining object memorability.

Therefore, these two visual factors are incorporated in determining object memorability in the proposed models. The initial stage of this section describes the details of input data preprocessing for object memorability prediction process. Later part describes the proposed deep learning based object memorability prediction models.

3.2.1 Input Data Preprocessing

In order to predict the object memorability score, the object segment needs to be extracted from the image. In the existing work [3], the input object segments are prepared by cropping and resizing the object segments, as shown in Figure 3.6. From this figure, it is evident that *Spatial-location* and *Spatial-size* information are distorted while extracting object segments. In contrast to the object segments



Figure 3.6: Examples of input object segments prepared according to the existing work [3]. (a) original image. (b), (c), and (d) are extracted object segments. The extracted object segments convey no *Spatial-location* and *Spatial-size* information.

of Figure 3.6, the images of Figure 3.7 retained *Spatial-location* and *Spatial-size* information as it is. In our proposed work, the input object segments are prepared to retain *Spatial-location* and *Spatial-size* information, as shown in Figure 3.7. It is carried out by assigning a value 128 to all the pixels belonging to RGB channels of the image except the object segment pixels. The nature of convolutional neural network architecture exploits the spatial locality information [4]. Due to this reason, the CNN model trained on object segments which are prepared, as shown in Figure 3.7 utilizes *Spatial-location* and *Spatial-size* details.

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

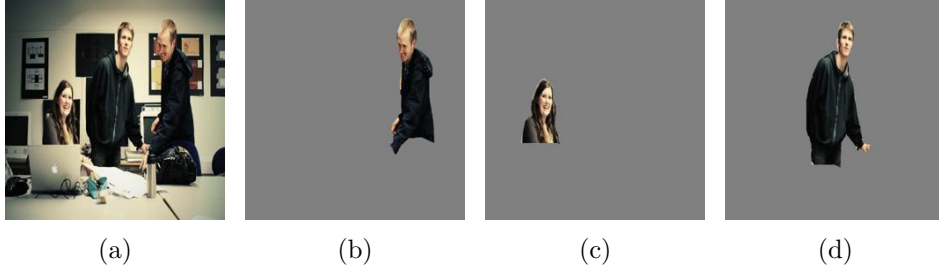


Figure 3.7: Examples of input object segmentation prepared according to the our proposed method. (a) original image. (b), (c), and (d) are extracted object segments. Images in (b), (c) and (d) represent object segments having *Spatial-location* and *Spatial-size* information.

3.2.2 Proposed Object Memorability Prediction Models

In this chapter, three models are proposed to predict the object memorability scores. The first model is named as *SVR for Object Memorability Prediction (SVR-OMP)*, which is similar to the existing model of [3] but differ in object segment preparation. This model is proposed to demonstrate the improvement of object memorability prediction due to the incorporation of *Spatial-location* and *Spatial-size*, which was not considered in [3]. Further, to provide an end-to-end object memorability prediction model, a deep *CNN* model is proposed, which is named as *Deep CNN for Object Memorability Prediction-I (DCNN-OMP_I)*. This model is extended to improve the performance using the modular approach of fine-tuning, which is named as *Deep CNN for Object Memorability Prediction-II (DCNN-OMP_II)*.

3.2.2.1 SVR-OMP Model

The *SVR-OMP* model is built by training the *SVR* on deep features. The architecture of *SVR-OMP* model is shown in Figure 3.8. The object segment is passed through the pre-trained *CNN* model [4] which is trained on *ImageNet* dataset [40], and deep features of 4,096 dimension are extracted from the last fully-connected layer FC7. These deep features are mapped to object memorability scores using *SVR*. To incorporate the *Spatial-location* and *Spatial-size* characteristics of the

object segment, the input image is prepared as proposed in Section 3.2.1.

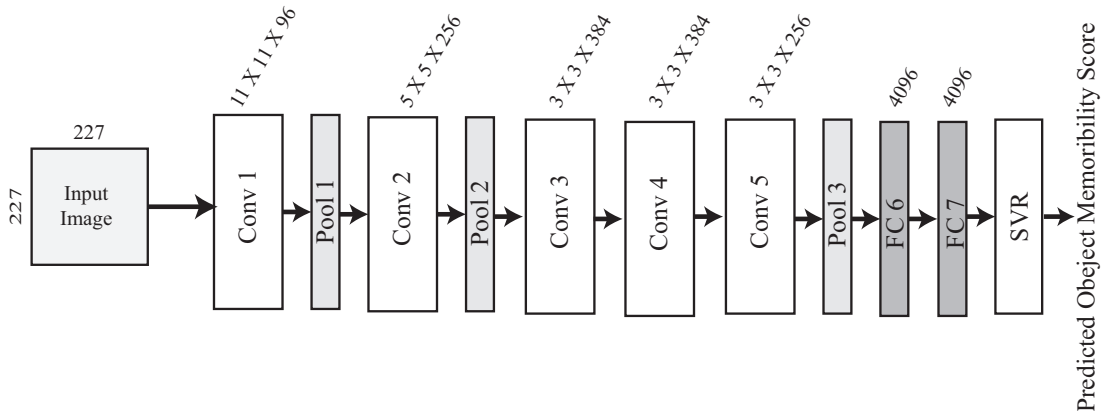


Figure 3.8: The architecture of the proposed SVR-OMP model.

3.2.2.2 DCNN-OMP_I

Though the *SVR-OMP* model can perform better than existing work [3] (Please refer to Section 3.3 for performance comparison), the deep features used to train the *SVR-OMP* are extracted from the model trained to perform object classification task [4]. Hence, the performance achieved through *SVR-OMP* can be further improved with a deep CNN model which is trained to perform object memorability prediction task. However, training a deep CNN model from scratch forces the use of a large amount of labeled dataset to get better performance. As the number of labeled object segments in the object memorability dataset [3] is very less, fine-tuning is employed to construct the deep CNN based object memorability prediction model. It is found from the literature that object memorability and image memorability are positively correlated [3]. Hence, it might be efficient to fine-tune a deep CNN model trained on image memorability prediction task. Recently, Khosla et al. [2] created a large volume of image memorability dataset and developed a deep CNN model on this dataset to predict image memorability scores. This model achieved state-of-the-art performance in image memorability prediction. In the proposed scheme, a deep CNN model for object memorability prediction is built by fine-tuning Khosla et al.’s [2] image memorability predic-

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

tion model. The architecture of the proposed model *DCNN-OMP_I* is depicted in Figure 3.9. The architecture is same as the deep CNN model proposed in [4] except the last layer. The number of output neurons are changed from 1000 to 1, and the output function is changed from $\text{Softmax}()$ to $\text{Sigmoid}()$ to predict object memorability score between 0 and 1.

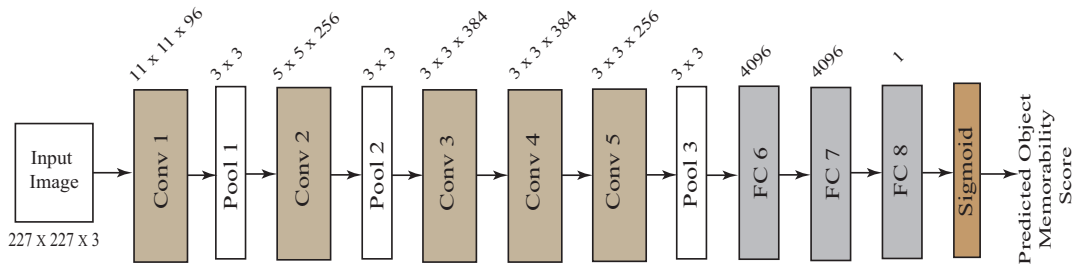


Figure 3.9: The architecture of the proposed Deep CNN model *DCNN-OMP_I*. The architecture is similar to the deep CNN model proposed in [4] except the last layer where the number of outputs is reduced from 1000 to 1 and Softmax layer is replaced by Sigmoid layer to predict object memorability score.

Recently, Anderson et al. [54] proposed a novel approach of fine-tuning, named as *modular approach*, to learn on small data. This approach ensures a better learning method even with the relatively low volume of training samples. In order to improve the performance of object memorability prediction further, the proposed model *DCNN-OMP_I* is extended by employing the *modular approach* of fine-tuning [54]. Accordingly, the architecture of the *DCNN-OMP_I* is modified. The modified model is named as *DCNN-OMP_II* and explained in the next subsection.

3.2.2.3 DCNN-OMP_II

In this section, the brief details of the modular approach [54] for fine-tuning are presented first. Then, based on the modular approach, the architecture of the *DCNN-OMP_I* is modified.

3.2 Object Memorability Prediction

The *modular approach* of fine-tuning [54] treats the entire network as a module during the fine-tuning process. The central idea of the *modular approach* is to merge the pre-trained and untrained modules to understand the shift in distribution between datasets. The effect of combining pre-trained and untrained modules adds new representations to the network model instead of replacing previously learned representations. Equation 3.6 and 3.7 shows the formal representation of regular fine-tuning and *modular approach* of learning from a small volume of data.

$$Y_{ft} = SF(Net(X, W = \{PT_Net\})) \quad (3.6)$$

$$Y_{md} = SF([Net(X, W = \{PT_Net\}), Net(X, W^* = \{PT_Net\})]) \quad (3.7)$$

$SF()$ in both the equations represents Softmax function. In Equation 3.6, Y_{ft} represents the class label predicted from the model built by the regular fine-tuning method. Net represents the deep network which is currently being learned. X is the input to the network. W is the set of trainable weights which are initialized with the weights of the pre-trained network denoted by PT_Net . In Equation 3.7, Y_{md} represents the class label predicted by the model built by the modular approach. W^* represents the non-trainable weights which are initialized with the weights of the pre-trained network denoted by PT_Net . Similar to the *modular approach* proposed by Anderson et al. [54], the *DCNN-OMP_I* model is extended to incorporate the *modular approach* of fine-tuning. The extended model is named as *DCNN-OMP_II*. The architecture of *DCNN-OMP_II* is depicted in Figure 3.10. The architecture contains a stack of two networks. Both these networks' architecture and parameters are similar to the image memorability model [2]. The upper branch of the network is initialized with the weights of the image memorability model [2], and it is frozen from being learned. Therefore, the previously learned feature representations of the image memorability model [2] are retained for object memorability prediction. The lower branch of the network is also initialized with the weights of the image memorability prediction model [2], but it

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

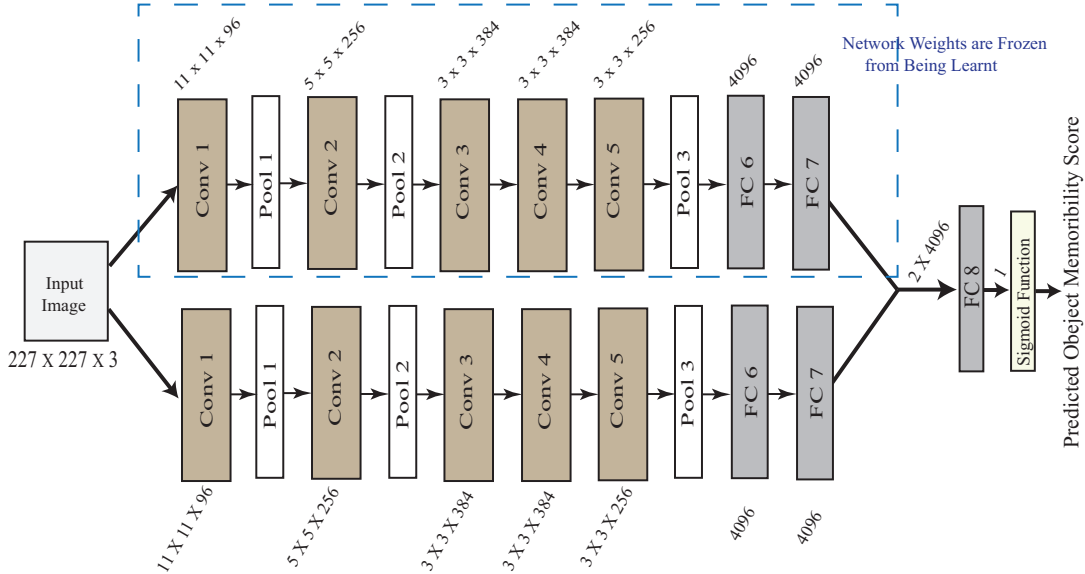


Figure 3.10: The architecture of the proposed Deep CNN model DCNN-OMP_II.

is allowed to learn during the fine-tuning process. This branch of the network learns the new feature representations for object memorability prediction. Both the upper branch and lower branch produces deep features of dimension 4,096. The last layer of the entire network takes these 2×4096 deep features as input and produces one output which is passed through the Sigmoid function to generate object memorability score between 0 and 1.

3.3 Experimental Results

All the experiments are conducted on object memorability dataset created by Dubey et al. [3]. The dataset contains ground-truth memorability scores for 3412 object segments extracted from 850 images. All of these 850 images are sampled from PASCAL-S dataset [47]. For more details of the object memorability dataset, kindly refer Section 2.3.

3.3.1 Experimental Set-up

The proposed models are trained on preprocessed ground-truth object segments. The preprocessing ensures the incorporation of the visual factors *Spatial-size* and *Spatial-location*, as explained in Section 3.2.1. To train the proposed models, the dataset is divided into training and testing sets in six-folds for cross-validation purpose. Further, proposed models are also evaluated on automatically generated object segments similar to Dubey et al. [3]. To generate object segments automatically, a state-of-the-art generic object proposal algorithm, MCG [5] is used.

The loss function, optimizer, and hyper-parameters are same for all the proposed deep CNN models. Object memorability prediction is essentially a regression task. For such tasks, $L2$ loss is the most widely used loss function. Equation 3.8 shows the employed $L2$ loss function.

$$L2 = \sum_j ||Y_j - y_j||_2^2 \quad (3.8)$$

where Y_j and y_j represent the predicted and ground-truth memorability scores of the j^{th} image. To minimize network loss, Ada-delta optimizer is used with an initial learning rate of 0.001. The models are trained with a batch size of 50 images. Preprocessed ground-truth object segments are used to fine-tune the proposed models.

3.3.2 Performance Evaluation

The Spearman’s rank correlation coefficient (ρ) is used to evaluate the performance of the proposed models. Table 3.1 shows the rank correlation coefficient values of existing [3] and proposed object memorability models. The models are evaluated on both ground-truth object segments as well as MCG [5] object segments.

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

Evaluation of the Proposed Models on Ground Truth Object Segments:

To evaluate the proposed models on ground-truth object segments, rank correlation is computed between predicted memorability scores and ground-truth memorability scores. Compared with the existing model [3], the *SVR-OMP* model shows improved performance of object memorability prediction with $\rho = 0.71$. This indicates that incorporation of *Spatial-location* and *Spatial-size* improved the object memorability prediction accuracy indicating *Spatial-location* and *Spatial-size* are two important factors in making object memorable or forgettable. Further, both deep CNN models *DCNN-OMP_I* and *DCNN-OMP_II* performed better with ρ value of 0.74 (almost near to human consistency $\rho = 0.76$). This indicates that deep features learned by fine-tuning of the image memorability model are good predictors of object memorability scores compared to the deep object features representing object category information. Though the *modular approach* of fine-tuning [54] is employed to improve the performance, the result did not improve further for ground-truth object segments. From this, it can be observed that the $\rho = 0.74$ (almost near to human consistency $\rho = 0.76$) may be an upper bound on object memorability prediction.

Table 3.1: Comparison of performance between the existing model [3] and the proposed models *SVR-OMP*, *DCNN-OMP_I* and *DCNN-OMP_II* on ground-truth object segments.

Models	Dubey et al.'s Model [3]	Proposed <i>SVR-OMP</i>	Proposed <i>DCNN-OMP_I</i>	Proposed <i>DCNN-OMP_II</i>
ρ	0.70	0.71	0.74	0.74

Evaluation of the Proposed Models on Automatically Generated Object Segments:

To completely automate the object memorability prediction, object segments are generated from each image present in the object memorability dataset using *MCG* [5] algorithm. For top $K = 20$ generated object segments, memorability scores are predicted using the proposed models. Then, the memorability map is generated for each image by computing average predicted scores

3.3 Experimental Results

of all the object segments belong to that image. Finally, the rank correlation between the average predicted memorability score inside each of the object segments and their ground-truth memorability scores is computed. Compared with the existing model [3], the *SVR-OMP* model shows improved performance of object memorability prediction for *MCG* object segments with $\rho = 0.40$. This proved again that the *Spatial-location* and *Spatial-size* play a role in determining object memorability prediction. However, the proposed deep *CNN* model, *DCNN-OMP_I* model performance ($\rho = 0.39$) is same as the existing model [3]. One of the major reasons for this performance is a relatively low volume of the training dataset. One possible solution to the low volume of the dataset is the *modular approach* of fine-tuning, as discussed in Section 3.2.2.3. Due to the incorporation of the *modular approach*, the extended model *DCNN-OMP_II* improved the performance ($\rho = 0.40$) for *MCG* object segments, as shown in Table 3.2. However, the improvement is still not significantly high for *MCG* object seg-

Table 3.2: Comparison of performance between the existing model [3] and the proposed models *SVR-OMP*, *DCNN-OMP_I* and *DCNN-OMP_II* on *MCG* [5] object segments.

Models	Dubey et al.'s Model [3]	Proposed <i>SVR-OMP</i>	Proposed <i>DCNN-OMP_I</i>	Proposed <i>DCNN-OMP_II</i>
ρ	0.39	0.40	0.39	0.40

ments. Another possible reason for this performance is the quality of the object segments generated using *MCG* [5] algorithms. The poor quality object segment not only disturbs the object category information but also the *Spatial-size* and *Spatial-location* information. Consider Figure 3.11, where the first row depicts the original image, second and third rows depict the ground-truth object segments, and fourth and fifth rows depict the *MCG* object segments. From this figure, it is observed that the *Spatial-size* and *Spatial-location* information are distorted in *MCG* object segments due to improper object proposals. Though the proposed models showed minor improvement for *MCG* object segments, the performance is much better (near to human consistency) than the existing model [3]

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

for ground-truth object segments.

