REAL-TIME ASSISTANCE IN PHOTOGRAPHY USING SOCIAL MEDIA

YOGESH SINGH RAWAT

NATIONAL UNIVERSITY OF SINGAPORE

2017

REAL-TIME ASSISTANCE IN PHOTOGRAPHY USING SOCIAL MEDIA

YOGESH SINGH RAWAT

(*B.Tech (Hons.), IIT(BHU), Varanasi, India*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF COMPUTER SCIENCE
SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE

2017

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has not been submitted for any degree in any university previously.

Yogesh Singh Rawat
Friday 27th January, 2017

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Mohan S Kankanhalli whose valuable guidance, suggestions, and motivation have helped me throughout the research. I could not have imagined having a better supervisor for my PhD research.

I would like to thank my thesis committee members, Professor Michael Brown and Professor Roger Zimmermann, for their constructive criticisms and insightful comments. I gratefully acknowledge the funding sources, Ministry of Education and the National University of Singapore for providing the financial support during the PhD research. I would also like to acknowledge the guidance of Professor Ramesh Chandra Jain and Professor Vivek Kumar Singh who's insightful comments and suggestions greatly assisted me in the research.

I would also like to thank all the colleagues: Dr. Wong Yong Kang, Dr. Christian Von Der Weth, Dr. Prabhu Natarajan, Dr. Gan Tian, Wang Yuhui, Zhang Yehong, Chen Xiang and many more for their help along the way. Finally, I take this opportunity to thank my family and all of my friends for their support. I would like to express the gratitude to my beloved parents, my brother and especially to my spouse for her continuous support and motivation.

Table of Contents

Summary

In the last decade, we have seen significant improvement in the ease and cost of capturing multimedia content. However, the aesthetic quality of the content captured by an amateur user still needs substantial improvement. Camera devices have intelligent features, such as automatic focus, face detection, etc, to assist users in taking better photos, however, it remains a challenge for an amateur user to capture high-quality photographs. The complex nature of photography makes it difficult to provide real-time assistance to a user for capturing high-quality images. However, advancement in digital photography, sensor technology, wireless networks and social media provides us an opportunity to enhance the photography experience of users.

This doctoral research aims at providing real-time photography assistance to users by leveraging on camera sensors and social media content. The research in this thesis is focused on two different aspects of user experience in photography. The first contribution focuses on camera guidance and the second contribution is focused on location recommendation for photography.

In our first contribution, we developed computational models based on machine learning which can provide real-time camera guidance to users for capturing high-quality photographs. The proposed models utilize publicly available photographs along with social media cues and associated metadata for photography learning. In the first part, we focus on landmark photography where a feedback regarding scene composition and camera parameter settings is provided to a user while a photograph is being captured. We propose the idea of computing the photographic composition basis, *eigenrules* and *baserules*, to support our composition learning. As context is an important factor from a photography perspective, we also explore the role of user-context in photography recommendation. In the second part, we focus on group photography where we use the idea of spring-electric graph model and augment it with the concept of color energy from the literature of visual arts. The proposed model is applied in group photography utilizing social media images to provide real-time feedback to the user regarding the arrangement of people, their position on image frame and relative size.

In the second contribution, we focus on location recommendation for photography to improve the experience of users at tourist locations. Firstly, we propose *ClickSmart*, a viewpoint recommendation system which can provide real-time guidance based on the preview on user's camera, current time and user's geo-location. It makes use of publicly available geotagged images along with associated metadata for learning a recommendation model. We define *view-cells*, macro blocks in geospace, and propose the idea of *popularity*, *quality* and *uniqueness* of view-cells from viewpoint perspective. Finally, we propose a photography trip recommendation method which guides a user in exploring any tourist location from the photography perspective. More specifically, a tour is recommended to the user based on Optimal Foraging Theory and social media images which provide a list of hot-spots to visit and corresponding stay time at each hot-spot for photography. We have conductive extensive experiments and user studies to demonstrate the effectiveness of the proposed methods.

List of Tables

List of Figures

Abbreviations

3D	**Three** **D**imensional
AoI	**A**rea **o**f **I**nterest
API	**A**pplication **P**rogram **I**nterface
AQS	**A**ctive **Q**uery **S**ensing
BIC	**B**ayesian **I**nformation **C**riterion
BoAP	**B**ag **o**f **A**esthetic **P**reserving feature
CBIR	**C**ontext **B**ased **I**mage **R**etrieval
CNN	**C**onvoluted **N**eural **N**etwork
DB-1	**D**ata **B**ase-1
DB-2	**D**ata **B**ase-2
EM	**E**xpectation **M**aximization
EV	**E**xposure **V**alue
Exif	**Ex**changeable image file format
GMM	**G**aussian **M**ixture **M**atrix
GPS	**G**lobal **P**ositioning **S**ystem
HOG	**H**istogram **o**f **O**riented **G**radients
HSV	Hue, Saturation and Value
IPC	**I**nternet **P**hoto **C**ollection
ISO	**I**nternational **O**rganization of **S**tandardization for sensitivity of camera sensor
lmo	**L**and**m**ark **o**bject
LDA	**L**atent **D**irichlet **A**llocation
MAP	**M**ean **A**verage **P**recision
MSE	**M**ean **S**quared **E**rror
micro-poi	**Micro**-**p**oint **o**f **i**nterest
nDCG	**n**ormal **D**iscounted **C**umulative **G**ain
NMF	**N**on-negative **M**atrix **F**actorization
MVT	**M**arginal **V**alue **T**heorem

ODM	**O**ptimal **D**iet **M**odel
OFT	**O**ptimal **F**oraging **T**heory
PCA	**P**rincipal **C**omponent **A**nalysis
PM	**P**opularity **M**ap
poi	**p**oint **o**f **i**nterest
PSM	**P**osition, **S**ize **M**odel
QM	**Q**uality **M**ap
RAR	**R**elative **A**rea **R**ank
RBF	**R**adial **B**asis **F**unction
RGB	**R**ed **G**reen **B**lue
RM	**R**ecommendation **M**ap
ROI	**R**egion **o**f **I**nterest
SIFT	**S**cale-**I**nvariant **F**eature **T**ransform
SLIC	**S**imple **L**inear **I**terative **C**lustering
SSIM	**S**tructural **Sim**ilarity
SURF	**S**peeded-**U**p **R**obust **F**eatures
SVM	**S**upport **V**ector **M**achine
UM	**U**niqueness **M**ap
WOA	**W**eighted **O**bject **A**rea

Symbols

α	The angle between the positions of two different configurations in a layout
β	Empirical weight for number-of-favorites in the evaluation of aesthetic score
γ	Empirical weight for number-of-comments in the evaluation of aesthetic score
Γ	Offset for diminishing gain curve
δ	Threshold value as a stopping criteria for energy optimization
Δ	Empirical constant for average gain in computing profitability
ε	Weight given to personal preference in computing profitability
ζ	Empirical constant used in computing sparseness of a view-cell
η	Empirical constant used in computing sparseness of a view-cell
ε	Empirical constant in diminishing gain curve
θ	Empirical constant as a weight for location popularity in computing profitability
ϑ	Empirical weight for interestingness in the evaluation of aesthetic score
Θ	Empirical weight for social media popularity in the evaluation of profitability
ι	Empirical constant in the evaluation of aesthetic score of an image
κ	Empirical constant in the evaluation of aesthetic score of an image
λ	The average rate of patch encounter
Λ	Variation in the stay time
μ	Mean
ν	Empirical constant for uniqueness in view-point recommendation
Ξ	Empirical constant for quality and popularity in view-point recommendation
π	Profitability
τ	Empirical constant as a weight for time in computing aesthetic score
υ	Empirical weight for number-of-views in the evaluation of aesthetic score
a_i	Aesthetic quality score of i^{th} visual object
A	Relative aperture value in camera
B	Brightness of a color
c_m	A 2-D representation (x,y) of the centre position in an image frame

C	Contrast of a visual object
d_{ij}	euclidean distance between point i and j
DCG_ρ	Discounted cumulative gain for the top ρ recommended items
E	Total energy of a spring-electric graph
$E_i{}^c$	Color energy of ith node in a graph
EV_s	exposure value at ISO S
f	Total number of user-favorites for an image
f_m	The median of number of user-favorites for an image
f^a	Attractive force between nodes
f^r	Repulsive force between nodes
h	Height of an image frame
H	Warmness of hue
H_s	Shanon's diversity index
I	Interestingness score of an image
lmo_i	The ith landmark object
N	The total number of images at a location
N_i	Total number of photographs at i^{th} point-of-interest
Num_i	Number of visual objects in a cluster
Num_{max}	The size of the largest cluster
Pop	The popularity score of a location
$q-rec$	Quality based view-point recommendation
R^2	The coefficient of determination
S	Saturation of a color
S_i	Sparseness of the i^{th} location
Sal_i	Saliency score of i^{th} visual object
T	Shutter speed of a camera
T^k	Total number of viewpoint samples
U_i	Uniqueness score of a view-cell
Q_i	Quality measure of i^{th} view-cell
w	Width of an image frame
x_i	The position of i^{th} node in a graph

t	Time in seconds
$t - rec$	Time based recommendation
$uq - rec$	Uniqueness and quality based view-point recommendation
$ut - rec$	Uniqueness and time based view-point recommendation
$uw - rec$	Uniqueness and weather based view-point recommendation
v	Total number of user views
v_m	The median of number of user views
$w - rec$	weather based view-point recommendation

To my wife and my family

Chapter 1

Introduction

We all love to capture important moments in our life and share them with our dear ones. With the advancement in technology, affordable mobile devices come with high-end embedded cameras and people have started taking more photos to capture their experiences. Developments in wireless technology also allow us to share our experiences on the move. Thus, technology has made it easier for us to capture and share our experiences with family and friends.

However, devices with advanced features cannot always guarantee high-quality multimedia capture. Although mobile cameras have advanced facilities like auto-focus, face detection, etc., for assisting users in capturing better photos, these are not sufficient considering the complex nature of photography. It still remains a challenge for an amateur user to take high-quality photos. Therefore, ease and cost of image capture have improved but the quality of capture needs substantial improvement.

1.1 Complex Nature of Photography

With a decent digital camera and a bit of practice, anyone can take photos with the camera set on *automatic* mode. We can even take average quality pictures and make them look good with image post-processing tools. But to capture truly beautiful photographs, we need to utilize every possible ability of the camera and for this, we need to acquire some knowledge of photography and learn the manual camera settings.

1.1.1 Camera Parameters

One of the most important camera settings includes exposure, in which we try to find the proper exposure for the subject and lighting conditions [60], [83]. Exposure is the amount of light hitting the camera's sensor when we capture an image. Generally, we want the exposure setting so that the image captured by the camera's sensor closely matches with what we see with our naked eyes. In the automatic mode, a camera tries to accomplish this, but it is not perfect, which is why professional photographers use manual settings to produce better photographs.

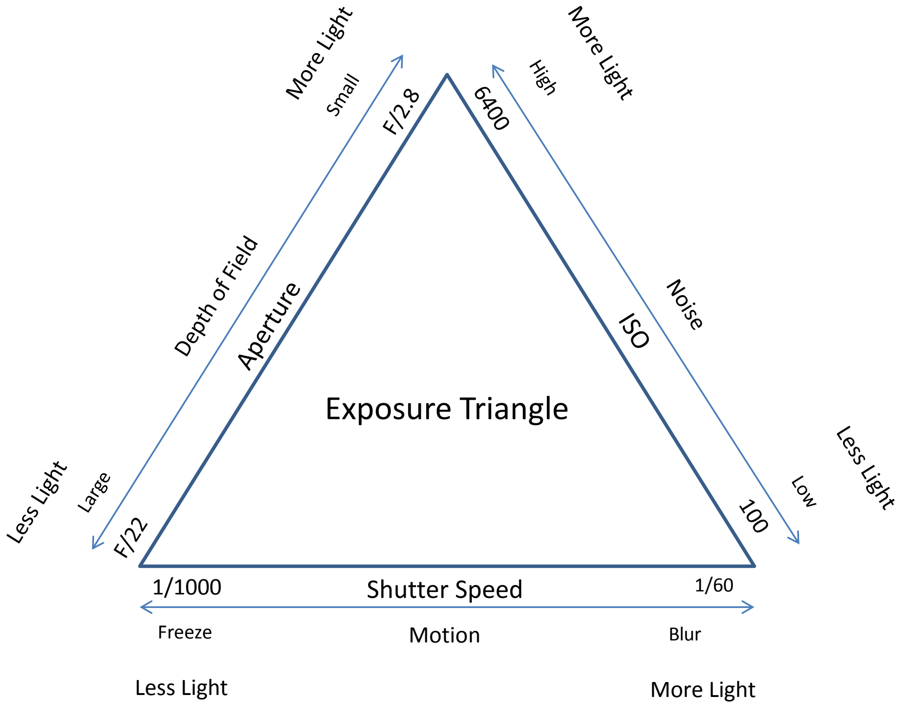


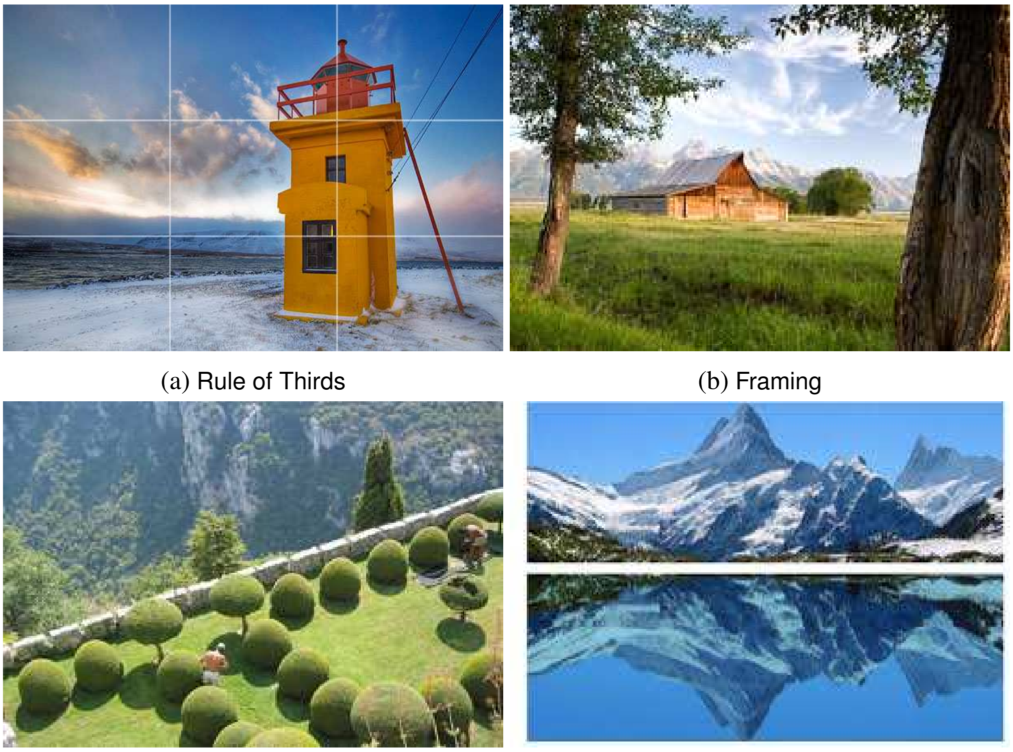
FIGURE 1.1: Exposure Triangle for Digital Photography [152]

Camera aperture, ISO sensitivity, and shutter speed are the parameters generally used to control the exposure (Figure 1.1). Aperture is the size of the opening in the lens when a picture is taken and ISO is the measure of a digital camera sensor's sensitivity to light. On the other hand, shutter speed controls the amount of time that the shutter is open. As we can see in figure 1.1, large aperture (f/2.8), high ISO (6400) and slow shutter speed (1/60) leads to high exposure and on the other hand small aperture (f/22), low ISO value (100) and high shutter speed (1/1000) will cause low exposure. Therefore, in bright light, a fast shutter speed and small aperture are used to control the amount of light that comes in and large aperture and lower ISO is used to get rich detail in low light conditions.

Adjustment of the above-mentioned parameters to capture a high-quality photograph is not so easy. Apart from controlling exposure, these settings depend on the type of photograph and they also affect other parameters. Aperture also controls the depth of field (Appendix B) and similarly shutter speed controls motion in the scene (Figure 1.1). For example, for portrait mode, we need large aperture which helps to keep the background out of focus (it sets a narrow depth of field to focus only on the main subject). Professional photographers practice and gain experience to figure out which combinations of aperture, ISO and shutter speed are best for different kinds of photos. Playing with ISO level and exposure can cause digital noise which degrades the image quality. For example, longer exposure heats up the camera sensor and this heat contributes to digital noise in the final image. Similarly, slower shutter speeds are often used to get enough light which can make it very difficult to take a photo without some blurring. Moreover, there are other parameters like depth of field, white balance, etc., which photographers have to decide based on the context (Appendix B).

1.1.2 Image Composition

Apart from the camera settings, the role of photography knowledge can not be ignored in capturing high-quality photographs. The composition of a photo is one of the essential factors in the art of photography. Appendix A provides a list of some of the important rules of composition followed by professional photographers. A good and balanced composition can make a photograph look more attractive even if the scene being shot is not appealing. In photography, widely accepted principles such as *rule of thirds, leading lines, repeating patterns, layering, horizon lines, relative scale* are useful guidelines for creating better photos [47], [83]. Photographers follow these photography rules and guidelines while they capture images.


(a) Rule of Thirds

(b) Framing

(c) Diagonal

(d) Symmetry

FIGURE 1.2: Rules of Composition [47] [1]

Figure (1.2) presents some example photographs where photographers have followed some of these composition rules. In figure 1.2a *Rule of Thirds* is followed. The basic principle behind this rule is to imagine breaking an image down into thirds (both horizontally and vertically) so that we have 9 parts. The theory [47] is that if we place points of interest in the intersections or along the lines then the photo becomes more balanced and will enable a viewer of the image

[1] Image source: www.digital-photography-school.com, www.photographymad.com

to interact with it more naturally. Similarly other example photographs also follow some rules (*Framing, Diagonal* and *Symmetry*), details of which can be found in appendix A.

In group photography, the complexity of image composition increases further. Apart from adjusting the background scene, the photographer also needs to determine the arrangement, position, and size of people in the photograph to achieve a balanced composition. Professional photographers use their experience and photography knowledge to determine these parameters as they compose a scene.

Amateur photographers might learn and know these rules, but it is still difficult to apply them while taking pictures. Some of the photography rules contradict each other, such as *Rule of Thirds* and *Rule of Center*, and therefore cannot be applied together. The user has to decide which rule is more appropriate for a given view, which is not trivial. It usually requires years of practice and experience to transform the rules into real-time intuitions so that a spontaneously taken photo can have a better quality.

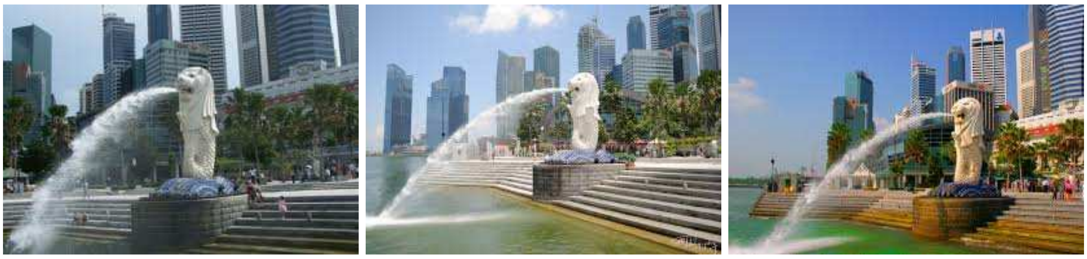
1.1.3 Viewpoint

In photography, viewpoint refers to the geo-location from where a photograph is captured and is considered as one of the essential factors in the art of photography [47]. It has a large impact on the composition of a photograph and as a result, it also affects the aesthetic quality of a captured image. Advanced digital cameras can provide features like auto-focus, face detection, etc., for assisting users in capturing better photos, however, it can be challenging for an amateur user to find a good viewpoint in any tourist location. It is a human tendency to follow others [127] and users generally follow the crowd to find a good viewpoint which can be misleading and therefore an amateur user may end up with bad quality photos. Although there are post-processing tools available to enhance photo quality, it is not an option for users who like to click and share their images instantly.

1.2 Motivation

Most people would agree that not everyone is a professional photographer and, capturing a good photograph is not always an easy task for amateur users. Figure 1.3 presents three different photographs of same location and view point. We can clearly see the difference between the average and the professional photograph. Apart from the difference in the color composition, the professional photograph is also following the *Rule of Thirds* and *Diagonal* in a more appropriate manner. Professional photographers usually take relevant training and practice for years to understand the photography knowledge. Sometimes they also use their intuitions depending upon the context to compose and capture a scene. Fortunately, we have social media sharing websites like *Flickr, Photo.net, Picasaweb, etc.*, with a large collection of photographs shared by professional and other users. With recent advances, these photographs also contain metadata in Exif (Exchangeable image file format), which provides us useful context information like, time of capture, geo-location, camera parameters, etc. Also, with advancement in sensor devices, it is easy to infer the current contextual information of the user. We can leverage the social media data for photography learning along with the user context information to assist amateur users during photo capture.

We have many post-processing tools like *Photoshop, Picasa, Inkscape, etc.*, which can be used for improving already captured photo quality. However, these are not an option for mobile users who love to click and share on the go as it delays the real-time sharing. These tools are meant for desktop PC's with high computational requirement thus not feasible for a real-time application. Moreover, post-processing tools can not do much in the case of poorly captured photos and not all the information lost during capture is recoverable using these tools. For example, it is always possible to use software to take certain areas of a photo out of focus, but it is very challenging to fix anything that's out of focus. Therefore, these photo editing tools will be of no use if the captured photo is not clicked properly. To avoid the use of any post-processing tool for enhancing the photo quality, most users try to capture as many photos as they can. Later, they browse through all the captured photos and select the best capture to share with their family and friends. Memory is becoming cheaper and digital photography allows us to capture as much as we can. But this is not desirable for a person who spends all his time in clicking photos rather than enjoying with his family or friends. Also, it is a very tedious task to look through the huge database of captured photos to find an attractive one. In this research, we propose to provide assistance during capture rather than post processing to avoid the unnecessarily captured data and improve the photography experience of the users at tourist locations.



(a) Average Photograph (b) Better composition (c) Professional photo

FIGURE 1.3: Same scene with different composition [2]

1.3 Contributions

In this thesis work we propose novel approaches for providing real-time assistance to users as they capture photographs. We mainly focus on camera guidance and location recommendation

[2] Image source: www.flickr.com

for photography. The camera guidance includes feedback regarding image composition, camera parameters, and group photography. And, the location recommendation includes viewpoint recommendation and tour recommendation from a photography perspective.

1.3.1 Camera Guidance

In camera guidance, we focus on landmark photography and group photography, which we will discuss in more detail in the following sub-sections.

1.3.1.1 Context-Aware Photography (Landmark Photography)

In this work, we propose a comprehensive system for photography assistance using crowd-sourced images which take into account available contextual information for image composition and camera parameter learning. We employ machine learning to build the models for image composition and camera parameters. The proposed framework has a real-time control feedback system where the input is sensed from the physical world (view, geo-context, etc.) and feedback is provided to the controller (photographer) to improve the image quality. This work has been published in [138, 139] and [137]. In summary, our primary contributions are:

- We propose a context based composition and camera parameter learning system which accounts for the presence of human objects in the photograph and is not limited by the number of main objects in the scene.

- We introduce the concepts of *eigenrules* and *baserules* to support composition learning. A detailed analysis is presented to understand their significance in image composition.

- Human position recommendation in a view is posed as an optimization problem and a run-time efficient solution is proposed which can provide position recommendation for multiple humans.

- We propose a camera motion feedback system which can guide the user to control the camera for pan, tilt and zoom motion to achieve a better composition of the scene.

1.3.1.2 Group Photography Assistance

In this work, we focus on group photography where we have multiple people standing in an image frame with a scenic view in the background. Obtaining visual balance in group photography is a challenging task as there are multiple parameters involved which affects the aesthetics quality of captured image. Some of the factors include the arrangement of people, their position in the image frame considering salient objects in the scene and distance how far they should stand from the camera. Professional photographers use their experience and knowledge to visualize how the visual elements in image frame could be better arranged, sized or positioned. However, it is not trivial for amateur users to obtain a visual balance in an image frame and capture a high-quality photograph.

We use the spring-electric model embedded with color energy to generate a real-time recommendation for users so that they can capture a visually balanced group photograph. The proposed method makes use of social media images to estimate a position and size where a group of people should stand in a photograph. The estimated position and size of people are then further optimized using a spring-electric model embedded with color energy which enables visual balance in the photograph. This work is currently under review [142]. We make the following novel contributions in this work.

- We introduce the idea of color energy from the art of composition and embed it in a spring-electric model to obtain a visual balance in an image frame.

- We present an application of this model in group photography where we leverage on social media images along with this model to produce real-time recommendation which can be used to capture high-quality group photographs. To the best of our knowledge, this is the first time the problem of group photography recommendation is being studied.

1.3.2 Location Recommendation

In location recommendation, we focus on viewpoint recommendation and trip recommendation from a photography perspective.

1.3.2.1 Context-Aware Viewpoint Recommendation

In this work, we propose *ClickSmart* which provides real-time viewpoint recommendation to the user based on the preview on the camera. It attempts to bridge the gap between view-based and location-based recommendation and also takes into account user contexts such as time and weather conditions. This work has been published in [140]. In summary, our primary contributions are:

- We propose a real-time viewpoint recommendation system for photography assistance which makes use of publicly available photographs via social media.
- We investigate the impact of context such as time and weather conditions on viewpoint recommendation.
- The proposed recommendation system also considers the presence of people in photographs, and finally
- We propose the idea of the uniqueness of geo-pixels which is further used in recommending rare but interesting viewpoints for photography.

1.3.2.2 Trip Recommendation

In this work, we focus on providing user recommendation to explore a tourist attraction from the photography perspective. We observe that each tourist attraction has multiple hots-spots (micro-poi: micro point of interest) which are visited by the tourists. There can be multiple ways to visit these micro-pois and searching for an optimal path is an NP-hard problem. We make use of social media images to learn previous patterns in the environment and employ Optimal Foraging Theory to determine an optimal path for exploring the attraction and capturing photographs. This work is currently under review [141]. Our contributions are,

- We propose a route recommendation method based on Optimal Foraging Theory which provides a path along with a list of micro-pois in a tourist attraction where the user should visit. In addition, we also recommend an optimal stay time for each of the micro-poi in the path which is determined based on the Marginal Value Theorem.
- The proposed method also takes into account personal preferences for providing a personalized recommendation.

1.4 Organization

The rest of the thesis is organized as follows. In chapter 2 we will discuss the related research work and present the state of the art in the relevant areas. Then in chapter 3, we will discuss our work on context-aware photography assistance. Chapter 4 will demonstrate the work on group photography assistance. In chapter 5, we will present the work on context-aware view-point recommendation. The work on route recommendation will be discussed in chapter 6. Finally, we will conclude the report and introduce future research directions in chapter 7.

Chapter 2

Literature Survey

In recent years, researchers have shown interest in photography assistance for amateur users. Although there are few works, which target to provide the assistance in real time, most of the studies focus on assistance provided after a photo has been captured. We will discuss the proposed methods and their limitations in the next section. Table 2.1 provides a brief overview of the presented literature review.

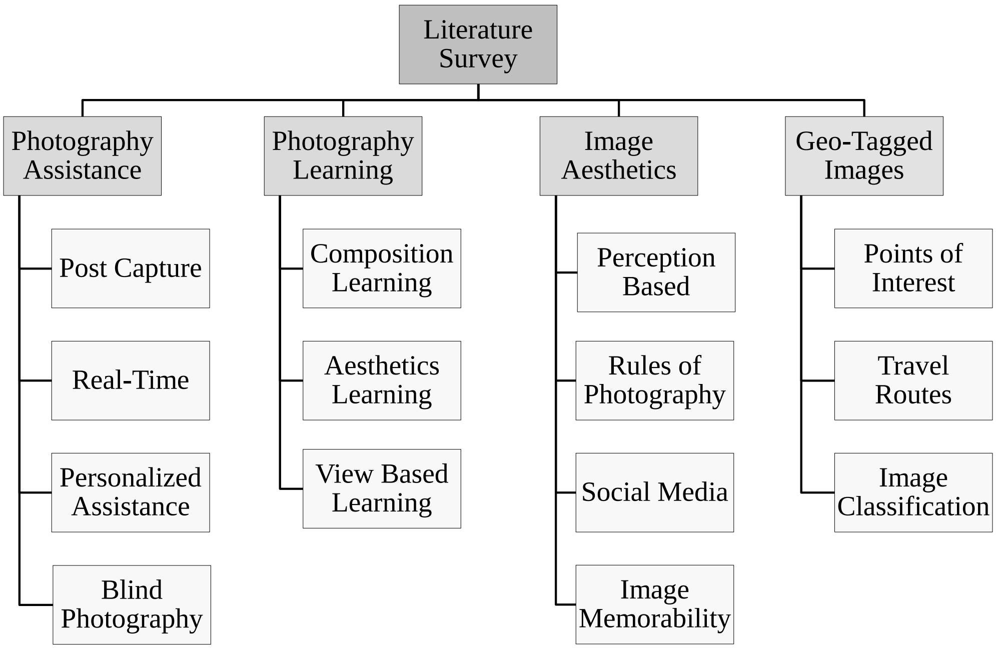


FIGURE 2.1: Literature Review Scheme

The main objective of photography assistance is to enable amateur users to capture a high-quality photograph without having any prior knowledge. During photography assistance, it is important to make sure that the user finally captures a high-quality photograph from the aesthetics perspective. Therefore aesthetic evaluation of images is of vital importance for this research as we may not be able to provide satisfactory assistance if we have no knowledge of image aesthetics. There has been a lot of work in image aesthetic evaluation in the past decade. Most of the work employs computationally expensive techniques which are not suitable for a real-time application. We will describe some of the important work in this area and discuss why they can not be directly used for our research.

In [4], Brett et al. proposed an integrated media creation environment for guiding amateur users in shooting videos. The presented approach takes into account various aspects of creating professional videos and provides a step by step guide to the user. However, the proposed method provides guidance only for creating videos.

Recently we have seen a lot of research in location-based services due to the widespread availability of GPS (Global Positioning System) sensors. Images with geo-location information can be utilized for a wide range of applications such as 3D model construction for a location. For this research, we want to utilize geo-tagged images from photography perspective for real-time photography assistance. We will discuss some of the important work in this area and then talk about the requirements for our research.

This chapter is organized as follows. In section 2.1 we will describe the proposed methods for photography assistance and their limitations. Thereafter we will discuss various computational photography learning techniques in section 2.2 followed by a review of work in image quality assessment in section 2.3. In section 2.4 we will discuss works related to geo-tagged images.

Finally, in section 2.5 we will conclude this chapter by presenting state of the art in this research area.

2.1 Photography Assistance

Photography assistance can be given either after capturing the image or as the image is being captured. There has been some work in photography assistance for already captured images, which either suggests similar scenes or do some post-processing to change the composition of the captured image. Researchers have employed both visual features based on photography rules and social media data. We will use the term post capture photography assistance for this kind of work. Also, some researchers have proposed real-time systems which can guide a user as they capture an image. They have also employed either social media data or photography rules to generate the suggestions. Apart from these two, we will also discuss work related to blind photography and personalized photography assistance.

2.1.1 Post Capture

Image processing and editing of captured photos to enhance aesthetics is a well-explored research area. In this section, we will discuss some of the important works which consider not only low-level visual features but also make use of the photography composition rules and social media data. Table 2.1 presents a detailed summary of the existing works.

Aesthetics and Photography Rules: Some of the early post-processing techniques include [123, 158, 178, 179, 201] for automatic image re-targeting, [29] image adaptation for smaller displays, [143, 146] image or video re-targeting. Suh et al. [158] developed an automatic image cropping method based on visual salient objects and face detection. Setlur et al. [146] employed segmentation and saliency to identify important regions in a photograph. The important regions are then re-organized in a resized background of the image. Zhang et al. [201] defined 14 templates based on photography composition rules and utilized them along with face detection for automatic photograph cropping.

Liu et al. [108] used certain photographic compositional rules such as *Rule of Thirds, Balancing Elements, Diagonal Rule, Size of Region*, to study the composition of photos. The proposed method employ a compound operator which re-targets a cropped part of the image into a target frame having a different dimension. The cropped frame with the highest aesthetic score is defined as the final re-targeted image. The aesthetic measurement is done on the basis of some heuristic photography rules. In the cropped image distortion of salient objects detected and patched back is quite possible.

In [18], the authors proposed a post-processing method in which they relocate the main object from the scene to a more pleasing location based on photography rules. The proposed method can also be used to crop or expand a given landscape image. The system segments main object area in a supervised way and then rearranges it to appropriate position using surface layout recovery [62] based on rule-of-thirds composition and visual weight ratio. Similarly, Mai et al. [120] presented a method to identify whether rule-of-thirds has been applied in capturing a photograph. The proposed method employ saliency map and object analysis for object localization and make use of power points A. However, these methods [18, 120] are based on only rule-of-thirds and there are many other popular heuristic rules which are used by photographers.

In [96], Lee et al. proposed an automatic approach for straightening up slanted man-made structures in an input image to improve its perceptual quality. Their method uses an energy minimization framework to compute an optimal homography that can effectively minimize the perceived distortion of slanted structures. All these proposed methods are computationally expensive and rely only on compositional heuristics.

Using Social Media: Chang et al. [26] proposed a system which finds a view enclosure in panoramic photographs with favorable compositions based on rules learned from exemplary photographs. The proposed system characterizes arrangement of structures and geometric patterns within an image using GIST descriptor [130] and use the saliency map using spectral residual approach [64] for the layout of salient elements. The system can suggest good view enclosure provided a panoramic scene and an exemplary photograph to learn from. The authors used a small set of exemplary photographs (around 100) in the learning dataset and this small dataset cannot be a holistic representation of all the photographic composition considering the wide range of possibilities. Their algorithm selects the initial set of exemplars randomly from the generated graph, therefore the system can have both scalability and performance issues as the initially selected exemplars may be totally unrelated to the input panorama.

Li et al. [100] proposed a system for photography learning using community contributed photo collection. The user captured image and a keyword, which describes the photo, is used to search the public image database and provides a user with similarly captured photographs along with the camera parameters (focus length, aperture, exposure time and ISO). The user can then explore these photos by changing the camera parameters and understand the effects of these on photo quality. The proposed system divides the photo type into three classes, close-up, mid-view and far-view and uses them for view-classification of photos. The system provides photographs similar to the user captured image based on the visual features and a description given by the user. The user has to manually provide a keyword description for each input image and will have to learn itself from the parameter settings of the retrieved photos. It might be helpful for a user

TABLE 2.1: Summary for Photography Assistance (Post Capture)

Approach	The work	Details	Aesthetic Assessment	Assistance	Dataset	Comments
Photography Knowledge and Perception Based	Setlur et al. 2005 [146]	Based on image segmentation and saliency map	none	Re-targeting salient objects in smaller frame	40	Detected salient objects are used re-compose the image in smaller frame size
	Zhang et al. 2005 [201]	Face detection and Region of Interest	Photographic rules: 14 defined templates	Main object relocation and image cropping	60 Images	Real time and also considered more than one person case
	Bhattacharya et al. 2010 [18]	Supervised segmentation for main object localization	Photographic rules: rule-of-thirds, visual weight ratio	Main object relocation and image cropping	632 Images	Supervised method (user involvement) for main object detection
	Liu et al. 2010 [108]	Based on saliency and prominent lines detection	Rule of Thirds, Balancing Elements, Diagonal Rule, etc.	Cropped image with best aesthetic score	900 Images	Employing heuristic photography rules for image re-targeting
	Mai et al. 2011 [120]	Saliency based object localization	Photographic rules: rule-of-thirds	Rule-of-thirds detection	4140 Images, 2089 positive and 2051 negative	Only determines whether a given photo follows rule-of-thirds
	Lee et al. 2013 [96]	Based on edge and corner point detection	Vertical alignment of objects	Straightening up slanted man-made structures in an input image to improve its perceptual quality.	[41] and [15]	Improving image aesthetics based on only vertical alignment of objects
Social Media and Data Driven Approach	Chang et al. 2009 [26]	Structural features (GIST descriptor [130]) and the layout of visual saliency learned from professional panoramic photographs	none	View enclosure in panoramic photographs	Set of 100 exemplar photographs	Small number of exemplars in the dataset (100), scalability and performance issues
	Li et al. 2011 [100]	Relevant images with metadata and view similarity are suggested to user from stored image database	none	Images based on view similarity suggested along with camera parameters	not stated	Only camera parameter suggested, not suitable for real-time application due to involvement of image search
	Yao et al. 2012 [186]	Edge detection and image segmentation along with photography heuristics are employed in the method	Photography heuristics and color composition	Provide aesthetic evaluation and list of high quality images based on view similarity	13,302	Run time of 7-8 seconds for processing a 256x256 image
	Zhang et al. 2013 [197]	Personal photograph enhancement using 3D models constructed employing public photos	none	Field of view expansion and image enhancement of person photographs	not specified	Quality of image enhancement will depend upon available public images

who intends to learn photography by spending some time, but it is not very intuitive for an amateur user who just wants to capture a good quality photo without spending much time on it. Also, the three class categorization of images is very crude which can result in a lot of retrieved images for a search query. Moreover, the authors have only considered the camera parameters ignoring the image composition which is important in learning photography.

In another work [186] Yao et al. presented a framework which provides a recommendation to the user based on composition and an aesthetic score of the captured image. Given a user captured image, as a feedback, the system provides a ranked list of high-quality images with a similar composition. Making use of photography heuristics on image composition, the authors categorized images into five groups, *diagonal, horizontal, vertical, texture, and center.* The proposed method employ segmentation and edge detection along with some photography heuristics for image classification. Apart from this it also utilizes color triplet sets as a basis for aesthetics computation. The color triplets are ranked based on the aesthetic scores of the images which are composed of those triplets. Although the authors tried to cover many aspects of aesthetic evaluation, the complete process is computationally expensive. In conducted experiments, the authors have shown a run time of 7-8 seconds for processing a 256x256 (image resolution) image which is not appropriate for a real-time application. Also, considering the fact that social image database is growing at a fast pace, the scene categorization is not very efficient for general photographs.

In this paper [197], Zhang et al. proposed personal photograph enhancement using IPCs (Internet Photo Collection). Their work leverages the 3D background models reconstructed from IPCs of the same landmark. Given a personal image, the system provides automatic field-of-view expansion, photometric enhancement and geo-tagging of the image.

Although we have powerful post-processing tools for improving image quality, there are some limitations and drawbacks. Once a photo has been-captured, it is not very easy to modify its composition. It is very time consuming for a user to select their best-captured photos and edit them using some post-processing. Also, it prevents the users from sharing their clicks right after they capture it. On the other hand, assisting users at the time of capture saves the post processing time and allows the user to share the photos as they click. Also, providing real-time assistance will enable users to capture high-quality photographs and there will be no need of taking multiple shots of the same view.

2.1.2 Real-Time

In this section, we will discuss the works which aim to provide photography assistance to a user at the time of capture. We have categorized the proposed methods based on the source of photography knowledge. First, we will discuss methods which make use of known heuristic rules of photography followed by methods which employ social media data for photography knowledge. A detailed summary of important works in this area is presented in table 2.2

Photography Rules and Aesthetics In one of the earliest work, Bares [14] presented an interactive camera system which follows photographic composition rules. The system provides framing suggestion to the user based on the composition objectives for balance, placement, and emphasis as indicated by the user. Similarly Banerjee and Evans [12] proposed a framework for in-camera automation of composition rules. The proposed method automates the placement of the main subject according to the *rule-of-thirds*, and background blur to improve the composition of a photograph. It employ the method from [13] which uses segmentation for main object detection. The proposed method is focused only on scenes with one main subject and it can not be used for wide range of general photographs. In another related work, Bae et al. [9]

proposed a computational rephotography system which enables users to capture a scene with exactly same view-point and composition as compared to some existing exemplary photograph. They presented a real-time pose estimation and visualization technique for rephotography that helps users reach the desired viewpoint during capture.

In [1] Abdullah et al. presented a camera system that automatically configures camera parameters in order to satisfy photographic compositional rules. The authors define rating functions for composition rules like the rule of thirds, diagonal dominance, visual balance, and depth of field, and used optimization to find best possible camera configuration. Although, their method can provide good accuracy but it is computationally expensive, hence not suitable for a real-time system. Gadde et al. [51] make use of robots for capturing images which follow photographic rules. The proposed method focus on the image aesthetics while capturing photos and represent an image using the spatial domain features [117]. In the proposed system, two aesthetic guidelines of professional photography, the rule of thirds and the golden ratio rule are used to assess the quality of the captured image.

In this work [116] the authors proposed a real-time image quality assessment system called PhotoGuide, for assisting users in mobile photography. The image view assessment is done on the basis of saliency features attempting to target composition and simplicity of the image. Users can decide based on the assessment whether to capture the current view or not. The proposed method use some heuristics to define the placement and arrangement of visual elements which are assumed to be pleasing for most of the users. The proposed quality assessment is merely based on some heuristics which only consider the composition for the placement of foreground object in the scene. The assessment is not computationally expensive for mobile devices but is meaningful for only portrait photographs where we can easily filter out the main object from the background. Image quality is not only about the composition and placement of the main object in the scene. This might be the reason why the presented results are better for portrait images as compared to landscape images.

Mitarai et al. [126] presented a system which used pre-defined photography rules of composition and assists a user in placing the main subject in the photo. The author defined some possible compositions which are used by professional photographers and guide users to follow these composition rules. The proposed method employ face and saliency detection along with the extraction of prominent lines in the photo and then use these features in assisting the user to compose and then capture a better photo. Although the composition is an important factor for image quality, there are many other factors like, color, texture, context, etc, which can not be ignored. Augmenting rule-based methods with scene understanding techniques such as, [105] and [172], can be utilized for incorporating complex photography rules.

In a more recent work, Baek et al. [10] proposed a system that implements image editing directly on a viewfinder even before it is captured. This provides the user a real-time interface for image editing and a depiction of the final image. This is useful for professional photographers who are well aware of what they want to capture but is of little value to amateur photographers as they do not give too much thought before capturing a photo.

Social Media - In one of the earliest work in photography assistance using crowd-sourced images [171], the authors proposed a view recommendation system making use of geo-referenced photos retrieved from the Internet. The proposed method utilize images of a particular location and classify them as object, scene, and object in scenes and then perform an image quality assessment based on the clear theme, main object and succinct composition. Finally, the current

TABLE 2.2: Summary for Photography Assistance (Real-Time)

Approach	The work	Details	Image Quality Assessment	Assistance	Dataset	Comments
Photography Knowledge and Perception Based	William Bares 2006 [14]	Based on image composition complying photography rules	Rule of thirds and image balance	Camera motion to achieve best composition	None	Based on two basic photography rules , images with single main object
	Banerjee and Evans, 2007 [12]	In-camera post image editing which follows photographic composition rules	Photographic rules: rule-of-thirds, background blurring, merger mitigation	None	Rule based framework	Focused on only three mentioned photographic rules
	Wang et al. 2008 [171]	Composition of photographs based upon salient features in crowdsourced images	Based on clear theme; positioning of main subject; being succinct without distractive subjects.	Geometric transformation for camera to capture best view in a wide view scene	Geo-Tagged images of three locations from Flickr	Based only on salient point matching between current view and corresponding exemplar view, over simplified image quality assessment
	Lujun et al. 2012 [116]	Used saliency map to extract Region of Interest (ROI) and used its (ROI) size to estimate image aesthetics	Based on size of ROI	Provide aesthetic score of view in real-time	None	Based only on positioning and size region of interest (ROI), over simplified image quality assessment
	Mitarai et al. 2013 [126]	Guides user in capturing images which comply with the rules of photography	Photography rules,	Real-time suggestions	None	Used only few photography rules for photo quality evaluation, presented only for indoor photography
Social Media and Data Driven Approach	Su et al. 2012 [156]	Used photographic rules to extract features for mapping image composition to aesthetic score	Social media (Image in the database is considered either high quality or low quality)	View enclosure with high aesthetic score	12000 highest-rated/lowest-rated photographs	Targeted for scenic images, only view enclosure suggestions, No use of geo-location
	Ni et al. 2013 [128]	Learning of positioning, shape and co-existence of visual words in highly rated images to infer photo quality	Images with more then 10 user likes on social media (Flickr) are considered as high quality images	View enclosure in a user provided wide view scene	80,000 landscape images from Flickr	Targeted only landscape images, considered co-relation between only two visual words, geo-location of images is not used
	Lo et al. 2013 [111]	Used color, composition, contrast, saturation and richness features for learning image aesthetic score mapping	Images in dataset marked either as high or low quality	Provides aesthetic score in real-time	9651 high/low quality crowdsourced photos	Targeted photos with single main subjects inside the scene, provides only aesthetic score
	Xu et al. 2014 [181]	Learn from crowdsourced images, the density distribution of position where people stand in an image	None	Suggests position on viewfinder, where a person should stand	Geo-Tagged images of ten landmark locations	Only works for single person photographs, no image quality assessment, suggestions made using earlier captured images
	Yin et al. 2014 [190]	Composition based upon main object positioning and learning camera parameters using crowdsourced images	Social media (count of user views and likes)	View enclosure with a high aesthetic score in wide view	Geo-Tagged images of eight hot spots from Flickr	Targeted images with some main object, only view enclosure suggestions, Photo quality evaluation based on pre-defined iconic images

view of the user is used to find a relevant high-quality image and a camera-motion is suggested to match the current view with the selected exemplary image. However, it will not always be possible to derive a camera transformation to match the expected exemplary view as it might have been taken from a different geo-location. Also, a user is expected to provide the wide view scene of the location. Otherwise, the consecutive snapshots of the user are employed to construct the wide view scene. The system only suggests a view enclosure to the user. The proposed techniques for mosaic generation, scene classification, and quality assessment are computationally expensive for a mobile device and it will be a challenge to build real-time system using these methods. For a geo-location, the system categorizes images into three themes, namely objects, scene and object in the scene. This organization of photos will lead to large size for each category and for big datasets it will not be feasible to do a real-time image search in such big dataset. Also, since a location can offer a wide variety and type of scenes, it is not very useful to have such broad categorization.

The authors in [32] employ community-contributed photos and trained a probabilistic model based on localization of visual words to assess the quality of a view. The system mine the underlying knowledge of professional photographers from massively crawled photos and learn the patch spatial distributions and correlation distributions of pair-wise patches to guide the photo composition. The proposed method employ image segmentation [46] to extract visual words in an image and then train Gaussian Mixture Model utilizing the positioning of individual visual words and co-occurrence of pair-of words in a scene. The set of images is categorized into 100 groups based on the histogram of visual words and each category is trained separately. Assuming a photo with more than 10 favours/likes as a high-quality photo, they use the trained model to assess the quality of the user view in a wide angle scene.

In an extension to [32], Cheng et al. proposed a more comprehensive model for encoding both spatial and geometric context of visual elements for mining professional photo compositional rules [128]. In this work, they also considered the geometry of visual word patches for individual words and geometric relationship in case of pair-of visual words. As stated by the authors, view recommendation takes around 5 seconds, it doesn't seem to be an optimal solution for a real-time application where the user love to capture and share images instantly. Also, the proposed learning method is not generic for all kind of scenes, as the position of various visual patches in an image is not always fixed or defined by some rule. The proposed method consider only a pair-of two patches for co-occurrence spatial learning. In general, we can have more than two patches occurring in pairs for an image composition. Moreover, considering the variety of photos available from social media, resizing every image into a fixed 500x333 resolution will affect the performance as, the proposed method use patch based visual words.

In [19], Steven et al. proposed a context aware image recommendation system. The proposed system makes use of information like current user location, time, compass direction and camera settings to identify the current context of the user. Relevant images with similar context are searched from an online database of images and based on their ranking they are recommended to the user. The user is expected to select one of the images and based on the selection, direction guidance is provided suggested to capture the desired scene. The recommendation is based on earlier captured images and the user is not assisted for what he wants to capture. Thus the system relies on the existence of already captured scenes at the location. The use of contextual information along with image content differentiate these methods from other image retrieval techniques such as, [65] and [115].

Su et al. [156] proposed a view recommendation system for scenic photos employing bag-of aesthetics preserving features. The proposed system utilizes personal favorite image database for an individual to train the model to ensure that their system provides personalized results. Employing color, texture, saliency and edge features they build a bag-of aesthetics library and use it for image representation. Instead of using vision based segmentation, they divide the image into pre-defined segments. Given a wide angle panoramic view, their model suggests a view enclosure which matches the taste of user which is inferred from his or her personalized image database. The proposed model consider only aesthetics and ignores camera parameters which are also very important for photography. Using personal favorite database for image selection is a wise choice for personalized results but this will lead to a smaller database size which will affect the learning of photography model. Another limitation of the proposed system is the availability of an appropriate personal image dataset. Also, there is no differentiation between various type of image scenes, which is an important factor for image quality. The aesthetic model of images will greatly vary with the scene type as different scene types show large variation in aesthetic composition.

Yin et al. [189] presented a crowdsourced learning system to assist in photography for mobile device users. The proposed system leverage the current scene context to search similar photos from social media and suggest optimal view enclosure to capture a high-quality photograph by learning the composition rule from the searched images. The system utilizes graph-based segmentation [46] to extract salient patches and their position in the image as features for learning the composition of a scene. It suggests optimal view to the user based on the aesthetic score of the view which is derived from the learned model using crowdsourced images. The system expects a user to capture an image with an aspect ratio of 3:2 or 2:3 with a resolution of 640x426 or 426x640 which is a very strong assumption. The system suggests only optimal view enclosure given wide view scene provided by the user. The proposed framework is not very efficient for real-time assistance, as it includes image search (based on visual features), image ranking (based on visual words), composition learning (dimensionality reduction) and generating suggestions (finding optimal view enclosure) in real-time. All these operations are computationally expensive and it doesn't seem feasible to provide a real-time assistance for large image database using this framework. For searching relevant images of a geo-location, the authors proposed a radius of 2 kilometers, which is not very intuitive as it represents a big geographical area and there can be many possible photographic views present in this area.

In an extension to [189], Yin et. al. [190] proposed a socialized mobile photography system which makes use of crowd-sourced images and suggests users the optimal view enclosure and camera parameters to capture a better photo. The author selected eight famous landmark locations and form view clusters based on visual features and geo-location of the images. Based on these clusters, photography rules are learned for image composition and camera parameters which are further used for finding the optimal view and camera parameters. Users current view is considered as wide view image and using the learned rules from crowd-sourced images with a similar view, optimal view enclosure and camera parameters are obtained which are then suggested to the user. Expecting to get a wide view image from the user is not very intuitive as the user may be trying to capture a specific portion of the wide angle. It is not always possible to construct the wide angle view using the previous clicks of the user, as is suggested by the author. Apart from the fact that only a few selected landmark locations are considered, the assistance is restricted to some geo-locations for which the proposed system is able to form clusters. Also, the proposed system is suitable for only landmark photographs with some main object in it. This limitation arises from their composition learning model which is based on object placement in the image and spatial distribution of saliency features. This was also evident from their evaluation where the performance was bad for images with no foreground objects and distinguished salient points. Moreover, the proposed system works only for landmark images and the presence of human objects is ignored.

Fu et al. [50] presented a system which guides a photographer in capturing portraits using earlier captured high-quality photos. The proposed system makes use of Kinect sensor for estimating the current pose of the target model and suggest pose modifications to the photographer based on stored portrait images. Since skeleton based representation of pose is compatible with Kinect sensor, they modeled the stored portraits in the form of skeletons. In similar effort [119] proposed a pose recommendation system for landscape photos. The proposed work is limited to portrait photos where only pose is important.

Xu et al. [181] proposed a framework in which they suggest a position in the viewfinder where a person should stand with the landmark in the background for a better quality photo. The proposed method employ internet photo collections of some landmark locations. Using this collection a 3D model of the location [154] is constructed and probability density of positions where people generally stand in the images is obtained. Based on the user's current view the system suggests a favorable location where a person should stand for a better quality capture. Although the proposed system suggests a favorable position in a frame but the authors have ignored many other factors which are important. First of all, the proposed system works for only single person images. Also, the system ignores context information like time of day, weather conditions, camera parameters, etc. Moreover, there should be some image quality assessment before making the suggestion to a user which can ensure a better quality. In a similar attempt, the authors of [174] proposed a method in which a position on image frame is recommended where a person should stand. However, this method is also limited to position recommendation for a single person in the scene.

In another work [144], San proposed mobile photography assistance tool which relates aesthetic visual features in an image with musical tones. The proposed system analyze current view on the camera to assess the composition and exposure features in the scene. As a feedback, a musical composition that maps visual features resulting from the real-time analysis of the image is composed and played back to the photographer. The authors have used only the pitch of tone for mapping image aesthetics to music quality, which is just a simplified heuristic assumption. Although relating musical tones with image aesthetics is a novel idea, different people can have different taste in music as is the case with image aesthetics. Therefore adding music as a feedback makes the image aesthetics evaluation more complex. Also, it is important to assess the usefulness and effectiveness of this approach in real-time photography assistance.

2.1.3 Personalized Assistance

In [156] Su et al. presented a personalized view recommendation system which is based on the aesthetic features of the image. The authors train their model using user favorite image database and suggest a view enclosure in a wide angle panoramic scene. One limitation of the proposed system is the availability of appropriate personal image dataset. Also, there is no differentiation between various type of image scenes, which is an important factor for image quality. The aesthetic model of images will greatly vary with the scene type as different scene types show large variation in aesthetic composition.

Lujun et al. [116] proposes a photography assistance system which is adaptive and show personalized results according to the users preference. It takes feedback from the user at real-time and adapts as per the user's personal taste. The preference mainly focuses on the position of the main object in the scene. For example, some users may prefer the target object in the center of the scene. Since their work performs well only for scenes with some main object, the personalization is also restrictive within this scope.

2.1.4 Blind Photography

Visually impaired people want to take photographs and share their experiences for the same reasons as others do. Here we discuss some of the existing work which aim at assisting visually impaired people in capturing better quality photos.

White et al. [176] presented a real-time application that enables visually impaired users to take high-quality photos by providing audio feedback as they point their camera on the target. The proposed method provides feedback for three different photo types: landscape, portrait, and documents. For landscape scenes, the system make use of gyroscope and accelerometer to help the user aim the device. In the case of portrait photos the system ensures the target object in the center of the image and for document style photos, the document is kept aligned with the frame of the scene. The captured image is also checked for appropriate exposure and sharpness. In an extension to this work, Jayant et al. [75] used face detection which can help in capturing solo or group photos.

Marynel et al. [169] proposed an interactive system to assist users with visual impairments in capturing street scenes. They used saliency map to find the region of interest (ROI) in the scene and suggests camera motion to place the ROI in center of the image. The proposed method considers only the saliency points in the image and aims to position the most important region of the scene at center of the image.

2.2 Photography Learning

There are many photographic rules and guidelines which are generally used by professional photographers as they capture images. These rules and guidelines are mostly context(time, weather conditions, etc.) dependent and vary with the type of scene. Also, apart from these general guidelines, professional photographers use their photographic knowledge and experience based on the context and at times use their intuitions by crossing these guidelines. Therefore, explicit rule-based suggestion systems can not possibly be holistic and capture all photography knowledge. Considering this, several approaches have been developed, attempting to discover the photography principles used by professional photographers through learning. With the availability of community contributed images from social media along with the meta data information, researchers have proposed different models for photography learning. These models try to capture information like, image composition, which leads to a high quality photograph. These are based on data driven approach where certain features are extracted from the image and corresponding model maps the image to a particular class or its aesthetic score. Based on the choice of features for learning, proposed methods can be broadly categorized into three groups, composition learning, low-level visual features based learning and view based learning. Table 2.3 presents a detailed summary of some important works in photography learning.

2.2.1 Composition Learning

Photographic composition is described as the positional layout of salient visual elements in an image [55]. It is believed that photographic composition is an important factor in image quality and contributes to its aesthetics [195]. In Park et al. [131] the authors proposed a composition learning model based on saliency. The method utilizes saliency map of professional photographs

TABLE 2.3: Photography Learning

The work	Approach	Features	Comments
Ni et al. 2013 [128] and Cheng et al. 2010 [32]	Composition learning method based on positioning and co-occurrence of visual words in an image using Gaussian Mixture Model (GMM)	Pair of visual words employing segmentation	Not scalable for considering co-occurrence of more than 2 visual words, type of image not considered
Park et al. 2012 [131]	Composition learning model based on saliency employing GMM	Saliency Map	Focused on re-arrangement of photo composition to improve aesthetics, Only single model for all types of images
Su et al. 2012 [156]	Aesthetic library is constructed using bag of aesthetic preserving features	Low level features, color, texture, saliency and edge	Generalized model, type of scene not considered
Yin et al. 2014 [190]	Proposed method based on the position of main subject in the image	SIFT points and saliency map	Presence of human objects in the view is ignored

and employed Gaussian Mixture Model (GMM) to represent photographic composition. The proposed method mainly focus on re-arrangement of photo composition to improve aesthetics.

In [128] and [32] the authors proposed a learning method based on positioning and co-occurrence of visual words in an image. Given an image dataset along with the aesthetic rating of the images, a model based on visual words is trained for the corresponding aesthetic score of an image. Using image segmentation, visual words are extracted from image and features corresponding to their position and co-occurrence in the image are generated. These features are then used to train a model which maps the positioning and co-occurrence of these visual words with the aesthetic score. The learning is based on only co-occurrence of pair of visual patches, which is a big limitation. Extending the proposed approach to more than two visual words will exponentially increase the run-time, making it unrealistic for real-time applications. Also, other important factors such as the presence of human objects and context information are not utilized in the proposed model.

2.2.2 Aesthetics Learning

To overcome the computational cost of image segmentation, Su et al. [156] make use of low level features, color, texture, saliency and edge, to derive *absolute (positioning in image)* and *contrast* information for an image and partition it into pre-defined segments. The image is partitioned into 2x2, 3x3 and 6x6 blocks to create segments and then using color, texture, saliency and edge features each image is encoded into a feature vector. The proposed inter-block operations are employed to construct the aesthetic library. The inter-block operations represent absolute and contrast feature of the image. Using a dataset of high-quality images, a set of features called *bag of aesthetics preserving features* are constructed. The extracted features of images from the dataset are employed to train a model which can further classify user images as good or bad. The proposed method does not consider context and also ignore possible variations in the type of scene.