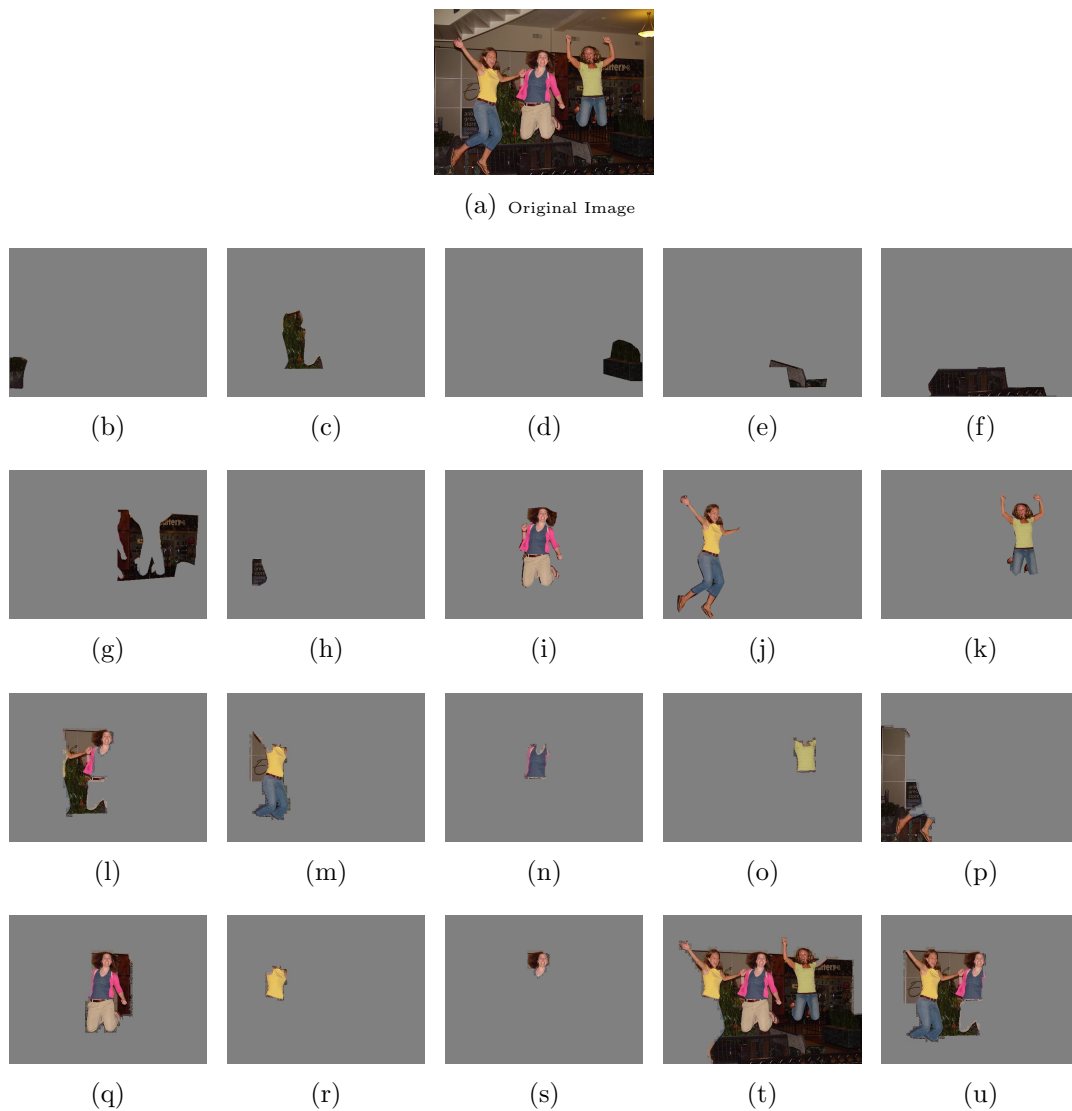


Figure 3.11: *Quality difference between ground-truth and MCG [5] object segments. (a) Original Image, (b-k) Ground-truth object segments and (l-u) Top 10 object segments generated using MCG algorithm.*

Object Memorability Prediction Models' Bias on the *Spatial-size* information: To demonstrate the proposed models' bias on the *Spatial-size* information, object segments are sorted based on predicted memorability scores.

3.3 Experimental Results

Various ranges of these sorted object segments are selected to examine the average *Spatial-size* of object segments on these ranges, as shown in Table 3.3. Object segments are sorted into sets according to predictions made by existing and proposed models (denoted by column headings of Table 3.3). From Table 3.3, it is evident that the proposed models assign the higher ranks to larger *Spatial-size* object segments and lower ranks to smaller *Spatial-size* object segments. It is also evident that *Spatial-size* bias is comparatively less in Dubey et al.’s model [3].

Table 3.3: Comparison of predicted memorability scores versus *Spatial-size* of object segments.

Range of Images	Dubey et al.’s [3]	SVR-OMP	DCNN-OMP_I	DCNN-OMP_II
Top 10	24%	27%	33%	36%
Top 25	26%	27%	33%	34%
Top 50	25%	26%	31%	32%
Top 100	24%	25%	28%	28%
Top 200	21%	22%	22%	22%
Bottom 200	6%	5%	5%	5%
Bottom 100	5%	4%	4%	4%
Bottom 50	5%	4%	4%	4%
Bottom 25	4%	4%	4%	4%
Bottom 10	4%	3%	3%	3%

Object Memorability Prediction Models’ Bias on the *Spatial-location* information: To demonstrate the *Spatial-location* bias of the proposed models, object segments are sorted based on predicted memorability scores and various ranges of these sorted object segments are selected to examine *Spatial-location* information of object segments. An average number of objects located at the center as well as at the corners are computed on these ranges, as shown in Tables 3.4 and 3.5 respectively. Object segments are sorted into sets according to predictions made by existing and proposed models (denoted by row headings of Tables 3.4 and 3.5). The % values of Top 10, 25, 50, 100, and 200 are computed for Tables 3.4 and 3.5 as shown in Equations 3.9 and 3.10 respectively. In the same manner, % values are also computed for the Bottom 10, 25, 50, 100, and

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

200 ranges for the Tables 3.4 and 3.5.

$$Center_Top_i_ \% = 100 * \frac{\sum_{k=1}^F center_obj_count}{F * i} \quad (3.9)$$

$$Corner_Top_i_ \% = 100 * \frac{\sum_{k=1}^F corner_obj_count}{F * C * i} \quad (3.10)$$

In Equations 3.9 and 3.10, variable i takes values from the set $\{10, 25, 50, 100, 200\}$. The $center_obj_count$ represents the total number of images containing objects at the center of the image. The $corner_obj_count$ represents the total number of images containing objects at the corners of the image, F represents the total number of folds that is six and C represents the total number of corners that is four. Further, each range’s result is included in all the other ranges’ results that are bigger than the current range. For example, Top 25 includes Top 10 results, Top 50 includes Top 25, and Top 10 results and so on. From Tables 3.4 and 3.5, it is evident that proposed models place object segments located at the center at higher ranks and object segments located at the corners at lower ranks. It is also evident that location-bias is comparatively less in Dubey et al.’s model [3].

Table 3.4: Comparison of predicted memorability scores versus the average number of objects located at the center of an image.

Range of Images	Dubey et al.’s [3]	SVR-OMP	DCNN-OMP I	DCNN-OMP II
Top 10	60%	74%	76%	77%
Top 25	60%	72%	76%	76%
Top 50	60%	68%	69%	72%
Top 100	59%	60%	61%	63%
Top 200	45%	50%	50%	50%
Bottom 200	11%	6%	4%	4%
Bottom 100	11%	5%	3%	2%
Bottom 50	12%	6%	2%	1%
Bottom 25	11%	7%	2%	1%
Bottom 10	8%	7%	0%	0%

Table 3.5: Comparison of predicted memorability scores versus the average number of objects located at the corners of an image.

Range of Images	Dubey et al.'s [3]	SVR-OMP	DCNN-OMP I	DCNN-OMP II
Top 10	2%	0%	1%	1%
Top 25	3%	1%	1%	1%
Top 50	3%	2%	2%	1%
Top 100	4%	3%	2%	2%
Top 200	5%	3%	3%	3%
Bottom 200	12%	15%	18%	16%
Bottom 100	10%	14%	16%	17%
Bottom 50	10%	13%	15%	17%
Bottom 25	10%	13%	15%	16%
Bottom 10	10%	13%	15%	15%

3.4 Summary

In this chapter, the influence of *Spatial-location* and *Spatial-size* of an object segment in predicting its memorability score is analyzed. A baseline model is developed to demonstrate the improvement of object memorability prediction due to the incorporation of these two characteristics in the prediction method. Next, a deep CNN model is devised for automatic feature learning on these two object characteristics to predict the object memorability scores. The proposed models utilized the *Spatial-location* and *Spatial-size* of an object segment in predicting memorability scores and performed better than the existing work. Further, the proposed deep CNN models showed their capability of scaling the object memorability prediction accuracy up to human consistency based on the quality of the input object segment.

This chapter addressed a few important visual factors which influence memorability at the object level and also devised deep learning models to utilize the proposed visual factors in determining memorability scores. In the next chapter, the influence of few important visual factors such as motion and depth cues on the memorability at image level will be analyzed. Further, deep learning based prediction models are proposed to incorporate the proposed visual factors to predict

3. OBJECT MEMORABILITY PREDICTION: LOCATION AND SIZE BIAS

image memorability.

Image Memorability: The role of Depth and Motion

In the last chapter, it has been shown that the relative size and location of an object within an image play important roles for understanding the corresponding object memorability. In this chapter, the memorability issue has been extended to image level, and it has been shown that in addition to the known visual characteristics such as object and scene semantics, emotion, saliency, etc., image depth and image motion are also playing important roles to understand the image memorability. It is observed from the computer vision literature that depth cues play a vital role in solving many computer vision related problems including image denoising [55, 56], visual saliency prediction [57, 58], image quality assessment [59], etc. Similarly, motion information is used in various vision applications, including video saliency detection [60], single image action recognition [61], etc. However, the influence of motion and depth cues has remained unexplored in case of image memorability. To summarize, the existing literature on image memorability suffer from the following limitations:

- Though depth and motion cues play a major role in solving many computer vision tasks, their influence in determining memorability of an image has remained unexplored.

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

- Most of the existing methods for image memorability predictions rely on hand-crafted features, which are manually engineered and based on peoples' observation. These features may not infer the image memorability related complex high-level visual factors, including motion and depth cues.

Motivated by these observations, this chapter explores the role of depth and motion cues in determining the memorability of an image. Due to the unavailability of a single dataset containing the annotations of depth, motion, and memorability altogether, current state-of-the-art prediction models are used to obtain optical flow (motion cues) and depth maps of still images of *LaMem* dataset [2]. The key contributions of this chapter are as follows:

- Explore the influence of depth and motion cues in determining memorability of an image through predicted optical flow (motion cue) and predicted depth map (depth cue) using state-of-the-art optical flow and depth map prediction models.
- Device deep learning based image memorability prediction models which utilize depth and motion cues along with object semantics to predict memorability scores for the given input image.

Rest of the chapter is organized as follows. The role of the motion and depth in image memorability prediction is presented in Section 4.1. The proposed image memorability prediction models are detailed in Section 4.2. Section 4.3 elaborates on experimental details, and finally, the summary of the chapter is presented in Section 4.4.

4.1 Role of Motion and Depth in Determining Image Memorability

This section explores the influence of motion and depth cues in determining the memorability of an image using publicly available large-scale image memorability dataset *LaMem* [2].

4.1 Role of Motion and Depth in Determining Image Memorability

4.1.1 Motion and Memorability

From the image memorability dataset LaMem [2] it can be observed that images containing objects with motion tend to be more memorable. In order to under-

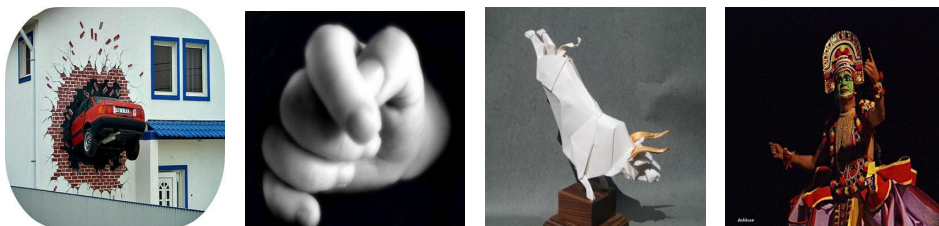


Figure 4.1: *Examples of high-memorable images containing objects with motion.*

stand this relationship further, motion cues of the single static image are essential. In the computer vision literature, a recent scheme [62] has been reported which developed a deep learning based optical flow prediction model to predict dense optical flow from a single static image. The model proposed in [62] takes a still image as input and predicts the optical flow to represent the future motion of every pixel. In this thesis, the motion cue is represented as the optical flow. Therefore, optical flow is obtained using the state-of-the-art deep learning model proposed in [62], and the same is used to study the relationship of memorability with motion cue of an image.

If the motion direction of a pixel in optical flow is different from its surrounding pixels' motion directions, then it is defined as *salient flow-pixel*. Heuristically, the existence of a *salient flow-pixel* may ensure the existence of object(s) with motion within an image. Consider the images in Figure 4.2, which shows the original images and corresponding predicted optical flow superimposed images. From Figure 4.2 it can be observed that images containing objects in motion tend to have more number of *salient flow-pixel* (depicted in the first row of Figure 4.2) than the images containing it can be observed that images containing still objects with less or no motion (depicted in the second row of Figure 4.2) In order to understand this observation further, a *Salient Motion Score (SMS)* is computed

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

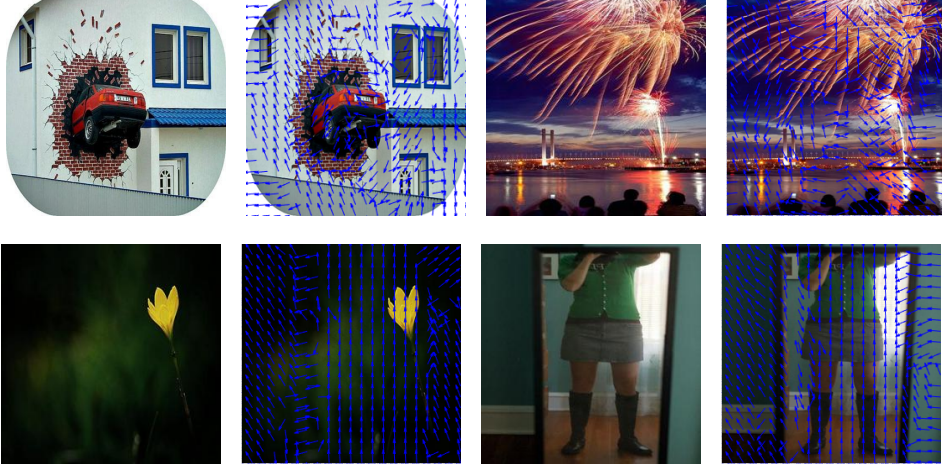


Figure 4.2: Examples of images containing objects with motion and no motion along with corresponding optical flow superimposed images.

for each image. The SMS of an image is an aggregate over pixel-level SMS, which is a measure of the degree to which a pixel’s direction is different from its 3×3 neighborhood. It is mathematically defined in the equation Eq.4.1.

$$SMS = \frac{1}{MN} \left(\sum_{i=1}^{i=M} \sum_{j=1}^{j=N} SMS_{i,j} \right) \quad (4.1)$$

$$SMS_{i,j} = \sum_{k=-1}^{k=+1} \sum_{l=-1}^{l=+1} |dir(i+k, j+l) - dir(i, j)|$$

where M and N are row and column of the image, $dir(i, j)$ is the pixel’s direction at i^{th} row and j^{th} column. The predicted pixel’s direction can take values in the range 1 to 40 i.e. the range 0 to 360 is normalized to the range 1 to 40. From the equation 4.1, difference between direction values 1 and 40 becomes 39 but it is wrong due to cyclicity. The correct difference is 1. To handle this cyclicity, the value of $|dir(i+k, j+l) - dir(i, j)|$ is set to $40 - |dir(i+k, j+l) - dir(i, j)|$ if $|dir(i+k, j+l) - dir(i, j)| > 20$. Interestingly, a meaningful rank correlation of 0.23 is found between image memorability score and SMS on *LaMem* dataset [2]. This correlation indicates that motion information positively influences the memorability.

4.1 Role of Motion and Depth in Determining Image Memorability

4.1.2 Depth and Memorability

Similar to motion information, depth is another vital image cue whose relationship with memorability is remained unexplored. This subsection sheds light on how depth cues affect the memorability of an image. The publicly available memorability datasets do not contain ground-truth depth information. Hence, the current state-of-the-art depth estimation model [63] is used to obtain depth maps for the memorability dataset. The depth map prediction model proposed in [63] takes an RGB image as input and produces the corresponding depth map. Consider Figure 4.3, which depicts high memorable images with corresponding

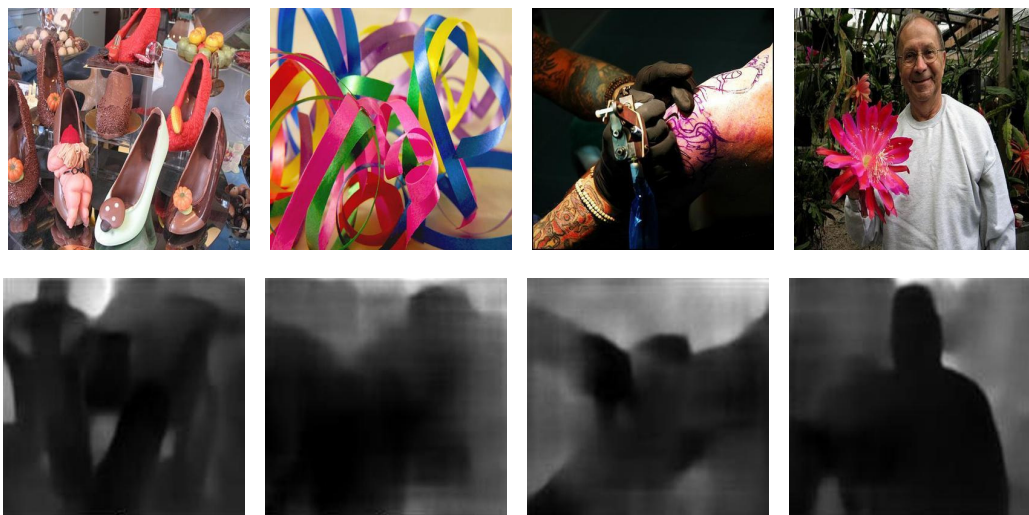


Figure 4.3: Examples of high memorable images (first row) with their corresponding predicted depth values (second row).

predicted depth maps. It is evident from Figure 4.3 that images having lesser depth values at the center tend to be more memorable. Similarly, consider Figure 4.4, which shows low memorable images with corresponding predicted depth maps. It is visible that images with higher depth values at the center tend to be less memorable. To further understand this observation, a *Weighted Depth Score* (WDS) is computed for each image using equation Equation 4.2.

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

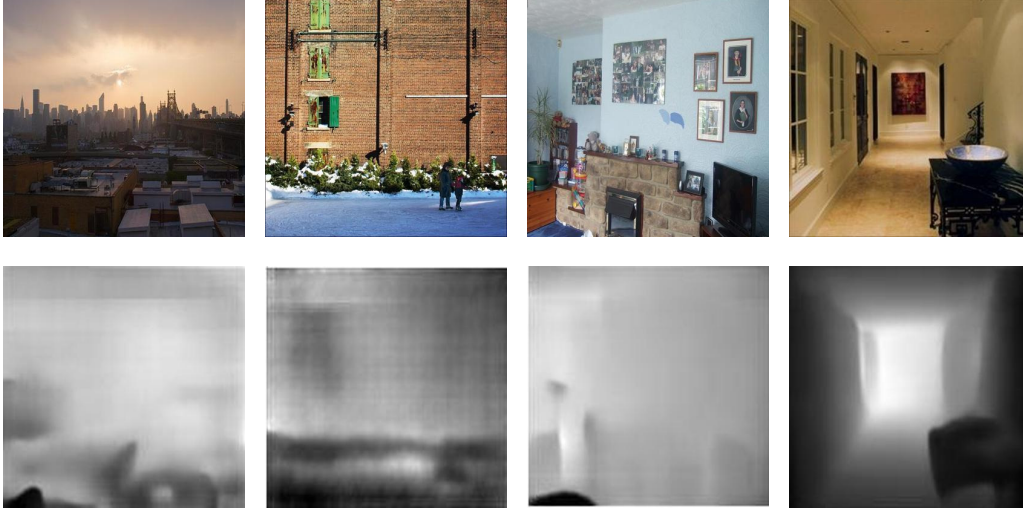


Figure 4.4: Examples of low memorable images (first row) with their corresponding predicted depth maps (second row).

$$\begin{aligned}
 WDS &= \sum_{i=1}^M \sum_{j=1}^N \frac{1}{d_{i,j}} (Pix(i, j)) \\
 \text{where, } d_{i,j} &= \sqrt{(i - i_{mid})^2 + (j - j_{mid})^2}, \\
 Pix(i, j) &= \text{current_pixel's_depth_values}, \\
 i_{mid} &= \text{Middle_pixel's_row_number}, \\
 j_{mid} &= \text{Middle_pixel's_column_number}, \\
 i &= 1, 2, \dots, M(\text{row_size}), \\
 j &= 1, 2, \dots, N(\text{column_size}).
 \end{aligned} \tag{4.2}$$

Higher **WDS** of an image indicates that it has farther depth values at the center and vice versa. On computing **WDS** for *LaMem* dataset [2], a reasonable rank correlation of 0.17 is observed between **WDS** and image memorability score. This analysis suggests that depth is one of the influential cues in determining image memorability.

4.2 Memorability Prediction

In this section, two deep learning based image memorability prediction models are proposed. In the first model (*OFD-MemNet-I*), the deep CNN models which were trained to predict depth map and optical flow are fine-tuned on image memorability dataset to utilize depth and motion cues. The fine-tuned models are ensembled with *MemNet* [2] to find the final predicted memorability scores. In the second model (*OFD-MemNet-II*), two copies of *LaMem* are created. Images belong to the first copy of *LaMem* dataset are superimposed with their corresponding depth maps. Images belong to the second copy of *LaMem* dataset are superimposed with their corresponding optical flow maps. To utilize depth cues to predict memorability scores, *VGG-16* is fine-tuned on depth imposed *LaMem* dataset. Similarly, to use motion cues in memorability prediction, another copy of *VGG-16* is fine-tuned on optical flow imposed *LaMem* dataset. Also, the third copy of *VGG-16* is fine-tuned on *LaMem* dataset without any super-imposition to utilize deep object features in memorability prediction. Finally, all three *VGG-16* are ensembled to obtain the final predicted memorability score.

4.2.1 Proposed OFD-MemNet-I

The proposed model, *OFD-MemNet-I* (Object Flow Depth-MemNet-I), devises multiple memorability scores based on motion, depth, and object features. These scores are ensembled to obtain the final memorability score for the given input image. The architecture of the proposed *OFD-MemNet-I* is shown in Figure 4.5, which consists of three deep CNN branches:

MemNet: To utilize the deep object features, the existing deep CNN model *MemNet* [2] is directly adopted. The model has acquired rich deep object feature representations which were learned on various datasets, including *ImageNet* [40], Places Database [39] and *LaMem* dataset [2]. As stated in [2], the *MemNet* predictions have a rank correlation of 0.64 with their corresponding ground-truth

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

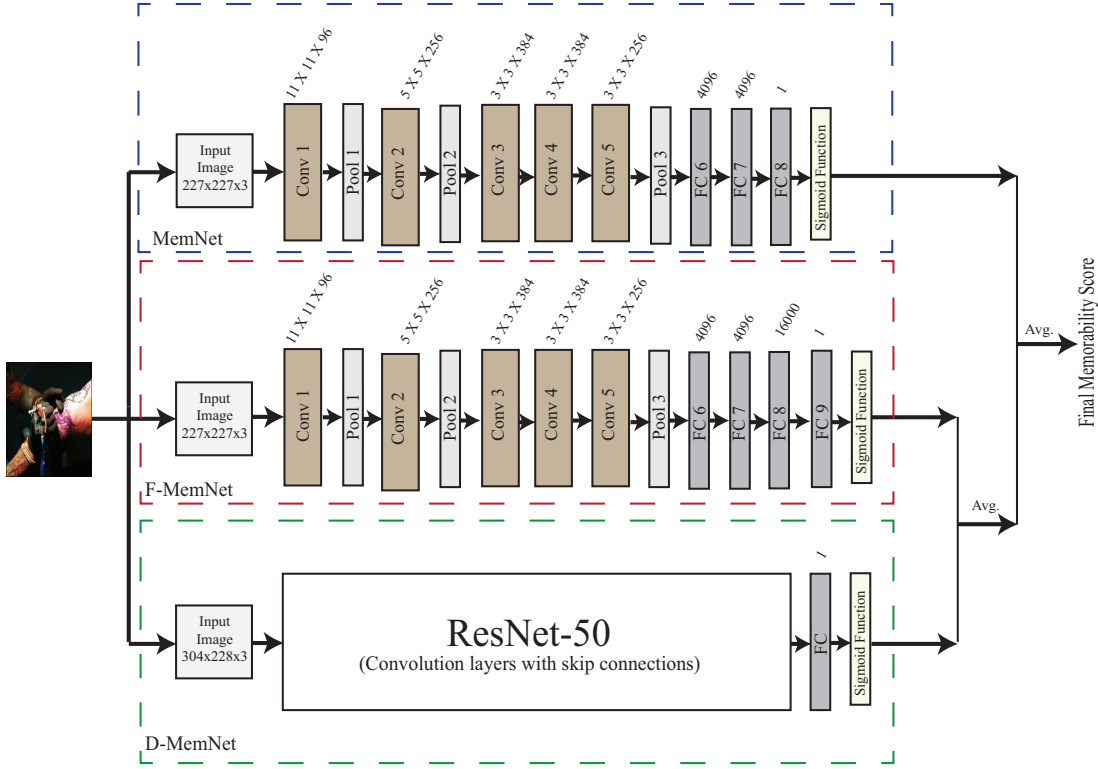


Figure 4.5: The architecture of the proposed Deep *CNN* model *OFD-MemNet-I*.

scores. The memorability score, say Y_{o-mem} predicted from *MemNet* for the input image, say X is defined in Equation 4.3.

$$Y_{o-mem} = \sigma(X_{o-FC}) \quad (4.3)$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad (4.4)$$

where X_{o-FC} is the single output value extracted from the last fully connected layer of the *MemNet* and $\sigma(\cdot)$ is the Sigmoid function defined as shown in Equation 4.4.

F-MemNet: To use the motion cues in predicting image memorability scores, the dense optical flow prediction model (*F-CNN*) [62] is fine-tuned on *LaMem* dataset [2]. The *F-CNN* model has rich deep motion feature representations

learned on realistic video frames. It takes a static image as input and predicts the optical flow representing future motion. The *F-CNN* is fine-tuned on *LaMem* dataset, termed as *F-MemNet* (Flow-MemNet), to map fine-tuned deep motion features to memorability scores. Predictions from the proposed *F-MemNet* model yield a rank correlation of 0.55 with their corresponding ground-truth scores. The memorability score, say Y_{f-mem} predicted from *F-MemNet* for the input image, say X is defined in Equation 4.5.

$$Y_{f-mem} = \sigma(X_{f-FC}) \quad (4.5)$$

where X_{f-FC} is the single output value extracted from the last fully connected layer of the *F-MemNet* and $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4.

D-MemNet: To incorporate depth cues in determining memorability of an image, the deep residual network proposed in [36] loaded with pre-trained weights from the state-of-the-art depth estimation model [63] is fine-tuned on *LaMem* dataset [2]. Due to pre-trained weights, the fine-tuned model, termed as *D-MemNet* (Depth-MemNet), is able to map the fine-tuned depth features to image memorability scores. Predicted memorability scores from *D-MemNet* model produced a rank correlation of 0.63 with their corresponding ground-truth scores. The memorability score, say Y_{d-mem} predicted from *D-MemNet* for the input image, say X is defined in Equation 4.6.

$$Y_{d-mem} = \sigma(X_{d-FC}) \quad (4.6)$$

where X_{d-FC} is the single output value extracted from the last fully connected layer of the *D-MemNet* and $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4.

All three models are trained independently, and their corresponding memorability scores are ensembled as shown in Figure 4.5 to obtain the final memorability score of the *OFD-MemNet-I*. Ensembling multiple deep network outputs not only

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

increases the model accuracy but also avoids model over-fitting [64]. *Basic Ensemble Method* (BEM) is the simple approach to combine network outputs by means of arithmetic mean [65]. Given n networks outputs $\{y_1, y_2, \dots, y_n\}$, the BEM output Y_{BEM} is defined in Equation 4.7.

$$Y_{BEM} = \frac{1}{n} \sum_i^n y_i \quad (4.7)$$

The individual performance of the *F-MemNet* and *D-MemNet* are relatively lower than the existing model *MemNet* [2] (refer Table 4.1). Therefore, the outputs of *D-MemNet* and *F-MemNet* are aggregated first and then the aggregated result is further aggregated with the output of *MemNet* to obtain the final memorability score say, Y_{OFD-I} of the *OFD-MemNet-I* as given in Equation 4.8.

$$Y_{OFD-I} = \frac{\frac{Y_{f-mem} + Y_{d-mem}}{2} + Y_{O-mem}}{2} \quad (4.8)$$

Experimental details are discussed in Section 4.3. The proposed model *OFD-MemNet-I* performed better than state-of-the-art image memorability prediction model *MemNet* by achieving a rank correlation of 0.655. However, the individual performance of the *F-MemNet* and *D-MemNet* are relatively lower than the existing model *MemNet* [2]. *F-MemNet* and *D-MemNet* are obtained by fine-tuning optical flow prediction model [62] and depth map prediction model [63], their performance is limited by the network architectures of the models proposed in [62] and [63] respectively. To address this limitation, we introduced one more model *OFD-MemNet-II*. The proposed *OFD-MemNet-II* will be explained in the next subsection.

4.2.2 Propose Model OFD-MemNet-II

The proposed model *OFD-MemNet-II* consists of three deep learning models, namely *VGG-FMemNet*, *VGG-DMemNet*, and *VGG-MemNet*, as shown in Figure 4.6. The first (*VGG-FMemNet*) model is trained to utilize motion cues to predict memorability scores. The second (*VGG-DMemNet*) model is trained

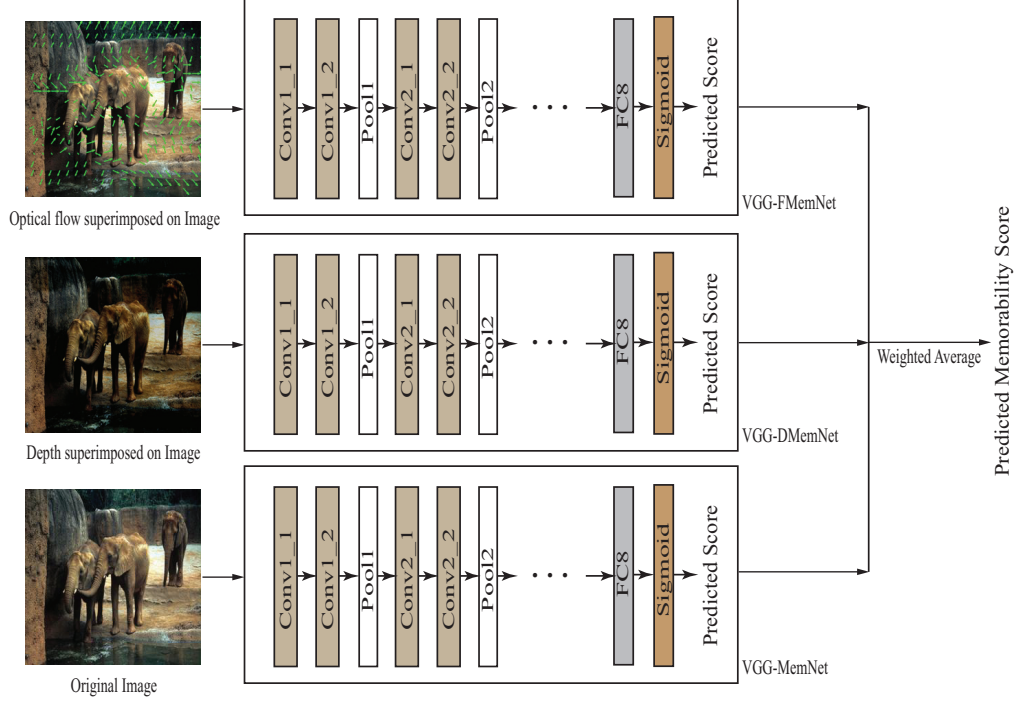


Figure 4.6: The architecture of the proposed model *OFD-MemNet-II*.

to utilize depth cues to predict memorability scores. Finally, the third (*VGG-MemNet*) model attempts to map fine-tuned deep object features to image memorability scores. In the end, the outputs of all three models are ensembled to obtain the final predicted memorability scores. All these three proposed models are fine-tuned on *VGG-16* model [34] and hence, share the same architecture, which is shown in Figure 4.6.

VGG-FMemNet: Instead of fine-tuning the dense optical flow prediction model on memorability dataset, a novel method is proposed which inputs motion cues explicitly to the deep CNN model. For this purpose, optical flow of the given image is predicted using the dense optical flow prediction model proposed in [62]. Then, the predicted sparse Optical flow vectors are super imposed on corresponding RGB image using the Matlab inbuilt function `quiver()`. A quiver plot displays vectors as arrows with components (u,v) at the points (x,y). In this manner,

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION



Figure 4.7: Examples of motion cues superimposed images. First row shows original images and the second row shows predicted optical flow superimposed images.

image memorability dataset *LaMem* is modified to incorporate motion cues. Figure 4.7 shows examples of super-imposition of predicted optical flow on original images. The deep CNN model, *VGG-16* [34] is fine-tuned on modified *LaMem* dataset to utilize motion cues in predicting image memorability scores. In this chapter, *VGG-16* architecture and its weights are employed for the fine-tuning purpose; hence, it is named as *VGG-FMemNet*. However, in place of *VGG-16*, other successful object or image classification models [35, 36], can also be fine-tuned on optical-flow superimposed *LaMem* dataset to utilize motion cues to predict memorability scores. In the fine-tuning process, the last fully-connected layer’s output, say X_{F-FC} is set to 1 to obtain a single score. The entire network’s weights are updated in the fine-tuning process. This model is named as *VGG-FMemNet* and its architecture is shown in Figure 4.8. The memorability score, say Y_{F-mem} predicted from *VGG-FMemNet* for the input image, say X is defined in Equation 4.9.

$$Y_{F-mem} = \sigma(X_{F-FC}) \quad (4.9)$$

where $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4.

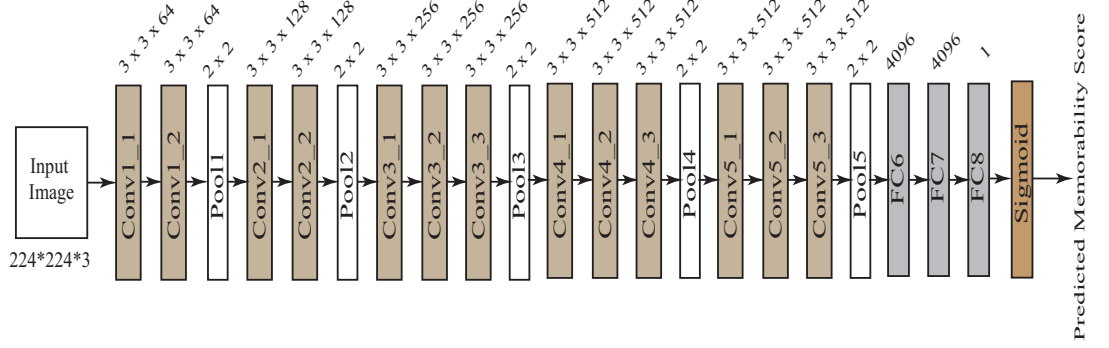


Figure 4.8: The VGG-16 architecture modified and employed to train VGG-FMemNet, VGG-DMemNet, and VGG-MemNet.

VGG-DMemNet: Similar to VGG-FMemNet, a novel method is proposed which inputs depth cues explicitly to the deep CNN model. For this purpose, the depth map for the given image is predicted using the depth prediction model proposed in [63]. Then, the predicted depth values are superimposed on the original image to provide depth cues to the deep CNN model. Consider I is an RGB image of dimension $M \times N \times 3$, D is the corresponding depth image of dimension $M \times N$. D contains values in the range of 0 to 1, where 0 represents the closest depth, and 1 represents the farthest depth. Input image X will be converted from RGB space to $YCbCr$ space and depth is imposed in Y -channel, and finally, the modified image is converted back from $YCbCr$ to RGB space as follows:

$$\begin{aligned}
 I_{YCbCr} &= RGB2YCbCr(X) \\
 DI_{YCbCr}(i, j, 0) &= I_{YCbCr}(i, j, 0) - D(i, j) \\
 DI_{YCbCr}(i, j, 1) &= I_{YCbCr}(i, j, 1) \\
 DI_{YCbCr}(i, j, 2) &= I_{YCbCr}(i, j, 2) \\
 DI &= YCbCr2RGB(DI_{YCbCr})
 \end{aligned} \tag{4.10}$$

where i varies from 0 to $M - 1$ and j varies from 0 to $N - 1$. $RGB2YCbCr()$ is a Matlab function used to convert an image from RGB to YCbCr space, $YCbCr2RGB()$ is a Matlab function used to convert an image from YCbCr to RGB space. This way of imposing depth map on its original image masks

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

farther objects and retains near objects within images. Figure 4.9 shows examples of depth-superimposed images obtained using Equation 4.10 in the second row, corresponding original images in the first row. In the fine-tuning process,



Figure 4.9: *Examples of depth superimposed images. The first row shows original images, and the second row shows the corresponding depth-superimposed images*

the last fully-connected layer’s output, say X_{D-FC} is set to 1 to obtain a single score. The entire network weights are updated in the process of fine-tuning. This model is named as *VGG-DMemNet* and its architecture is same as shown in Figure 4.8. The memorability score, say Y_{D-mem} predicted from *VGG-DMemNet* for the input image, say X is defined in Equation 4.11.

$$Y_{D-mem} = \sigma(X_{D-FC}) \quad (4.11)$$

where $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4.

VGG-MemNet: From the existing literature on image memorability, it can be observed that object semantics are more important in making an image memorable. Therefore, *OFD-MemNet-II* includes object related features along with depth and motion cues to determine memorability score of an image. To utilize object related features in predicting memorability scores, object classifi-

cation model *VGG-16* is fine-tuned on large-scale image memorability dataset *LaMem* [2]. The fine-tuned model is named as *VGG-MemNet*. The proposed model *VGG-MemNet* is similar to *MemNet* proposed in [2]. However, *MemNet* is obtained by fine-tuning *AlexNet* [33] on *LaMem*, but the *VGG-MemNet* is obtained by fine-tuning *VGG-16* on *LaMem* and hence, it is as named *VGG-MemNet*.