2.2.3 View Based Learning

For images with any main subject, it's positioning in the image is a crucial factor which determines the quality of captured scene. In [190], Yin et al. proposed a method based on the position of the main subject in the image frame. The proposed method employs image transformation based on detected SIFT points in the user image. The method transforms the user image to a predefined iconic image which is selected as target composition for the corresponding cluster. Saliency map and detected SIFT points are utilized to find an aesthetic score for the user image which is based on a classifier trained using dataset from community contributed images. Number of user favors and shares are used to derive the aesthetic score of a given view. The authors also proposed a separate model for learning camera parameters like, exposure, ISO and aperture, using time and weather conditions. The proposed aesthetic model is trained based only on the position of main subject in the image and it is limited for scenes with some main subject. Also, presence of human objects in the view is ignored in the proposed technique. Moreover, type of image view and objects present in the image are ignored for camera parameter learning.

2.3 Image Aesthetics

A photography assistance system guides a user to capture high-quality photographs. To build such a system, we first need to understand the idea of a high-quality photograph. We should then have the capability to assess the quality of an image. The objective of an image quality assessment is to design methods which can predict the perceived quality of an image. The problem of evaluating the quality of photos is considered to be quite complex due to its subjective nature. General image quality assessment methods are computationally expensive and can not be directly employed for real-time assessment of images in photography assistance. There has been a lot of work in image quality assessment in past few years. We will first describe some of the important works in this area and then discuss their limitations.

2.3.1 Visual Features - (Perception Based)

Tong et al. [166] employed low-level features derived from computer vision techniques to classify photos into those taken by professional photographers or home users. In [33], the authors presented a method which enhances the harmony among the colors of a given photograph to improve its aesthetics. Harmonic colors are sets of colors that are aesthetically pleasing in terms of human visual perception.

Datta et al. [37] proposed a quality evaluation technique based on low-level visual features such as, color intensity, texture and region composition, to infer numerical aesthetic ratings for a given image employing linear regression. In [38], the authors extended their work from [37] and included a weighted learning procedure to improve the photo quality prediction performance using the same set of low-level features. In a similar approach, Wong et al. [177] proposed an aesthetic evaluation method where they separate the main subject from the background and use the global set of features along with a relationship between foreground and background objects.

In [82], the authors make use of high level features such as spatial distribution of edges, color distribution, and blur for image quality classification. The proposed method employ mathematical formulations to calculate the features from user photographs. Using Naive Bayes, the method exhibits that high-level features can be used for classification of professional photographs. In another work, Luo et al. [117] focused on the main subject in the image and assess the quality of image based on blur detection and clarity of contrast.

Sun et al. [159] employed visual attention and saliency map for assessment of photo quality. The proposed method detect salients region in a photo and analyze the aesthetics of the photo with the relative position of the subject region. Dhar et al. [42] proposed the use of high-level describable image attributes with scene composition for photo quality prediction. The proposed method tries to encode interestingness in the image using high-level features and uses them to predict image quality. In [99], the authors employ features like color, lighting, and composition to evaluate the aesthetic quality of images and provide cropping suggestion to improve image quality.

In [39] the authors employ visual features to develop an aesthetic quality inference engine which allows a user to upload their photos and rate their aesthetic quality. The rating is calculated based on the distance from the hyperplane obtained by an earlier trained Support Vector Machine(SVM) classifier. Similarly, Yeh et al. [187] proposed an online ranking system for personal photo collections based on aesthetic rules. In the proposed system, features are computed for composition, color, and intensity. User provides their own images, adjust weights for the importance of different features, and can search for similar photographs. In another work, Marchesotti et al. [124] also proposed a photo evaluation scheme based on aesthetic quality making use of generic image descriptors.

Su et al. [157] defined Bag of Aesthetics Preserving Feature (BoAP) based on visual features without the use of formalized rules of photography. The proposed method employed color, texture, saliency and edges from segmented images to form feature vectors and a library is created using a dataset of high-quality photographs. However, the composition is also important for image aesthetics and the proposed method does not consider image composition.

The authors in [186] used color triplets to model the aesthetic quality of an image. The proposed method employs histogram for dominant color triplets present in images from the database and based on the aesthetic rating of the corresponding images it generates ratings for the color triplets. The authors also proposed an aesthetic evaluation of black and white photographs which was based on contrast, shape and saliency of the image. This rating was used as a feedback to the user during photography.

Yin et al. [188] proposed a scene dependent aesthetic model to assess photo quality which leverages both geo-context and visual content. The proposed method groups images based on geo-location and content in the image and a separate model is built for each group to assess the image quality. The method employs GIST descriptors and HSV color space to represent an image and employs social media ratings for aesthetic scores. Context is not only about geo-location and visual content. Image quality may vary with the time as well for the same view and geo-location. Also, image quality is a subjective evaluation and may vary with person to person. Social media meta data is noisy sometimes and using it with some other evaluation strategy can help us produce better results. Also, the proposed method employs only low-level features (GIST and HSV color space) for image representation which is not a comprehensive approach.

Tang et al. [160] proposed a framework which makes use of a set of the foreground, background, and global features for assessment of photo aesthetics. The images are classified into different categories and a different set of visual features are extracted corresponding to each scene type.

In [111], the authors proposed an efficient on-device aesthetic quality assessment method. A rich set of low-level features with low computational overhead are defined to represent aesthetic characteristics of an image. Color, composition, richness, contrast and saturation are modeled employing low-level features and a SVM model is trained to assess the image quality. For color combination, top five dominant colors are extracted in the image. The composition is represented using edge features and HSV color channels. For contrast, the width of dominant range in color histograms of the image are utilized. Richness is defined using variety in color composition and for spatial richness, the image is divided into regions and difference in edge intensity maps of adjacent segments are computed. A dataset of images with both bad and good quality is used to train the defined features for rating an image as good or bad. As shown in experimental results, proposed method is feasible for real-time mobile applications, but for experiments they used a limited dataset with a low-resolution images. Also, context, view or any other information is not utilized which makes the system too generalized and it is difficult to model all the photography rules in one simple model using only low-level features. Moreover, the proposed method assumes prior knowledge about the quality of images in the dataset.

In a recent study, [8] Aydin et al. proposed an aesthetic-attributes based automatic evaluation method. The proposed method employs sharpness, colorfulness, tone, clarity, and depth as aesthetic attributes and corresponding features are extracted to find an image aesthetics score. The proposed approach outperforms existing methods, however, it is based on only five aesthetic attributes defined by authors. There can be other possible attributes which can affect image aesthetics, like image composition and also the influence of attributes might be dependent on the type of image.

More recently, researchers have also explored the features extracted using Deep Networks for predicted aesthetic quality of images [81, 113, 164], and [200]. These methods utilize deep networks to extract the aesthetic features which are difficult to design manually and improve the predictions as compared with the other hand crafted features. However, using deep features it is difficult to understand the aesthetics of photographs as visualization of the deep features is not very trivial.

2.3.2 Rules of Photography

The work in [54] uses rules of thirds and fifths to enhance the compositional aesthetics of a 3D model. The employed features tries to encode the attractiveness in image and the system finds corresponding format(image size, shape, and orientation), viewpoint, and layout for an image of a 3D object. Similarly, in [22] the authors used rules of third to position the main object in a scene for an automatic robot camera system. Zhang et al. [201] defined 14 templates based on photography composition rules and utilized them along with face detection for automatic photograph cropping. The authors in [116] presented a real-time photo quality assessment method for mobile photography. The proposed method is not computationally expensive and is suitable for mobile devices. But the assessment methodology considers only composition of the scene and is merely based on the position of main object in the image frame.

The authors in this work [171] use three basic principles for image quality assessment: i) having a clear theme, ii) viewer's attention on main subject, iii) being succinct without distractive subjects. For clear theme they use blur estimation to identify 'in focus' and 'out of focus' patches. The proposed method employ *rule of thirds* to assess the placement of main object in the scene. To

TABLE 2.4: Image Aesthetics Evaluation

Approach	**The work**	**Details**	**Features for Aesthetic Assessment**	**Comments**
Visual Features and Perception Based	[33, 37–39, 166, 186, 187, 204]	Image quality prediction based on low level features	features like color intensity, color composition, texture and region composition	Difficult to interpret the mapping of low level features to image quality
	[82, 117]	Aesthetic quality prediction based on defined high level features	Spatial distribution of edges, color distribution, clarity and blur	Proposed high level features are based on heuristics
	Sun et al. 2009 [159]	Based on visual attention and saliency map	composition based on saliency map	Can not be generalized for various types of photographs
	Dhar et al. 2011 [42]	Based on defined high level describable image attributes	Attributes tries to encode interestingness	Consider only defined set of attributes
	Su et al. 2011 [42]	Define bag of aesthetics preserving features	Color, texture, saliency and edge features	Image composition and type of image is not considered
	Yin et al. 2012 [188]	Scene dependent aesthetic evaluation which also considers geo-context	GIST descriptors and HSV color space	Employed only two set of features and image categorization is only based on geo-location and view
	Lo et al. 2013 [111]	Real-time aesthetic evaluation based on low-level features	Color, composition, richness, contrast and saturation	Proposed system is too generalized considering the wide variety of image categories
	Aydin et al. 2014 [8]	Aesthetic attribute based evaluation	sharpness, colorfulness, tone, clarity and depth	Image composition and type of images not considered
Heuristic Rules of Photography	[22, 54, 108, 116, 201]	Known heuristic rules for feature extraction	Rule of Thirds, Balancing Elements, Diagonal Rule, Size of Region	List of heuristic rules is not exhaustive
	Wang et al. 2008 [171]	Proposed approach based on clear theme, viewer's attention and image focus	Rule-of-thirds, motion blur and focus on main subject	Computationally expensive for real-time applications
Social Media	Yin et al. 2012 [189]	Employed social media cues available with crowd-sourced images	Using number of user-views and user-favorites	Based on only two cues (number of user views/favorites)
	Yin et al. 2014 [190]	Employed image interestingness and social media cues available with crowd-sourced images	Image interestingness [21], number of user-views and user-favorites	Ignored other cues like user comments/tags/attributes

ensure succinct composition it is checked whether the main object in the scene is in focus or not. The proposed method tries to cover some of the photography principles if not all. However, it will be computationally expensive for a mobile device to use it for a real-time application. Also, apart from these three principles, there are many other photography rules which should be considered to assess a photo quality.

Liu et al. [108] used certain photographic compositional rules such as *Rule of Thirds, Balancing Elements, Diagonal Rule, Size of Region*, to study the composition of photos. Based on these heuristics, aesthetic score of an image is measured and it is further utilized for image retargeting. Segmentation is employed for detecting salient regions and prominent lines in the image. Mathematical expressions are formulated to determine if the extracted salient regions and prominent lines are following the defined rules of photography. Based on these formulations numerical value for the aesthetic score is generated for an image. Each image pixel is assigned a saliency value based on a low-level saliency score of Itti et al. [70]. This saliency value is then propagated with the salient features and contribute to the actual aesthetic score generation. One limitation of this approach is that only heuristic photography guidelines are not sufficient for photo quality evaluation. Some pre-defined rules can not capture the various composition possibilities of general photography. Also, the described rules are not exhaustive and the rules are scene dependent, so it is unwise to apply all the defined rules in a generalized fashion.

2.3.2.1 Visual Balance

Visual balance is considered an important factor in the art of image composition. This fact has been widely studied by researchers in predicting and improving the quality of image from aesthetics perspective [14, 18, 82, 108, 118, 129, 151, 184]. Bares *et al*. [14] proposed to use the concept of visual weights of elements to balance frame composition in virtual camera environment. They consider factors such as size, brightness, and position of objects to compute their visual weights and move the center of visual weights to the frame center. The proposed method does not consider the color component and is limited to a virtual environment.

In Bhattacharya *et al*. [18] the authors proposed to balance the sky and ground region in a photograph according to golden ratio for image cropping. Similarly, the authors of [108] also propose image cropping to improve aesthetics by obtaining visual balance considering the size and the saliency of important objects in the image. Ma *et al*. [118] extended this idea to find where a person should stand in a given image. One major limitation of existing methods is that they all ignore the involvement of color composition in attaining visual balance. They mainly rely on automatic saliency detection of visual elements and computing visual saliency accurately in photographs can be very challenging.

The concept of visual balance to improve aesthetics is not limited to photographs and it applies equally to other areas of visual art where human perception is involved such as information presentation [112], web-page layout [135], magazine covers [73], etc. Purchase *et al*. [135] proposed a metric based on the area of important objects to evaluate the quality of visual balance in the layout of a webpage. In Jahanian *et al*. [73], the visual balance of colors based on a scale [89] was utilized to identify a position where text in a magazine cover should be placed. In Lok *et al*. [112] the authors studied the idea of visual balance for visual layouts in information presentation. They proposed weightmaps and make use of color histogram as the visual weight of objects to balance the layout. The proposed method requires human intervention and the position of one visual element is changed at a time iteratively to achieve balance. Therefore the final layout highly depends upon the order in which the visual elements are added to the layout.

In another interesting application of visual balance force-directed algorithms were proposed for aesthetic drawing of planar graphs [16, 45, 49, 80]. A graph is represented as a model where the nodes are connected by springs and force of attraction and repulsion act between the nodes. Energy minimization is performed to bring this system of forces into an equilibrium where the edge tend to have uniform length and the nodes that are not connected tend to be drawn apart. The proposed methods assume all nodes and edges in the graph to be similar.

The existing methods for obtaining visual balance have the following limitations: 1) They assume that the position of the visual elements are either fixed or known beforehand and therefore they cannot be used for a balanced layout of dynamic visual elements. Methods which have addressed the dynamics of visual elements considered only one visual element and placed it to balance rest of the layout. 2) Most of the methods are focused on evaluating the quality of visual balance and then recommend image cropping to improve the aesthetics which however does not change the arrangement of visual elements in the layout. 3) The existing methods have only explored factors such as area, position, and color for visual balance, however, there are many other factors such as visual direction, contrast, etc. which have an impact on visual balance [6].

2.3.3 Social Media

In this [32] photography learning system, the authors assume a community contributed photo with more than ten favours/likes as a high quality image and used this to train probabilistic model which serves as a basis for user image quality evaluation. Similarly, in [189] the authors make use of number of user views and number of favours to assess the image quality. They approximated the aesthetic score by

$$S = 100 \times (1 - \exp^{-(\upsilon.views + \beta.favours)}), \tag{2.1}$$

where *views* and *favours* are the number of user views and user likes. They used υ = 0.1 and β = 1 for their experiments.

In this work [190] the authors proposed a photo quality assessment using social media data. They make use of interestingness factor which is provided by Flickr API [21] and is based on the quantity of the user-entered metadata, such as tags, comments and annotations, and similar other factors. Apart from this, the number of user-favourites and user views are also employed. Based on all the factors aesthetic score for each image in the database is generated. The aesthetic score for an image is approximated by

$$S = 100 \times (1 - \exp^{-(\upsilon.views+\beta.favours)}) \times \exp^{-I/N}, \tag{2.2}$$

where *views* and *favours* are the number of user views and user likes. *I* is the interestingness factor and *N* is the total number of images crawled in the location scope. For constant values, υ = 0.2 and β = 1 are used in the experiments.

The assessment is merely based on social media data and does not take into account any other factor which can be computationally evaluated. The social media data can be noisy and a hybrid approach based on both social media and computational methods can provide more accurate aesthetic scores.

2.3.4 Image Memorability

Like many image properties such as image quality, aesthetics, saliency, attractiveness, and composition, image memorability is another criteria for image quality evaluation. Isola et al. [69] conducted a set of experiments using visual memory game to measure memorability of images and find out whether it depends upon the context and vary with users. Statistical results showed that image memorability can be context independent and persistent within a different set of users and time. The authors also conclude that image memorability is not related to image aesthetics and attractiveness. On the basis of results from memory game, SVR model is trained to map from image features to memorability scores [68]. The proposed method employs GIST [130], spatial pyramid histograms of SIFT [93], HOG [36], and SSIM [148] features to train the model. The authors also came to a conclusion that most landscape images have low memorability.

In another related work [67], Isola et al. investigated image features which contributed to memorability. The proposed method make use of conceptual attributes like spatial layout, aesthetics, etc, and used related features to represent an image for memorability computation. The authors concluded that contrary to popular belief, unusual and aesthetically pleasing images are not predominantly the most memorable ones.

One important thing to consider is that, their proposition [68], [69] and [67] is completely based on the crowd-sourced feedback of 665 users with a dataset of only 2222 images. The fact that one second was given to the user per photograph, it is arguable whether this is sufficient for an average user for analyzing the complete image. Although there has been some earlier research in face photo memorability [11], and the memorability of facial caricatures [168], this is a first attempt towards research in this direction for general photographic images.

In an extension to this work [69], Khosla et al. [85] proposed a probabilistic framework which maps memorability to image regions. In this work, the authors presented two different representation of an image, external and internal. External view is the original image and internal view is the representation in memory which accounts forgotten image regions and hallucinations. Based on the distance between these two representations and known memorability of images they tried to find our the contributions of different regions in image memorability. In continuation to this work, authors in [84] presented the idea of visual inception and image modification to change the memorability of an image based on the memorability contribution of each region individually.

Bora et al. [24] and Mancas et al. [122] explored the role of visual attention in understanding memorability of images. Bora et al. [24] used visual saliency map along with low-level features (GIST [130], SIFT [93], HOG [36], and SSIM [148]) for predicting the memorability of an image. Mancas et al. [122] also conducted an eye-tracking experiment on images from the memorability database [69] to find out the relationship between image memorability and visual attention. The authors find out that there is a long eye fixation time for images with higher memorability. The authors also proposed two attention-related features (RARE saliency map coverage and structures visibility) for predicting image memorability and showed that these features performed better than earlier suggested features.

In another related work, Kim et al. [86] investigated spatial features that are correlated with the image memorability. The authors proposed Weighted Object Area (WOA) that jointly considers the location and size of objects and the Relative Area Rank (RAR) that captures the relative unusualness of the size of objects. These features gave similar results as compared to earlier works without the use of low-level features.

In the proposed methods, only low-level features (GIST, SIFT, HOG and SSIM) [69] and high-level describable attributes related to spatial layout, aesthetics, emotions [67], and attention [24], [122] are utilized. Context also plays an important role in memorability of an image which is not considered in any of the earlier works. Also, we believe its not the objects, which contribute to memorability but their composition combined with many other factors. Otherwise, all images with faces should be memorable which is not the case in general. Although [86] studied relative spatial features, but their approach is not holistic and only consider the size and position of

TABLE 2.5: Memorability

The work	Approach	Features	Conclusion
Isola et al. [68]	Used low level features to predict memorability of images	GIST [130], SIFT [93], HOG [36], and SSIM [148]	They concluded image memorability is not related to aesthetics and attractiveness
Khosla et al. [85]	Maps memorability score to segmented image regions	Color, texture and gradient	They tried to find out individual contribution of segmented image regions to memorability of an image
Bora et al. [24], Mancas et al. [122]	Explored role of visual attention in image memorability	Saliency map along with GIST, SIFT, HOG, SSIM	Showed positive co-relation between saliency and memorability
Kim et al. [86]	Explored role of spatial features in image memorability	Location and size of objects, relative size of salient object for unusualness	The spatial features directly corelates to memorability

objects. Moreover, there can be other aspects, like image composition, photographic rules, color composition, etc, which can be explored for their role in image memorability.

2.4 Geo-Tagged Images

The sharing of geo-tagged photographs by users on social media platforms such as, Flickr, FourSquare, etc. has increased tremendously in the last decade. This has motivated researchers to exploit this data for various kinds of recommendations in location based services. This includes, finding location using images [79], [28], point of interest for photo capture [97], [71], [150], [107], travel route recommendation [207], [191], [30], [106], 3D scene reconstruction [153], [101] and event recognition using image database [43]. In this section we will discuss some of the important work in location based recommendations and image classification which are closely related to our research.

2.4.1 Points of Interest Identification and Recommendation

Zhang and Kosecka [202] used image retrieval to find similar images for user query and then determined the landmark location information for the query. Hays et al. [59] adopt image-based matching to locate user photos on the world map. Joshi et al. [78] proposed a location identification method which also uses user tags along with geo-tags to infer the geo-location of images using geo-tagged images.

In another interesting work for location identification, Yu et al. [192] proposed Active Query Sensing (AQS) framework which aims to help the mobile user take a successful second query once the first query fails. It informs the mobile user how to sense his/her surrounding environment so that the captured image is most distinctive and can be used to recognize the location. The main idea behind their proposed work is that each location has a unique subset of preferred views for recognition and it varies from location to location. Based on the view of a location, they divided it into *View-Independent Confident Location, View-Dependent Searchable Location* and *Difficult Location*. Once the first user query is failed, the system finds the most salient view in the top returned results and use polling to find the next best view direction which is suggested to the user. The performance results are based on well organized New York city street view dataset with 0.3 million images, which comprises of 50,000 locations with six views for each location. However, the availability of such an organized dataset for other locations can be argued.

Crandall et al. [34] proposed a classification method to identify the taken location of a photo using its textual-tags and visual and temporal features. The authors of [66] proposed a method to find popular viewing directions for capturing a landmark object. However, the proposed method does not take into account the user context and also it is limited to photographs with a single landmark object.

Point-of-interest (poi) recommendation to users is another area which has recently received a lot of attention from the community. In [56, 98, 205], the authors employed collaborative filtering for providing personalized location recommendation to the users. To make the recommendation process more personalized, the authors of [90, 194] presented interactive framework where the user can provide real-time feedback regarding his or her choices. To further improve the relevance of recommendation, the authors of [103, 110, 193, 198, 199] also take into account the geographical, temporal and sequential influence of location and user movements.

As the semantic information of the location also plays an important role in users preferences, the authors of [77, 173] incorporate venue semantics into user recommendation. The user preference usually changes with his or her context, therefore to take this into account authors in [57, 182] proposed context-aware recommendation which incorporate factors such as weather and season into account or a feedback from the user for making the recommendation.

Most of the methods discussed so far assume that the poi recommended to the user are known beforehand which may not be always true. Therefore, to overcome this problem researchers have proposed methods to automatically identify points of interests in a location utilizing user contributed photographs. In [97, 107, 150, 185] the authors proposed to used clustering based algorithms to identify the points-of-interest which helped in detecting popular locations. Rattenbury et al. [136] applied clustering algorithms to identify landmarks and event names by extracting semantics from geo-tags of photos available on Flickr.

In [150], Shirai et al. proposed a method to discover multiple hotspots, where many photos have been taken, using geo-tagging photos posted on social media sites. Additionally, the authors infer range and shape of hotspots based on the deviation of places where photos have been taken. Liu et al. [107] presented an approach to discover Areas of Interest (AoI) by analyzing both geo-tagged images and check-ins. The approach exploits travelers flavors as well as the preferences of daily-life activities of local residents to find AoI in a city. The authors devise a density-based clustering method to discover AoI, mainly based on the image densities but also reinforced by the secondary densities from the images of neighboring venues.

In [97], Lee et al. segment geo-tagged photos into different categories to find categorized poi(Point of Interest) using density-based clustering. They apply two levels of clustering to detect global level(GPS based) and local level categorized poi. In the local level, they categorized the images into following groups, long-term (more than 9 days) vs. short-term (1-9 days), summer season (September to February) vs. winter season (March to August), peak period (May and October) vs. non-peak period (all other months), day-time (6am to 6pm) vs. night-time (6pm to 6 am), weekdays (Monday to Friday) vs. weekends. In a similar effort the authors of [92] proposed an approach to discover area-of-interests using social media images.

Zhuang et al. [209] argue that non-popular locations can also be preferred by some users and therefore proposed a method to discover obscure sightseeing spots for recommendation. In [161, 162] Thomee et al. proposed to find the actual location of point-of-interest which may be different from location of the point of photo capture.

Recently, researchers have focused their interest on identifying points-of-interests which are good from the photography perspective. Kimura et al. [88] proposed a clustering-based approach to identify hot-spots for photography. As time also plays an important role in the photography Ying et al. [204] also considers time dimension in discovering photo capturing locations. The proposed method is based on learning from crowd-sourced images and Gaussian Mixture Model is employed to find hotspots for aesthetically good images which also considers the time factor. In a more recent study, the authors of [132] use collaborative recommendation algorithms in order to produce personalized suggestions for photo-taking spots. The photographs captured by a user in the past are used for making the recommendation.

2.4.2 Travel Route Identification and Recommendation

The movement of tourists from one location to another is captured as a footprint in the geo-tags associated with the photographs shared on social media. This has been exploited by the researchers to identify various travel patterns followed by the tourists for recommendation. Zheng et al. [206], [208] applied a statistical approach to extract tourist movement trajectories from geo-tagged photos and analyzed corresponding travel patterns. Liu et al. [109] employed collaborative filtering to recommend inter-city travel packages to users based on their past travel records. In [23, 52, 121, 134, 175], the authors identify the popular routes followed by tourist as they travel from one attraction to another at a tourist location.

In [74], Jain et al. proposed a method which recommends a route passing through locations with a large number of captured photographs with given starting location and travel distance. In a similar effort, Zheng et al. [207] proposed a driving route for users which passes through scenic locations. The proposed system adapted an attention-based approach to exploit GPS-tagged photos for discovering scenic roadways and formulated the scenic driving route planning as an optimization task towards the best trade-off between sightseeing experience and traveling distance. One of the major limitations of these methods for intra-city recommendation is that they provide generalized recommendation which is independent of user preferences.

To overcome this limitation, the authors of [30, 31, 53, 104, 114] also incorporated user preferences for making route recommendation. Lu et al. [114] proposed a method to identify tourist locations and then generate a personalized trip route to travel between identified attractions. Chen et al. [30] proposed a probabilistic personalized travel recommendation model which exploits the automatically mined knowledge from the travel photo logs. Chen et al. [30] proposed to use people attributes (gender, age, and race) extracted from captured photos for personalized route recommendation. The proposed method identifies popular travel routes and recommends next best location to visit based on current user location. Yamasaki et al. [183] proposed Markov model approach for route recommendation and incorporates personalization using collaborative filtering.

2.4.3 Image Classification

Cristani et al. [35] presented a framework for geo-located image categorization. The proposed system aims at grouping together visually similar images with close geo-location. The method also focus on visual content management and visualization of large geo-located image databases. Snavely et al. used internet photo collections to construct a 3D model of a location [154]. The proposed system employ structure-from-motion and image-based rendering algorithms that operate on set of images to construct 3D model of a location.

In this [32] photography learning system, the authors make use of image segmentation and derive visual word patches in a photo. Based on histogram of these visual words the complete image dataset is classified into 100 groups using clustering. For large datasets this is computationally expensive and inefficient. Also, there was no use of geo-location and the formed clusters does not have any meaning in photography sense.

Ahern et al. [5] suggested context-aware tagging of photos at the time of capture before they are uploaded to any social media. It is easy to identify the context and find meaningful tags at the time of capture and it also saves the tagging time which is usually done at a later stage. The authors make use of current location and user profile for tag suggestion. Annotation of images with semantic level tags is also useful for image classification as it provides additional information apart from geo-location.

In [171] the authors used *focus of attention* to find the theme of a photo. The proposed method defines three basic categories for a photo from a particular location: *objects, scene* and *objects in a scene*. They used bottom-up saliency map using center-surround principle on color, intensity and orientation feature for attention modeling and identify the theme of a photo. Considering the fact that a particular location can have a large variety of possible views, diving photos from a location into these three categories is not very intuitive.

In [101] the authors presented modeling of landmark sites based on large-scale contaminated image collections gathered from the Internet. The proposed system builds an iconic scene graph using view clusters to construct the 3D model. Li et al. [102] presented a landmark classification system based on multi-class Support Vector Machine trained using geo-tagged Flickr Images. Yao et al. [186] identify five different forms of compositions namely, textured images, diagonally, vertically, horizontally and centered composed images. They employed image segmentation and edge detection to extract features to model these compositions. These compositions are mutually exclusive, based on photography heuristics and also not exhaustive enough to model all possible scene types.

In [189] the authors used a 2 kilometers radius for searching relevant images from the community contributed images. This is a very naive way to search geo-tagged images from the community contributed database when we can get the geo-location of captured images to an accuracy of few meters. The authors in this work [190] proposed a view based cluster formation method. For a specific geo-location they form clusters of similar views and based on the view of user image they find the cluster to which it belongs. In this work the authors have used it for photography learning and assistance. Clusters are formed using the visual features (SIFT saliency feature) and geo-location. The proposed view clustering is based on important object location and distinguished salient points on the image. This choice of features restricts the system to work only for images with foreground objects. Since the geo-location is also used for clustering, similar views can formed different clusters due to change of geo-location. Also, more than one clusters can be form for one geo-location based on the view.

2.5 State of the Art Summary

In this section we will summarize the state of the art in photography assistance and related areas. A detailed summary is presented in table 2.6.

Photography Assistance: There are only few methods which claim to provide a real-time photography assistance during image capture [156], [128], [190], [171], [111], [116] and [126]. Some of the authors have also developed an application to demonstrate the real-time feasibility of the proposed method [111], [116] and [126]. However, the proposed methods are not generalized and are focused on only few of the aspects among many. Some of the limitations include smaller datasets, few selected geo-locations and specific image types. Table 2.1 and 2.2 provides a more detailed summary of the literature survey.