In the fine-tuning process, the last fully-connected layer’s output, say X_{O-FC} is set to 1 to predict memorability score. The entire network weights are updated in the fine-tuning process. This model is named as *VGG-MemNet* and its architecture is the same as shown in Figure 4.8. The memorability score, say Y_{O-mem} predicted from *VGG-MemNet* for the input image, say X is defined in Equation 4.12.

$$Y_{O-mem} = \sigma(X_{O-FC}) \quad (4.12)$$

where $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4. The final predicted memorability score Y_{OFD-II} for the given input image X is obtained by combining the three networks outputs *VGG-FMemNet*, *VGG-DMemNet*, and *VGG-MemNet* using **BEM** in Equation 4.13.

$$Y_{OFD-II} = \frac{Y_{F-mem} + Y_{D-mem} + Y_{O-mem}}{3} \quad (4.13)$$

where Y_{F-mem} , Y_{D-mem} , and Y_{O-mem} are the memorability scores generated by *VGG-FMemNet*, *VGG-DMemNet*, and *VGG-MemNet*, respectively. Though each of the networks *VGG-FMemNet*, *VGG-DMemNet*, and *VGG-MemNet* are intended to utilize different features, they are performing equally well, as shown in Table 4.1. Hence, ensembling is carried out with equal weights.

4.3 Experiments and Results

This section details the experiments and corresponding results to demonstrate the performance of the proposed model on image memorability prediction. Three publicly available datasets (*LaMem* [2], Isola et al.’s memorability dataset [1] and

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

Table 4.1: Performance comparison of the existing and proposed models along with Human Consistency (ground-truth).

Dataset	MemNet [2]	OFD-MemNet-I (Proposed)	F-Memnet (Proposed)	D-MemNet (Proposed)	VGG-MemNet (Proposed)	VGG-FMemNet (Proposed)	VGG-DMemNet (Proposed)	OFD-MemNet-II (Proposed)	Human Consistency (ground-truth)
LaMem [2]	0.640	0.655	0.55	0.63	0.650	0.652	0.654	0.671	0.68
Isola [1]	0.61	0.63	0.53	0.59	0.638	0.638	0.641	0.667	0.75
Dubey [3]	0.450	0.47	0.36	0.45	0.492	0.495	0.497	0.515	0.76

Dubey et al.’s memorability dataset [3]) are used in the experiments. The Spearman’s rank correlation coefficient (ρ) is employed to evaluate the performance of the proposed model.

4.3.1 Experimental Set-up

The training process of *VGG-MemNet*, *F-MemNet*, and *D-MemNet* are similar to that employed in [2]. However, all these models are initialized with different pre-trained models weights. For example, *VGG-MemNet* is loaded with *VGG-16* model, *F-MemNet* is loaded with *F-CNN*. All these models have fine-tuned on LaMem [2] dataset. In the fine-tuning process, all the layers are allowed to learn, and the output of the last fully-connected layer of all these models are set to one for single score prediction. *VGG-MemNet* has fed with an image of $224 \times 224 \times 3$ dimension, *F-MemNet* is fed with an image of $227 \times 227 \times 3$ dimension, and *D-MemNet* is fed with an image of $304 \times 228 \times 3$ dimension. The training process of *VGG-FMemNet* is same as *VGG-MemNet*. However, the input image is modified to superimpose optical flow, as explained in Section 4.2.2. Similarly, the training process of *VGG-DMemNet* is conducted with the depth superimposed RGB input.

The loss function, optimizer, and hyperparameters are same for all the networks. Image memorability prediction is essentially a regression task. For such tasks, L2 loss is the most widely used loss function. The equation Eq.4.14 shows the loss function employed in this work.

$$L2 = \sum_j \|Y_j - y_j\|_2^2 \quad (4.14)$$

where Y_j and y_j represent the predicted and ground-truth memorability scores of the j^{th} image. To minimize the network loss, Ada-delta optimizer is used with an initial learning rate of 0.001. The models are trained with a batch size of 20 images.

4.3.2 Performance Evaluation

The *LaMem* contains 60,000 images and divided into five sets for cross-validation purpose. Each set contains 45,000 training samples, 10,000 testing samples, and 3,741 validation samples. Accordingly, five models are trained, tested, and results are averaged. Further, the trained models are tested on other publicly available memorability datasets. Table 4.1 presents the performance comparison of the existing and proposed models along with Human Consistency (ground-truth) on three publicly available memorability datasets. The performance is represented by means of rank correlation (ρ) computed between ground-truth, and predicted memorability scores.

From Table 4.1 it is evident that the proposed *OFD-MemNet-I* performs better than *MemNet* [2] by achieving a rank correlation of 0.655. However, the individual performance of *F-MemNet* and *D-MemNet* (which are part of the *OFD-MemNet*) are relatively lower to the state-of-the-art image memorability prediction model *MemNet* [2]. However, *F-MemNet* and *D-MemNet* improve performance when aggregated with *MemNet* [2]. The reason behind individual lower performance of *F-MemNet* and *D-MemNet* may be the underlying architecture, i.e., the ρ value of 0.63 and the ρ value of 0.55 is the maximum performance which can be achieved by means of fine-tuning the depth prediction model [63] and optical flow prediction model [62] respectively. One more possible reason for the curtailed individual performance of *F-MemNet* may be the absence of an abundant number of images which contains objects with motion in *LaMem* dataset.

Interestingly, the proposed *OFD-MemNet-II* outperforms both *OFD-MemNet-*

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

I and *MemNet* [2] by achieving a rank correlation of 0.671, which is near to human consistency ($\rho = 0.68$). From Table 4.1, it can be observed that *VGG-MemNet* is ranking the images better than *MemNet*. This improvement in *VGGMemNet* is expected as underlying architecture of *VGG-MemNet* (*VGG-16* [34]) has shown better accuracy in image classification task [40] than the underlying architecture of *MemNet* (*AlexNet* [33]). From the results, it is also evident that the performances of *VGG-FMemNet* performed better than *F-MemNet* and *VGG-DMemNet* performed better than *D-MemNet*. This improvement indicates that explicit incorporation of depth and motion cues enabled the underlying architecture to learn motion and depth related features to predict image memorability scores more accurately.

In order to analyze how well the image memorability prediction models are ranking the images, we employed the variation of precision-recall task defined by Isola et al. [1]. In this analysis, images are arranged in descending order of predicted memorability scores. Various ranges (such as “Top 10”, “Top 100”, “Bottom 25”, etc.) of these sorted images are selected and examined the average ground-truth memorability scores on these ranges. Table 4.2 shows this analysis on *LaMem* dataset [2]. Images are sorted into sets according to predictions made by existing and proposed models (denoted by column headings of Table 4.2). The same analysis is also carried out on the other two datasets. Table 4.3 shows corresponding results on Isola et al. dataset [1] and Table 4.4 shows corresponding results on Dubey et al. dataset [3]. From Tables 4.2, 4.3, and 4.4, it is evident that the proposed models (*VGG-MemNet*, *VGG-FMemNet*, *VGG-DMemNet*, and *OFD-MemNet-II*) ranks the images better than the existing model (*MemNet*) based on image memorability property. For example, Table 4.2 shows that “Bottom 10” images ranked based on *MemNet* prediction are 58.06% memorable. Whereas “Bottom 10” images ranked based on *OFD-MemNet-I* prediction are 45.94% memorable. This observation indicates images which are ranked with lower values by *OFD-MemNet-I* are least memorable to human visual system than the images which are ranked with lower values of *Mem-*

Net. Similarly, from Table 4.3 and Table 4.4, it is evident that top ranked images of *OFD-MemNet-II* model are more memorable than the top ranked images of *MemNet* model. Also, the bottom ranked images of *OFD-MemNet-II* model are less memorable than bottom ranked images of *MemNet* model.

Table 4.2: Comparison of predicted versus ground-truth image memorability scores on *LaMem* dataset [2]. Images are arranged in descending order of predicted memorability scores. Various ranges of these sorted images are selected. The average ground-truth memorability scores are shown for each set in each row. The reported results are averaged over 5-fold cross-validation models.

Range of Sorted Images	MemNet [2]	OFD-MemNet-I (Proposed)	VGG-MemNet (Proposed)	VGG-FMemNet (Proposed)	VGG-DMemNet (Proposed)	OFD-MemNet-II (Proposed)	Ground-truth
Top 10	91.70%	91.90%	90.85%	92.44%	92.86%	92.86%	100%
Top 25	90.40%	90.99%	90.22%	91.34%	91.88%	91.92%	100%
Top 50	89.57%	90.29%	90.18%	90.50%	91.02%	91.10%	99.35%
Top 100	89.17%	89.50%	89.99%	89.59%	90.05%	90.15%	98.45%
Top 200	88.91%	89.05%	89.52%	88.63%	88.99%	89.07%	97.57%
Bottom 200	55.06%	54.40%	54.94%	51.70%	51.52%	50.82%	42.16%
Bottom 100	54.35%	53.50%	53.7%	50.09%	49.93%	49.12%	39.01%
Bottom 50	54.20%	54.04%	52.02%	48.88%	48.62%	47.80%	36.3%
Bottom 25	54.44%	53.60%	50.88%	47.91%	47.50%	46.85%	34.41%
Bottom 10	58.06%	55.51%	48.76%	49.97%	46.30%	45.94%	33.57%
ρ	0.64	0.655	0.650	0.652	0.654	0.671	0.680

Table 4.3: Comparison of predicted versus ground-truth image memorability scores on *Isola et al.* dataset [1]. Uses same measures as detailed in Table 4.2.

Range of Sorted Images	MemNet [2]	OFD-MemNet-I (Proposed)	VGG-MemNet (Proposed)	VGG-FMemNet (Proposed)	VGG-DMemNet (Proposed)	OFD-MemNet-II (Proposed)	Ground-truth
Top 10	80.16%	81.11%	81.23%	83.33%	82.57%	82.75%	96.54%
Top 25	75.46%	79.3%	80.09%	79.85%	80.59%	81.83%	94.39%
Top 50	75.13%	77.89%	78.96%	78.84%	80.09%	80.18%	92.24%
Top 100	74.32%	76.7%	77.13%	77.42%	77.87%	78.65%	89.59%
Top 200	73.58%	75.36%	75.75%	75.25%	76.43%	76.51%	85.33%
Bottom 200	35.91%	35.46%	35.23%	36.09%	35.63%	34.58%	22.85%
Bottom 100	32.8%	31.94%	31.64%	33.20%	32.51%	31.17%	18.66%
Bottom 50	30.14%	29.23%	30.06%	31.20%	30.01%	29.07%	14.93%
Bottom 25	28.81%	28.39%	28.69%	30.60%	28.03%	27.64%	10.95%
Bottom 10	28.29%	27.83%	26.59%	28.63%	23.81%	23.33%	5.69%
ρ	0.610	0.63	0.638	0.638	0.641	0.667	0.750

4.4 Summary

This chapter made the first attempt to understand the influence of motion and depth cues on image memorability and found that these two cues have a positive role in determining memorability. This chapter also proposed two novel image

4. IMAGE MEMORABILITY: THE ROLE OF DEPTH AND MOTION

Table 4.4: Comparison of predicted versus ground-truth image memorability scores on Dubey et al. dataset [3]. Uses same measures as detailed in Table 4.2.

Range of Sorted Images	MemNet [2]	OFD-MemNet-I (Proposed)	VGG-MemNet (Proposed)	VGG-FMemNet (Proposed)	VGG-DMemNet (Proposed)	OFD-MemNet-II (Proposed)	Ground-truth
Top 10	84.72%	85.76%	86.84%	85.52%	85.14%	85.76%	92.23%
Top 25	83.79%	84.43%	85.21%	84.02%	84.39%	84.68%	91.23%
Top 50	83.69%	84.27%	84.69%	82.88%	83.30%	83.54%	90.13%
Top 100	83.36%	83.58%	83.61%	81.77%	82.11%	82.31%	88.65%
Top 200	82.17%	82.35%	82.50%	80.54%	80.75%	80.96%	86.94%
Bottom 200	71.8%	71.56%	71.28%	67.89%	67.46%	67.35%	65.17%
Bottom 100	69.85%	69.76%	68.91%	64.90%	64.57%	64.33%	60.87%
Bottom 50	68.01%	67.95%	66.33%	62.55%	62.43%	61.87%	57.15%
Bottom 25	66.36%	66.62%	65.74%	61.39%	61.18%	60.51%	54.2%
Bottom 10	63.98%	64.49%	66.00%	60.39%	59.57%	59.22	50.92%
ρ	0.450	0.47	0.492	0.495	0.497	0.515	0.76

memorability prediction models, which exploits motion and depth cues along with object features. The proposed models outperform the state-of-the-art model, *MemNet* [2], by achieving a near human consistency rank correlation.

In the next chapter, another important visual factor, visual emotion is utilized to predict image memorability scores. Visual emotion is found important in making an image memorable from the existing literature. However, emotion features are not considered in the existing deep learning based image memorability prediction models. In the next chapter an MIL based deep learning model is developed which utilizes emotion cues from multiple local salient regions as well as from single global region of an input image to predict memorability score.

Visual Emotion based Image Memorability Prediction using Multiple Instance Learning

In the last chapter, it has been shown that how image motion and depth influence the overall image memorability. In this chapter, image memorability has been studied with more high-level features such as visual emotion.

Visual emotion is one of the important visual factors which play a vital role in solving many computer vision tasks, including eye fixation prediction [66], image retrieval [67], cloud gaming framework [68], and many more. From the existing literature on image memorability discussed in Section 1.2 of Chapter 1, it is found that visual emotions are important in making an image memorable or forgettable. However, the existing image memorability prediction models fall short in utilizing visual emotion cues in predicting memorability scores and suffer from the following limitations:

- Most of the existing methods for image memorability predictions rely on hand-crafted features, which are manually engineered and based on peoples' observation. These features may not be able to completely infer the image memorability related complex high-level visual factors, including object semantics, and visual emotions.

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

- Recently, few attempts have been made to learn the memorability related features automatically using deep learning based image memorability prediction models [2, 25]. However, these models are obtained by fine-tuning object classification models to utilize object and scene semantics. Hence, the learned features are limited to object and scene semantics and not considered emotion cues in particular.

These shortcomings are addressed in this chapter with the following contributions:

- A novel deep learning based image memorability prediction model is proposed, which learns automatically various image memorability related complex high-level visual factors and utilizes the same in predicting memorability scores.
- The proposed deep learning model employs multiple instance learning framework specifically for the utilization of emotions evoking from global and multiple salient local regions of an image to predict image memorability scores.

The overview of the proposed deep learning model *Ens-MemNet* is depicted in Figure 5.1. The *Ens-MemNet* architecture contains two main branches. The upper branch (*VGG-MemNet*) is a deep CNN which is fine-tuned on *VGG-16* [34] to map deep object features to memorability scores. The lower branch (*MCDR-MemNet*) maps deep emotion features (extracted from a single global view and multiple salient local regions) to memorability scores by means of MIL. The final memorability score of the given input image is obtained by ensembling the scores generated from the upper and lower branch of the *Ens-MemNet*. An extensive set of experiments is conducted on publicly available image memorability datasets to evaluate the performance of the proposed model.

Rest of the chapter is organized as follows. Section 5.1 presents the proposed image memorability prediction model. Experimental details and corresponding

5.1 Proposed Model for Image Memorability Prediction

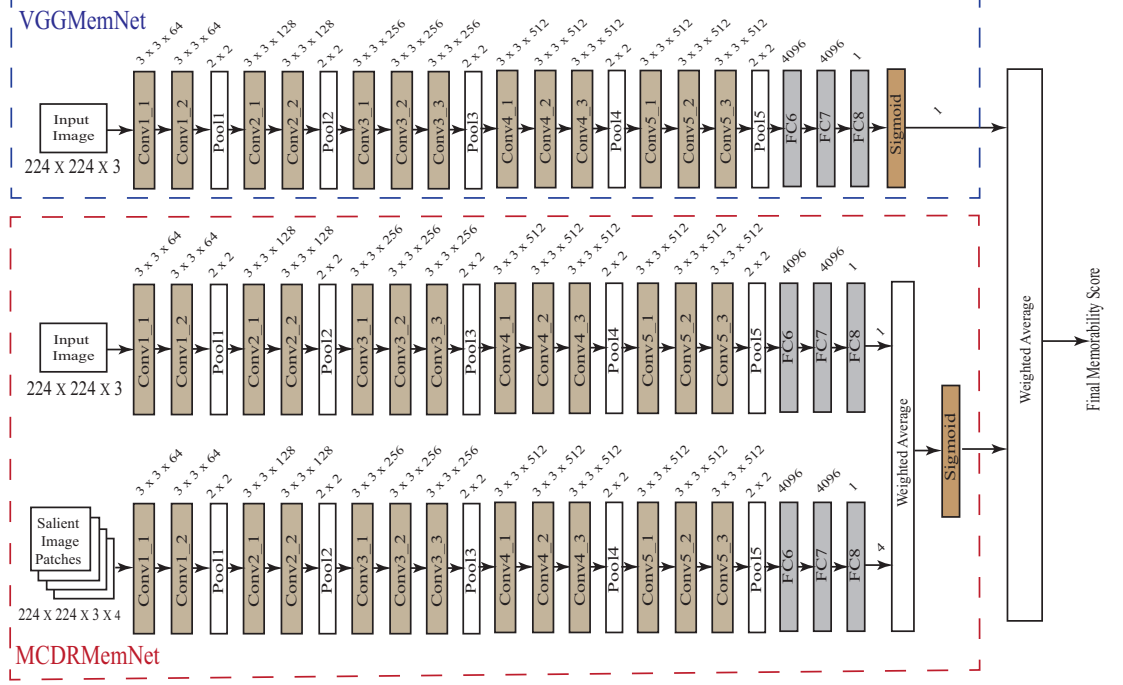


Figure 5.1: The architecture of the proposed model *Ens-MemNet*.

results are discussed in Section 5.2. Finally, the summary of the chapter is presented in Section 5.3.

5.1 Proposed Model for Image Memorability Prediction

This section presents a detailed description of the proposed model *Ens-MemNet* with its two branches, *MCDR-MemNet* and *VGG-MemNet*. The *MCDR-MemNet* attempts to extract and utilize the emotion cues from a single global region and multiple salient local regions by means of MIL framework to predict memorability scores. The *VGG-MemNet* attempts to map fine-tuned deep object features to image memorability scores. Finally, the *Ens-MemNet* ensembles the outputs of *MCDR-MemNet* and *VGG-MemNet* to obtain the memorability scores which are predicted based on object, saliency, and emotion cues.

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

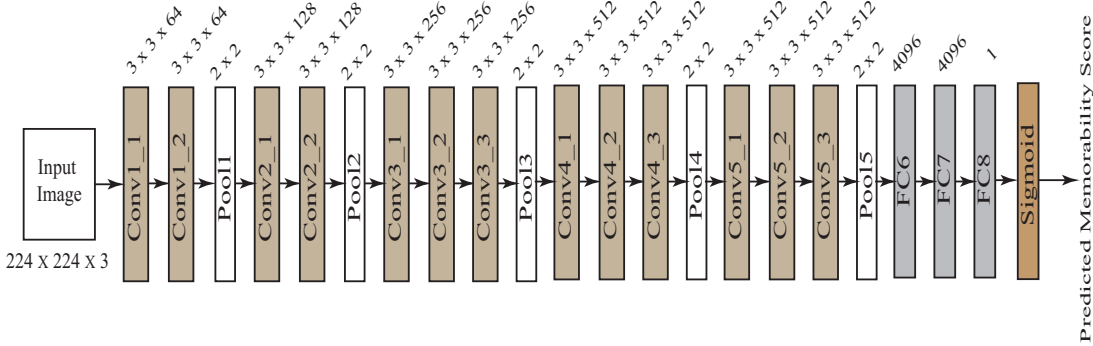


Figure 5.2: The architecture of the proposed VGG-MemNet model.