In [171], Wang et al. used a dataset with images from only three locations and the proposed system is based on salient point matching between the current view and corresponding exemplar view from the dataset. In [116] Lujun et al. used only *region of interest* (ROI) and the proposed method guides the user to adjust ROI in the view to improve aesthetics. In [156], Su et al. used a dataset with only 1020 photographs and their method was targeted for only scenic images. Lo et al. [111] used describable features to predict the aesthetic quality of an image in real-time, but apart from aesthetic score, the proposed system didn't provide any other feedback for photography assistance. In [126], Mitarai et al. used image composition in their system which guides a user to adopt relevant photographic rules. Similarly in [128], Ni et al. proposed a view enclosure suggestion based on the aesthetic score which is predicted on the basis of spatial position and co-occurrence of visual elements in an image. The proposed method is not suitable for real-time assistance as the authors stated a run-time of 4-5 seconds. Yin et al. [190] proposed a more comprehensive solution to photography assistance in which camera parameters are also considered. However, the proposed method provide only view enclosure suggestions to the user based on learning from the dataset for eight tourist locations. Moreover, they targeted only scenic views with some main object which can be separated from the background. In a more recent study [181], Xu et al. proposed a position recommendation system which can suggest where a person should stand in a photograph. However, the proposed method does not evaluate image quality and is limited to a recommendation for a single person.

The methods proposed so far make suggestions about optimal view enclosure possible in the wide view scene provided by the user. The user cannot always provide a wide angle view using his mobile camera and also it is not a user-friendly approach to ask him for a wide angle view to suggest the best view enclosure. Also, it is not always possible to construct the wide angle view using the previous clicks of the user, as is suggested by many researchers. A better image might be lying outside the current view of the user which is another limitation of the proposed systems. Another approach can be to build a wide angle view offline and rather than just suggesting a view enclosure, we can have other types of suggestions like pan/tilt/zoom to the user.

Existing methods are not generalized for the different types of a photograph (landscape, scenic, portrait, etc.) and target some specific range of image types. Like, in most of the proposed methods it is assumed that the target view has one main object which can be separated from the background. This is a big limitation as the users do not always capture landmark photos with some main objects. We can have many other categories of photographs apart from landmark photos like portrait, scenic beauty, etc. Also, the proposed methods have not considered image composition with human objects (single or multiple) along with the background view. Moreover, the proposed methods are focused only on few well-known landmark locations. A more general solution will be to provide assistance for all types of images and also for locations where no photographs have been captured before.

As discussed earlier, camera parameters play an important role in the quality of captured image. In [190], the authors proposed to learn camera parameters such as exposure, ISO, shutter speed and aperture from the metadata available with crowd-sourced images. However, the proposed method consider only environment conditions for parameter learning. These camera parameters also depend on factors such as the type of image and its content which is ignored by the authors.

Image quality evaluation is subjective in nature and the perception of quality may vary from person to person. In [156] and [116], authors presented image quality evaluation system which tries to make personalized recommendations. The method proposed in [156] make view enclosure suggestions based on some heuristics and [116] is merely based on size and position of region-of-interest (ROI). Therefore, a comprehensive method which can consider other important factors is required for providing personalized assistance to the users.

Viewpoint is another factor which can affect the quality of captured image. The proposed methods do not make suggestions for viewpoint recommendation, which is sometimes important. Also, none of the methods is utilizing the direction of image capture, which can be computed by making use of inbuilt sensors. Moreover, the existing methods do not employ built-in camera sensors for assisting in photography. A hybrid system utilizing both the camera and social sensors is needed which can provide better results.

Photography Learning: The existing photography learning techniques make use of social media data along with photography rules and visual features to infer aesthetic scores of images. The proposed methods employ low-level features, visual words or salient regions for image representation based on its composition. In the learning phase, the extracted features are mapped to the corresponding aesthetic score of the image. The aesthetic score is generally inferred from the social media metadata (number of user likes) attached with the corresponding image and it is considered as a ground truth [190], [128], [156]. Social media data is noisy and alone it can not be relied upon for aesthetic quality. Also, all of the photography rules are not utilized in any of the learning models. Photography rules are important but not sufficient for image quality evaluation. Similarly, visual features also provide only partial low-level information which is insufficient for aesthetic score interpretation. Therefore we need a hybrid approach which can leverage both these aspects together for photography learning. Table 2.3 presents the summary of literature survey for composition learning.

Most of the existing methods build a general system which tries to capture all the photography knowledge in one model. Considering the complex nature of photography and knowing that photography rules are scene dependent, it is not possible to capture all the photography knowledge in a single model. Different rules may apply in different locations based on the view and context, therefore some categorization is required based the context and the type of photographs. Classification based only on geo-location will be not good for locations with sparse datasets and view based categorization (as suggested in many works) will be difficult to scale with growing data. Another interesting direction for research will be to transfer photography learning from locations with dense datasets to locations with sparse datasets.

TABLE 2.6: State of the Art Summary. This thesis is focused on the boxed items in the missing list.

Aspect	What is possible	What is missing
Image Quality Evaluation	• Separate methods based on visual features, composition, heuristics and social media • Offline, computationally intensive methods • Preliminary work on image memorability • Generic notion of image quality	• Comprehensive model for all aspects • Real-time online evaluation • Real-time evaluation of memorability • Personalization • Context based evaluation
Photography Assistance (In-Camera Guidance)	• Post capture image enhancement • Heuristic rules based composition assistance • Real-time aesthetic evaluation based on few heuristics • Generalized models for composition evaluation • Assistance regarding placement of main object • Human position recommendation (geo-based, single person, without aesthetic evaluation) • View recommendation based on geo-location • Camera parameter suggestion based on geo-context	• Comprehensive model (photography rules, composition, social media, etc.) for real-time recommendation • Context based (type of image, view, etc.) real-time aesthetic evaluation • Human position recommendation (single/multiple person, with aesthetic evaluation) • Personalized recommendation • Photography knowledge transfer based on context similarity • Context based (objects in scene, type of image, etc.) camera parameter suggestions • Camera motion (pan/tilt/zoom) suggestions
Photography Assistance (Location Based Recommendation)	• Points of Interest identification • City-scale tour recommendation • Personalized Point of Interest recommendation • Personalized tour recommendation	• Viewpoint recommendation for photography based on context • Context-aware within attraction tour recommendation • Personalization in within attraction tour recommendation • Real-time sensing for adaptive tour recommendation

Image Quality Evaluation: There has been a lot of work in image quality assessment in recent years. Researchers have proposed methods for image quality assessment based on low-level features, high-level semantic concepts, photographic principles, saliency, composition, social media data, etc. The proposed methods focus on different aspects of image quality and make use of a different set of features, but they all share the common aim of assessing the quality of an image. Every photographer desires to cover all the aspects to make their captured images look good. Therefore it is required to understand the correlation between different aspects of image quality and come up with a hybrid approach which can holistically cover all possible directions. A summary of literature survey is presented in table 2.4 and table 2.5

Existing photo quality evaluation techniques are computationally expensive and are not suitable for a real-time application. Also, none of the methods use visual features, photography rules, and social media altogether for aesthetic computation. Considering the various types of images (portrait, scenic, landmark, etc.), it is difficult to use a general model for aesthetic evaluation. Therefore, some form of image type categorization and separate aesthetic models for each category can be one way to improve aesthetic evaluation.

There can be many other factors involved which affect the image quality. A photo may be considered more or less appealing depending on the person in the photograph. A badly composed photo of some celebrity might be popular and in contrast a well-composed photo of an unknown person may be ignored by the viewers. This can be important when we are trying to learn from social media data.

The place where a photo is captured also plays an important role in image aesthetics. A badly composed photograph of a popular location might attract the viewer's and a well-composed photo of unknown location may not be appreciated by many users. Apart from these factors personal preferences also play a crucial role in image aesthetics. Some people might prefer a photo which does not obey any composition rules.

More recently, there has been some work related to the memorability of images (Table 2.5). But there is no work on photography assistance for capturing memorable images. Also, it will be interesting research direction to understand the relation between various image qualities like saliency, attention, composition, etc, with each other and image memorability.

Location Based Recommendation: The existing works in location recommendation relevant to photography mainly focus on points of interest identification and city-scale tour recommendations. The proposed methods utilize social media images to identify interesting photo-shooting locations for photography. There are methods which first identify these interesting locations and then determine a route through these locations for a recommendation. There are some recent works which also incorporate personal preference in providing the recommendation. The past travel behavior or the past photography behavior is utilized to determine the personal preference for a recommendation. However, the existing works do not focus on how a tourist location should be explored by the users. As a tourist location may have multiple hot-spots which are interesting from the photography perspective, it is important for the users to determine how to explore a tourist location.

2.6 Summary

In this chapter, we discussed research work relevant to real-time photography assistance and also pointed out some of the limitations of the existing methods. We also presented a state of the art summary of the related work. Photography assistance is a new research area and not much work has been done. In the next chapter, we will present our work in photography assistance for a real-time recommendation.

Chapter 3

Context-Aware Photography Assistance

In this chapter, we will discuss a photography model based on machine learning which can assist a user in capturing high-quality photographs. As scene composition and camera parameters play a vital role in aesthetics of a captured image, the proposed method addresses the problem of learning photographic composition and camera parameters. Further, we observe that context is an important factor from a photography perspective, therefore we augment the learning with associated contextual information. The proposed method utilizes publicly available photographs along with social media cues and associated meta information in photography learning. We define context features based on factors such as time, geo-location, environmental conditions and type of image, which have an impact on photography. The metadata available with crowd-sourced images is employed for context identification and social media cues are used for photo quality evaluation. We also propose the idea of computing the photographic composition basis, *eigenrules* and *baserules*, to support our composition learning. The proposed system can be used to provide feedback to the user regarding scene composition and camera parameters while the scene is being captured. It can also recommend a position in the frame where people should stand for better composition. Moreover, it also provides camera motion guidance for pan, tilt,

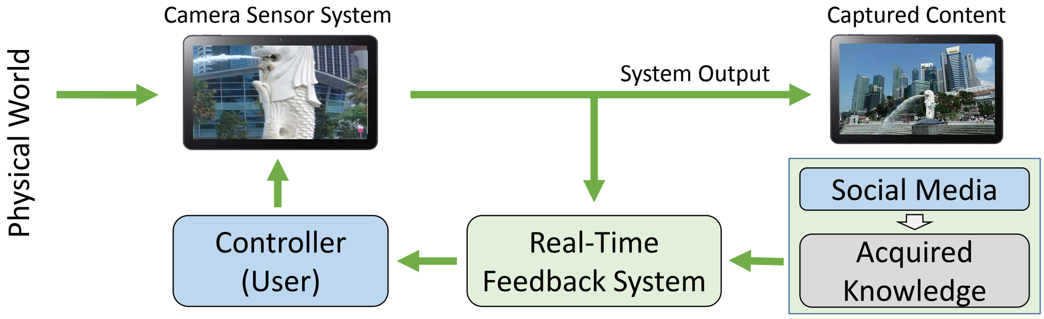


FIGURE 3.1: Real-time Feedback Control System

and zoom, to the user for improving scene composition. The work presented in this chapter was published in [138, 139] and [137].

3.1 Introduction

With a high-quality camera and a bit of practice, it is possible to take average quality photos with the camera set on the *automatic* mode. But in order to capture truly beautiful photographs, we need to leverage every possible ability of the camera and for this, we need to acquire some knowledge of photography and learn the manual camera settings. One of the most important camera settings includes exposure, in which we try to find the proper exposure for the subject and lighting conditions [83]. The automatic mode is not perfect, which is why professional photographers prefer manual settings to produce better photographs. They practice and gain experience to understand which combinations of aperture, ISO and shutter speed are best for different kinds of photos.

Professional photographers use their experience and knowledge to capture high-quality images based on the context. Motivated by this fact, we propose a comprehensive system for photography assistance using crowd-sourced images which take into account available contextual information for composition and camera parameter learning. We employ machine learning to build the models for image composition and camera parameters. Figure 3.1 shows an overview of the proposed approach. The framework shows a real-time control feedback system where the input is sensed from the physical world (view, geo-context, etc.) and feedback is provided to the controller (photographer) to improve the image quality.

The remaining sections are organized as follows. Section 3.2 presents an overview of the proposed system. Technical details of the proposed system are presented in section 3.3 and section 3.4. Section 3.5 presents the experiments conducted for the evaluation of this work. Finally, section 3.6 concludes the chapter with possible future research direction.

3.2 Proposed Framework

The proposed approach consists of two phases, photography learning, and real-time assistance. The first phase is an offline process which trains the photography model. In the second step, the learned photography knowledge is utilized to provide real-time assistance to the user.

In the photography learning phase, models for photographic composition and camera parameters are trained using the crowd-sourced images. Figure 3.2 presents an overview of the proposed photography learning model. The crowdsourced image database is augmented with social media cues such as user likes, shares, views, etc., which are used for image aesthetic quality evaluation. Each image is also associated with *Exif* information which provides details like, geo-location, time of capture, camera parameters, etc. The time stamp can be used to identify the environmental conditions of a location at the time of capture. Popular landmark objects are identified for each location and their position in the image frame is utilized along with the context for composition learning. Probabilistic generative models are built for landmark objects based on their position in the image frame which are further used in the second phase for providing camera motion recommendation to the user. As context information such as geo-location, lighting conditions,

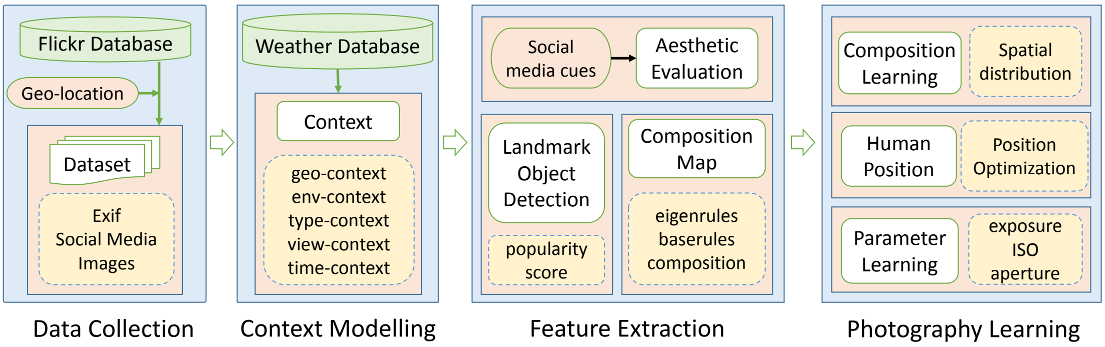


FIGURE 3.2: Proposed Framework for Photography Learning

etc, are important for setting camera parameters, context features are utilized for learning camera parameter models.

In the feedback phase, assistance regarding image composition and camera parameters is provided to the user in real-time based on the current user context. The geo-location of the user and time of capture are used to get the environment conditions and derive associated context features. The current view on the user camera and context features are fed into the composition model and an aesthetic score is predicted for the view based on its composition. Similarly, camera parameter values are predicted using the model for camera parameters. If there are faces detected in the current view then the composition model is also used for recommending a position in the frame where people should be placed. The generative models build for landmark objects are used for recommending camera motion such as pan, tilt, and zoom to improve the scene composition.

3.3 Photography Learning

3.3.1 Context Features

Image composition and camera parameters are the two most important factors affecting image aesthetics. However, both these factors are context dependent and therefore context plays an important role in photography. We define five context features for photography, *time-context, geo-context, env-context, view-context* and *type-context*, which have impact on image composition and camera parameters.

Time-Context Time is an important factor for learning camera parameters as it has a direct impact on lighting conditions. Time of image capture is used along with sunrise, sunset and sunpeak time to define *time-context*. It is described as a three-dimensional vector and the values are calculated by finding the time difference between time-of-capture and each of sunrise time, sunset time and sunpeak time.

Geo-Context Different geo-locations may have different visual elements, therefore, geo-location is important from a composition perspective. Geo-location also affects camera parameters as lighting and weather conditions may also vary from location to location. To define the *geo-context* we use the latitude and longitude of the location.

Env-Context Using date, time and geo-location information of captured image, we extract the environment conditions from the weather database. We utilize factors such as temperature, visibility, humidity, haze, rain conditions, the month of capture, dew, mist and cloud details to define *env-context*. Cloud conditions are defined as no-clouds (0), scattered clouds (1), partly cloudy (2), mostly cloudy (3) and overcast (4). Rain is defined as no-rain(0), light rain(1) and heavy rain(2). Detailed haze and mist levels are encoded as 0 or 1 based on whether haze/mist is present or not. All these factors are combined to form a 9-dimensional feature descriptor for *env-context*.

View-Context It is introduced to differentiate between various possible views at a particular geo-location. It basically refers to the viewing direction at any geo-location, but compass information is not available for the crowd-sourced images. However, at any geo-location, the viewing direction can also be inferred from the content of the image captured at that point. Images captured at any geo-location will have varying image composition and we assume that different views will have different color composition. Therefore, we make use of the color composition of an image to infer the viewing direction and use localized color (RGB) histogram to define *view-context*. The image is divided into small cells of uniform size using a grid of N x N cells. A block is formed by grouping M x M spatially connected cells and, a color histogram is extracted for each block. To form the descriptor for *view-context*, histogram from all the blocks are combined together. RGB color space is used for the histogram with 32 bins for each color. For our experiments we used N=4 and M=2 to form a 864 (3x3x96) dimensional feature descriptor to represent *view-context*.

Type-Context A photograph with and without people will have different image composition as well as camera parameters. To differentiate between images with and without people, we define the *type-context*. Feature descriptor for *type-context* is formed using face recognition on the image. The number of people present in the image is used along with the size of the largest face to form a 2-dimensional normalized feature vector. The size of a face is important because it reflects the distance of a person from the camera and will affect the aperture setting in the parameters.

Context information like time of capture and geo-location can be extracted from the Exif metadata of the image. Using the time-stamp and geo-location of the captured photograph we can find out environmental conditions from online weather service providers. We obtained the historical sunrise, sunset and sunpeak time from [165] by specifying the geo-location and date for the

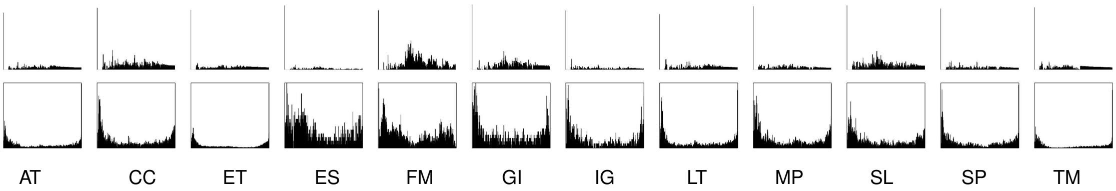


FIGURE 3.3: Distribution of aesthetic scores for different locations in our dataset. First row shows the distribution using method proposed by [190] and second row shows distribution using our method. Please refer to table 3.1 for details of the acronyms used for locations.

crowd-sourced images. The historical hourly weather information is obtained from [180] given the time, date, and geo-location of the crowd-sourced image.

3.3.2 Aesthetic Score Evaluation

We make use of social media cues for finding the aesthetic quality of the crowd-sourced images. The number of user-favorites and user-views of any photographs on social media indicates its popularity among social media users. Apart from social media cues, we also take into account Flickr's '*interestingness*' score. It is based on the quantity of user-entered metadata, the number of users who have assigned metadata to the media object and access patterns related to the media object [21].

Using this meta-data information we assign an aesthetic score to every image in the dataset which ranges from 0-1. Here a value close to 0 indicates a badly composed image and a value close to 1 indicates a good photograph. We used the logistic function to compute the aesthetic score using the number of user-views (v), number of user-favorites (f) and *'interestingness'* score (I). Earlier studies have employed exponential function for aesthetic score evaluation [190]. The motivation behind using the logistic function is that it will well separate the good and bad photographs and the aesthetic score will not increase exponentially with the increasing number of user-likes and favorites. We use equation 3.1 to calculate the aesthetic score for an image,

$$aesthetic\_score(v,f,I) = \frac{1}{1+e^{-F(v,f,I)}}, \tag{3.1}$$

where,

$$F(v,f,I) = \upsilon \log(v+1) + \beta f + \frac{\iota f}{v+1} + \vartheta I - \kappa. \tag{3.2}$$

Here, $\upsilon, \beta, \iota, \vartheta$ and κ are constants which are empirically determined such that photographs with median values of v, f and I for any geo-location get an aesthetic score of 0.5. The value of κ is set to 6 as for $v = f = I = 0$ aesthetic score will be close to 0 which is the desired value. $\upsilon = \frac{2}{\log(v_m)}, \beta = \frac{1}{f_m}, \iota = \frac{v_m}{f_m}, \vartheta = 4$ and $\kappa = 6$ where v_m is the median number of user-views and f_m is the median number of user-favorites from the dataset. Figure 3.3 presents the distribution of aesthetic scores for all the images in our dataset for different locations. The x-axis represents aesthetic score (0-1 from left to right) and the y-axis represents number of photos with that aesthetic score. The distribution shown in the first row is computed employing the method proposed by [190] and the second row presents the distribution using equation 3.1. The datasets are constructed using the ranked list (based on the interestingness score [21]) of bad and high quality images which is why we have large number of images with low and high aesthetic scores. However, this is not evident in the distribution shown in first row of figure 3.3 where exponential function is used for computing the aesthetic scores.

3.3.3 Photographic Composition

Photographic composition is the organization of important objects in the image. Position, size, texture, color and shape of the object in an image are some of the factors which define the quality of a photograph [47]. In crowd-sourced image dataset for any geo-location, we have images with similar views but different aesthetic score based on user ratings. One of the reasons for varying user-ratings of these set of images is different photographic composition. We aim to learn the best photographic composition rules for any geo-location making use of crowdsourced images.

For composition learning, we first extract the most popular landmark objects for a given geo-location. It is important because of the following reasons. In crowd-sourced image database, there can be some random images clicked by users (like a self-portrait, image of some animal, etc.) which can contribute to noise in the learning process. Also, popular landmark object detection is an offline process and it avoids saliency detection in real-time, thus reducing the run-time overhead.

3.3.3.1 Popular Landmark Object Detection

Popular landmark objects of any geo-location will occur more often in the captured images. We make use of image segmentation, saliency detection and clustering to identify these landmark objects. For image segmentation we use SLIC (Simple Linear Iterative Clustering) [2] approach for fine grain segmentation of an image into superpixels. The obtained superpixels are then merged to form bigger segments based on color similarity. Saliency map provides information about salient regions in an image. For saliency map, we used visual attention based saliency proposed by Achanta et al. [3]. Image segmentation and saliency map of an image are combined to extract salient objects from a photograph. A saliency score is assigned to each segment computed by taking the average of saliency values of all the pixels which form the segment.

The obtained segments are termed as visual words and represented using a set of visual features. For feature extraction we use Histogram of Oriented Gradients (HOG) [36], Speeded-Up Robust Features (SURF) [17] and RGB color histogram. These three features are combined to form a normalized 904-dimensional feature vector which comprises of 72 dimensions for HOG,

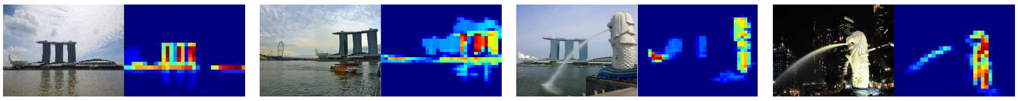


FIGURE 3.4: Composition map for some sample images

64 dimensions for SURF and 768 dimensions for RGB. Image segments with low saliency are dropped from the pool and then clustering (k-means) is performed to group the identical visual words. Each cluster represents a popular landmark object and is represented using the mean value (features) of visual words which belong to the same cluster.

Visual words for a geo-location are also assigned a saliency value. We compute the average saliency for each cluster using saliency of visual words which belong to that cluster. Popular landmark objects will occur more often in captured photographs and corresponding visual words will usually have higher saliency value. Therefore, the popularity of a landmark object is computed as,

$$popularity(V_i) = average(\frac{Num_i}{Num_{max}}, \frac{\sum_{j=1}^{Num_i} s_j}{Num_i}), \tag{3.3}$$

where, V_i is the i^{th} landmark object, Num_i is the cluster size ,Num_{max} is size of the largest cluster and s_j is the saliency value for visual word j for cluster i.

3.3.3.2 Composition Learning

We define the composition of an image in the form of feature descriptor which is further used for composition learning. For each image, visual words extracted after segmentation are classified into popular landmark objects from similar geo-context employing the Nearest Neighbor approach. Composition map for each image is formed using detected popular landmark objects and their corresponding pre-computed popularity value. The value of a pixel in the composition map is the popularity score of the salient objects which is present in that pixel location. There can be faces present in the image which may change the composition of a view. To take this into account we make use of face detection and all the detected faces are marked as salient and are assigned highest saliency. The composition map is divided into fixed number of cells using a grid (NxN). Cell size (num. of pixels per cell) will depend upon the size of an image. Spatially connected cells (MxM) are merged to form blocks. A histogram of popularity is extracted for each block using B bins. Blocks can have overlapping cells and composition descriptor is formed by combining histogram of popularity for all the blocks. Thus we have a $B(N-M+1)(N-M+1)$ dimension feature descriptor for composition. This composition descriptor is inspired by HOG feature descriptor [36], which has shown great success in human detection in images. Further details of feature extraction are presented in the experiment section. Figure 3.4 shows some sample images with corresponding composition map.

Each image is assigned an aesthetic score based on the social media cues. We consider images with aesthetic score >0.6 as good and <0.4 as bad images. Images with aesthetic scores between 0.4 and 0.6 are ignored to omit images with ambiguous aesthetics. Viewpoint and type of view are two factors which can bring variation in scene composition. Therefore, we use the composition descriptor along with *view-context* and *geo-context* to train a classifier for composition learning. We employ Support Vector Machine [25] for training a classification model for composition.

3.3.3.3 Composition Learning and Photography Rules

Different photography rules may apply for different geo-locations and views based on image composition. To find out popular photographic compositions for any geo-location we employ matrix decomposition on the composition feature of images. We take images with a good aesthetic score (>0.6) and form a matrix where columns of the matrix represent composition feature for each image. The composition map is divided into a grid with m rows and n columns with m×n rectangular cells and, an average popularity score is calculated for each cell. A feature vector is formed by combining the average popularity score of all the cells in the grid. We employ Principal Component Analysis (PCA) to find the basis for the composition rules. The idea is inspired by [167] where *eigenfaces* are used as the basis for human faces. Here we use a similar technique for computing the basis for photographic composition rules. The components which contribute most towards the variance in the composition are chosen as the basis vectors and are called *eigenrules*.

To explore photographic composition further, we employ Non-Negative Matrix Factorization (NMF) to find a popular position in the image where objects are placed by the photographers. NMF has been used for learning parts of faces and semantic features of text by researchers [95]. Here we utilize NMF factorization to learn the composition basis for images and termed them as *baserules*. The motivation behind *eigenrules* and *baserules* is to visualize and support the proposed composition learning. Since NMF decomposition only gives the positive basis, complex compositions can be formed by adding different combinations of *baserules*. On the other hand *eigenrules* can have negative components as well, therefore inferring complex compositions involve addition as well as subtraction of *eigenrules*. Therefore, from the users perspective, *baserules* are more important and easy to visualize as they indicate positions where salient objects should be placed.

3.3.4 Spatial Distribution Modeling

Placement of popular landmark objects in the image frame at different positions may lead to varying compositions. Some of these positions will lead to high-quality photographs based on the composition. Also, there can be more than one position on the image frame for a landmark object which may lead to a high-quality image. For any landmark object, if we find out the favorable positions in the image frame which lead to high-quality photograph, then we can guide a user in

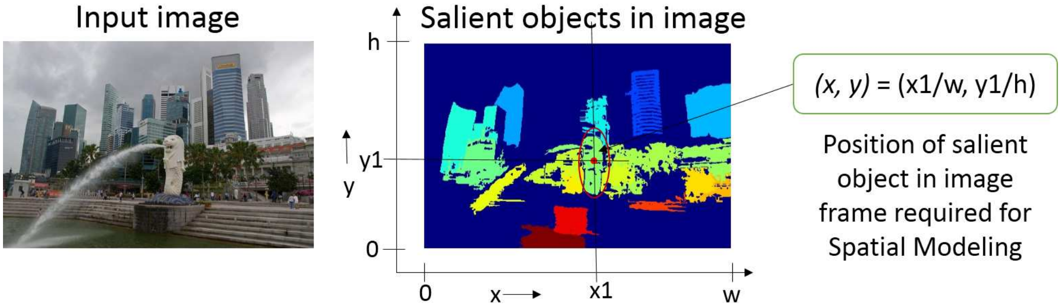

Input image
Salient objects in image
h
y1
y
0
0
x
x1
w
(x, y) = (x1/w, y1/h)
Position of salient object in image frame required for Spatial Modeling

FIGURE 3.5: Salient object detection and object position for spatial distribution modeling.

camera motion as a scene is being composed. Based on the current position of the landmark objects in the image frame, camera motion such as pan, tilt, and zoom can be recommended which will improve the current scene composition.