5.1.1 Deep CNN for mapping object features to memorability scores

It is evident from the existing literature that object semantics play a major role in determining the image memorability scores. Therefore, the proposed model considers object features, along with the emotion and saliency cues to predict image memorability scores. To utilize the advantage of object features in image memorability prediction, the framework proposed in [2] is employed. In [2], Khosla et al. employed the deep CNN model, *AlexNet*, proposed in [33] to map deep object features with memorability scores. Their model was pre-trained on two datasets: *ILSVRC 2012* [40] and *Places* [39]. Recently, *AlexNet* accuracy for image classification task [40] is extended by other successful deep learning models, including *VGG* [34], *GoogLeNet* [35] and *ResNet* [36] and their variants. In this chapter, the *VGG-MemNet* proposed in Section 4.2.2 of Chapter 4 is utilized to predict memorability scores. The *VGG-MemNet* is obtained by fine-tuning the *VGG-16* model [34] on *LaMem* dataset [2]. In place of *VGG-16*, any other deep learning based image classification model can be employed. In the fine-tuning process, the last fully-connected layer's output, say X_{or} is set to 1 to predict memorability score. The entire network weights are updated in the fine-tuning process. This model is named as *VGG-MemNet* and its architecture is shown in Figure 5.2. The memorability score, say Y_{O-mem} predicted from *VGG-MemNet*

5.1 Proposed Model for Image Memorability Prediction

for the input image, say X is defined as given in Equation 5.1.

$$Y_{O-mem} = \sigma(X_{or}) \quad (5.1)$$

where X_{or} is the output of the last fully-connected layer of *VGG-MemNet* and $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4.

5.1.2 MIL based Deep CNN for mapping emotion features to memorability scores

In this subsection, a novel deep learning model, *MCDR-MemNet* is proposed to utilize emotion and saliency information in predicting image memorability scores. The *MCDR-MemNet* learns multi-context deep emotion feature representations from multiple salient local regions within the image as well as from the global view of the entire image. These multi-context deep emotion feature representations are aggregated under MIL framework to predict memorability score. Before presenting *MCDR-MemNet*, a brief description of MIL is presented.

5.1.2.1 Multiple Instance Learning (MIL) Framework

MIL is a weakly supervised learning framework. In this framework, the learner receives bags, each containing multiple training instances. However, the ground-truth label is provided for the entire bag, and no label information is provided for the instances within each bag [69]. The basic underlying assumption of the MIL framework states that all the positive bags must contain at least one positive instance, but all the negative bags must contain only negative instances [69]. MIL framework is initially proposed for drug design [70], and later it is used in computer vision domain to address various kinds of problems including object detection [71], visual categorization [72], image retrieval [73]. Most of these works have applied MIL on hand-crafted features to achieve the given task. Recently, MIL framework is also applied on deep representations to solve various computer vision problems. Song et al. [74] proposed a weakly-supervised framework which combines deep representations and MIL to perform object detection. In [75],

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

deep feature representations and MIL are combined to analyze medical images. In [76], the authors proposed a MIL based deep learning model to perform image segmentation at pixel-level. In [77], deep representations learned from local and global views are combined with MIL to analyze image aesthetics. In recent years, notable success has been achieved using the combination of MIL and deep learning representation. However, no existing methods are explicitly intended for image memorability prediction where MIL is employed. Hence, a MIL framework has been employed in this chapter to utilize efficiently the emotion cues to predict image memorability scores.

5.1.2.2 Multi-context patch extraction for MIL based memorability prediction

In a recent work [78], the authors shown that different regions within an image may evoke different emotions. Due to this effect, different regions within an image may yield different memorability scores. Similarly, Dubey et al. [3] showed that not all objects within an image are equally memorable. This study also indicates that different regions within an image can yield different memorability scores. Therefore, predicting the memorability score for a given image from a single global view may not always be accurate. It is also observed from the image memorability datasets that some of the images contain a single focused object. Such kind of images may not evoke multiple emotions from different regions; instead, the entire image evokes a single emotion. Hence, a single global view of the input image is sufficient to determine the memorability score for such images. These observations suggest that global and local contexts are essential to utilize visual emotion information in image memorability prediction. To predict the memorability score based on the emotion evoked from the global view, an entire image is treated as a global image patch (x_g) as shown in the first column of Figure 5.3. To predict memorability score based on multiple local emotions, four patches are extracted from four salient local regions (x_{l1} , x_{l2} , x_{l3} , and x_{l4}) within an image as shown in second, third, fourth and fifth columns of Figure 5.3. Since

5.1 Proposed Model for Image Memorability Prediction

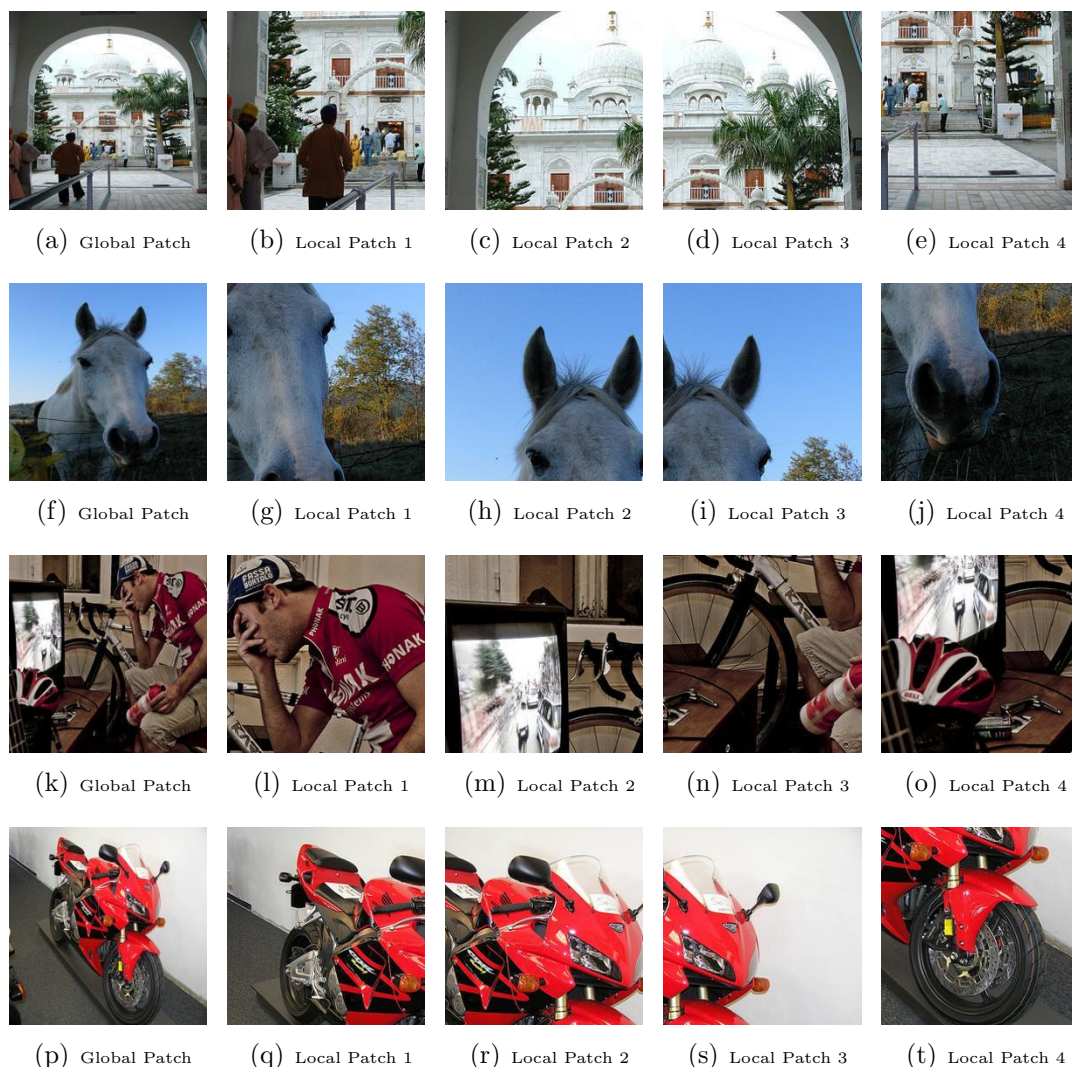


Figure 5.3: Example images showing global and local patches. Local patches are extracted from top four salient regions.

the memorability dataset does not contain memorability score at pixel level but image level, it is not possible to employ a fully-supervised learning framework. Therefore, a weakly supervised learning framework MIL is employed. The MIL assumption can be easily fit by considering the input image X as bag and the multi-context patches x_{l1} , x_{l2} , x_{l3} , x_{l4} and x_g as multiple instances related to bag X . The feature representations generated from these multi-context patches are

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

combined through MIL to predict the memorability scores. In order to generate local image patches, top four salient regions are selected. The saliency map is generated by means of the saliency prediction model proposed in [6].

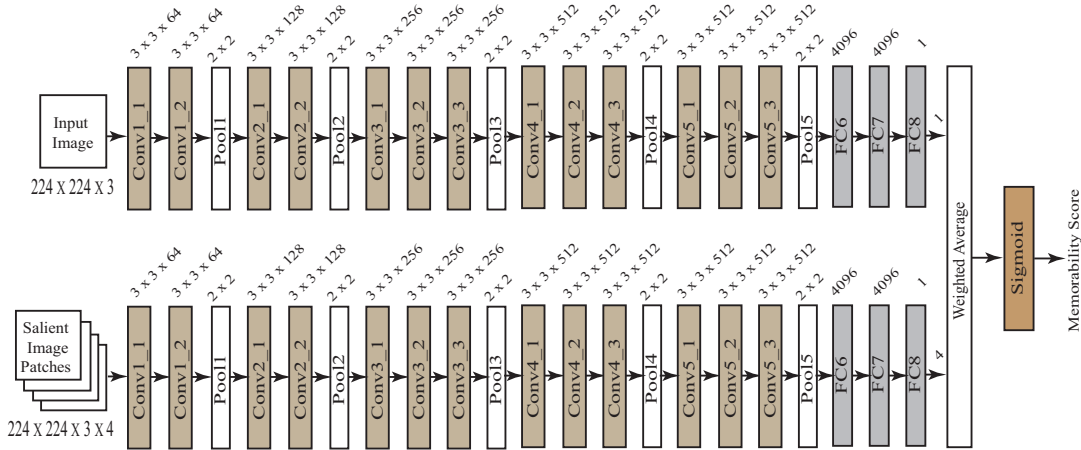


Figure 5.4: The architecture of the proposed MCDR-MemNet model.

5.1.2.3 MCDR-MemNet: MIL Based Multi-Context Deep Representation Network for emotion-based image memorability prediction

A simple way to utilize emotion information in image memorability prediction is fine-tuning one of the existing deep CNN image classification models [33–36] in two stages. In the first stage, deep emotion representations are learned by means of transfer learning technique. In this process of transfer learning technique, an image classification model is fine-tuned on visual emotion dataset to perform the emotion classification task. In the second stage, the fine-tuned model is further trained on image memorability dataset to map learned deep emotion representations to memorability scores. In this work, VGG-16 [34] is fine-tuned in the same fashion to utilize emotions in image memorability prediction. In the first stage, VGG-16 is fine-tuned on the publicly available large-scale visual emotion classification dataset [78] to learn deep visual emotion features. The fine-tuned model is named as VGG-Emo. In the second stage, VGG-Emo is further

5.1 Proposed Model for Image Memorability Prediction

fine-tuned on *LaMem* dataset to map fine-tuned deep emotion features with image memorability scores. The *VGG-Emo* model fine-tuned on *LaMem* is named as *VGG-EmoMemNet*. Fine-tuning details are discussed in section 5.2.1. However, *VGG-EmoMemNet* can utilize the emotion cue evoking from the global view of an input image and may not utilize the multiple emotions evoking from various local regions within an image. In order to utilize emotions evoking from both local and global contexts, an MIL based Multi-Context Deep Representation Memorability Network (*MCDR-MemNet*) is devised using *VGG-EmoMemNet*. The proposed *MCDR-MemNet* architecture is shown in Figure 5.4, which contains two branches. Both of these branches are initialized with weights of the *VGG-EmoMemNet*. The upper branch takes global patch (x_g) as input, and the lower branch takes local patches (x_{l1} , x_{l2} , x_{l3} , and x_{l4}) as inputs. In order to effectively utilize the learned emotion features, the proposed deep network, *MCDR-MemNet* is trained to learn multi-context deep representation by means of MIL framework.

Training *MCDR-MemNet* using MIL framework: On the basis of MIL framework property, input image X is defined as the bag. The global and local image patches (x_{l1} , x_{l2} , x_{l3} , x_{l4} , and x_g) extracted from the input image (as explained in section 5.1.2.2) are defined as instances. The last fully-connected layer’s output of the upper branch of the *MCDR-MemNet* provides single global representation: x_{gr} . Similarly, the last fully-connected layer’s output of the lower branch of the *MCDR-MemNet* provides four local representations: x_{lr1} , x_{lr2} , x_{lr3} , and x_{lr4} . The representation of the entire bag is produced by aggregating the global and local representations using the aggregating function $AGG(\cdot)$ as given in Equation 5.2. The aggregate function $AGG(\cdot)$ performs a linear combination of its inputs with equal weights. The predicted memorability score $Y_{MIL-mem}$ of the entire bag is defined as given in Equation 5.2.

$$Y_{MIL-mem} = \sigma(X_{er})$$

where,

$$X_{er} = AGG(x_{gr}, x_{lr1}, x_{lr2}, x_{lr3}, x_{lr4})$$
(5.2)

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

where $\sigma()$ is the Sigmoid function defined as shown in Equation 4.4.

5.1.3 *Ens-MemNet*: Ensemble of Memorability Networks

Final predicted memorability score for the given input image is obtained by ensembling two deep networks: *VGG-MemNet* and *MCDR-MemNet*, as shown in Figure 5.1. Ensembling multiple network outputs not only increases the model accuracy but also avoids model over-fitting [64]. **BEM** is a simple approach to combine network outputs by means of arithmetic mean [65]. Given n networks outputs $\{y_1, y_2, \dots, y_n\}$, the **BEM** output Y_{BEM} is defined as given in Equation 5.3.

$$Y_{BEM} = \frac{1}{n} \sum_i^n y_i \quad (5.3)$$

The final predicted memorability score Y_{Ens} for the given input image X is obtained by employing **BEM** as given in Equation 5.4. Memorability scores generated by *VGG-MemNet* and *MCDR-MemNet* are represented as Y_{O-mem} and $Y_{MIL-mem}$, respectively. in Equation 5.4.

$$Y_{Ens} = \frac{Y_{O-mem} + Y_{MIL-mem}}{2} \quad (5.4)$$

5.2 Experiments and Results

This section details about the experimental set-up and the corresponding results to demonstrate the superiority of the proposed model over the existing model. Three publicly available datasets (*LaMem* [2], Isola et al.’s memorability dataset [1] and Dubey et al.’s memorability dataset [3]) are used in the experiments. The Spearman’s rank correlation coefficient (ρ) is employed to evaluate the performance of the proposed model.

5.2.1 Experimental Set-up

The training process of *MCDR-MemNet* is carried in two phases. In the first phase, *VGG-16* is fine-tuned twice on two different datasets. First, it is fine-

tuned on image emotion classification dataset [78] to learn emotion features. In this process, the model is trained to classify an input image into one of the eight emotion classes (*Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sad*). The fine-tuned model is named as *VGG-Emo*. During the fine-tuning process, all the layers are allowed to learn, and the last fully-connected layer output neurons of *VGG-16* are varied from 1000 to 8. The *VGG-Emo* is further fine-tuned on *LaMem* dataset to predict image memorability scores. The fine-tuned model is named as *VGG-EmoMemNet*. In this fine-tuning process, all the layers are allowed to learn, and the number of output neurons of the last fully connected layer of *VGG-Emo* model is varied from 8 to 1.

In the second phase, two copies of *VGG-EmoMemNet* are created as the upper and lower branches of *MCDR-MemNet* as shown in Figure 5.5. Then *MCDR-MemNet* is fine-tuned under MIL framework. *MCDR-MemNet* is fed with one global and four local patches. The global patch has the dimension of $224 \times 224 \times 3$ which is resized, and center cropped from the original image and local patches are extracted from top four salient regions from the original image with each patch has the dimension of $224 \times 224 \times 3$.

The training process of *VGG-MemNet* is carried out by fine-tuning the *VGG-16* on *LaMem* [2]. In this fine-tuning process, all the layers are allowed to learn, and the output of the last fully-connected layer of *VGG-16* is set to 1 for memorability prediction. *VGG-MemNet* is fed with an image of $224 \times 224 \times 3$ dimension, which is resized and center cropped from the original image. Image memorability prediction is essentially a regression task. For such tasks, L2 loss is the most widely used loss function [2]. The L2 loss function used to train the proposed models is shown in Equation 5.5.

$$L2 = \sum_j \|Y_j - y_j\|_2^2 \quad (5.5)$$

where Y_j and y_j represent the predicted and ground-truth memorability scores of the j^{th} image. To minimize network loss, *Ada-delta* optimizer is used with an initial learning rate of 0.001. The models are trained with a batch size of 50

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

images.

5.2.2 Results

The *LaMem* dataset contains 60,000 images and is divided into five sets for cross-validation purpose. Each set contains 45,000 training samples, 10,000 testing samples, and 3,741 validation samples. Accordingly, five models are trained, tested, and the results are averaged. Further, the trained models are tested on other publicly available memorability datasets. Table 5.1 presents the performance of *MemNet* [2], *VGG-MemNet*, *MCDR-MemNet*, and *Ens-MemNet* on three publicly available memorability datasets. The performance is represented by means of rank correlation (ρ) computed between ground-truth and predicted memorability scores. From Table 5.1, it can be observed that *VGG-MemNet* is ranking the images better than *MemNet*. This improvement in *VGG-MemNet* is expected as the underlying architecture of *VGG-MemNet* (*VGG-16* [34]) has shown better accuracy in image classification task [40] than the underlying architecture of *MemNet* (*AlexNet* [33]). From the results, it is also evident that the performance of *VGG-EmoMemNet* is relatively better than *VGG-MemNet*, indicating that deep emotion features are important in determining image memorability. Interestingly, *MCDR-MemNet* performed much better than *VGG-MemNet*. This result suggests that the utilization of local emotions evoking at the salient regions within an image and global emotion evoking from the entire image can help the deep learning model to rank images more accurately. The results also prove that the performance of *Ens-MemNet* model is better than that of all the other models, including *MemNet*. The better performance of *Ens-MemNet* shows that the combination of object, emotion, and saliency information can make the deep learning model better in image memorability prediction.

In order to analyze the performance of the proposed models further, images are arranged in descending order of predicted memorability scores. Various ranges of these sorted images are selected and examined the average ground-truth memora-

5.2 Experiments and Results

Table 5.1: Performance comparison of the existing (*MemNet* [2]) and proposed models (*VGG-MemNet*, *VGG-EmoMemNet*, *MCDR-MemNet* and *Ens-MemNet*).