To estimate the favorable positions of a landmark object in a given image frame we associate each object with a spatial probabilistic distribution model, which is its probability over different possible positions. This model is used to estimate the best possible position for extracted landmark object in an image frame based on the current view, and camera motion is predicted for recommendation. The spatial distribution for each landmark object over image frame is assumed to be a Gaussian Mixture Model (GMM). For any landmark object L, we denote $\mathbf{x}(L) = (x,y)^T$, where (x,y) are the normalized center of mass coordinates for landmark object on image frame with respect to frame size (see Figure 3.5 for details). Therefore for a given landmark object L_k, the probabilistic distribution of $\mathbf{x}(L_k)$ is expressed as:

$$p(\mathbf{x}(L_k)) = \sum_{i=1}^{N_k} w_i^k \mathcal{N}(\mathbf{x}|\mu_i^k, \Sigma_i^k), \tag{3.4}$$

where, w denotes the prior, $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ denotes a Gaussian component and N_k is the number of Gaussian components in the mixture. We employ Bayesian information criterion (BIC) to set the number of Gaussian components ranging from 1-20. The Gaussian mixture model parameters

(w^k, μ^k, Σ^k) can be estimated for all the extracted landmark objects (L_k) for a given location using expectation-maximization (EM) algorithm [40]. In the E-step we compute the expected class probability for each of the visual word ($\mathbf{x}_t$) corresponding to a landmark object L^k:

$$P(i|\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t|\mu_i^k, \Sigma_i^k)w_i^k}{\sum_{l=1}^{N_k} \mathcal{N}(\mathbf{x}_t|\mu_l^k, \Sigma_l^k)w_l^k}. \tag{3.5}$$

In the M-step, mean ($\hat{\mu}$), covariance ($\hat{\Sigma}$) and priors for each class ($\hat{w}$) are updated:

$$\hat{\mu}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}, \tag{3.6}$$

$$\hat{\Sigma}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)[\mathbf{x}_t - \hat{\mu}_i^k][\mathbf{x}_t - \hat{\mu}_i^k]^T}{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}, \tag{3.7}$$

$$\hat{w}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}{T_k}, \tag{3.8}$$

where, T_k is the number of samples for a given landmark object L_k. We will discuss later how this model can be used for camera motion recommendation.

3.3.5 Camera Parameters

Along with image composition, setting camera parameters is another challenging task for amateur users. ISO, aperture, and shutter speed are most important among other camera parameters and these three also play a vital role in adjusting the exposure. These parameters can be extracted from *Exif* metadata for the captured images. Since these parameters highly depend upon the scene composition and lighting conditions, we developed a system where we make use of environmental factors along with image composition to learn the camera parameters. As different combinations of aperture, ISO and shutter speed can lead to same exposure value [72], we propose to learn these parameters independently based on relevant context information.

3.3.5.1 Exposure Value Learning

Exposure value (EV) represents a combination of camera's shutter speed and f-number (aperture), such that all combinations that yield the same exposure have the same EV value. Although all camera settings with the same EV give the same exposure, they do not necessarily give the same picture. The f-number determines the depth of field, and the shutter speed (exposure time) determines the amount of motion blur leading to the differences in captured photograph. Exposure value is a base-2 logarithmic scale defined as [72],

$$EV = log_2 \frac{A^2}{T}, \tag{3.9}$$

where, A is the relative aperture, T is the exposure time ('shutter speed') in seconds. ISO setting of the camera also affects the exposure value. The relation between exposure value and ISO value is given by [72],

$$EV_S = EV_{100} + log_2 \frac{S}{100}, \tag{3.10}$$

where, S is the ISO value, EV_S is the exposure value at ISO S and EV_{100} is the exposure value at ISO 100. From 3.9 and 3.10 we get,

$$EV_{100} = log_2 \frac{A^2}{T} - log_2 \frac{S}{100}. \tag{3.11}$$

Based on the aesthetic score we select good quality photographs (score >0.6) and calculate the exposure value (EV_{100}). For each photograph we define a context feature vector using the information described in section 3.3.1. *Time-context, geo-context, env-context* and *view-context* are used to define feature vector for exposure learning. Geo-context will not have much effect on exposure learning as we have geo-localized datasets, but it is used to make the system generalized. Good quality images are used for exposure learning and regression analysis is done using the context information with calculated exposure (EV_{100}) as target values.

3.3.5.2 Aperture Learning

The aperture value of a camera lens is used to control the amount of light reaching the image sensor. In combination with variation of shutter speed and ISO, the aperture size regulates the image sensor's degree of exposure to light. Apart from controlling the amount of exposure, the aperture is also used to control the depth of field in the photograph [83]. Smaller aperture size will bring all foreground and background objects in focus, while a larger aperture size will isolate the foreground from the background by making the foreground objects sharp.

As the aperture setting of a camera lens depends upon the type of photographic view and its composition, we will make use of the image composition to learn the corresponding aperture value from crowdsourced images. Photographs with human objects will have greater depth of field as compare to images without human objects. Also, the size of human objects will affect depth of field as bigger human objects will be closer to the camera leading to a larger depth of field as compare to smaller human objects. The aperture value will depend upon the view, presence and absence of human object and the size of human object if present. Therefore, we use *time-context, geo-context, view-context, env-context* and *type-context* for aperture learning. Regression analysis is performed using these features to train a model for aperture value.

3.3.5.3 Shutter Speed Learning

Shutter speed controls the exposure time of the lens and thus affects the exposure value. Shutter speed is also controlled to capture dynamic scene content. Finding scene dynamics from already captured image is not easy. However, for a given geo-location *view-context* can be used to differentiate between various views and we assume that scene dynamics does not vary much for landmark locations. Therefore we utilize the *view-context* along with *time-context, env-context* and *geo-context* to learn shutter speed. Regression analysis is done for training a model using context as the feature and shutter speed as target value.

3.3.5.4 ISO Estimation

ISO controls the exposure value and has a direct affect on the quality of image. ISO is the level of sensitivity of the camera's sensor to available light. The lower the ISO number, the less sensitive it is to light, while a higher ISO number increases the sensitivity of the camera. But higher sensitivity comes at an expense and it adds grain or 'noise' to the photograph. After learning the exposure, aperture and shutter speed we can estimate ISO value of the camera using equation 3.12.

$$ISO = \frac{100 \times A^2}{T \times 2^{EV_{100}}}, \tag{3.12}$$

where, A, T and EV_{100} are the corresponding predicted values for aperture, shutter speed and exposure respectively.

3.4 Real-time Feedback

In the feedback phase, a recommendation regarding the image composition and the camera parameters is provided to the user. The geo-location, using GPS of the smart device, and the time information are used to obtain the environmental conditions of the location using weather forecasting services [180], [165]. The obtained information is further utilized to derive the context features as described in section 3.3.1. The current view on the camera device is segmented to find visual objects in the scene. The extracted visual objects are classified as popular landmark objects using a Nearest Neighbor approach and a composition map is constructed with associated popularity score of objects. Thereafter, a composition feature descriptor is extracted for the view as described in section 3.3.3.2. The feature descriptor is used along with the context features to predict an aesthetic score for the current view using the trained composition model. Similarly, camera parameters are also predicted based on the context features using trained parameter models.

3.4.1 Human Position Recommendation

In the case of portrait or group photos, we can make suggestions about the location where a person or group of people should be in the image. The composition learning phase takes into account the faces occurring in the image and therefore the trained model for composition can be used to find out a position of faces in the image frame which will lead to better photographs. The detected faces in the scene are used to find a bounding box for all the faces. The composition grid is searched for a position which gives the best aesthetic score based on the composition. In our earlier work [138], the complete composition grid was searched with a step size of one cell for an optimal solution. This approach can provide an optimal solution but it is not suitable for a real-time system as searching the complete grid will be computationally expensive.

We propose a more efficient method to solve this problem by employing some well-known photography rules along with a Hill Climbing approach. Finding the best position for people to stand in an image frame is posed as an optimization problem. Instead of searching the complete grid we employ the Hill Climbing algorithm and the starting points for the algorithm are chosen based on photography rules. The *rule-of-thirds* and *rule-of-center* are used to determine the initial points from where the search begins. These rules are used as guidelines by photographers to place the salient objects in a scene. *Rule-of-center* states that the salient object should be placed at the center of the image frame and *rule-of-thirds* define four power points and four power lines and states that the salient objects should be placed either close to these points or along these lines [47]. Based on this we define nine positions in the image frame, one at the center, four at power points and four at the center of power lines, to start the search.

The position of a person can be defined as a two-dimensional point (x,y) in an image frame. A person can also move towards or away from the photographer which will have a zoom (in/out) effect and will change the image composition. To take this into account we introduce another dimension (z) to the position and it can be defined as a three-dimensional point (x, y, z). Therefore, the search for an optimal position is performed in three-dimensional space using Hill Climbing algorithm. The center of the bounding box for all the detected faces in the scene is set as the initial position. The predetermined nine positions along with initial position are used as a starting point for the search algorithm. The value of initial points in the z dimension is set to 0, which represents no zoom factor in the initial configuration. The search space is explored in steps of (*x_step, y_step, z_step*) at a time in the composition map with *x_step = cell_x/10, y_step = cell_y/10* and *z_step = (cell_x+cell_y)/20* where, *cell_x* and *cell_y* are the width and height of a cell in the composition map. The details of human position recommendation are presented in algorithm 1. Six point connectivity is assumed as neighbors ($Neighbors$) of a point are analyzed in the search process. The composition map is modified ($Modify_{map}$) using the bounding box as the search space is explored. An aesthetic score is evaluated ($Score$) for the modified composition map and search continues as long as a better composition is found in the neighborhood. After all the positions are explored, the position with the best aesthetic score is selected as the final recommendation.

ALGORITHM 1: Human Position Recommendation

```
Input: Composition map ($CM_0$), bounding box for detected faces ($BB$), initial position of $BB(pos_0)$, list of
       predefined starting positions $(pos\_list_0)$
Output: Best position for recommendation $P_{best}$.
$P_{best} := pos_0$; $best\_score$ := 0; $PosList$ := {$pos_0, pos\_list_0$};
for each position $pos$ in $PosList$ do
    $current\_pos$ := $pos$; $change$ := 0;
    $CM$ := Modify_Map($CM_0$, $BB$, $current\_pos$); $score$ := Score($CM$);
    repeat
        $N$ := Neighbors($current\_pos$);
        for each position $p$ in $N$ do
            $CM$ := Modify_Map($CM_0$, $BB$, $p$); $s_p$ := Score($CM$);
            if ($s_p > score$) then
                $score$ := $s_p$; $current\_pos$ := $p$; $change$ := 1;
            end
        end
    until $change = 1$;
    if ($score > best\_score$) then
        $best\_score$ := $score$; $P_{best}$ := $current\_pos$;
    end
end
```

3.4.2 Camera Movement Recommendation

The spatial distribution model for landmark objects are trained in learning phase. The obtained model is used to predict the most favorable position in an image frame for the landmark objects extracted from a given scene. Each image has some set of landmark objects $I(L_1, L_2, ..., L_n)$, and each object can be represented as $\mathbf{x}(L) = (x, y)^T$, where (x, y) are the normalized center of mass coordinate for landmark object on image frame with respect to frame size, where n is the number of landmark objects in the image. For each landmark object we predict a target position on the image frame using the probabilistic model defined earlier. Therefore, for each object we have an initial position (i_x, i_y) and a target position (t_x, t_y). The set of initial positions in an image $I(I_x, I_y)^T$ can be represented using the affine motion model [87] based on set of corresponding target positions $I(T_x, T_y)^T$:

$$\begin{pmatrix} I_x \\ I_y \end{pmatrix} = \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} T_x \\ T_y \end{pmatrix} + \begin{pmatrix} a_1 \\ a_4 \end{pmatrix}, \tag{3.13}$$

where, $\Phi = (a_1, a_2, a_3, a_4, a_5, a_6)$ is the parameter vector which is estimated using least square (LS) method. After the estimation of affine model parameters, the camera motion operations can be obtained as follows [87]:

$$pan = a_1, tilt = a_2, zoom = \frac{1}{2}(a_2 + a_6). \tag{3.14}$$

The obtained camera operations will transform the initial image composition to the predicted composition. These operations can be recommended to the user for improving the scene composition as a photograph is being captured.

3.5 Experiments and Results

3.5.1 Dataset

We used Flickr's API to download crowd-sourced images along with social mediadata. Each image is associated with Exif meta data and social media cues such as user likes, user favorites and user comments. We collected around 62K images for 12 different tourist locations including *Merlion Park* (Singapore), *Esplanade* (Singapore), *Float at Marina Bay* (Singapore), *Eiffel Tower* (Paris, France), *Statue of Liberty* (New York, USA), *Taj Mahal* (Agra, India), *India Gate* (Delhi, India), *Gateway of India* (Mumbai, India), *Leaning Tower of Pisa* (Italy), *Arc de Triomphe* (Paris), *Cologne Cathedral* (Germany) and *St. Peter's Basilica* (Vatican City). We make use of Flickr's *photos.search* API which allows a search of geo-tagged images in order of *interestingness* score and gathered top ranked and bottom ranked images for each location. Apart from this, we also collected popular photographs from Flickr for learning composition basis. We constructed two independent datasets using '*interestingness*' API with 21K (DB-1) and 51K (DB-2) images. For the first dataset we crawled images from the year 2008-12 and for the second dataset, we crawled images from the year 2013-14. We chose two different periods to make sure we do not have common images in the two datasets.

3.5.2 Landmark Object Identification

Using image segmentation and saliency detection we extract salient objects from images to generate a pool of visual words for each geo-location. We chose images with an aesthetic score above 0.80 for landmark object detection to reduce the time complexity as well to reduce noise from images with bad aesthetic scores. Another reason for choosing only good quality images is that the camera focus is usually on salient objects in high-quality images and therefore saliency detection provides better results. Features are extracted for each image segment as described in section 3.3.3.1. Similar segments are grouped together using K-means algorithm with 200 clusters. The number of clusters is heuristically chosen based on the observation that for each image we extract around 20-50 visual words and each geo-location can have around 5-10 views.

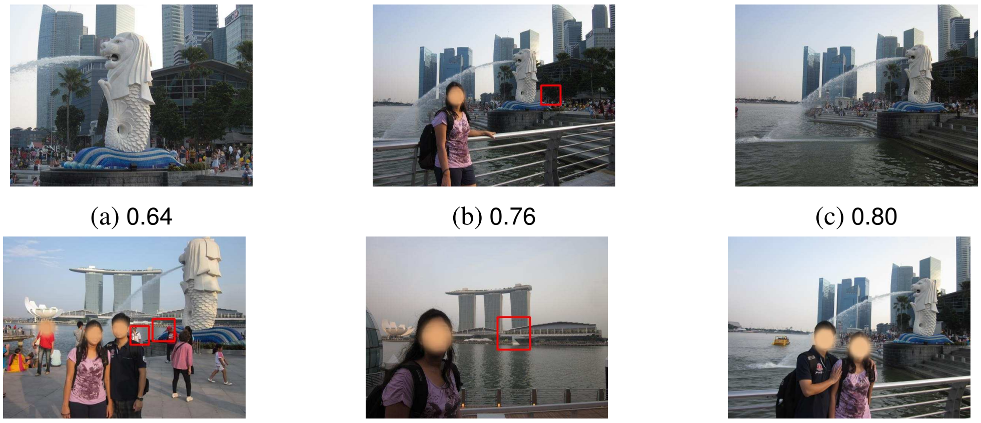
3.5.3 Composition Learning

In composition learning, we aim to train a classification model which can differentiate between good and bad compositions. We consider images with aesthetic score >0.6 as good and <0.4 as bad. Composition feature descriptor is extracted for each image as discussed in section 3.3.3.2, with N=9, M=3 and B=20. Thus we have 980 dimensional feature descriptor representing image composition which is used along with *view-context* and *geo-context* to train a classification

TABLE 3.1: Classification Results for Composition Learning

Location	Dataset Size	*linear* Kernel		*rbf* Kernel	
		Accuracy	Precision	Accuracy	Precision
Arc de Triomphe (AT)	5500	0.65	0.65	0.68	0.68
Cologne Cathedral (CC)	6056	0.68	0.57	0.68	0.60
Eiffel Tower (ET)	17093	0.64	0.63	0.69	0.69
Esplanade (ES)	1837	0.78	0.67	0.81	0.72
Float at Marina Bay (FM)	3862	0.68	0.60	0.72	0.69
Gateway of India (GI)	1536	0.65	0.51	0.66	0.55
India Gate (IG)	1498	0.73	0.57	0.76	0.70
Leaning Tower of Pisa (LT)	5133	0.64	0.56	0.76	0.61
Merlion Park (MP)	3385	0.70	0.44	0.72	0.71
Statue of Liberty (SL)	5667	0.69	0.67	0.69	0.67
St. Peter's Basilica (SP)	4964	0.64	0.71	0.69	0.75
Taj Mahal (TM)	6118	0.66	0.52	0.70	0.62

model. To test our method, we employ binary Support Vector Machine (SVM) [25] with both *linear* and *rbf* kernel to train separate classification model for each location. We use 5 fold cross validation to determine the accuracy rate and precision of the classifier. Table 3.1 presents the classification accuracy and precision score for all the locations in our dataset. The average accuracy for all the datasets is around 71% with 67% average precision. The probabilistic score from classification is used to evaluate the aesthetic score of the image. For images with people, position recommendation is made to obtain a maximum aesthetic score. Figure 3.6 shows sample images with predicted aesthetic scores along with a position recommendation which is shown in red blocks.



(a) 0.64

(b) 0.76

(c) 0.80

(d) 0.84

(e) 0.81

(f) 0.86

Image	Aesthetic Score	Actual Values/Predicted Values			
		Exposure	Aperture	Shutter Speed	ISO
a	0.64	11.96/12.37	4.0/5.05	0.0050/0.0026	80/180
b	0.76	12.26/11.75	2.8/5.70	0.0020/0.0063	80/147
c	0.80	12.26/12.43	2.8/4.06	0.0020/0.0016	80/176
d	0.84	12.26/13.0	2.8/5.18	0.0020/0.0016	80/196
e	0.81	9.61/13.44	2.8/5.99	0.0125/0.0023	80/134
f	0.86	12.26/13.40	2.8/5.6	0.0020/0.0028	80/103

FIGURE 3.6: Sample images from ‘Merlion Park’ location with predicted aesthetic scores and camera parameters

3.5.4 Analysis of Eigenrules and Baserules

To explore composition learning we extracted the basis for photographic composition as discussed in section 3.3.3.3. For the experiments, we use m=24 and n=32, with a 3:4 aspect ratio, and extracted a 768-dimensional feature vector for each image. We used the two datasets with popular photographs for this experiment as these photographs are mostly captured by good photographers. We extract top *Eigenrules* and top *Baserules* for the two datasets separately. As discussed earlier these two datasets are independent and we tried to avoid having any common images in these datasets. The number of *eigenrules* are selected based on the total variance contributed by these basis in the data and the total number of *baserules* are selected by minimizing the reconstruction error after NMF decomposition. Based on these criteria we chose 48 *eigenrules*, with more than 80% variance contribution, and 24 *baserules*.

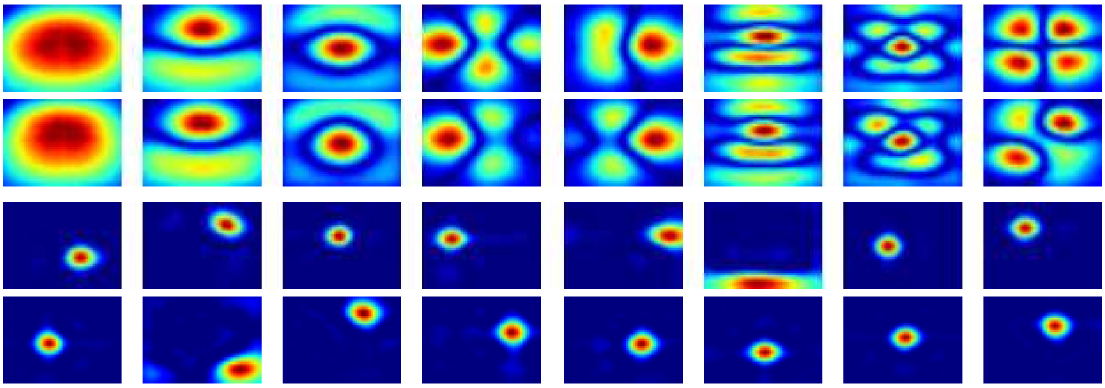


FIGURE 3.7: Visualization of *Eigenrules* and *Baserules*. Row 1 - *Eigenrules* for dataset DB-1, row 2 - *Eigenrules* for dataset DB-2, row 3 - *Baserules* for dataset DB-1 and row 4 - *Baserules* for dataset DB-2

Figure 3.7 shows top *eigenrules* and *baserules* extracted for the two datasets (DB-1, DB-2). Some of the *eigenrules* reflect popular heuristic rules of photography which are used by photographers. For example, the first *eigenrule* corresponds to 'rule of center', the second *eigenrule* corresponds to 'rule of symmetry' and the third *eigenrule* corresponds to 'rule of framing'. Similarly, some of the *baserules* shown in figure 3.7 corresponds to popular photography rules used by the photographers. For example, 'rule of thirds' is a popular and mostly utilized composition rule which recommend four positions (power points) in the frame to place the salient objects. All these four power points are found in top *baserules* for both the datasets. Also, we observed that both the set with high quality images share most of the *eigenrules* and *baserules*. Moreover, similar *eigenrules* and *baserules* are observed for all the locations in our dataset. Based on this observation we can infer that various photographic compositions can be expressed as combinations of some basic rules.

Figure 3.8 illustrates how image composition can be understood using *eigenrules* and *baserules*. The first and the sixth column shows sample images followed by *baserule* map, top *baserule*, *eigenrule* map and top *eigenrule*. *Baserule* map and *eigenrule* map shows the contribution of corresponding rule basis in the image composition. Each block in the *baserule* and *eigenrule*

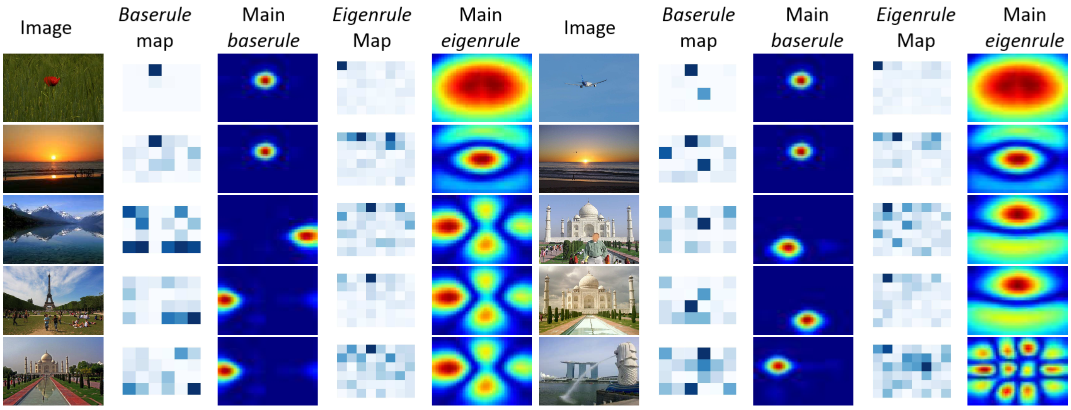


FIGURE 3.8: *Eigenrule* and *Baserule* Analysis. *Baserule* map shows the contribution of *baserules* in the image composition. Each cell in the matrix represents a *baserule*. The main *baserule* is the visualization of basis which has the maximum contribution. Similarly, *eigenrule* map shows the contribution of *eigenrules* in the image composition and the main *eigenrule* is the visualization of the *eigenrule* with maximum contribution.

map corresponds to a basis rule where color intensity maps to the contribution of corresponding rule. The first image in row 1 has only one salient object, and we can see in the *baserule* map that there is mainly one *baserule* (dark cell at location (1,3) in the matrix) contributing to the composition. The corresponding *eigenrule* is visualized next to it and it indicates the position of the salient object in the image frame. Now consider the first image in row 5 which has a complex composition and we can see in the *baserule* map that multiple *baserules* are involved. However, in *eigenrule* map, we can observe that one of the *eigenrule* is contributing more than the others (dark cell at location (1,4) in the matrix). Corresponding visualization of the *eigenrule* is shown next to it and indicates how salient objects are organized in the image frame. It can be seen that images with simple composition (first two rows) can be easily categorized using *baserules*. On the other hand, images with complex composition (rows 3-5) are easier to understand using *eigenrules* and it can be observed that images with similar composition have similar *eigenrule* maps.

In this work, we make use of the composition learning to differentiate between good and bad images based on the context. The motivation behind exploring *eigenrules* and *baserules* is to understand and visualize the composition learning. However, they can be utilized for other aspects of photography as well, like understanding the combination which leads to various known rules of photography. Different combination of *eigenrules* and *baserules* may lead to varying compositions which can be applied for more complex compositions.

3.5.5 Parameter Learning

We employ ε-Support Vector Regression (ε-SVR) with RBF kernel for training separate models for the three camera parameters. For exposure learning, *time-context, geo-context, env-context* and *view-context* are utilized to train a regression model with EV_{100} as target value. Similar context features are used for shutter-speed learning with shutter-speed (in seconds) as a target value. However, for aperture learning *type-context* is also employed along with these features for the regression model with aperture value as a target. Images with aesthetic score >0.6 are used for parameter learning. Table 3.2 shows mean squared error and R^2 score (coefficient of determination) for various locations with ten-fold cross-validation. It can be observed that the results for exposure learning are better as compared to aperture and shutter speed learning. This is because the various combinations of aperture, ISO and shutter speed can produce similar exposure and thus can lead to similar image quality. However, getting the exposure correct is more important than other parameters as the right amount of light entering the camera lens is the major factor affecting image quality.

Sample images from '*Merlion Park*' location are shown in figure 3.6 with position recommendation (red region) and corresponding actual/predicted camera parameters. Figure 3.6b and 3.6c shows similar views with and without a person standing in the foreground. As we can see from the predicted values of aperture, the image with face requires a larger aperture value as compared

TABLE 3.2: Regression Results for Parameter Learning

Location	**Exposure**		**Aperture**		**Shutter Speed**	
	R2 Score	MSE	R2 Score	MSE	R2 Score	MSE
Arc de Triomphe	0.63	0.065	0.34	0.036	0.39	0.041
Cologne Cathedral	0.58	0.092	0.41	0.018	0.41	0.018
Eiffel Tower	0.70	0.049	0.47	0.047	0.32	0.037
Esplanade	0.70	0.171	0.20	0.021	0.53	0.085
Float at Marina Bay	0.65	0.047	0.33	0.038	0.59	0.004
Gateway of India	0.45	0.090	0.43	0.033	0.59	0.016
India Gate	0.64	0.065	0.51	0.039	0.47	0.017
Leaning Tower of Pisa	0.70	0.048	0.48	0.011	0.56	0.025
Merlion Park	0.76	0.062	0.42	0.036	0.45	0.194
Statue of Liberty	0.51	0.028	0.31	0.050	0.66	0.014
St. Peter's Basilica	0.54	0.105	0.36	0.030	0.32	0.044

to the image without a face. A sample snapshot frame is shown in Figure 3.9 to demonstrate the feedback provided to a user in real-time.

3.5.6 User Study

To further evaluate the proposed system we conducted a user study and invited 8 skilled photographers and 38 amateur users. In the study, we evaluated the system for composition learning, position recommendation, and camera parameter prediction. Amateur users were invited to

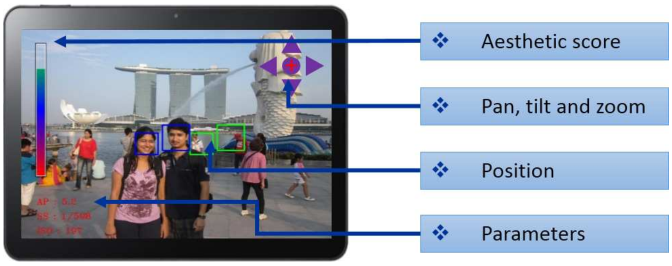
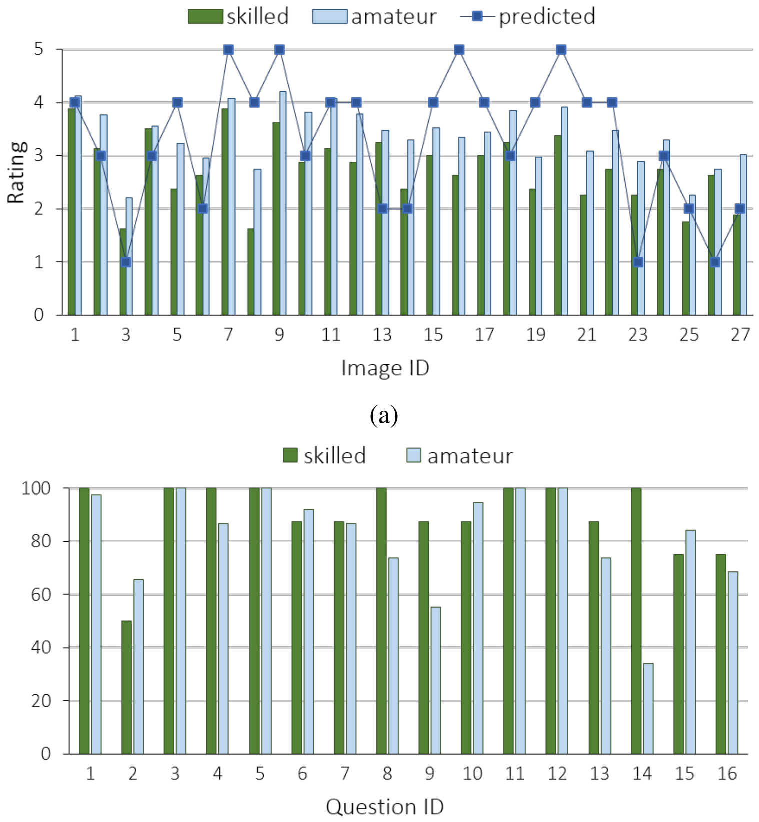


FIGURE 3.9: Demonstration of how a real-time feedback will be given to a user. The left vertical bar represents a predicted quality score of the current composition and varies from 0-1 as the color changes from red (bottom position) to green (top position). Below the bar are the predicted camera parameters (aperture, shutter-speed, and ISO values) for the current user-context. The blue boxes are the identified faces in the frame and the green boxes are corresponding boxes for recommended position. On the top right we have the recommendation for a camera motion to improve the composition. The visual buttons are for horizontal motion (left and right buttons), vertical motion (top and bottom buttons), and the center visual button is for zooming-in and zooming-out. A recommendation will be given to a user by blinking these visual buttons for corresponding motion. A plus sign (shown in sample snapshot) will be used for zooming-in and a minus sign will be used for zooming-out.

evaluate image composition and position recommendation, and skilled users evaluated camera parameter prediction as well.