Dataset	MemNet [2]	VGG-MemNet	VGG-EmoMemNet	MCDR-MemNet	Ens-MemNet
LaMem [2]	0.640	0.650	0.655	0.663	0.671
Isola [1]	0.61	0.638	0.633	0.638	0.664
Dubey [3]	0.450	0.492	0.483	0.497	0.511

bility scores on these ranges. Table 5.2 shows this analysis on *LaMem* dataset [2]. Images are sorted into sets according to predictions made by existing and proposed models (denoted by column headings of Table 5.2). This analysis is also carried out on the other two datasets. Table 5.3 shows corresponding results on Isola et al. dataset [1] and Table 5.4 shows the results on Dubey et al. dataset [3]. From Tables 5.2, 5.3, and 5.4, it is evident that the proposed models (*VGG-MemNet*, *VGG-EmoMemNet*, *MCDR-MemNet*, and *Ens-MemNet*) ranks the images better than the existing model (*MemNet*) based on image memorability property. For example, Table 5.2 shows that “Bottom 10” images ranked based on *MemNet* prediction are 58.06% memorable. Whereas “Bottom 10” images ranked based on *Ens-MemNet* prediction are 48.41% memorable. Similarly, from Tables 5.3 and 5.4, it is evident that top ranked images of *Ens-MemNet* model are more memorable than top ranked images of *MemNet* model. Also, the bottom ranked images of *Ens-MemNet* model are less memorable than bottom ranked images of *MemNet* model.

5.2.3 Emotion bias in existing and proposed models

In [2], the authors collected the ground-truth memorability scores for the labeled emotion dataset (affective images dataset [44]) and analyzed the relationship between visual emotions and memorability. From their analysis, they discovered that images portraying negative emotions such as *disgust*, *anger*, and *fear* tend to be more memorable (except for *amusement*) than those projecting positive emotions such as *contentment* and *awe*. Based on their analysis, they ranked the emotions in decreasing order of ground-truth memorability scores as follows:

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

Table 5.2: Comparison of predicted versus ground-truth image memorability scores on LaMem dataset [2]. Images are arranged in descending order of predicted memorability scores. Various ranges of these sorted images are selected. The average ground-truth memorability scores are shown for each set in each row. Reported result are averaged over 5-fold cross-validation models

Ranges of Sorted Images	MemNet [2]	VGG-MemNet (Proposed)	VGG-EmoMemNet (Proposed)	MCDR-MemNet (Proposed)	Ens-MemNet (Proposed)	Ground Truth
Top 10	91.70%	90.85%	92.42%	93.15%	91.89%	100%
Top 25	90.40%	90.22%	91.72%	91.76%	91.36%	100%
Top 50	89.57%	90.18%	90.71%	91.27%	90.85%	99.35%
Top 100	89.17%	89.99%	90.35%	90.42%	90.38%	98.45%
Top 200	88.91%	89.52%	89.64%	90.00%	90.02%	97.57%
Bottom 200	55.06%	54.94%	54.69%	54.57%	54.23%	42.16%
Bottom 100	54.35%	53.7%	53.33%	52.79%	52.74%	39.01%
Bottom 50	54.20%	52.02%	51.90%	51.65%	51.25%	36.3%
Bottom 25	54.44%	50.88%	51.46%	51.20%	50.34%	34.41%
Bottom 10	58.06%	48.76%	49.69%	50.94%	48.41%	33.57%
ρ	0.64	0.65	0.655	0.663	0.671	0.680

Table 5.3: Comparison of predicted versus ground-truth image memorability scores on Isola et al. dataset [1]. Uses same measures as detailed in Table 5.2.

Ranges of Sorted Images	MemNet [2]	VGG-MemNet (Proposed)	VGG-EmoMemNet (Proposed)	MCDR-MemNet (Proposed)	Ens-MemNet (Proposed)	Ground Truth
Top 10	80.16%	81.23%	80.78%	81.75%	82.43%	96.54%
Top 25	75.46%	80.09%	79.8%	79.77%	81.41%	94.39%
Top 50	75.13%	78.96%	77.86%	78.57%	79.48%	92.24%
Top 100	74.32%	77.13%	77.23%	76.64%	77.63%	89.59%
Top 200	73.58%	75.75%	75.21%	74.9%	76.29%	85.33%
Bottom 200	35.91%	35.23%	34.81%	34.83%	34.08%	22.85%
Bottom 100	32.8%	31.64%	32.64%	31.64%	31.66%	18.66%
Bottom 50	30.14%	30.06%	30.64%	29.41%	28.94%	14.93%
Bottom 25	28.81%	28.69%	28.91%	26.6%	29.47%	10.95%
Bottom 10	28.29%	26.59%	28.52%	26.23%	26.86%	5.69%
ρ	0.610	0.638	0.633	0.638	0.664	0.750

Table 5.4: Comparison of predicted versus ground-truth image memorability scores on Dubey et al. dataset [3]. Uses same measures as detailed in Table 5.2.

Ranges of Sorted Images	MemNet [2]	VGG-MemNet (Proposed)	VGG-EmoMemNet (Proposed)	MCDR-MemNet (Proposed)	Ens-MemNet (Proposed)	Ground Truth
Top 10	84.72%	86.84%	86.11%	85.91%	86.69%	92.23%
Top 25	83.79%	85.21%	84.8%	84.95%	85.23%	91.23%
Top 50	83.69%	84.69%	84.43%	84.25%	84.24%	90.13%
Top 100	83.36%	83.61%	83.6%	83.16%	83.71%	88.65%
Top 200	82.17%	82.50%	82.14%	82.25%	82.53%	86.94%
Bottom 200	71.8%	71.28%	70.69%	70.58%	70.72%	65.17%
Bottom 100	69.85%	68.91%	68.7%	68.29%	68.12%	60.87%
Bottom 50	68.01%	66.33%	66.35%	66.17%	65.72%	57.15%
Bottom 25	66.36%	65.74%	65.15%	64.77%	65.51%	54.2%
Bottom 10	63.98%	66.00%	64.6%	63.47%	63.78	50.92%
ρ	0.450	0.492	0.483	0.497	0.511	0.76

$\{Disgust, Amusement, Fear, Anger, Excitement, Sad, Awe, Contentment\}$. This finding is similar to [23], where authors showed that attributes like ‘peaceful’ are negatively correlated with memorability. In this chapter, the degree of emotion bias of the existing and proposed models in determining image memorability is also analyzed. For the analysis purpose, 400 random images labeled with eight emotion classes (*Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sad*) are sampled from the emotion dataset created by Rao et al. [78]. These 400 images are sampled with equal distribution of emotion classes (50 images for each emotion class). For these images, memorability scores are predicted from the existing as well as proposed models. Based on the predictions, these images are sorted. Further, various ranges of these sorted images are selected, as shown in Figures 5.5 and 5.6. Distribution of emotion class is computed to understand ranking behavior of the existing and proposed models for each emotion category. In each sub-figures of Figures 5.5 and 5.6, emotion categories are mentioned in decreasing order of ground-truth memorability scores from left to right.

From Figures 5.5 and 5.6, it is evident that proposed emotion based models (*VGG-EmoMemNet, MCDR-MemNet, and Ens-MemNet*) tend to put images which evoke more memorable emotions such as *disgust* and *amusement* in higher ranks and images which evoke less memorable emotions such as *contentment* and *awe* in lower ranks. The degree of this emotion bias is higher in *MCDR-MemNet* compared to other proposed and existing models. It is evident from this observation that MIL based multi-context deep representation model (*MCDR-MemNet*) can effectively capture the emotion cues which are important to determine the image memorability scores.

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

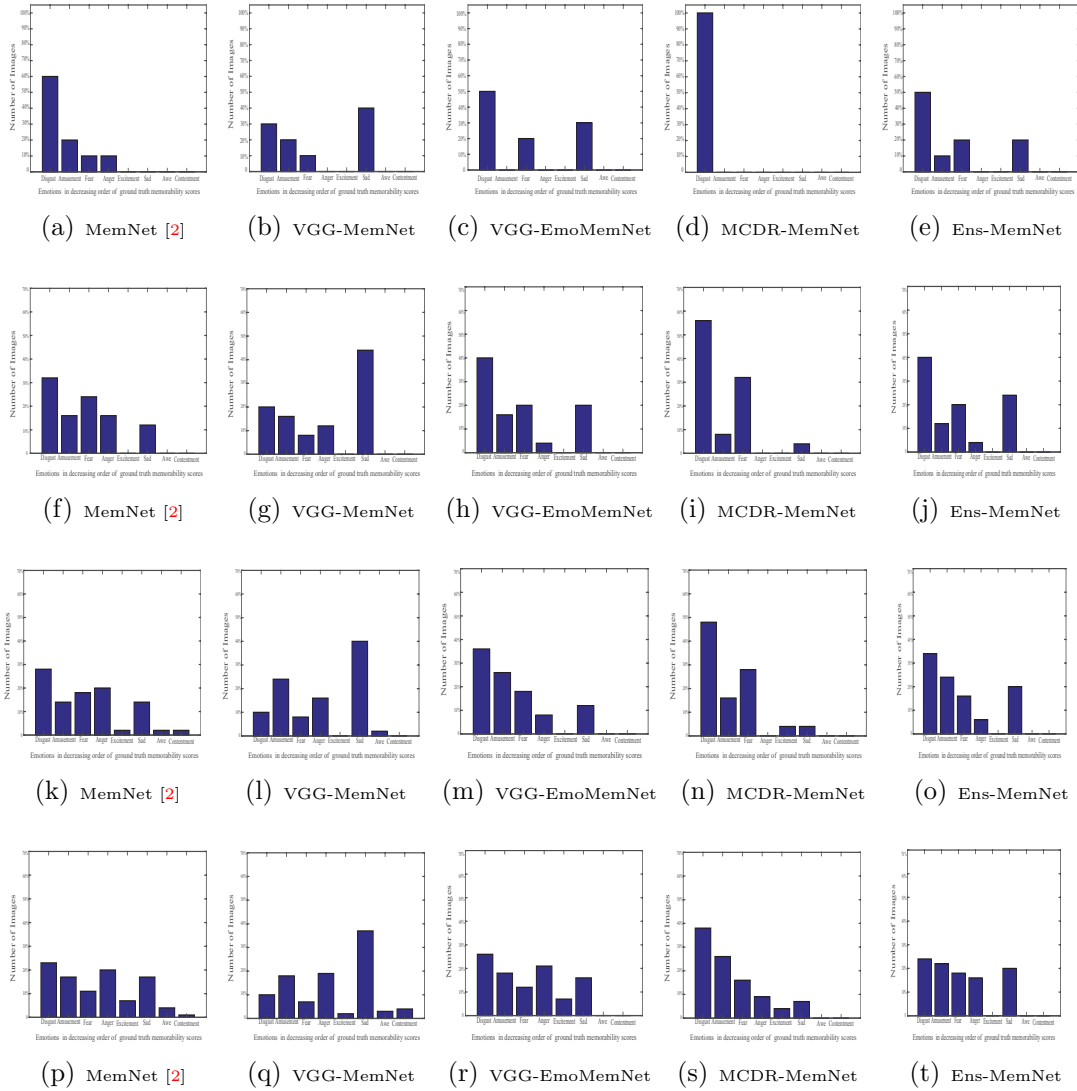


Figure 5.5: Emotion distribution on top ranked images according to the predictions of existing and proposed models (denoted by each sub-figure title). Images are sorted according to the predictions made by existing and proposed models and chosen sets of “Top 10”, “Top 25”, “Top 50”, and “Top 100” images. Emotion distributions are reported for these sets of “Top 10”, “Top 25”, “Top 50”, and “Top 100” images in the first, second, third and fourth rows of the image respectively.

5.2 Experiments and Results

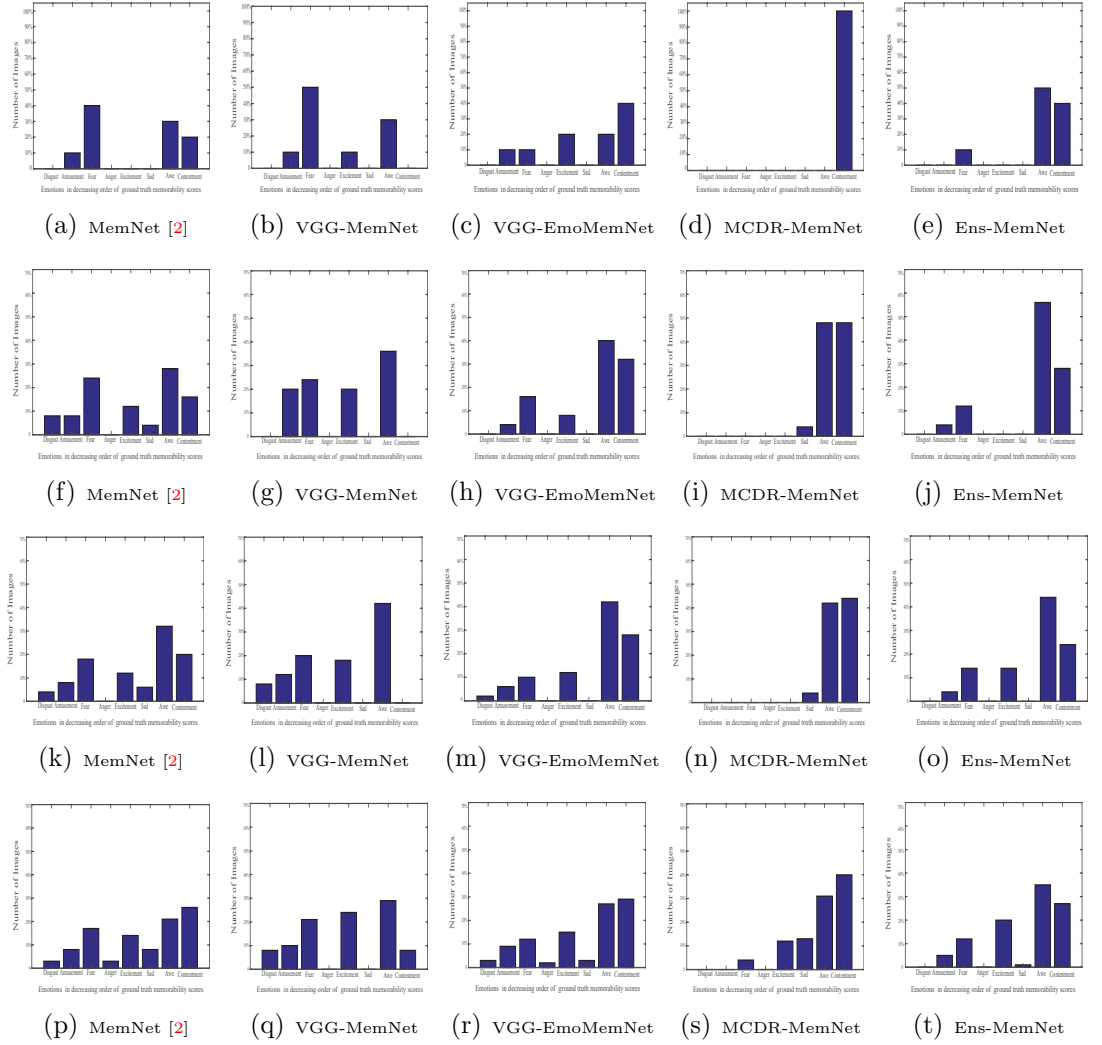


Figure 5.6: Emotion distribution on least ranked images according to the predictions of existing and proposed models (denoted by each sub-figure title). Images are sorted according to the predictions made by existing and proposed models and chosen sets of “Bottom 10”, “Bottom 25”, “Bottom 50”, and “Bottom 100” images. Emotion distributions are reported for these sets of “Bottom 10”, “Bottom 25”, “Bottom 50”, and “Bottom 100” images in the first, second, third and fourth rows of the image respectively.

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

5.2.4 Ensembling of VGG-FMemNet, VGG-DMemNet, MCDR-MemNet, and VGG-MemNet

In order to verify the influence of motion, depth, emotion, and other object related features on image memorability prediction process, *VGG-FMemNet* (proposed in Chapter 4), *VGG-DMemNet* (proposed in Chapter 4), *MCDR-MemNet*, and *VGG-MemNet* ensembled. Model ensembled in this way is named as *Final-Ens-MemNet*. The corresponding results are shown in Table 5.5 along with the results of *OFD-MemNet-II* (proposed in Chapter 4) and *Ens-MemNet*.

Table 5.5: Performance of the proposed models, *VGG-FMemNet*, *VGG-DMemNet*, *MCDR-MemNet*, *VGG-MemNet*, *OFD-MemNet-II*, and *Ens-MemNet* along with *Final-Ens-MemNet*.

Dataset	VGG-MemNet	VGG-FMemNet	VGG-DMemNet	MCDR-MemNet	OFD-MemNet-II	Ens-MemNet	Final-Ens-MemNet
LaMem [2]	0.650	0.652	0.654	0.663	0.671	0.671	0.676
Isola [1]	0.638	0.638	0.641	0.638	0.667	0.664	0.675
Dubey [3]	0.492	0.495	0.497	0.497	0.515	0.511	0.519

From Table 5.5, it is visible that ensembling emotion, depth, motion and object related features enhances the accuracy of the memorability prediction process. The *Final-Ens-MemNet* yielded a rank correlation co-efficient value of 0.676 for *LaMem* dataset which is very close to human consistency (0.68).

5.3 Summary

In this chapter, a novel deep learning model *Ens-MemNet* is proposed to predict image memorability scores. The proposed model is designed to learn and utilize various high-level visual factors, including object semantics, and visual emotions evoking at single global and multiple salient local image regions. The *Ens-MemNet* is obtained by ensembling two networks: *MCDR-MemNet* and *VGG-MemNet*. In order to utilize multiple emotions evoking from various salient regions within an image, MIL based deep learning model, *MCDR-MemNet* is devised. *VGG-MemNet* is obtained by means of fine-tuning a deep learning based object classification model, *VGG-16* [40], on image memorability dataset to utilize object semantics. The *MCDR-MemNet* model is ensembled with *VGG-*

MemNet model to obtain final predicted memorability scores. Through an extensive set of experiments, it is shown that the proposed image memorability prediction model *Ens-MemNet* performed better than the state-of-the-art model [2]. The proposed model, *Ens-MemNet* yielded a rank correlation of 0.67, which is very close to human consistency ($\rho = 0.68$) on large-scale image memorability dataset *LaMem* [2].

Until now, in this dissertation, the influence of some visual factors are analyzed for determining object and image level memorability. In the next and the final contributory chapter of this dissertation, an application of image memorability prediction is addressed where it has been discussed how memorability of image can be increased while retaining most of its high-level contents. To achieve this application, an end-to-end deep learning model is devised which takes a natural image as input and modifies it in such a way that its memorability score is increased while retaining its most of its high-level contents.