For composition evaluation, users were asked two types of questions. In the first type of question, they were asked to rate an image from 1-5 based on its aesthetics quality and assign one of the value as 1 (Poor), 2 (Below Average), 3 (Average), 4 (Good) and 5 (Excellent). A total of 27 images were presented to the users for evaluation. Figure 3.10a shows the bar plot of average ratings assigned to each image by the amateur and skilled users. The point plots are the ratings predicted by our system and we can observe in figure 3.10a that most of the times ratings predicted by our system are close to the ratings assigned by the users. In the second type of question, users were provided two images of the same view but with different compositions. We collected images with wide view angle for the locations from our dataset and used our system to find good compositions for that view. An image frame of size 640x480 pixels is slid in the wide angle image with a step size of 50 pixels. For each image frame, camera motion is detected using our system and a target image is generated. The target image with the best aesthetic score is



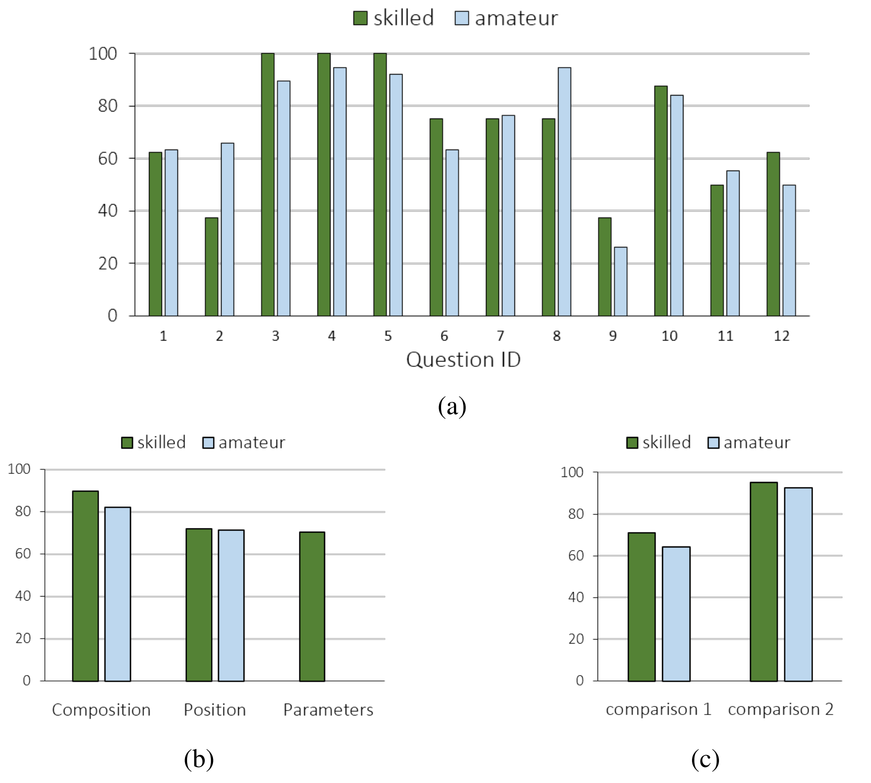
(a)

(b)

FIGURE 3.10: User study results for composition learning. (a) The bar plots are average rating for each image assigned by amateur and skilled users and the point plot is the predicted rating. (b) It shows the percentage of users who made the same choice as made by our system for amateur and skilled users.

used for the user study. Now, two images with similar views are chosen from the dataset, one with a low aesthetic score (<0.5) and another with a high aesthetic score (>0.7). The generated target image is then compared with these two images separately in the user study. The order of the images and order of the options were generated randomly. Figure 3.10b presents the bar plots for the percentage of users who preferred the image produced by our system for both amateur and skilled users. The overall consensus for composition evaluation between our system and the users was around 90% for skilled users and 82% for amateur users (figure 3.11b).

For position recommendation, the users were provided images with same scene but people standing at different positions. A total of 24 images were presented to the users and they were asked to choose the one with better quality. Figure 3.11a presents the plot for % of users who chose the position same as our system. Overall, around 72% of the choices made by skilled users and 71% of the choices made by amateur users were similar to the choices predicted by our system. For comparison with bad quality images, the consensus between our system and the users (comparison 2 in figure 3.11c) is around 95% for skilled users and 92% for amateur users, on the other hand, comparison with high-quality images (comparison 1 in figure 3.11c) have the consensus of around 71% for skilled users and 64% for amateur users. Only skilled users were invited for the evaluation of camera parameter prediction. They were shown an image along with



(a)

(b)

(c)

FIGURE 3.11: User study results for position recommendation, composition learning and camera parameter prediction. (a) It shows the percentage of users who made the same choice as made by our system for better position. (b) The overall consensus between users and our system for composition, position recommendation and parameter prediction, and (c) The overall consensus between our system and the users for different type of comparison (comparison 1 - with high quality images and comparison 2 - with low quality images).

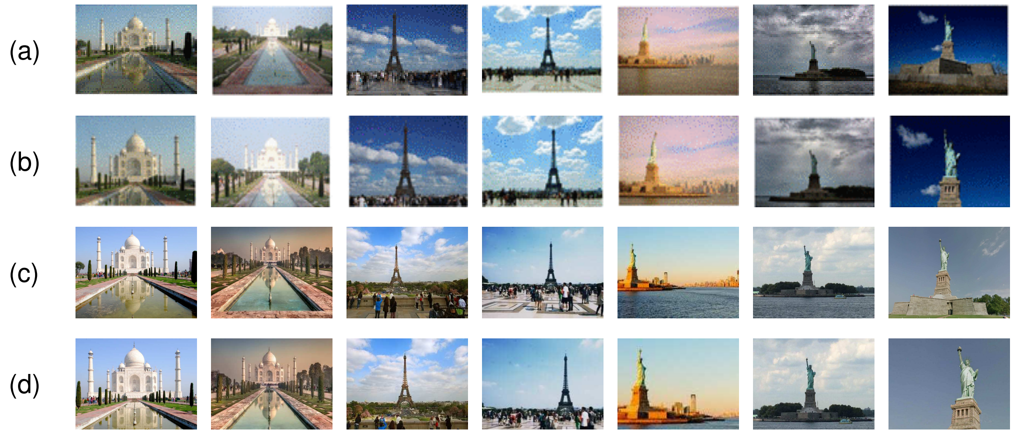


FIGURE 3.12: Comparison of composition recommendation results. Row (a) and (b) shows the wide angle view and recommended view respectively using [190]. Row (c) and (d) shows similar wide angle view and recommended view respectively using our approach.

the camera parameters which were used to capture it. With each image, two set of camera parameter options were provided along with original parameters, one of which was suggested by our system. The users were asked to choose the set of parameters which could have used to capture a better image. Around 71% of the times, choices made by the skilled users were similar to the what our system predicted (figure 3.11b).

3.5.7 Comparison

The proposed method focuses on image composition for finding the aesthetic quality of photographs. However, there are many other factors such as color composition, lighting, image content, etc. which also affects image aesthetics. Researchers have proposed various methods to compute image aesthetics which employ different types of low-level and high-level features. In [124], the authors have compared a different set of features and their combinations for computing image aesthetics and the highest accuracy rate of 89.9% is achieved for a combination of features. Our proposed method for aesthetics evaluation achieves an accuracy of 71%, which is reasonable as we are considering only composition factor. We have compared the qualitative results of our composition recommendation with state of the art method in photography assistance

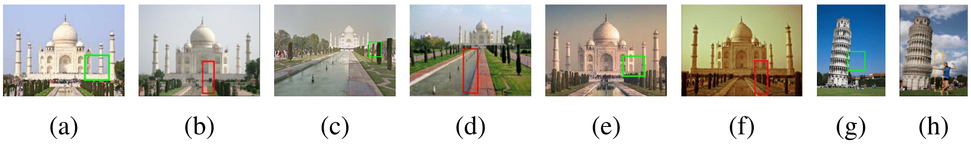


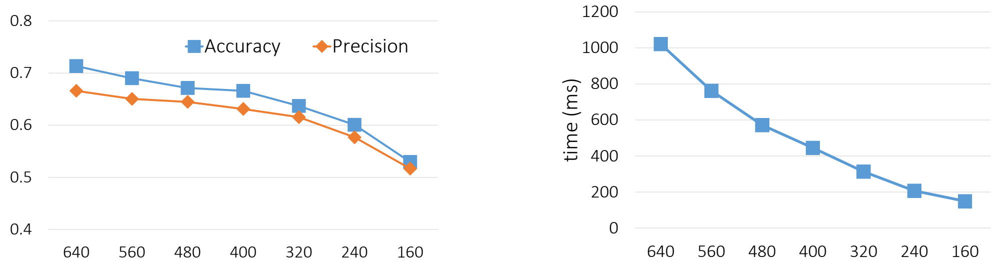
FIGURE 3.13: Comparison of position recommendation. (b), (d) (f) and (h) shows position recommendation using [181] and (a), (c), (e) and (g) shows position recommendation for similar views using our approach.

[190]. The comparison is shown in figure 3.12 and we can observe that for a similar type of view the recommended results for both the methods are almost similar. As the method proposed by [190] requires a wide angle view for making a recommendation, we have also presented a wide angle view image in our results for comparison purpose. However, it is important to note that our proposed method does not require a wide angle view for making composition recommendation. Another major limitation of [190] is that it does not take into account the presence of people in photographs and also the composition recommendation is based on some preselected exemplar view. One the other hand, our proposed method considers the presence of people in the photograph and also makes position recommendation.

We have compared our position recommendation results with state of the art method for human position recommendation [181]. The comparison is shown in figure 3.13 and we can observe that some of the results are similar. In figure 3.13d, the position recommended by [181] was not preferred by users in the survey, which was also discussed by the authors. However, for similar view 3.13c, the position recommendation proposed by our method was preferred by almost 95% of the users. The method proposed by [181] does not consider overall composition of the image and is limited to the presence of a single person. On the other hand, our method also provides composition guidance and is not limited by the number of people in the image.

3.5.8 Running-time Analysis

The experiments for the proposed system were performed on a 8 core Intel processor running at 3.40 GHz and 8 GB of RAM using unoptimized python code. The average time to process a 640×480 pixel image for predicting an aesthetic score, camera parameters and camera motion is around 1 second and position recommendation takes around 3 seconds. As the proposed method does not impose any restriction on the input image size, we further investigated the performance trade-off of the system for varying image size. 10% of the images from each location are left out for testing and rest of the images are used for training the models. The testing images are resized into different image resolutions with the longest dimension of an image ranging from 640 to 160 pixels. Figure 3.14 shows the plot of average accuracy and precision of composition learning vs. image size (fig. 3.14a) and running-time vs image size (fig. 3.14b). The running-time includes the overall time required for predicting an aesthetic score, camera parameters and camera motion for an input image. The most time-consuming process in the pipeline is image segmentation and therefore we can see a significant gain in running-time as we decrease the image size. The average R2 score for parameter learning dropped from 0.62 to 0.56 as we changed the maximum dimension of an image from 640 to 160, which is not very significant.



(a)

(b)

FIGURE 3.14: Performance evaluation of the system for varying image size. x-axis refer to the maximum image dimension in number of pixels. (a) Plot of accuracy and precision of composition learning as we vary the image size. (b) Plot of running-time with varying image size.

This is because RGB histogram of an image does not change much on image rescaling and the parameter model also depends on other factors such as *type-context, geo-context, time-context* and *env-context*. With the increasing processing power, we believe that the proposed system can be further optimized and efficiently implemented into portable smart devices with embedded cameras.

3.5.9 Limitations

Since we utilize crowd-sourced data along with social media, there are some drawbacks despite the advantages. Social media data is susceptible to noise and it can be inaccurate as people do not always express what they actually feel. We also make the assumption that popular photographs will have better compositions. However, there can be other possible reasons for the popularity of crowd-sourced photographs in social media, like it can a photograph of some celebrity. Also, we can have tourist locations with a limited number of images and the photography model learned from a sparse dataset might not be that effective.

3.6 Summary

We have presented a context based photography learning method which utilizes crowd-sourced images and associated social media cues. The proposed method can provide composition and camera parameter guidance to the user based on context. It can also provide human position and camera motion guidance to improve the image composition. We also presented the idea of photographic composition basis, *eigenrules* and *baserules* to substantiate the proposed composition learning. The idea of *eigenrules* and *baserules* can be further exploited to better understand photographic composition. Using crowd-sourced images for learning photographic rules is a promising way to capture the knowledge and intuition of professional photographers.

The problem is that most of the places will have sparse datasets. The existing methods possibly cannot provide similar performance for sparse locations as they do for dense locations. Therefore, there is a need of some kind of knowledge transfer so that the knowledge acquired from dense locations with similar context can be applied to sparse locations. In our future work, we plan to explore these ideas further.

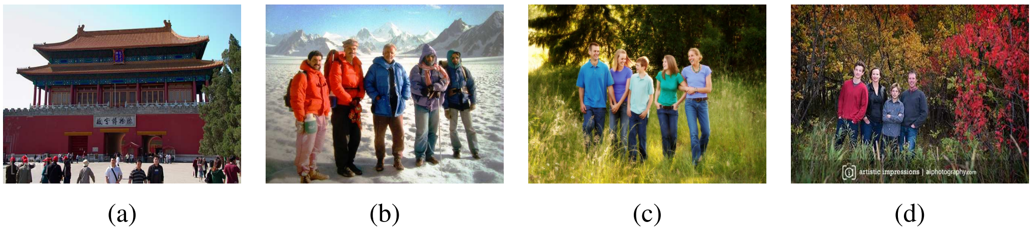
Chapter 4

Group Photography Assistance

Visual balance is considered as one of the important factors in defining the aesthetic quality of visual arts. In this work, we propose a novel method to obtain visual balance in a layout with dynamic visual elements. We use the idea of spring-electric graph model and augment it with the concept of color energy from the literature of visual arts. We also present an interesting application of the proposed model in photography assistance. We mainly focus on group photography and utilize social media images along with proposed spring-electric model for providing a recommendation to the user. The proposed method can provide real-time feedback to the user regarding the arrangement of people, their position and relative size on the image frame. We conducted qualitative experiments along with user studies to evaluate the proposed method. Experimental results and user studies show the effectiveness of the proposed model in obtaining visual balance and group photography recommendation.

4.1 Introduction

A work of art is considered aesthetically pleasing to human eye if elements within the work are arranged in a balanced compositional way [44]. Painters or still photographers also try to arrange the static pictorial elements in a picture such that they look and feel inevitably balanced [7, 196]. Therefore visual balance in an image is considered as one of the important factors in the art of composition. There are several aspects which accounts for balance such as color, size, shape,



(a) (b) (c) (d)

FIGURE 4.1: Sample amateur (a & b) and professional (c & d) photographs. (a) size of people too small and not balanced with the background view, (b) visually imbalanced, as viewers attention is focused on the leftmost two people, (c) visually balanced arrangement of people, and (d) the arrangement of people is well balanced with the background.

etc. of visual elements and they can be used individually or in a combination to obtain a balanced composition. Color energy is one such measure which is used by artists to obtain visual balance in a design. Color energy of a visual element indicates the relative aesthetic impact a color has on a viewer [196].

The interplay of screen forces among visual elements in structuring a two-dimensional field is also considered important in the art of visual design [196]. This idea of presence of virtual forces between visual elements has been widely studied in drawing aesthetic planar graphs [16, 49, 80]. In these proposed methods, a graph is represented as a spring-electric system in which visual balance is achieved by balancing mechanical and electric forces acting on the nodes of the graph. These works are mainly focused on drawing aesthetic representation of graphs where all nodes are similar and therefore factors such as color, size, shape etc. are not considered for obtaining a visual balance. In this work, we modify and extend the spring-electric model by embedding color energy to obtain a visual balance which can have a wide range of applications in the field of computational media aesthetics.

We further apply this spring-electric model embedded with color energy in real-time photography assistance. In particular we focus on group photography where we have multiple people standing in an image frame with a scenic view in the background. In group photography, obtaining visual balance can be a challenging task as there are multiple parameters involved which affect the aesthetics quality of the captured image. Some of the factors include arrangement of people, their position and distance, i.e. how far they should stand from the camera. Professional photographers use their experience and knowledge to visualize how the visual elements in image frame could be better arranged, sized or positioned. However, it is not trivial for amateur users to estimate these parameters as there can be multiple possibilities. Fig. 4.1 shows some sample group photographs captured by amateur as well as professional photographers.

We use the spring-electric model along with color energy to generate real-time recommendation for users so that they can capture a visually balanced group photograph. The proposed method makes use of social media images to estimate an initial position, where a group of people should stand, and their relative size in the photograph. The estimated position and size of the people are further optimized and their arrangement is determined using a spring-electric model which enables visual balance in the image.

We make the following novel contributions in this work. We introduce the idea of color energy from art of composition and embed it in a spring-electric model to obtain a visual balance in a layout with dynamic visual elements. We present a novel application of this model in group photography where we leverage on social media images along with this model to produce real-time recommendation which can be used to capture high-quality group photographs. To the best of our knowledge, this is the first time the problem of group photography recommendation is being studied.

The rest of the chapter is organized as follows. In section 4.2 we will present an overview of the proposed method. Section 4.3 presents the concept of spring-electric graph model and the idea of embedding color energy in the graph. In section 4.4, we present the application of spring-electric graph model in group photography. The experimental results are presented in section 4.5. We have developed a mobile application for group photography recommendation which is discussed in section 4.6. Finally, we will conclude the paper in section 4.7.

4.2 Overview

This work focuses on two different problems. In the first problem we propose a spring-electric graph model embedded with color energy to solve the problem of visual balance in layouts with dynamic visual elements. Each visual element is assigned a color energy based on its color, size and surroundings which indicate its aesthetic impact on a viewer [196]. A graph is created with visual elements as its nodes which can be static as well as dynamic. An attractive as well as repulsive force act on the nodes which is computed using the color energy of the nodes. A energy term is defined for the graph which is based on the forces acting on the nodes. The energy of the graph is then minimized to obtain visual balance in the system.

In the second problem, we use this model to obtain a visual balance in a photograph with a group of people. Here the people are considered dynamic visual elements and the objects in the scene are treated as static visual elements. We leverage on social media images for estimating the initial position and size for the group of people which is further optimized using the proposed graph model. After obtaining visual balance, a recommendation is provided regarding arrangement, position, and size of people in the image frame so that a user can capture a high-quality photograph.

4.3 Spring-Electric Graph Model

Force-directed graph drawing algorithms are a well studied class of algorithms for drawing graphs in an aesthetically pleasing way [16, 45, 49, 80]. In this work we improve the spring-electric graph model proposed by Fruchterman *et al*. [49]. In this model spring-like attractive force (f^a) based on Hooke's law attract connected pairs of nodes towards each other, while a repulsive force (f^r) based on Coulomb's law separate all pairs of nodes. The forces are defined as,

$$f^a = d_{ij}^2/K, \tag{4.1}$$

$$f^r = -K^2/d_{ij}. \tag{4.2}$$

Here, d_{ij} is the euclidean distance between nodes i and j, and K is a constant.

When this system of forces is in equilibrium, the edges tend to have a uniform length (because of the spring forces), and nodes that are not connected tend to be drawn further apart (because of the electrical repulsion). Equilibrium is achieved in the system by using the attractive and repulsive forces either to simulate the motion of the nodes or to minimize the system energy.

Visual balance in a work of art depends on two major factors, visual weights, and visual direction [6]. The visual weight of an element depends on factors such as its size, position, color, texture, orientation, etc. and visual direction is the force exerted by the weights of neighboring elements. We represent the visual weights with color energy and the visual direction is induced in the system with the help of forces acting on nodes in the spring-electric model.

4.3.1 Color Energy

Color energy is defined as the relative aesthetic impact a color has on a viewer [196]. The energy of a color depends on (1) hue, saturation and brightness attributes of a color; (2) size of the colored area; and (3) relative contrast between foreground and the background colors. Table 4.1 shows how these factors affect the color energy of a visual element. We compute the color energy of a visual element as,

$$E^c = \frac{w_1 H + w_2 B + w_3 S + w_4 A + w_5 C}{\sum_{i=1}^{5} w_i}, \tag{4.3}$$

where w_1, w_2, w_3, w_4 and w_5 are mixing weights for relevant factors. H, B, S, A and C are warmness of hue, brightness, saturation, area and contrast respectively corresponding to the visual element and E^c the computed color energy. The warmness of hue is computed using hue wheel with red (0°/360°) as warmest and blue (180°) as coolest. Area of a visual element is normalized by the area of the largest visual element present in the layout. The contrast of the visual element is computed using Michelson formula [125].

$$C = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}, \tag{4.4}$$

where L_{max} and L_{min} are the maximum and minimum luminance values in the visual element and its adjacent visual elements. We also compute hue contrast using the Michelson formula and the average of hue and luminance contrast is utilized for color energy. Fig. 4.3 shows the variation of color energy with hue, saturation and value on hue wheel and HSV cylinder.

The effects of the color composition can be readily integrated with aesthetic elements if the colors are translated into color energies. One of the key principle in aesthetics says that the

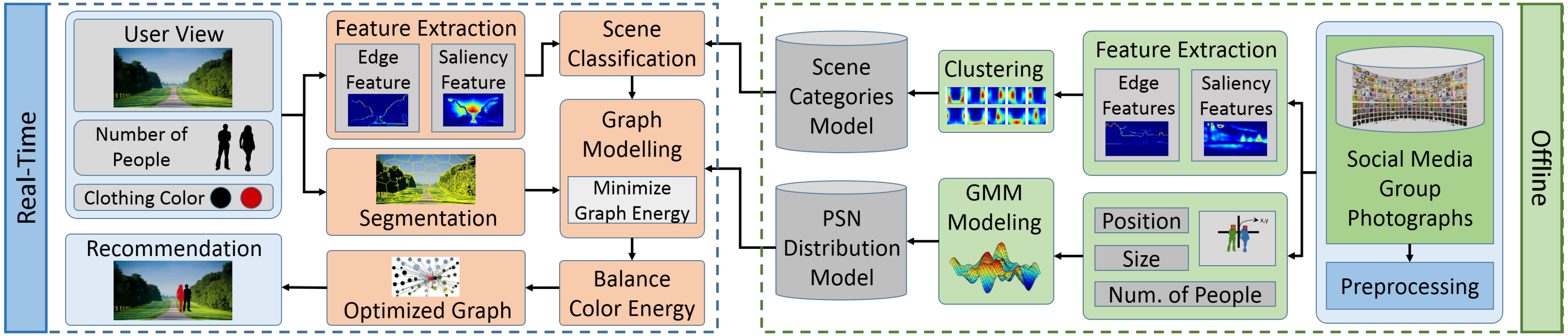


FIGURE 4.2: A detailed outline of the proposed socialized group photography method.

ALGORITHM 2: $MIN\_GE(G)$

```
Input: Graph G(V, E) with set of nodes (V) and set of edges (E).
Output: Updated graph G(V, E) where the nodes in the graph (V) are placed to minimize the total graph
        energy.
$C := 1.0$          // to determine relative strength of attractive and repulsive forces
$K1 := 1.0$                                   // to determine strength of spring force
$K2 := 0.01$                               // to determine strength of electrical force
$t := 0.5$                               // temperature to limit the node displacement
$\delta := 0.00001$            // threshold value as stopping criteria for optimization
$max\_iter := 100$, $num\_iter := 0$           // maximum number of iterations, step
$E^g := compute\_energy(G)$                                  // using equation 4.3
repeat
    // displacement due to repulsive forces acting on u
    for each node u in V do
        $u.disp := 0$                                    // displacement vector for u
        for each node v in V do
            if $u \neq v$ then
                $dist := u.pos - v.pos$                   // distance between u and v
                // $f^r$ computed using equation 4.2
                $u.disp := u.disp + (dist/|dist|) * f^r_{uv}$
            end
        end
    end
    // displacement due to attractive forces acting on nodes
    for each edge e in E do
        $dist := e.u.pos - e.v.pos$                 // each edge has two set of nodes
        // $f^a$ computed using equation 4.1
        $e.u.disp := e.u.disp - (dist/|dist|) * f^a_{uv}$
        $e.v.disp := e.v.disp + (dist/|dist|) * f^a_{uv}$
    end
    // update node positions
    for each node u in V do
        if u.fixed $\neq$ true then
            // limit the maximum displacement with temperature t
            $u.pos := u.pos + (u.disp/|u.disp|) * min(u.disp, t)$
        end
    end
    $update\_color\_energy(V)$
    $E^g_o := compute\_energy(G)$
    $\delta := |E^g - E^g_o|$, $E^g := E^g_o$
    $cool(t)$, $num\_iter := num\_iter + 1$
until $\delta \geq threshold$ and $num\_iter < max\_iter$
```

various color energies should be balanced in a composition and the areas of high-energy colors should be set-off against background areas of low-energy color [196]. Based on these principles we define the following objectives for obtaining a visual balance,

1. High color energy visual elements should not be close to each other.

Attribute	Variable	Color Energy
Hue	Warm	High
	Cold	Low
Brightness	High	High
	Low	Low
Saturation	High	High
	Low	Low
Area	Large	High
	Small	Low
Contrast	High	High
	Low	Low

TABLE 4.1: Aesthetic Energy of Colors [196].

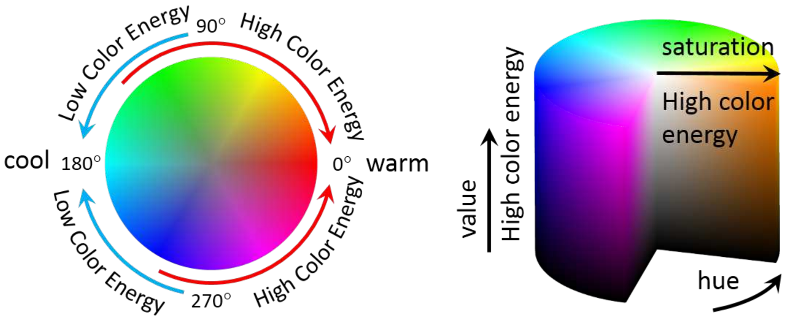


FIGURE 4.3: Variation of color energy with Hue, Saturation and Value shown on hue wheel and HSV cylinder.

2. High color energy visual elements should be close to low color energy elements.

3. The overall color energy should be balanced at the center of the layout.

To achieve the above-mentioned objectives we modify the spring-electric model by embedding color energy into graph nodes. Color energy is introduced to determine the magnitude of forces acting on the nodes. In the resultant system, the high energy nodes repel each other with a greater force and the nodes with a high energy attract the nodes with a low energy. The forces acting on nodes in this updated model are defined as,

$$f^a = d_{ij}^2 |E_i^c - E_j^c| / K, \tag{4.5}$$

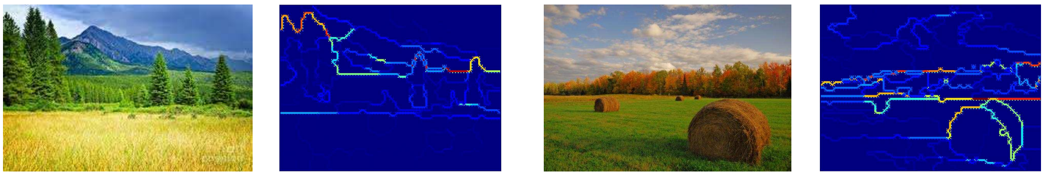


FIGURE 4.4: Visualization of extracted color based edge features.

$$f^r = -K^2 |E_i^c + E_j^c| / d_{ij}. \tag{4.6}$$

Here, E_i^c is the color energy of the i^{th} visual element. To balance the acting forces in this model energy minimization is performed which is discussed in the next section. After balancing the forces the overall color energy is balanced at the center of the layout. We will discuss this in detail when explaining the application in section 4.4.3.3.

4.3.2 Energy Minimization

The various forces acting on the nodes of the graph can be used to estimate its energy [49]. The energy of spring-electric graph model is defined as,

$$E^g = \sum_{i=1}^{N} f_i^2, \tag{4.7}$$

where, N is the total number of visual elements and,

$$f_i = C \sum_{j \neq i} f_{ij}^r \frac{(x_j - x_i)}{d_{ij}} + \sum_{i \leftrightarrow j} f_{ij}^a \frac{(x_j - x_i)}{d_{ij}}. \tag{4.8}$$

Here, C is a constant that determines the strength of the attractive and repulsive forces, f_{ij}^r and f_{ij}^a are the repulsive and attractive forces between nodes i and j, and x_i and x_j are the position of the nodes i and j in the layout. We minimize this energy function by the method proposed by [49]. In this iterative approach, each node of the graph is displaced according to the effective

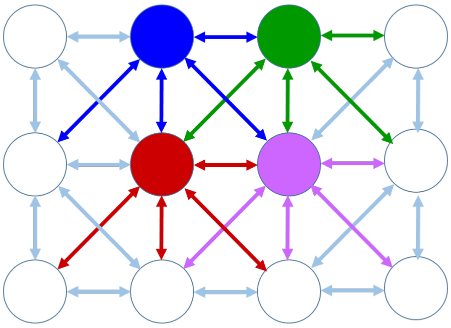
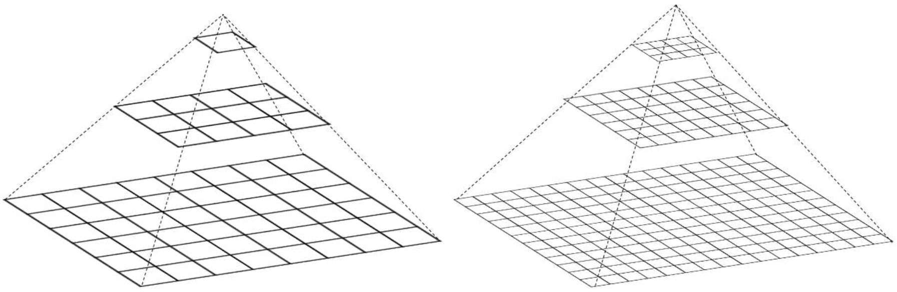


FIGURE 4.5: Association of edges with nodes in image graph.

force acting on it in each iteration. The details of the optimization process are presented in algorithm 2.

4.4 Group photography

In group photography, the main challenge for a photographer is to determine the arrangement, position, and size of the people standing in the image frame. This task can be easy for professional photographers but it can be very challenging for an amateur user to identify these parameters. In this work we propose the use of the principle of visual balance and estimate these parameters as a user is trying to capture a group photograph. We leverage on social media images to first estimate the position and size of people and then make use of an electric-graph model to find the arrangement and optimize the position and size of people in the image frame. We collected a dataset of high-quality group photographs from social media for our experiments (section 4.5.1). These images are used to develop a computation model which can predict an initial configuration for group photography. We first perform scene categorization to differentiate between different scene types and then a probabilistic model is trained for the position and size of people in different scene types. The next subsections discuss this in detail. The Fig. 4.2 shows the outline of the proposed method for generating group photography recommendation.



(a) Edge Features

(b) Saliency Features

FIGURE 4.6: Spatial pyramid layers for extracting edge and saliency features for structure learning.

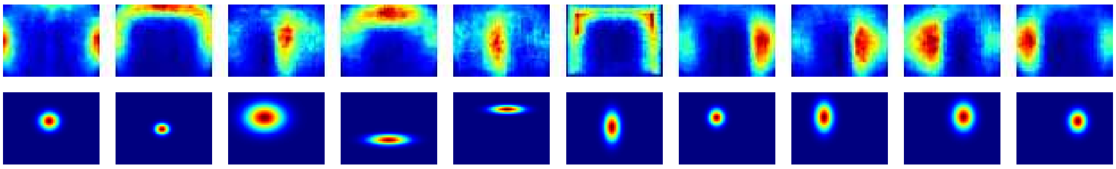


FIGURE 4.7: The first row shows the visualization of saliency map for identified scene categories. In the second row we have the visualization of distribution of position predicted for two people with average face within each category.

4.4.1 Scene Categorization

Scene structure plays an important role in finding the position where a person or group of people should stand in an image frame. We performed a scene categorization to distinguish between different types of scenes based on the scene structure. Scene structure is represented using a combination of edge-based and saliency-based features. These features are further utilized to perform clustering and identify a set of scene categories.

We propose a graph-based approach for edge detection to extract the color based edges in an image. We first perform segmentation on the image to find the superpixels using SLIC approach [2]. The superpixels are then utilized to create a graph representation of the image and weighted edges are added to the graph for neighboring superpixels based on the color similarity. If all the superpixels are utilized for graph creation then different images will have a varying graph structure. Therefore we follow a grid approach which allows us to represent different images with a consistent graph structure.

Each image is divided into a grid with size $N \times N$ and each cell in the grid is assigned the superpixel with highest number of pixel contribution. Now each cell in the grid is used to form a node in the graph and an edge is added between neighboring cells with color similarity as edge weight. For each node in the graph we use eight cell connectivity and, the edge weights are utilized to extract a feature representation for scene structure. To remove redundancy of edges from this representation each edge is only considered once and thus eventually we have four edges corresponding to each node in the graph.

We employ spatial pyramid approach [93] on this graph model to form a feature descriptor for representing scene structure. We consider three levels in spatial pyramid with $(1 \times 1), (3 \times 4)$ and (6×8) blocks at each level of hierarchy. For each block we aggregate the edge weights for the cells in that block and normalize by number of cells in corresponding block. The features extracted from each level are used to form a feature descriptor as defined by [93].

Apart from the edge information, saliency map of an image is also utilized for scene categorization. We employ [3] for saliency map detection of an image and spatial pyramid approach similar to edge features but with different grid size is used for feature extraction. In the spatial pyramid we consider three levels with (3×4), (6×8) and (12×16) blocks at each level of hierarchy.

We use the face detection approach proposed in [170] to detect faces in the image. Presence of people in photographs can change the scene structure and therefore in order to represent the scene structure for images with people we use the method proposed by [203] to modify the saliency map. Similarly, we also modify the edge graph of an image by removing all the edges in the region where people are present before performing feature extraction.

Finally, we get a 244-dimensional feature descriptor for edges and a 252-dimensional feature descriptor for saliency. The edge and saliency feature descriptor are then combined together to represent the scene structure of an image. We employ k-means algorithm for clustering to identify different scene categories. Figure 4.7 shows the mean saliency map for identified categories with 10 clusters.

4.4.2 Position and Size Modeling

The position where a group of people should stand in the image frame and the size of standing people relative to the image frame impacts the aesthetics of an image. Estimating the position and corresponding size for given scene and the number of people in the group can be challenging for amateur users. We leverage on the crowd-sourced images to estimate these parameters using a generative probabilistic distribution model. The position where a group of people should stand in an image frame will vary with the scene structure and therefore we perform separate position modeling for different scene categories.

For each image in our dataset, we perform face detection to find the faces in the image. Then we compute a mean position in image frame based on all the detected faces which represent the position of the people. The mean position is computed as a weighted mean of the center of detected faces where the weights are the size of the detected faces. We also compute average face size for each of the image using the size of detected faces. We build a generative probabilistic model to train a distribution of position, size and the number of people for each of the identified scene category. We employ Gaussian Mixture Model (GMM) for learning the probabilistic distribution. For any scene category I, we denote $\mathbf{x}(I) = (x,y,s,n)^T$, where (x,y) represents the mean position of people in the image, s represents the mean size of faces and n denotes the number of people present in the photograph. Therefore, the probabilistic distribution of $\mathbf{x}(I_k)$ for a given scene category I_k can be expressed as,

$$p(\mathbf{x}(I_k)) = \sum_{i=1}^{N_k} w_i^k \mathcal{N}(\mathbf{x}|\mu_i^k, \Sigma_i^k), \tag{4.9}$$

where, w denotes the prior, $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ denotes a Gaussian component and N_k is the number of Gaussian components in the mixture. We use Bayesian information criterion [145] to estimate the number of Gaussian mixture components. The parameters (w^k, μ^k, Σ^k) of Gaussian mixture model are estimated for all the identified scene categories (I_k) using expectation-maximization (EM) algorithm [40]. The EM algorithm maximizes the log-likelihood of class probability for each configuration associated with a scene category. In the E-step we compute the expected class probability for each of the configuration ($\mathbf{x}_t$) corresponding to a scene category I^k:

$$P(i|\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t|\mu_i^k, \Sigma_i^k) w_i^k}{\sum_{l=1}^{N_k} \mathcal{N}(\mathbf{x}_t|\mu_l^k, \Sigma_l^k) w_l^k}. \tag{4.10}$$

In M-step the latent variables, mean ($\hat{\mu}$), covariance ($\hat{\Sigma}$) and priors for each class ($\hat{w}$) are updated as follows:

$$\hat{\mu}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}, \tag{4.11}$$

$$\hat{\Sigma}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)[\mathbf{x}_t - \hat{\mu}_i^k][\mathbf{x}_t - \hat{\mu}_i^k]^T}{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}, \tag{4.12}$$

$$\hat{w}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}{T_k}, \tag{4.13}$$

where, T_k is the number of photographs for a given scene category I_k. We term this distribution model as PSN (position, size and number of people) and it is used to predict the initial position and size of a group of people based on the scene category and the number of people present in the image.

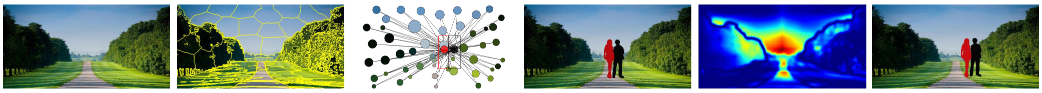


FIGURE 4.8: A sample example of recommendation. a) Input image, b) segmented image, c) spring-electric model, d) recommendation using color energy, e) saliency map, and f) recommendation using both color energy and saliency.

4.4.3 Real-Time Recommendation

In the first step, we categorize the image view as one of the identified scene categories. Edge and saliency features are extracted for the view as described in section 4.4.1. We employ the Nearest Neighbor approach for category identification using the extracted features. Thereafter, the user is asked to provide the number of people in the photograph and a position in the image frame (x,y) and the size of people is predicted using the trained PSN model based on maximum posterior probability. The recommended position and size are estimated based on learning from social media images which are further improved with spring-electric graph model. A sample example is shown in fig. 4.8.

4.4.3.1 Graph Modeling

In the first step, the user view is represented as a graph where the nodes represent the visual objects in the image. We employ the method proposed by Achanta et al. [2] for fine grain segmentation. The identified superpixels are further merged based on color similarity to obtain visual objects in the view and small sized visual objects (<1% of image size) are discarded. The nodes corresponding to people are also added to this graph. Their initial position and size are determined using the PSN model. We term the nodes corresponding to people as p-nodes and nodes corresponding to image segments as s-nodes for further discussion. In a given view, the position of objects in the image will be fixed. However, people standing in the frame can change their position. Therefore p-nodes are kept fixed and s-nodes are dynamic. Edges are created between p-nodes and s-nodes in the graph, however, the p-nodes are not connected with each other so that they can move independently in the layout. We will see later in the next section (4.4.3.2) how this is useful for estimating the arrangement of people in the photograph.

As described in section 4.3.1, color energy is computed for each node in the graph. The dress color of the people is used for computing color energy for p-nodes. For dresses with multiple colors, top 5 dominant colors are used and 5 (or less) p-nodes are created for each person depending on the number of colors detected. During energy optimization, a collective force is computed for the p-nodes corresponding to a person based on individual forces. The contrast value of a node is computed with all the adjacent nodes in the image frame and then an average value of contrast is assigned to the node.

In a photograph standing in front of any salient visual element may block the object and is not recommended. Color energy is not an alternative for saliency and we came across many such cases in our experiments. To take this into account we compute saliency of each visual element using the method proposed by [3]. Also, the p-nodes in the graph are assigned the maximum saliency value computed for visual elements in the view. The repulsive force in the graph model is updated as follows,

$$f^r = -K^2|E_i^c + E_j^c||S_i^s + S_j^s|/d_{ij}. \tag{4.14}$$

Here, S_i^s is the saliency value of i^{th} node in the graph. This repulsive force between nodes ensures that people are not standing in front of the salient objects.

4.4.3.2 Formation Estimation

The recommendation is generated in three steps. In the first step, the group formation of people is estimated. In the second step, a position is determined and finally, in the third step, the size

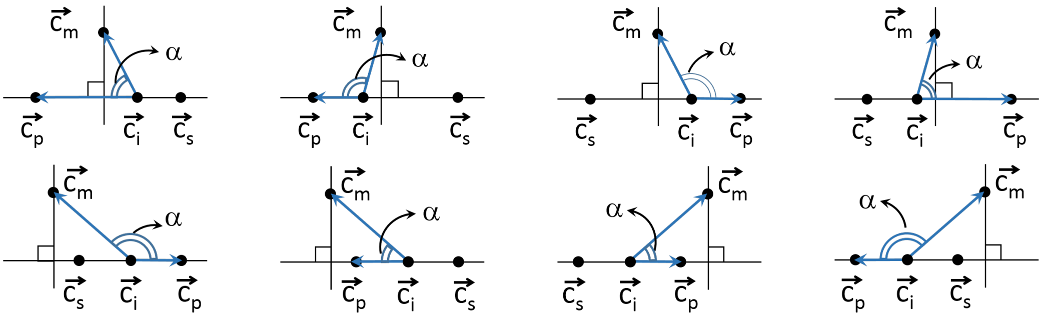


FIGURE 4.9: Possible configurations of $\vec{c}_m, \vec{c}_i, \vec{c}_s$ and $\vec{c}_p$ on an image frame for obtaining visual balance.

of people is optimized. The details of the complete process are described in algorithm 3. The p-nodes are initially placed at the position predicted using the PSN model. Energy minimization is performed on this initial graph configuration and an unorganized position of p-nodes is obtained. This intermediate graph configuration is then used to obtain a group formation of people. There can be multiple ways for a group of people to stand which we term as group formation. The study of various group formations will be the focus of our future work where recommendation regarding group formation will also be included. In this work, we consider only the horizontal linear formation where the people are standing side by side in a line.

A mean position is computed using the intermediate positions of p-nodes and these nodes are moved towards this position in small steps along the vertical axis. The motion in the horizontal axis is determined by the spring and electrical forces on the nodes. This step is iterated until all the p-nodes are in the same horizontal position to form a linear group formation. The motivation behind performing this in a step-wise iterative process is to maintain the visual balance between p-nodes.

4.4.3.3 Position and Size Optimization

The formation estimation may lead to increase in graph energy and therefore we need to further optimize the position of people. After formation estimation, all the p-nodes are considered as a

ALGORITHM 3: $FPS\_EST(G)$

```
Input: Graph G(V, E) with set of nodes (V), which includes p-nodes corresponding to people, and s-nodes
       corresponding to image segments, and set of edges (E) connecting p-nodes with s-nodes.
Output: Updated graph G(V, E) where the position and size of p-nodes indicates the arrangement,
        position and size for recommendation.
$t := 0.005$                                   // temperature to limit the node displacement
Step 1: Formation Estimation
$MIN\_GE(G)$                                   // Algorithm 1:  optimize the graph energy
$mean\_pos := find\_mean\_position()$          // mean position for p-nodes
$step\_size := find\_step\_size()$             // find step size for each p-node
repeat
    // pos.x for horizontal axis and pos.y for vertical axis
    for each node u in p-nodes do
        $u.pos.y := u.pos.y + step\_size(u)$
        // u.disp is computed as described in algorithm 1
        $u.pos.x := u.pos.x + (u.disp/|u.disp|) * min(u.disp, t)$
    end
    $update\_color\_energy(V)$
until all p-nodes are in line formation
$make\_dist\_equal()$                          // make p-nodes equidistant from each other
Step 2: Position Estimation
// First the p-nodes are combined together for uniform displacement
$combine\_pnodes()$                            // p-nodes are combined together
$MIN\_GE(G)$                                   // Algorithm 1:  optimize the graph energy
Step 3: Size Estimation
$\vec{c}_m := (0.5, 0.5)$                      // center of image frame
// compute $\vec{c}_p, \vec{c}_s$ and $\vec{c}_i$ position using color-energy as weights
$\vec{c}_p := compute\_weighted\_mean(p\text{-}nodes)$    // people nodes
$\vec{c}_s := compute\_weighted\_mean(s\text{-}nodes)$    // image segments
$\vec{c}_i := compute\_weighted\_mean(V)$                 // all the nodes
$enlarge\_face := true$
// compute the angle $\alpha$ as defined in figure 4.9
$\alpha := angle(\vec{c}_m - \vec{c}_i, \vec{c}_p - \vec{c}_i)$
if $\alpha > 90^o$ then
    $enlarge\_face := false$
end
$\delta := 0, flag := true$                    // face size should be further increase/decrease
$Cost^{ce} := compute\_cost()$                 // using equation 4.15
repeat
    $flag := modify\_face\_size(enlarge\_face)$
    $update\_color\_energy(V)$
    $Cost_o^{ce} := compute\_cost()$           // using equation 4.15
    $\delta := Cost^{ce} - Cost_o^{ce}, Cost^{ce} := Cost_o^{ce}$
until $\delta \geq 0$ and $flag == true$
```

single node and the forces on this node is determined by aggregating the forces acting on all the p-nodes. With this constraint the energy minimization algorithm described in section 4.3.2 is again applied on the graph model.

The overall color energy of the system is balanced at the center of image frame to optimize the size of people. As size also plays a role in color energy, change in size will also change the color energy. For balancing the color energy of the system at the center we minimize the following cost function,

$$Cost^{ce} = ||\vec{c}_m - \vec{c}_i||, \tag{4.15}$$

where $\vec{c}_m$ denotes the center position of the image frame and $\vec{c}_i$ is the current position on image frame where the color energy (both the p-nodes and s-nodes) is centered. It is defined as,

$$\vec{c}_i = \frac{\sum_{i=1}^{N} E_i^c \vec{x}_i}{\sum_{i=1}^{N} E_i^c}. \tag{4.16}$$

Here, N is the total number of nodes in the graph, E_i^c is the color energy of i^{th} node, and $\vec{x}_i$ is the position of the i^{th} node in the frame. We also define $\vec{c}_p$ and $\vec{c}_s$ which indicates the center of color energy for p-nodes and s-nodes respectively. Figure 4.9 shows different possible configurations for $\vec{c}_m$, $\vec{c}_i$, $\vec{c}_p$ and $\vec{c}_s$ on an image frame. The size of p-nodes is either increased or decreased in small steps to minimize the cost function defined in equation 4.15. Since changing the size of a node also changes its color energy, increasing or decreasing face sizes will move the center of color energy ($\vec{c}_i$) towards the center of image frame ($\vec{c}_m$).

To determine whether the face size should be increased or decreased we compute angle α between vectors $(\vec{c}_m - \vec{c}_i)$ and $(\vec{c}_p - \vec{c}_i)$. It can be observed from figure 4.9 that if the angle $\alpha > 90^o$ (b, c, e & h) then reducing the color energy of p-nodes will move $\vec{c}_i$ towards $\vec{c}_m$ which will reduce the cost $Cost^{ce}$ and therefore the face size should be decreased. And, if $\alpha < 90^o$ (a, d, f and g), then increasing the color energy of p-nodes will move $\vec{c}_i$ towards $\vec{c}_m$ and therefore face sizes should be increased. It can also be observed from fig. 4.9 that increasing or decreasing the size of p-nodes will keep on reducing the cost $Cost^{ce}$ for configurations (e-f). We set an upper as well as a lower limit on the face size based on social media images to resolve this problem. A lower and upper limit on face size is determined for each of the scene category based on the number of people and minimum and maximum average face sizes in each category. However, it is important to note that the configurations from e-f were not observed in our experiments because the image is already visually balanced using spring-electric graph model before optimizing the face size.

4.5 Experiments and Results

Experiments were performed to evaluate three things: 1) Visual balance obtained by spring-electric model, 2) Recommendation provided for group photography regarding the arrangement, position and size of people, and 3) Real-time performance analysis of the proposed method.

4.5.1 Dataset

We used *Flickr* images to build our dataset for performing evaluation experiments. We utilized Flickr's *photos.search* API and used image-tags to find images related to group photography. We used tags such as 'group photography', 'family portrait', and 'group portrait' for retrieving relevant images. Using these tags we gathered around 24K images in order of *interestingness* score and then performed a post processing to clean the dataset.

For dataset cleaning, we performed face detection and removed the images with no face or one face from the dataset. Also, we ignore images with no scenic view present in the background. For this, the images in which people cover most of the image frame are ignored. We keep images in which the area covered by people is less than 30%. The method proposed by Wang *et al*. [174]

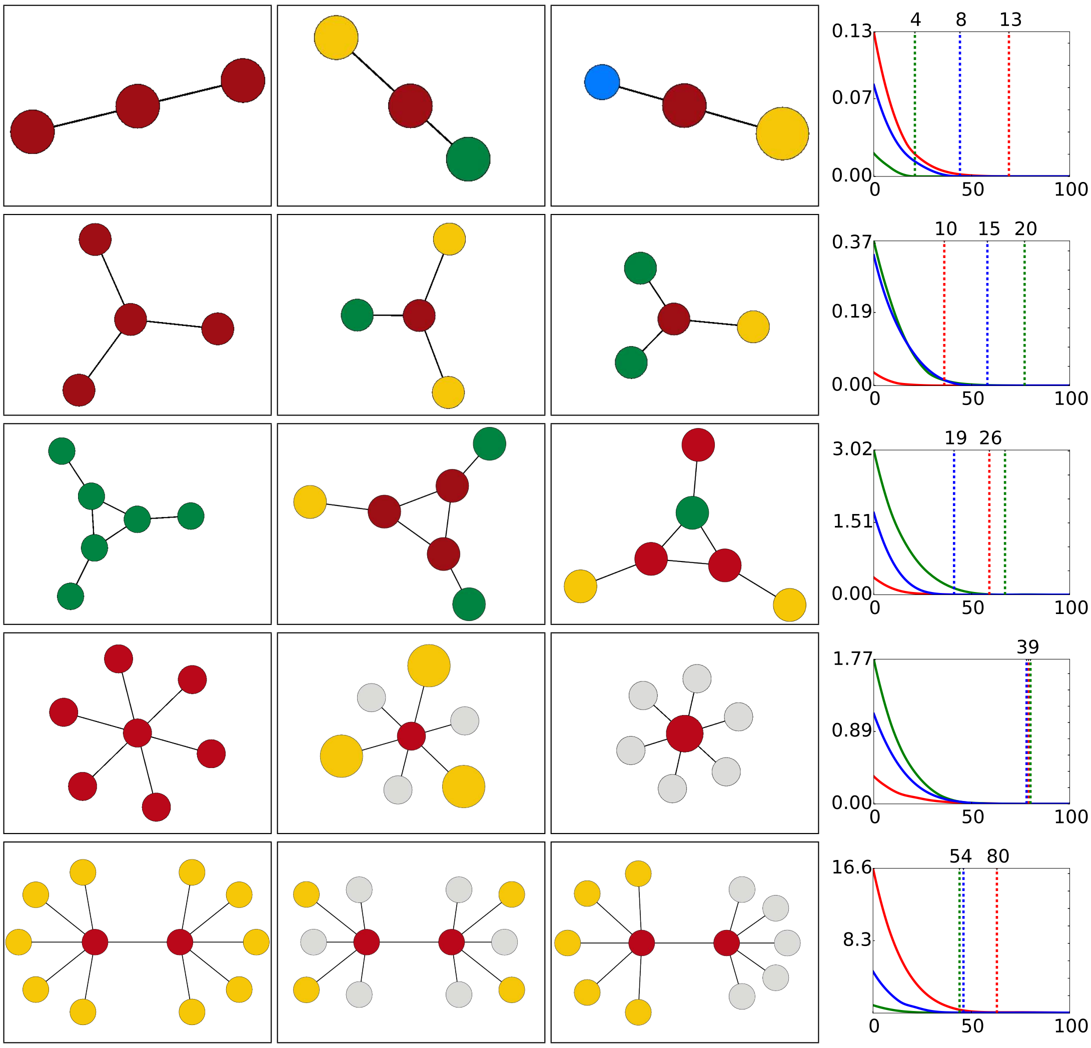


FIGURE 4.10: Visual balance results obtained for different graphs using the proposed spring-electric graph model. From left to right (row 1-3), we can observe how the layout configuration for similar graphs changes with variation in color and size of nodes. The last column shows the plot of energy as final configuration is obtained using algorithm 2. The y-axis is the graph energy, the bottom x-axis is number of iterations and the top x-axis is time in milliseconds in which the layout was obtained. (— first column, — second column and — third column.)

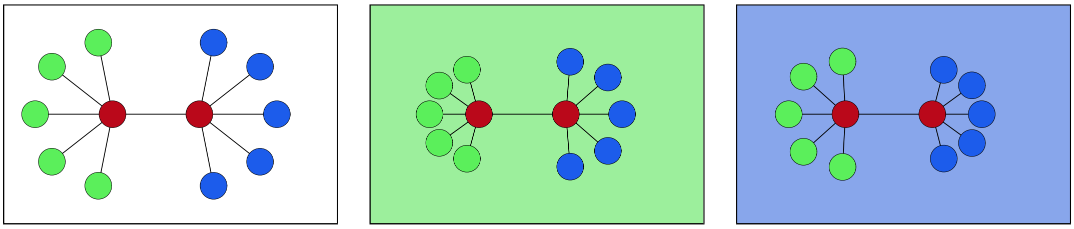


FIGURE 4.11: Effect of contrast on visual balance. The blue and green nodes have similar color energy and therefore the first layout is symmetrical. However, changing the background color changes the color energies of these nodes leading to a different layout.

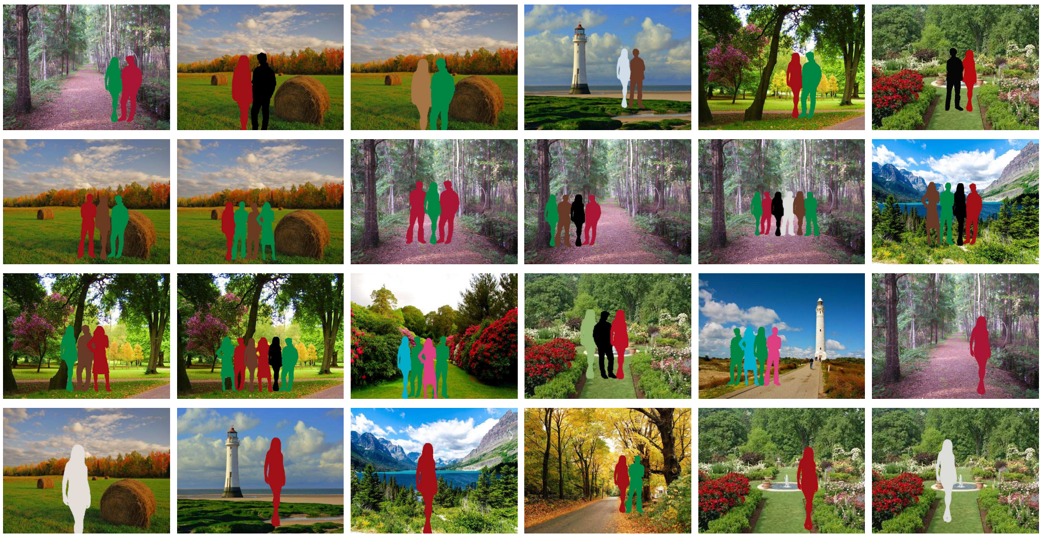
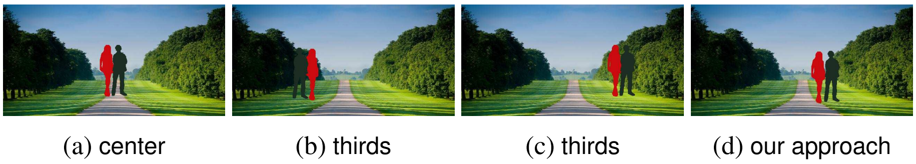


FIGURE 4.12: Recommended results. A colored silhouette is rendered at the recommended position with corresponding size and clothing color.

is utilized to compute the area covered by people in the image frame. After this post processing, we get a dataset of 5941 images and it is used for performing our experiments.

4.5.2 Visual Balance

We use the proposed spring-electric graph model to obtain visual balance in drawing graphs with colored nodes. Figure 4.10 shows some of the results for graphs with varying structure, node color and node size. The initial position of all the nodes in the graph are randomly generated. The positions of the graph nodes are updated iteratively using algorithm 2 and, the layouts presented in figure 4.10 are obtained after minimizing the graph energy. We can observe that the layouts are symmetrical and, changing the node color and size affect the layout. Increasing the color energy of a node by changing its color and size increases its distance from a high color energy node



(a) center (b) thirds (c) thirds (d) our approach

FIGURE 4.13: Comparison with known rules of photography.

and decreasing the color energy attracts it towards a high color energy node. This is consistent with the objectives we defined for obtaining visual balance in a layout. The last row in figure 4.10 shows the variation of graph energy with each iteration along with the time required for obtaining visual balance in each of the graphs. The initial graph energy is different for graphs with similar structure because the position of graph nodes are randomly initialized. It can also be observed that time required for obtaining visual balance increases as we increase the number of nodes in the graph.

Figure 4.11 shows the effect of contrast on visual balance. The first layout shown in figure 4.11 is symmetrical as the green and blue nodes have similar color energy. Changing the background color changes the contrast and therefore the color energy of nodes. In the second layout, we can observe that the green background reduces the color energy of green colored nodes and therefore their position is closer to the red colored node. A similar change in the layout can be observed for blue colored nodes in the third figure where we have a blue colored background.

4.5.3 Group Photography

For qualitative evaluation, we predict arrangement, size, and position recommendation for randomly generated number of people and clothing color for a set of scenic images which are not in our dataset. Fig 4.12 shows some of the results we obtained using our proposed system. We can see how the recommendation changes to achieve visual balance as the clothing color is changed (Fig. 4.12, last row, column 5 and 6). In the case of the arrangement of people, it can be observed that persons with high energy clothing are never adjacent to each other which makes them distinctly visible (Fig. 4.12, row 2, column 2, 3, and 5). Also, in fig. 4.8 we can observe how the recommendation position changes when we incorporate saliency of visual elements into consideration.