5. VISUAL EMOTION BASED IMAGE MEMORABILITY PREDICTION USING MULTIPLE INSTANCE LEARNING

Chapter 6

Memorability based Image to Image Translation

In the initial three contributory chapters of this dissertation, it has been shown how different visual factors like object location and size, image motion, depth, and emotions help to understand and predict the object or image level memorability. In this last contributory chapter, it has been addressed how image memorability can be increased. Towards this goal, a deep learning model has been devised which can enhance the memorability of an image. Generating memorable images is essential in many practical applications, including the creation of a memorable logo, magazine cover photo, user interface design, academic materials, and much more. In recent literature, Khosla et al. [79] showed that memorability of face images could be modified without disturbing the properties like identity, age, attractiveness, and emotional magnitude of the person. However, their model is limited to face images. Along the same line of thought, image color and texture features are modified in [80] to change the emotions evoked by an image. These findings boosted our intuition that it may be possible to modify an image to make it more memorable while retaining its high-level content. This chapter attempts to alter the given image to make it more memorable while retaining its high-level contents (see Figure 6.1). Since the proposed scheme aims to translate an input image to another image having higher memorability, the underlying problem can

6. MEMORABILITY BASED IMAGE TO IMAGE TRANSLATION

be treated as memorability based image-to-image translation.

Recently, Isola et al. [7] proposed a conditional generative adversarial network based method to perform automatic image-to-image translation. Their proposed model can successfully translate an image from one representation of a given image to another, e.g., labels to street scenes, black and white to color, edges to photo, etc. This approach also supported the proposed intuition that images can be modified to increase their memorability score. However, incorporation of this approach for memorability based image translation requires paired image dataset. Generating paired image dataset is not only expensive and time-consuming but also practically not feasible because it requires manual image manipulation and verification of memorability property. Furthermore, image memorability is influenced collectively by various visual factors including emotions, saliency, object and scene semantics [2] and hence, it is difficult to change these factors individually to increase the image memorability. Image-to-image translation technique without paired examples is proposed in [81] to translate, for instance, day to night, horse to zebras, photo to monet, and vice-versa. For this kind of translation, the need of target domain is necessary, and in most of the cases, it is known. In the case of memorability based image translation, the target domain is unknown.

To address the aforementioned problems, a memorability based image-to-image translation is proposed in this chapter by defining image translation as the mapping $F : I \rightarrow I'$ between two image domains I and I' . Here, I corresponds to input image domain, and I' is the unknown image domain which contains the modified versions of the images present in I . Also, every image in I' is more memorable than its corresponding image in I . To achieve the proposed translation, a novel deep learning model is devised, which is trained in the absence of paired examples using mean-squared error and memorability loss between I and $F(I)$. The mean-squared error is employed to enforce minimal changes in the translated image. Whereas, memorability loss is used to implement the modifications such that the translated image is more memorable than its input counterpart. To the

best of our knowledge, this is the first end-to-end deep learning model that translates a generic image (not specific to face images) to its modified version to make it more memorable without using paired image dataset. The detailed description of the proposed memorability based image-to-image translation approach is presented in the next section.

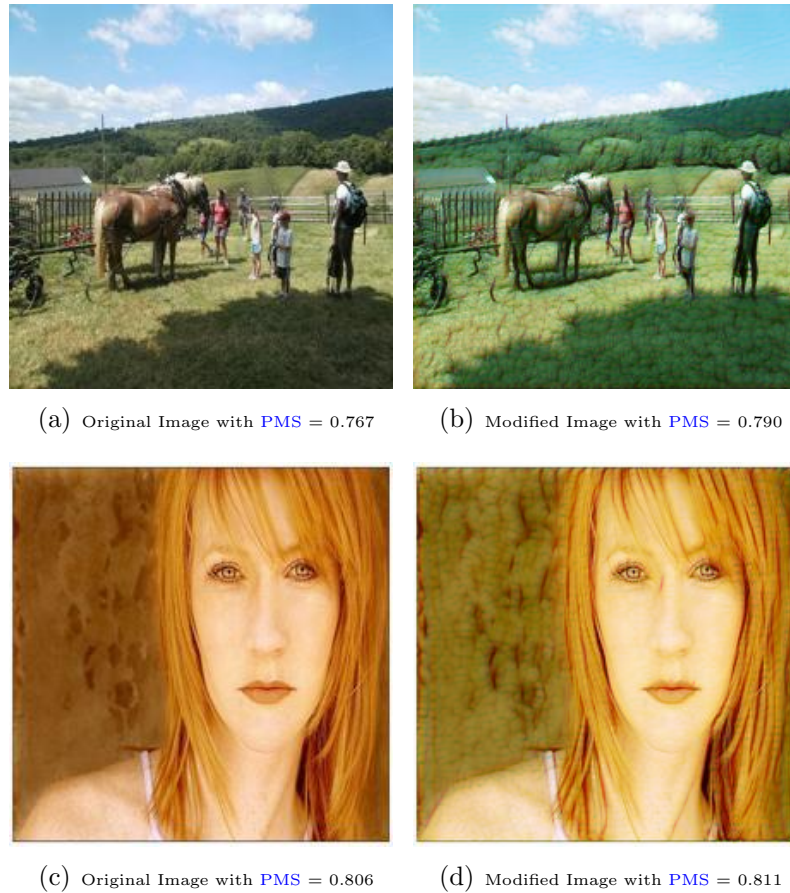


Figure 6.1: *Examples of images with their modified versions to increase the memorability score. The Predicted Memorability Score (PMS) is reported for each image.*

Rest of the chapter is arranged as follows. In Section 6.1, the proposed approach to increase image memorability is explained. Section 6.2 details about the experimental set-up and corresponding results. Finally, the summary of the chapter is presented in Section 6.3.

6. MEMORABILITY BASED IMAGE TO IMAGE TRANSLATION

6.1 Proposed Model

The architecture of the proposed model is shown in Figure 6.2. The model has two networks: *Translator Network* and *Memorability Prediction Network*. The *Translator Network* learns to modify the given input image to increase its memorability score in such a way that high-level contents of the input image are retained. The *VGG-MemNet* is used as *Memorability Prediction Network* to evaluate the modifications with respect to the memorability of an image. Further details of both the networks are presented in the following subsections.

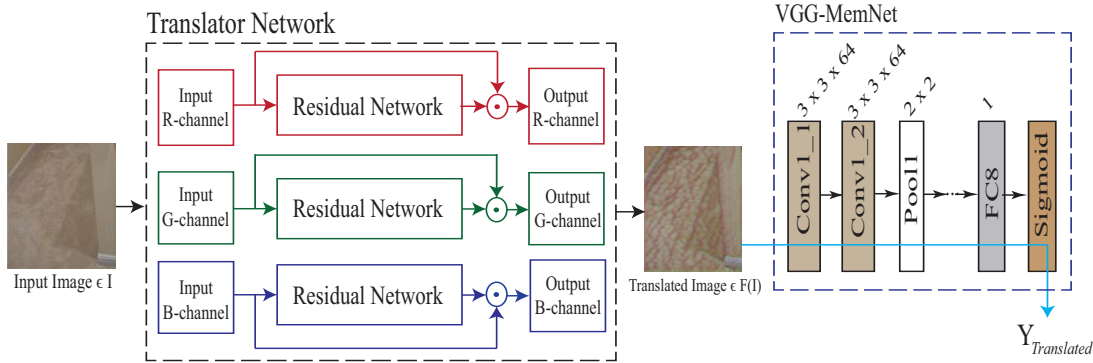


Figure 6.2: Framework of the Proposed Method.

6.1.1 Translator Network

The proposed *Translator Network* takes an RGB image as input and outputs its modified version in RGB space. As shown in Figure 6.2, *Translator Network* contains three branches. Each branch is a *Residual Network* dedicated to processing a color channel. All these branches have the same architecture, which is shown in Figure 6.3. The circle with a dot symbol in Figure 6.2 represents a linear combination of input color channels and residual information. The proposed *Residual Network* contains a skip connection between layer 1 and layer 4 to retain the structural information of the given color channel. Residual values generated from each channel represents different style information. This style information

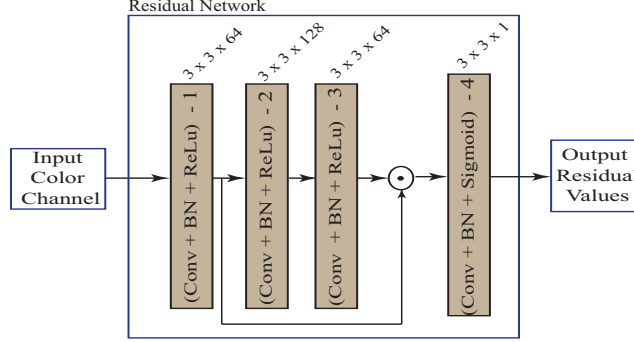


Figure 6.3: *Residual Network: Generates residual values which are later used to modify the image to increase its memorability. BN represents Batch Normalization.*

is incorporated into the input image using pixel-wise weighted addition as given in Equation 6.1.

$$C_{Output}^i = \alpha.C_{Input}^i + \beta.C_{Residual}^i \quad (6.1)$$

where, C_{Output}^i , C_{Input}^i , and $C_{Residual}^i$ represent modified color channel, input color channel and predicted residual channel for channel $i \in \{R, G, B\}$. Each layer of the *Residual Network* has 3×3 size kernels. Batch normalization is added for each layer for faster convergence. *ReLU* is used as activation function for all the layers except the last layer. The output of the last layer is normalized between 0 and 1 by means of sigmoid function. The number of output channels are fixed to 64, 128, 64 and 1 for the layers 1 to 4 respectively.

6.1.2 VGG-MemNet

In order to verify the modifications performed by the proposed *Translator Network*, a memorability prediction model is required. For simplicity, the *VGG-MemNet* proposed in Section 4.2.2 of Chapter 4 is utilized to predict memorability scores. The *VGG-MemNet* is obtained by fine-tuning the *VGG-16* model [34] on *LaMem* dataset [2]. The scores generated by *VGG-MemNet* are in the range of 0 to 1, where 0 indicates the least memorable image, and 1 indicates the most memorable image. The image generated by *Translator Network* is fed to the

6. MEMORABILITY BASED IMAGE TO IMAGE TRANSLATION

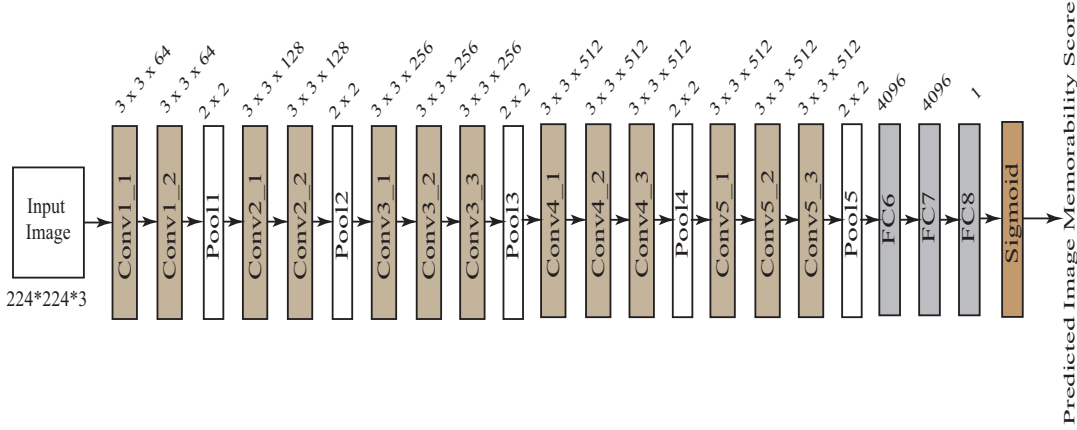


Figure 6.4: *VGG-MemNet architecture*

VGG-MemNet for memorability prediction to evaluate the modifications with respect memorability.

6.1.3 Loss Function

The objective of the proposed memorability based image to image translation model is to modify the given input image such that the modification has to satisfy two criteria. The first criterion is that memorability of the modified image must be more memorable than the input image. The second criterion is that the modified image must contain most of the high-level contents of the input image. Based on these two criteria, the loss function $L_{translator}$ is defined as shown in Equation 6.2. The loss function $L_{translator}$ for the proposed approach is comprised of two loss functions: L_{mse} and L_{mem} . While L_{mse} tries to retain the contents of the input image to achieve the second criterion, L_{mem} enforces modifications such that the modified image’s predicted memorability score should increase to reach 1 for achieving the first criterion. In order to define the loss function $L_{translator}$, let I and I' be the domains of input and translated images respectively, and the one-to-one mapping function is to be learned by the *Translator Network* be $F : I \rightarrow I'$. Also, let $p \in I$ be a given input image, and $F(p)$ be the output generated by the *Translator Network*. Then, the loss function $L_{translator}$ is defined

as

$$L_{translator} = \lambda_{mse} \cdot L_{mse} + \lambda_{mem} \cdot L_{mem} \quad (6.2)$$

where λ_{mse} and λ_{mem} are hyper-parameters, the mean-squared error L_{mse} is written as

$$L_{mse} = \frac{1}{M \cdot N \cdot D} \sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^D \|F(p)^{i,j,k} - p^{i,j,k}\|_2^2 \quad (6.3)$$

where, M , N , and D are the image height, width, and channels respectively. The memorability loss L_{mem} is given by

$$L_{mem} = \frac{1}{t} \sum_{i=1}^t \|1 - Y_{F(p_i)}\| \quad (6.4)$$

where, t is the total number of training samples and $Y_{F(p_i)}$ refers to the predicted memorability score of the i^{th} translated image $F(p_i)$. With this, the objective of the proposed framework is to minimise the $L_{translator}$ loss.

6.2 Experiments and Results

This section details the experiments and corresponding results. The proposed model is trained and tested on image memorability dataset *LaMem* [2].

6.2.1 Training of Memorability prediction model VGG-MemNet

VGG-MemNet is fine-tuned on *LaMem* dataset [2] by varying the number of outputs of last fully connected layer to 1. For training purpose, 45000 images are used. The trained network is tested on 10000 images. Image memorability prediction is basically a regression task. $L2$ loss is the most commonly employed loss function [2] for regression tasks. Therefore, the $L2$ loss function is used to train the proposed model, which can be represented mathematically as shown in Equation 6.5.

$$L2 = \sum_j \|Y_j - y_j\|_2^2 \quad (6.5)$$

6. MEMORABILITY BASED IMAGE TO IMAGE TRANSLATION

where Y_j is the memorability score obtained from the proposed *VGG-MemNet* and y_j is the ground-truth memorability scores for the j^{th} image. Adam [82] optimiser is employed to reduce network loss. The initial learning rate is set to 0.001.

6.2.2 Training of Translator Network

The *Translator Network* is trained on 5000 images which are not used to train the *VGG-MemNet*. To ensure the diversity in terms of memorability, images with the following range of memorability scores are chosen equally (i.e., 1000 from each range): 0 to 0.6, 0.6 to 0.7, 0.7 to 0.8, 0.8 to 0.9 and 0.9 to 1.0. The proposed model is tested on 2000 images which are not used to train the *VGG-MemNet* as well as *Translator Network*. Diversity in terms of memorability is ensured in the testing dataset also. Adam [82] optimizer is used to reduce the $L_{translator}$ loss with 0.001 learning rate. Hyper-parameters are experimentally adjusted with the following values: $\alpha = 0.7$, $\beta = 0.3$, $\lambda_{mse} = 10$, and $\lambda_{mem} = 1.0$.

6.2.3 Performance Evaluation

To evaluate the performance of the proposed *VGG-MemNet*, the Spearman’s rank correlation coefficient (ρ) is employed. The *VGG-MemNet* yielded a ρ value of 0.65 that is closer to human performance ($\rho = 0.68$ [2]) in memorability prediction. The proposed memorability based image to image translation model is evaluated with respect to two aspects: (1) the increase in **PMS** obtained from *VGG-MemNet* and (2) the *retention of structural similarity*. In order to quantify the increase in **PMS**, a new measure *Mean Memorability Score Difference (MMSD)* is defined as

$$MMSD = \frac{1}{s} \sum_{i=1}^s (Y_{F(p_i)} - Y_{p_i}) \quad (6.6)$$

where s is the total number of image samples. p_i and $F(p_i)$ are the i^{th} input and translated image samples. To quantify the retention of structural similarity, **SSIM** [83] is employed.

Table 6.1: Comparison of performance between proposed memorability based image-to-image translation method and the style transfer methods [7].

Methods	Cezanne Style Transfer [7]	Monen Style Transfer [7]	Ukiyoe Style Transfer [7]	Vangogh Style Transfer [7]	Proposed Method
MMSD in %	0.2548	0.3453	0.5453	0.7981	2.0298
Average SSIM	0.6668	0.7219	0.5351	0.5647	0.7972

To the best of our knowledge, this is the first end-to-end deep learning model that translates a generic image (not limited to face images) to its modified version to make it more memorable without using paired image dataset or using any additional style information. Therefore, the proposed model is compared with an unpaired image to image translation model [81], which transfer the given input image to styles like *Cezanne*, *Monet*, *Ukiyo-e*, and *Van Gogh*. Table 6.1 shows the MMSD and mean SSIM, which are computed between input images and the corresponding translated images. From Table 6.1, it is visible that style transfer methods proposed by [81] increased the memorability scores but not better than the proposed method. It is also evident that the proposed model has been learned to modify the input image to increase its memorability (visible from the MMSD values) by preserving most of the image contents (evident from the SSIM values). Further, Table 6.2 shows qualitative results with a few example images translated using proposed and existing models. From Table 6.2, it is visible that the images generated from the proposed model modified in terms of color and texture to increase the memorability property (indicated from PMS values). From the translated images, it can also be noticed that the proposed method retains most of the high-level contents of the input image compared to other methods.

6.3 Summary

In this chapter, a novel memorability based image-to-image translation method is proposed using deep learning approach. The proposed method modifies the given

6. MEMORABILITY BASED IMAGE TO IMAGE TRANSLATION

Table 6.2: *Examples of style transfer methods along with the proposed method. PMS is reported for each image.*






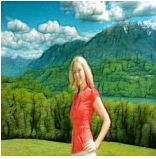







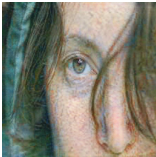
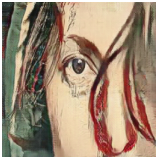
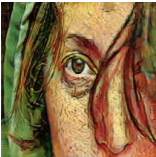
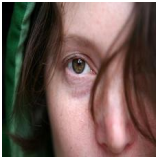

Cezanne [81]	Monet [81]	Ukiyo-e [81]	Van Gogh [81]	Original	Proposed Method
					
PMS = 0.764	PMS = 0.774	PMS = 0.766	PMS = 0.778	PMS = 0.762	PMS = 0.788
					
PMS = 0.749	PMS = 0.750	PMS = 0.753	PMS = 0.755	PMS = 0.746	PMS = 0.789
					
PMS = 0.794	PMS = 0.797	PMS = 0.794	PMS = 0.805	PMS = 0.784	PMS = 0.810

image to increase its memorability score while retaining its high-level contents. Also, the developed method learned the mapping between two image domains without using paired (input, label) image dataset like conventional image translation techniques. Experimental results showed that the proposed method increases the memorability score of a given image higher than that of the state-of-the-art image-to-image translation techniques.

The next chapter concludes the thesis by briefly summarizing the work presented in the dissertation and explaining the future research directions.

Conclusion and Future Works

7.1 Summary of the Contributions

In this dissertation, major part of the work is motivated to understand and predict the memorability at object and image levels. Towards this goal, the relationship of memorability with different visual factors which influence the memorability at object and image levels are analyzed. Also, various deep learning models are developed to predict memorability scores at image and object level. Further, an application of image memorability prediction model is devised to increase the memorability of an image using an end-to-end deep learning model. A brief summary of these contributions is narrated in the following subsections.

7.1.1 Object Memorability Prediction: Location and Size Bias

In the first contributory chapter, the relationship between object memorability and its two spatial characteristics, such as *Spatial-location* and *Spatial-size*, is explored. Various experiments are conducted to understand the influence of these two spatial characteristics on object memorability. From the experimental results, it has been shown that (a) objects of larger size tend to be more memorable than objects of smaller size, and (b) objects present at the centre of the image tend to be more memorable than the objects present at the corners. Further, a deep

7. CONCLUSION AND FUTURE WORKS

learning based object memorability prediction model is proposed to utilize the proposed spatial characteristics along with other object features. Experimental results highlight that the *Spatial-location* and *Spatial-size* of an object play a significant role in object memorability prediction and the proposed deep learning model performs better than the existing object memorability prediction model.

7.1.2 Image Memorability: The Role of Depth and Motion

The second contributory chapter extended the study of memorability concept to image level and explored the relationship between image memorability and two important image features: motion and depth. Various experiments have been conducted to understand the influence of these two features in making an image memorable or forgettable. From the experimental results, it has been shown that (a) images containing objects in motion tend to be more memorable, (b) images containing objects nearer to the camera at the center tend to be more memorable, and (c) images containing objects farther from the camera at the center tend to be less memorable. Further, deep learning based image memorability prediction models are proposed which utilize motion and depth cues along with the object features to predict memorability scores. Experimental results demonstrated that the proposed models perform better than the current state-of-the-art model indicating depth and motion are two important visual cues which need to be considered in image memorability prediction.

7.1.3 Visual Emotion based Image Memorability Prediction using Multiple Instance Learning

From the existing literature, it is evident that visual emotions have a significant role in making an image memorable or forgettable. However, the existing image memorability prediction methods have not been considered emotion cues in predicting memorability scores. In the third contributory chapter, a multiple instance learning based deep CNN is proposed to utilize visual emotion cues along

with other deep object features to predict image memorability scores. Experimental results depicted that incorporation of emotion cues through MIL framework improved the memorability prediction task and the proposed model performed better than the current state-of-the-art model by achieving a rank correlation close to human consistency.

7.1.4 Image Memorability Enhancement using Memorability based Image-to-Image Translation

In the fourth and final contributory chapter, an end-to-end deep learning model is proposed to enhance the memorability of a generic image. Since the aim of the proposed scheme is to translate an input image to another image having higher memorability, the underlying problem has been considered as memorability based image-to-image translation. The proposed model modifies the given input image to increase its memorability score while retaining its high-level contents. Also, the developed method learned the mapping between two image domains without using paired (input, label) image dataset. To the best of our knowledge, the proposed model is the first of its kind. Experimental results showed that the proposed model increases the memorability score of the given image higher than that of the state-of-the-art general image-to-image translation techniques.

7.2 Future Scope

The present study of this dissertation can be extended further in several directions, as listed below:

- The proposed memorability prediction models are limited to object and image levels. Therefore, the study of memorability can be extended for the video sequence. This extension may open many opportunities to understand the relationship of memorability at video level, including various temporal dynamics. Also, it may enable to build video memorability prediction models to predict memorability scores for the given video.

7. CONCLUSION AND FUTURE WORKS

- Existing and proposed research works on image and object memorability have shed light on various visual characteristics which influence memorability. All these works are with respect to 2D images. The effect of these visual factors on memorability with respect to 3D images is yet to be discovered, and the proposed memorability prediction models can be extended accordingly.

References

- [1] P. Isola, J. Xiao, A. Torralba, and A. Oliva, “What makes an image memorable?” in *CVPR , 2011 IEEE Conference on*. IEEE, 2011, pp. 145–152. [Pg.xix], [Pg.xxiii], [Pg.xxiv], [Pg.1], [Pg.2], [Pg.3], [Pg.4], [Pg.5], [Pg.6], [Pg.7], [Pg.9], [Pg.10], [Pg.24], [Pg.27], [Pg.28], [Pg.30], [Pg.65], [Pg.66], [Pg.68], [Pg.69], [Pg.80], [Pg.83], [Pg.84], [Pg.88]

- [2] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, “Understanding and predicting image memorability at a large scale,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015. [Pg.xix], [Pg.xxiii], [Pg.xxiv], [Pg.1], [Pg.2], [Pg.3], [Pg.6], [Pg.8], [Pg.9], [Pg.11], [Pg.24], [Pg.27], [Pg.39], [Pg.41], [Pg.52], [Pg.53], [Pg.54], [Pg.56], [Pg.57], [Pg.58], [Pg.59], [Pg.60], [Pg.65], [Pg.66], [Pg.67], [Pg.68], [Pg.69], [Pg.70], [Pg.72], [Pg.74], [Pg.80], [Pg.81], [Pg.82], [Pg.83], [Pg.84], [Pg.86], [Pg.87], [Pg.88], [Pg.89], [Pg.92], [Pg.95], [Pg.97], [Pg.98]

- [3] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem, “What makes an object memorable?” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1089–1097. [Pg.xx], [Pg.xxiii], [Pg.xxiv], [Pg.2], [Pg.3], [Pg.9], [Pg.10], [Pg.25], [Pg.27], [Pg.28], [Pg.29], [Pg.37], [Pg.38], [Pg.39], [Pg.42], [Pg.43], [Pg.44], [Pg.45], [Pg.47], [Pg.48], [Pg.49], [Pg.66], [Pg.68], [Pg.70], [Pg.76], [Pg.80], [Pg.83], [Pg.84], [Pg.88]

REFERENCES

- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105. [Pg.xx], [Pg.19], [Pg.21], [Pg.23], [Pg.28], [Pg.37], [Pg.38], [Pg.39], [Pg.40]
- [5] J. Pont-tuset, “Multiscale combinatorial grouping,” in *In CVPR*. Citeseer, 2014. [Pg.xx], [Pg.xxiii], [Pg.28], [Pg.43], [Pg.44], [Pg.45], [Pg.46]
- [6] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O’Connor, “Shallow and deep convolutional networks for saliency prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606. [Pg.78]
- [7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *arxiv*, 2016. [Pg.xxiv], [Pg.92], [Pg.99]
- [8] M. Hilbert, “How much information is there in the information society?” *Significance*, vol. 9, no. 4, pp. 8–12, 2012. [Pg.1]
- [9] A. F. Blackwell, “Correction: A picture is worth 84.1 words,” in *Proceedings of the first ESP student workshop*, 1997, pp. 15–22. [Pg.1]
- [10] L. Standing, “Learning 10000 pictures,” *The Quarterly journal of experimental psychology*, vol. 25, no. 2, pp. 207–222, 1973. [Pg.1]
- [11] I. Rock and P. Englestein, “A study of memory for visual form.” *The American Journal of Psychology*, 1959. [Pg.1], [Pg.2]
- [12] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, “Visual long-term memory has a massive storage capacity for object details,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 325–14 329, 2008. [Pg.1], [Pg.2], [Pg.5]

- [13] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva, “Conceptual distinctiveness supports detailed visual long-term memory for real-world objects.” *Journal of Experimental Psychology: General*, vol. 139, no. 3, p. 558, 2010. [Pg.1], [Pg.2]
- [14] —, “Scene memory is more detailed than you think the role of categories in visual long-term memory,” *Psychological Science*, vol. 21, no. 11, pp. 1551–1556, 2010. [Pg.2]
- [15] A. M. Turk, “Amazon mechanical turk,” *Retrieved August*, vol. 17, p. 2012, 2012. [Pg.2]
- [16] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3485–3492. [Pg.3], [Pg.24]
- [17] R. R. Hunt and J. B. Worthen, *Distinctiveness and memory*. Oxford University Press, 2006. [Pg.3], [Pg.5]
- [18] C. Spearman, “The proof and measurement of association between two things,” *American journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. [Pg.4], [Pg.23]
- [19] S. Maren, “Long-term potentiation in the amygdala: a mechanism for emotional learning and memory,” *Trends in neurosciences*, vol. 22, no. 12, pp. 561–567, 1999. [Pg.5]
- [20] A. K. Anderson, P. E. Wais, and J. D. Gabrieli, “Emotion enhances remembrance of neutral events past,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 5, pp. 1599–1604, 2006. [Pg.5]

REFERENCES

- [21] E. A. Phelps, “Human emotion and memory: interactions of the amygdala and hippocampal complex,” *Current opinion in neurobiology*, vol. 14, no. 2, pp. 198–202, 2004. [Pg.5]
- [22] M. M. Bradley, M. K. Greenwald, M. C. Petry, and P. J. Lang, “Remembering pictures: pleasure and arousal in memory.” *Journal of experimental psychology: Learning, Memory, and Cognition*, vol. 18, no. 2, p. 379, 1992. [Pg.5]
- [23] P. Isola, D. Parikh, A. Torralba, and A. Oliva, “Understanding the intrinsic memorability of images,” in *Advances in NIPS*, 2011. [Pg.5], [Pg.27], [Pg.85]
- [24] M. Mancas and O. Le Meur, “Memorability of natural scenes: The role of attention,” in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013. [Pg.6], [Pg.7], [Pg.27]
- [25] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet, “Deep learning for image memorability prediction: the emotional bias,” in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 491–495. [Pg.6], [Pg.9], [Pg.27], [Pg.72]
- [26] M. E. Gist and T. R. Mitchell, “Self-efficacy: A theoretical analysis of its determinants and malleability,” *Academy of Management review*, vol. 17, no. 2, pp. 183–211, 1992. [Pg.7]
- [27] D. G. Lowe *et al.*, “Object recognition from local scale-invariant features.” in *iccv*, vol. 99, no. 2, 1999, pp. 1150–1157. [Pg.7]
- [28] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893. [Pg.7]

- [29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402. [Pg.7], [Pg.24]
- [30] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, “Memorability of image regions.” in *NIPS*, vol. 2, 2012, p. 4. [Pg.7], [Pg.9], [Pg.28]
- [31] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, “Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis,” *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013. [Pg.7]
- [32] H. Peng, K. Li, B. Li, H. Ling, W. Xiong, and W. Hu, “Predicting image memorability by multi-view adaptive regression,” in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1147–1150. [Pg.7]
- [33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105. [Pg.8], [Pg.10], [Pg.17], [Pg.19], [Pg.65], [Pg.68], [Pg.74], [Pg.78], [Pg.82]
- [34] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. [Pg.8], [Pg.17], [Pg.19], [Pg.21], [Pg.61], [Pg.62], [Pg.68], [Pg.72], [Pg.74], [Pg.78], [Pg.82], [Pg.95]
- [35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, “Going deeper with convolutions.” *Cvpr*, 2015. [Pg.8], [Pg.9], [Pg.62], [Pg.74], [Pg.78]
- [36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pat-*

REFERENCES

- tern recognition*, 2016, pp. 770–778. [Pg.8], [Pg.17], [Pg.19], [Pg.22], [Pg.59], [Pg.62], [Pg.74], [Pg.78]
- [37] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. [Pg.8]
- [38] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017. [Pg.8]
- [39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in *Advances in neural information processing systems*, 2014, pp. 487–495. [Pg.9], [Pg.57], [Pg.74]
- [40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” vol. 115, no. 3. Springer, 2015, pp. 211–252. [Pg.9], [Pg.10], [Pg.28], [Pg.38], [Pg.57], [Pg.68], [Pg.74], [Pg.82], [Pg.88]
- [41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014. [Pg.20]
- [42] M. J. Huiskes and M. S. Lew, “The mir flickr retrieval evaluation,” in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 39–43. [Pg.25]
- [43] N. Murray, L. Marchesotti, and F. Perronnin, “Ava: A large-scale database for aesthetic visual analysis,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415. [Pg.25]

- [44] J. Machajdik and A. Hanbury, “Affective image classification using features inspired by psychology and art theory,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 83–92. [Pg.25], [Pg.83]
- [45] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2106–2113. [Pg.25]
- [46] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, “An eye fixation database for saliency detection in images,” in *European Conference on Computer Vision*. Springer, 2010, pp. 30–43. [Pg.25]
- [47] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, “The secrets of salient object segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287. [Pg.25], [Pg.42]
- [48] M. Everingham, “The pascal visual object classes challenge 2007,” in <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>, 2009. [Pg.25]
- [49] W. A. Bainbridge, P. Isola, and A. Oliva, “The intrinsic memorability of face photographs.” *Journal of Experimental Psychology: General*, vol. 142, no. 4, p. 1323, 2013. [Pg.27]
- [50] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, “Intrinsic and extrinsic effects on image memorability,” vol. 116. Elsevier, 2015. [Pg.27]
- [51] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva, “Image memorability and visual inception,” in *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 2012. [Pg.27]

REFERENCES

- [52] J. Kim, S. Yoon, and V. Pavlovic, “Relative spatial features for image memorability,” in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013. [Pg.27]
- [53] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos *et al.*, “Understanding and predicting importance in images,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3562–3569. [Pg.29], [Pg.30]
- [54] A. Anderson, K. Shaffer, A. Yankov, C. D. Corley, and N. O. Hodas, “Beyond fine tuning: A modular approach to learning on small data,” *arXiv preprint arXiv:1611.01714*, 2016. [Pg.40], [Pg.41], [Pg.44]
- [55] F. Cozman and E. Krotkov, “Depth from scattering,” in *cvpr*. IEEE, 1997, p. 801. [Pg.51]
- [56] W. Lu, J. Qi, Q. Liu, Z. Zhou, and J. Yang, “Depth estimation for image dehazing of surveillance on education,” *Journal of Intelligent & Fuzzy Systems*, vol. 31, no. 5, pp. 2629–2636, 2016. [Pg.51]
- [57] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, “Depth really matters: Improving visual salient region detection with depth.” in *BMVC*, 2013. [Pg.51]
- [58] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, “Depth matters: Influence of depth cues on visual saliency,” in *Computer vision–ECCV 2012*. Springer, 2012, pp. 101–115. [Pg.51]
- [59] L. Shen, R. Fang, Y. Yao, X. Geng, and D. Wu, “No-reference stereoscopic image quality assessment based on image distortion and stereo perceptual information,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, no. 99, pp. 1–14, 2018. [Pg.51]

- [60] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *Proceedings of the 14th ACM international conference on Multimedia*. ACM, 2006, pp. 815–824. [Pg.51]
- [61] M. Safaei and H. Foroosh, “Single image action recognition by predicting space-time saliency,” *arXiv preprint arXiv:1705.04641*, 2017. [Pg.51]
- [62] J. Walker, A. Gupta, and M. Hebert, “Dense optical flow prediction from a static image,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2443–2451. [Pg.53], [Pg.58], [Pg.60], [Pg.61], [Pg.67]
- [63] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 239–248. [Pg.55], [Pg.59], [Pg.60], [Pg.63], [Pg.67]
- [64] I. Maqsood, M. R. Khan, and A. Abraham, “An ensemble of neural networks for weather forecasting,” *Neural Computing & Applications*, vol. 13, no. 2, pp. 112–122, 2004. [Pg.60], [Pg.80]
- [65] M. P. Perrone and L. N. Cooper, “When networks disagree: Ensemble methods for hybrid neural networks,” in *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*. World Scientific, 1995, pp. 342–358. [Pg.60], [Pg.80]
- [66] Y. Niu, R. M. Todd, M. J. Kyan, and A. K. Anderson, “Visual and emotional salience influence eye movements.” *TAP*, vol. 9, no. 3, pp. 13–1, 2012. [Pg.71]
- [67] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, “Affective image retrieval via multi-graph learning,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1025–1028. [Pg.71]

REFERENCES

- [68] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, “Audio–visual emotion-aware cloud gaming framework,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 2105–2118, 2015. [Pg.71]
- [69] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, “Multiple instance learning: A survey of problem characteristics and applications,” *Pattern Recognition*, 2017. [Pg.75]
- [70] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997. [Pg.75]
- [71] C. Zhang, J. C. Platt, and P. A. Viola, “Multiple instance boosting for object detection,” in *Advances in neural information processing systems*, 2006, pp. 1417–1424. [Pg.75]
- [72] S. Vijayanarasimhan and K. Grauman, “Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8. [Pg.75]
- [73] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang, “Text-based image retrieval using progressive multi-instance learning,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2049–2055. [Pg.75]
- [74] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, “Weakly-supervised discovery of visual pattern configurations,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1637–1645. [Pg.75]
- [75] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, “Deep learning of feature representation with multiple instance learning for medical image analysis,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1626–1630. [Pg.75]

- [76] P. O. Pinheiro and R. Collobert, “From image-level to pixel-level labeling with convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721. [Pg.76]
- [77] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, “Deep multi-patch aggregation network for image style, aesthetics, and quality estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 990–998. [Pg.76]
- [78] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, “Multi-scale blocks based image emotion classification using multiple instance learning,” in *Image Processing (ICIP), 2016 IEEE International Conference on.* IEEE, 2016, pp. 634–638. [Pg.76], [Pg.78], [Pg.81], [Pg.85]
- [79] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva, “Modifying the memorability of face photographs,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3200–3207. [Pg.91]
- [80] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, “A mixed bag of emotions: Model, predict, and transfer emotion distributions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 860–868. [Pg.91]
- [81] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232. [Pg.92], [Pg.99], [Pg.100]
- [82] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014. [Pg.98]
- [83] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, “Complex wavelet structural similarity: A new image similarity index,” *IEEE transactions on image processing*, vol. 18, no. 11, pp. 2385–2401, 2009. [Pg.98]

Appendix: A

Journal Publications:

- [1] **Sathisha Basavaraju**, Sibaji Gaj, Arijit Sur. “*Object Memorability Prediction using Deep Learning: Location and Size Bias*”. **Journal of Visual Communication and Image Representation - Elsevier.**, vol 59, pp. 117-127. DOI: <https://doi.org/10.1016/j.jvcir.2019.01.008>
- [2] **Sathisha Basavaraju**, Arijit Sur. “*Multiple Instance Learning based Deep CNN for Image Memorability Prediction*”. **Multimedia Tools and Applications - Springer**, vol 78(24), pp. 35511-35535. DOI: <https://doi.org/10.1007/s11042-019-08202-y>
- [3] **Sathisha Basavaraju**, Arijit Sur. “*Image Memorability Prediction using Depth and Motion Cues*”. **IEEE Transactions on Computational Social Systems**. 2020, Early Access. DOI: <https://doi.org/10.1109/TCSS.2020.2973208>

Conference Publications:

- [1] **Sathisha Basavaraju**, Paritosh Mittal, Arijit Sur. “*Image Memorability: The Role of Depth and Motion*”.: 25th **IEEE International Conference on Image Processing (ICIP)**, Athens Greece, Oct, 2018. DOI: <https://doi.org/10.1109/ICIP.2018.8451334>

- [2] **Sathisha Basavaraju**, Prasen Kumar Sharma, Arijit Sur. “*Memorability Based Image to Image Translation*”. **Twelfth International Conference on Machine Vision (ICMV 2019)** 11433, 114331G.
DOI:<https://doi.org/10.1117/12.2556543>