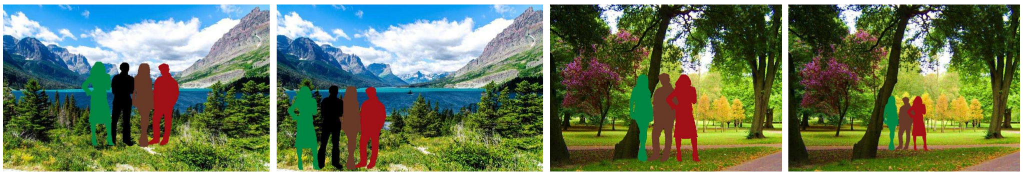


FIGURE 4.14: Recommendation results with user disagreement. Our recommendations (First: 0.36% and third: 0.2% user agreement) compared with second and fourth image with different position and size.

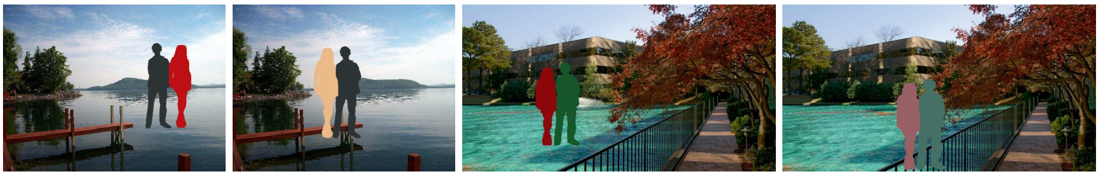


FIGURE 4.15: Images with non-symmetrical view and limitations of visual balance.

We also compare our approach with known rules of photography. Fig. 4.13 shows a sample example where our approach performs better than existing photography rules. We can observe that when rule-of-center is applied the vanishing point is obstructed and using rule-of-thirds one of the people is not visible due to a low contrast. However, with our approach, the group is placed at a position such that both of the above-mentioned issues are avoided.

4.5.4 User Study

To further evaluate the recommendation, we conducted a user study in which 25 users participated including 20 amateur users and 5 skilled photographers. Users with at least 5 years of experience in the single-lens reflex camera were considered skilled photographers. The average age of the amateur users was around 30 years with a range of 21-58 years and for the skilled users, the average age was around 31 years with a range of 27-34 years.

In the survey, there were four set of questions to evaluate different types of recommendation generated by our proposed method. We evaluated our method for the position, arrangement, size and overall aesthetics. For each question, a pair of images with a similar view with different recommendations were shown to the user. In one of the option recommendations were generated

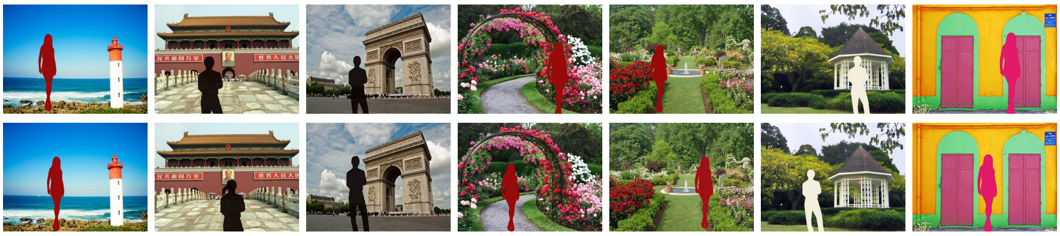


FIGURE 4.16: Comparison of recommendations for single person photography with Wang et al [174]. The first row shows results obtained using [174] and the second row shows the results obtained with our proposed method.

using our approach and for the other option, our proposed recommendations were modified to generate a different recommendation. For the arrangement of people, a random arrangement is generated and in the case of position a displacement vector (dx, dy) is randomly generated to move the position of people. Here, dx is the displacement in the horizontal direction with (0.1*w <dx <0.2*w) and dy is the displacement in the vertical direction with (0 <dy <0.1*h), where w and h are the width and height of the image frame respectively. Similarly, for the size of people, it was either increased or decreased by a random amount within the range (10%-20%) of the recommended size.

The users were asked the following question, '*Which of two positions do you think is better from aesthetics perspective?*'. For each type, 10 set of questions were presented to the user. The options in a question and the questions itself were randomly placed in the survey.For each set of questions the users were asked to consider the following factors,

1. Clothing color and position of the person/group.
2. Clothing color and size of the person/group.
3. Clothing color and arrangement of the group.
4. Clothing color, size, position and arrangement of the person/group.

The overall percentage of the user agreement with our proposed recommendation was 75.6%.

This user agreement comprised of 72.6% agreement with amateur users and 78.5% with skilled users. Figure 4.17 shows the distribution of the percentage of the user agreement with our proposed recommendation for the different type of recommendations. It can be observed that agreement percentage is higher with skilled users as compared to amateur users. We also observed that the user agreement level for position and size recommendation is lower as compared to the arrangement and overall aesthetics. Fig 4.14 shows some of the cases when users were in disagreement with our recommendation. We can see in Fig. 4.14 (first image) that the image is symmetrical and the group is placed in the center which hides the lake view. This makes the second image (Fig. 4.14) more attractive as the lake is visible if we move the group to another position. The reason for this failed case can be attributed to saliency detection as the lake was not marked as salient. However, when the arrangement of people was also considered in aesthetics evaluation the user agreement level was improved.

4.5.5 Comparison

To the best of our knowledge, this is the first attempt to study the problem of group photography recommendation. The proposed method can also be used to generate a recommendation for single person photography. To validate the effectiveness of our proposed method we compare our results with one of the state-of-the-arts work in single person photography [174]. The initial position and size are estimated using the PSN model for 2 people. For many cases, our proposed method generated almost similar results as produced by Wang et al. [174]. Some of the sample results for comparison are shown in fig. 4.16. However, the method proposed by Wang et al. does not consider the color of objects and the clothing color of people for generating the recommendations. Therefore, we observed different recommendation results generated using their approach and our proposed method for sceneries where color plays an important role in

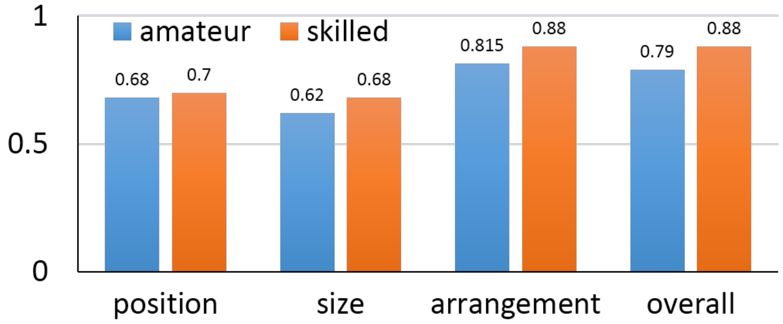


FIGURE 4.17: User study results. The bar graph shows percentage of user agreement with our proposed recommendation.

the placement of people. The difference in generated recommendations is shown in fig. 4.16 (column 4-7). We can observe that the recommendations generated using [174] may be good for certain clothing colors but not good for the color shown in the image as the person is not distinctly visible due to low contrast. On the other hand, since our approach also considers the clothing color as well as the color of objects in the scene, a better position is recommended where the person is distinctly visible.

4.5.6 Limitations

The position recommendation is generated only based on visual balance and therefore there are some limitations of the proposed approach. For non-symmetrical views when a visual balance is achieved, it might not always be feasible for the group to stand in the recommended position (Fig. 4.15). In the first and the third image high energy colors are used and we can observe how the recommendation changes when we have low color energy clothes (second and fourth image). This limitation can be overcome either by considering scene semantics or taking feedback from the user in an interactive way.

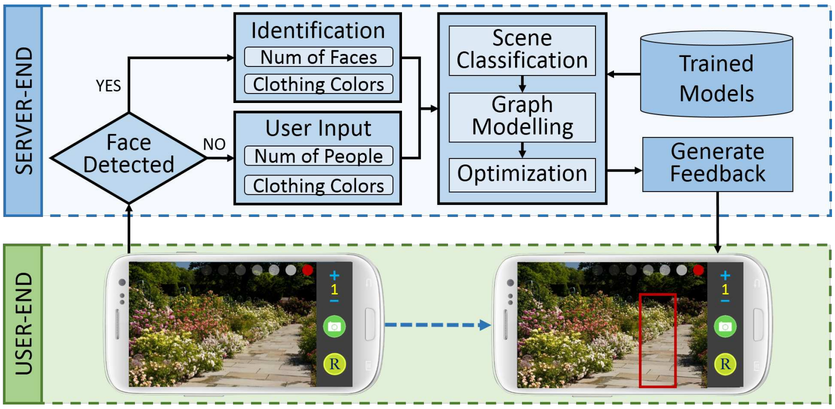


FIGURE 4.18: System overview of the developed mobile application for proposed group photography recommendation method.

4.5.7 Computation Time Analysis

The algorithm used for optimizing graph energy has a time complexity of $O||V|^2 + |E||$ for each iteration, where $|V|$ is the total number of nodes and $|E|$ is the total number of edges in the graph. Therefore, the running-time of the optimization algorithm will increase as we increase the number of nodes in the graph. Figure 4.10 (fourth row) shows the plot of energy minimization with the number of iterations and running-time. We can observe that the running-time increases from around 4-12 milliseconds to 50-80 milliseconds as we increase the number of nodes in the graph from 3 to 12. The experiments were conducted on a desktop machine with 8 GB of RAM using an unoptimized python code.

The graph constructed for group photography recommendation for a given image has around 40-60 nodes and it takes around 500 ms (mean for 100 images) for optimizing graph energy using algorithm 3. We also tested the running-time of algorithm 3 on a graph with around 100 nodes by over-segmenting the image. It took around 800 ms (average for 50 images) to optimize the graph energy. The average time to process a 640×480 pixel image for generating the position, size, and arrangement recommendation is around 1.5 sec. Rest of the processing time is spent on image segmentation and saliency map detection. Image segmentation and saliency map

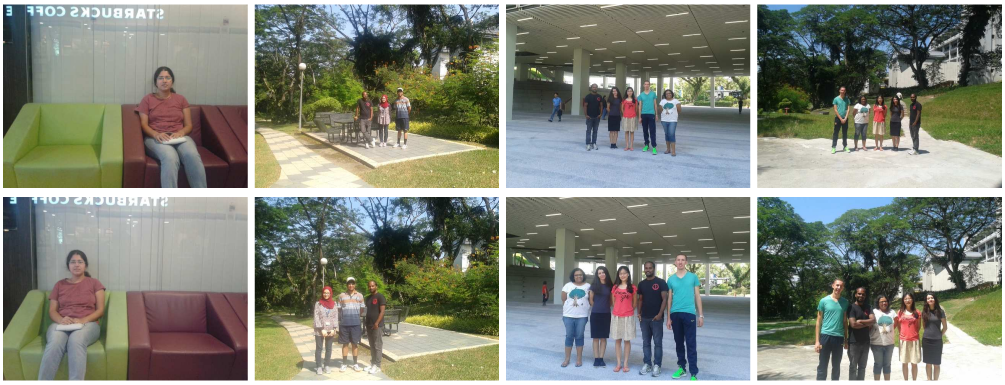


FIGURE 4.19: Some sample photographs captured using the developed mobile application. The first row shows the photographs captured without any recommendation and the second rows shows photographs captured using the recommendation.

detection can be performed in parallel to improve the runtime. Also, the proposed method does not impose any restriction on the size of the input image and therefore the runtime can be further optimized by reducing the image resolution.

4.6 Mobile Application

We have developed a mobile application in Android platform for the proposed group photography recommendation. The developed application is a cloud-based system in which the user view and user input regarding the number of people and clothing preference is sent over a network to the cloud server. If faces are detected in the user view, the number of people and their clothing color is identified from the image, otherwise, the user-input is considered. A recommendation regarding the arrangement of people, their position, and relative size is generated on the server and a feedback is sent to the user. A system overview of the mobile application is presented in fig. 4.18. Some sample photographs captured using the developed mobile application are shown in fig. 4.19. We can observe that the photographs captured using the recommendation are visually balanced and of better quality.

4.7 Summary

In this work, we proposed a novel method to obtain visual balance in a layout with dynamic visual elements. We extended the idea of spring-electric graph model and augmented it with the concept of color energy. We mainly focused on obtaining a visual balance in an image frame and providing real-time assistance to users for capturing high-quality group photographs. We leveraged on social media images and make use of proposed spring-electric model to provide a recommendation to the user for capturing visually balanced photographs. Experimental evaluations showed that the proposed method can provide effective real-time feedback to the user regarding the arrangement of people, their position on image frame and relative size. The concept of spring-electric graph modeling can be further explored by bringing in more aesthetic principles such as shape, texture, etc. in the model and can be used for a variety of other applications in computational media aesthetics.

Chapter 5

Context-Aware Viewpoint Recommendation

In this chapter, we will discuss a novel viewpoint recommendation system which can assist a user in capturing high-quality photographs at well-known tourist locations. The proposed system, *ClickSmart*, can provide real-time viewpoint recommendation based on the preview on user's camera, current time and user's geo-location. It makes use of publicly available geo-tagged images along with associated metadata for learning a recommendation model. We define view-cells, macro blocks in geo-space, and propose the idea of popularity, quality and uniqueness of view-cells from viewpoint perspective. Viewpoint recommendation is generated at the granularity of a view-cell and is based on its popularity, quality and uniqueness, which are estimated using social media cues associated with images. We further observe that contextual information such as time and weather conditions play an important role in photography, and therefore augment the recommendation system with associated context. *ClickSmart* also takes into account presence of people in the view for making the recommendation. It can provide two kinds of recommendations, quality-based and uniqueness-based. Although, both were found effective in the experimental evaluation, user study showed that uniqueness based recommendation was preferred more by

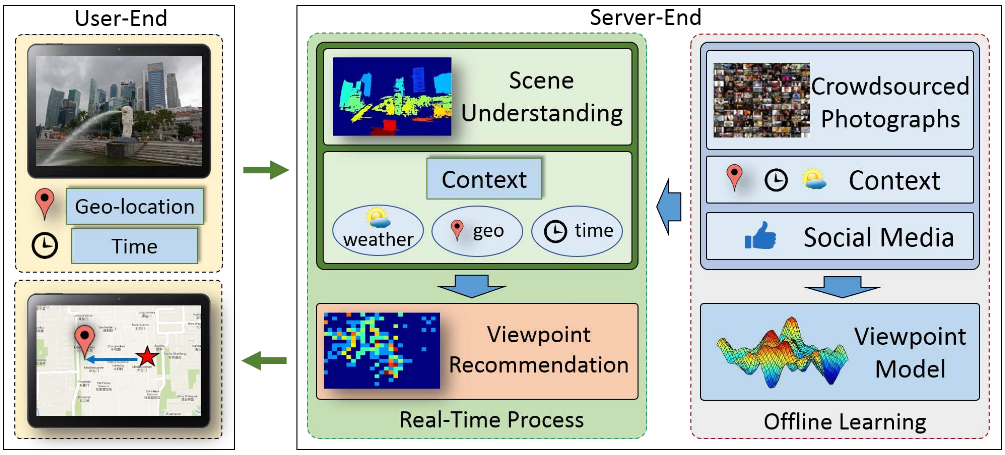


FIGURE 5.1: Overview of the framework for *ClickSmart*

skilled photographers as compared to amateurs. The work presented in this chapter has been published in [140].

5.1 Introduction

In photography, viewpoint refers to the geo-location from where a photograph is captured and is considered as one of the essential factors in the art of photography [47]. It has a large impact on the composition of a photograph and as a result, it also affects the aesthetic quality of a captured image. We also observe that context (time and weather conditions) also play an important role in viewpoint selection for landmark photography. For example, it is difficult to get a good quality image when the camera lens is facing the sun. Now, as the sun moves during the day, viewpoints for photography will also change with time for a view at a given location (Fig. 5.2). Similarly, weather conditions also affect viewpoint as factors like visibility, clouds, etc., have an impact on lighting conditions, which is known to play an important role in photography [83].

In the last decade, we have seen an increasing trend in people's photo taking and sharing behavior. There are many social media services such as *Flickr, Panoramio* and *Photo.net*, with a large collection of photos shared by professional and other users. These photos have *Exif* data, which

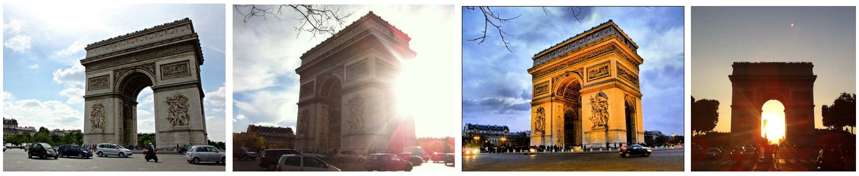


FIGURE 5.2: Photographs of same monument with change in viewpoint and time.

provide us context information like, time of capture and geo-location of the captured image. Using these details we can infer the photo-taking behavior of people for popular tourist locations. Also, the shared photos are augmented with social media cues such as the number of user views, likes, and comments. In this work, we integrate the photo-taking behavior with social media cues to develop a recommendation system which can provide real-time viewpoint recommendation to the user for taking better photos. The proposed method, which we call *ClickSmart*, is focused on landmark photography and aims to provide real-time guidance to the user before an image is captured. Fig. 5.1 presents an overview of the proposed framework.

In recent years, researchers have shown interest in the field of photography assistance. They have proposed methods which provide assistance to the user for capturing high quality images using publicly available photographs [111, 128, 138, 139, 156, 181, 190]. The proposed methods focus on improving the image composition based on the preview on user's camera. However, the recommendation provided by these methods assumes that the user is already standing in a good viewpoint. Besides view-based photography assistance, there are methods which provide location recommendation using publicly available photographs. These methods are mainly focused on studying the photo-taking behavior of people at well-known tourist locations and providing photo-shooting location recommendation to the user [149, 204] and [132]. However, the recommendation provided by these methods is not generated at the time of capture and is static irrespective of the view a user is trying to capture.

In this work, we attempt to bridge the gap between view-based and location-based recommendation. *ClickSmart* provides viewpoint recommendation based on the preview on user's camera. The rest of the chapter is organized as follows. We begin by discussing the technical details of the proposed offline learning in section 5.2. We then delve into the real-time recommendation phase in section 5.3. Detailed experimental evaluation and results are presented in section 5.4. Finally, section 5.5 concludes with a summary of our work and a discussion on possible future work.

5.2 Offline Learning

The framework of *ClickSmart* consists of two phases, offline learning and real-time viewpoint recommendation. In offline phase, publicly available images are utilized along with associated meta-data information to train a viewpoint recommendation model. The number of possible views at any tourist location can be numerous and therefore the problem of scene based viewpoint recommendation is challenging. Bringing in the time and weather parameters into consideration makes the problem even more difficult. To solve this problem, we follow a bottom-up approach and instead of focusing on the complete view we focus on the landmark objects present in the view. The photo-taking behavior of users corresponding to each landmark object is modeled using a generative (Gaussian Mixture Model) approach. An overview of the offline learning process is outlined in figure 5.3.

5.2.1 Landmark Objects

In any tourist location, there are usually multiple landmark objects which the users capture in their photographs. We first extract visual words from the images captured at a location using image segmentation technique. We use SLIC (Simple Linear Iterative Clustering) [2] method

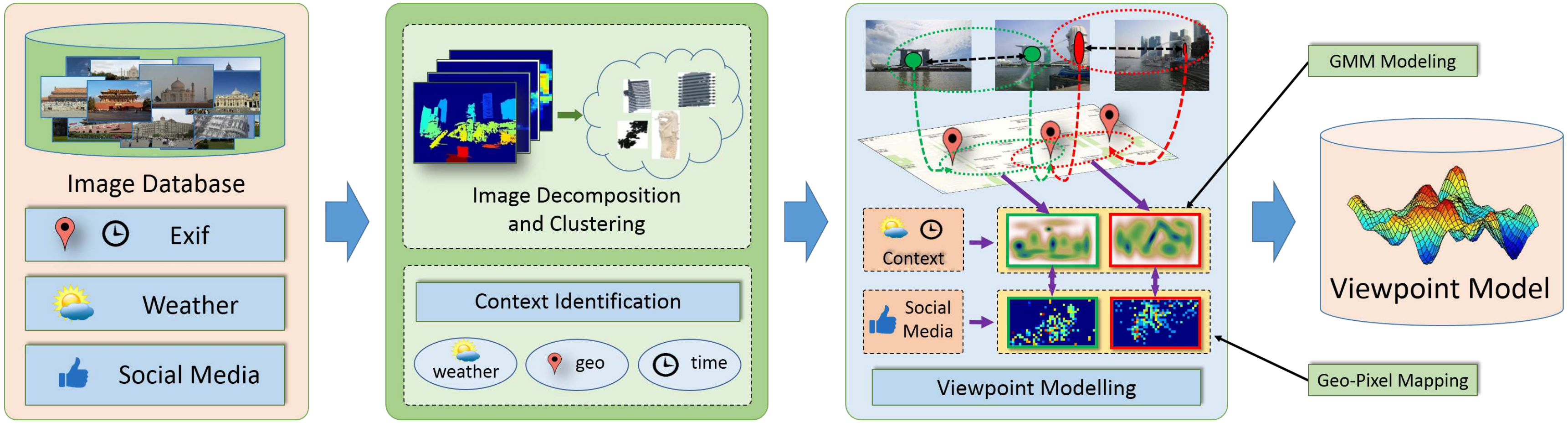


FIGURE 5.3: Overview of the offline learning phase of *ClickSmart*

for segmentation of an image which generates small superpixels. The obtained superpixels are merged based on their color similarity to form segments termed as visual words. Thus, a pool of visual words is created for each tourist location.

To identify landmark objects, clustering is performed on the pool to group the identical visual words. We use Affinity Propagation [48] algorithm to perform clustering. A visual word is represented by a set of features. We employ RGB color histogram, Histogram of Oriented Gradients (HOG) [36] and Speeded-Up Robust Features (SURF) [17] for feature extraction. Each visual word is represented with a 896-dimensional feature vector. It comprises of 3×256 dimensions for RGB color histogram, 64 dimensions for SURF features and another 64 dimensions for HOG features. After performing clustering a set of clusters are formed and each cluster is represented by an exemplar visual word denoted as a landmark object.

5.2.2 Popularity of Landmark Objects

For a tourist location, all landmark objects may not be equally important from the photography perspective. To find the importance of a landmark object we propose a hybrid approach which utilizes image saliency and a variant of *tf-idf*. We make the assumption that popular landmark objects detected in an image will be more salient. For saliency map, we use visual attention based saliency proposed by Achanta et al. [3]. We extract saliency map for each image and then saliency value of each pixel is used to evaluate saliency of each visual segment obtained from image segmentation. A normalized saliency value is evaluated for all the visual words extracted for a tourist location. Then, average saliency value is computed for visual words which belong to the same cluster and this average value is assigned as a saliency score to corresponding landmark object.

We further observe that popular landmark objects will occur more often in captured photographs. The size of a cluster indicates the occurrence of that landmark object in the photographs. However, multiple occurrences of a visual word does not always indicate its importance. A visual word corresponding to the sky and trees may be found multiple times in an image and may not be important. To distinguish between popular and common landmark objects, we employ a variant of *tf-idf*. We use term-frequency (*tf*) to represent the number of times a visual word occurs in the same image and similarly, document-frequency (*df*) represent the number of times a visual word occurs in different images. Now, a popular landmark object will have a low *tf* and a high *df*. Therefore, to estimate the popularity of a landmark object we compute *itf-df*, the product of inverse term-frequency and document-frequency, where *itf* is the inverse term-frequency.

We use a weighted combination of saliency and *itf-df* to estimate the popularity of a landmark object (*lmo*).

$$Pop(lmo_i) = a_1(Sal_i) + a_2(itf_i \times df_i). \tag{5.1}$$

Here, $Pop(lmo_i)$ is the popularity, Sal_i is the saliency, itf_i is the inverse term-frequency and df_i represents document-frequency of i^{th} *lmo*. a_1 and a_2 are weights for saliency and *itf-df*. Substituting the values of Sal_i, itf_i and df_i we get,

$$Pop(lmo_i) = a_1 \frac{\sum_{j=1}^{T_i} sal_j}{T_i} + a_2 \frac{Num_i}{\sum_{j=1}^{Num_i} num_j} \log Num_i, \tag{5.2}$$

where, T_i is the total number of occurrence of i^{th} *lmo* in all images, sal_j is the saliency value of lmo_i in its j^{th} occurrence, Num_i is the number of images in which i^{th} *lmo* was present and num_j is the number of times lmo_i was present in the j^{th} image.

5.2.3 Viewpoint Modeling

In viewpoint modeling, we utilize the photo-taking behavior of people and try to model the popular viewpoint locations for identified landmark objects. We associate each landmark object with a spatial probabilistic distribution model which illustrates the popularity of different possible viewpoints. After performing clustering for landmark object detection, the visual words belonging to the same cluster represents similar landmark object. Thus, we have multiple geo-locations for a landmark object which represents different viewpoints from where it was captured. We use the latitude and longitude coordinates of the viewpoint location of a landmark object to build a generative model. The spatial distribution of the viewpoints for each landmark object is assumed to be a Gaussian Mixture Model (GMM). For any landmark object l, we denote $\mathbf{x}(l) = (lat, lon)^T$, where (lat, lon) represent the latitude and longitude coordinates of the viewpoint for a landmark object respectively. Therefore, the probabilistic distribution of viewpoints $\mathbf{x}(l_k)$ for a given landmark object l_k can be expressed as,

$$p(\mathbf{x}(l_k)) = \sum_{i=1}^{N_k} w_i^k \mathcal{N}(\mathbf{x}|\mu_i^k, \Sigma_i^k), \tag{5.3}$$

where, w_i^k denotes the prior, $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ denotes a Gaussian component and N_k is the number of Gaussian components in the mixture. To estimate the number of Gaussian mixture components, we use Bayesian information criterion (BIC) [145]. The parameters (w^k, μ^k, Σ^k) of Gaussian mixture model are estimated for all the extracted landmark objects (l_k) for a given location using expectation-maximization (EM) algorithm [40]. The EM algorithm maximizes the log-likelihood of class probability for each of the viewpoints associated with a landmark object. In the E-step we compute the expected class probability for each of the viewpoint ($\mathbf{x}_t$) corresponding to a landmark object l^k:

$$P(i|\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t|\mu_i^k, \Sigma_i^k)w_i^k}{\sum_{l=1}^{N_k} \mathcal{N}(\mathbf{x}_t|\mu_l^k, \Sigma_l^k)w_l^k}. \tag{5.4}$$

In the M-step latent variables, mean ($\hat{\mu}$), covariance ($\hat{\Sigma}$) and priors for each class ($\hat{w}$) are updated as follows:

$$\hat{\mu}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}, \tag{5.5}$$

$$\hat{\Sigma}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)[\mathbf{x}_t - \hat{\mu}_i^k][\mathbf{x}_t - \hat{\mu}_i^k]^T}{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}, \tag{5.6}$$

$$\hat{w}_i^k = \frac{\sum_{t=1}^{T_k} P(i|\mathbf{x}_t)}{T_k}, \tag{5.7}$$

where, T_k is the number of viewpoint samples associated with a given landmark object l_k.

5.2.3.1 Photographs With People

We define a virtual landmark object representing a human object to model the photo-taking behavior of users for photographs with people. The geo-locations from all the images which have people in them are considered as viewpoints for a virtual human object. For a virtual human object l_v, we denote $\mathbf{x}(l_v) = (lat, lon, num)^T$, where (lat, lon) represent the latitude and longitude of the viewpoint and num denotes the number of people present in the image. The probabilistic distribution of viewpoints $\mathbf{x}(l_v)$ for a virtual human object l_v in 3 dimensional space can be expressed using equation 5.3.

5.2.3.2 Role of Context

We add another parameter to the probabilistic distribution model of landmark objects to incorporate the effect of time on viewpoint. Now, for a landmark object l we define $\mathbf{x}(l) = (lat, lon, time)^T$, where (lat, lon) represent the latitude and longitude coordinates of the viewpoint for a landmark object and $time$ denotes the time of capture. This 3-dimensional spatial and time distribution for a landmark object is expressed using the equation 5.3 and the corresponding parameters are evaluated using EM algorithm.

Apart from time, we also consider weather conditions for viewpoint recommendation. The time and geo-location of the captured image can be used to find the weather condition at the time of capture. We consider visibility, haze, temperature, sunrise time, sunset time, sunpeak time, cloud conditions, rain conditions and month of capture to define a 9-dim feature vector. A discrete feature value is used to define the cloud and rain conditions (overcast-4, mostly cloudy-3, partly cloudy-2, scattered clouds-1, no-clouds-0, no-rain-0, light rain-1 and heavy rain-2). We make use of weather forecast services [165, 180] to obtain the details.

Now, for each landmark object, we have three types of features, location, time and weather-conditions. This gives us a 12-dim feature space corresponding to each visual word. Since, the weather conditions such as visibility, haze, cloud conditions, etc, are correlated, the dimensionality of weather feature space can be reduced. Therefore, we employ Principal Component Analysis (PCA) to reduce the dimensionality of weather features. The weather feature space is reduced from 9 to 4 which is based on the variance contribution score after performing PCA. Using equation 5.3 a probabilistic distribution is expressed for each landmark object which considers both time and weather conditions.

5.2.4 View-Cell (Popularity vs Quality)

According to a recent study [58], the latitude and longitude coordinates of the geo-location of *Flickr* images is accurate up to 10 meters for popular locations. Taking this into account, we divide the geographical area of a tourist location into equal sized macro blocks, termed as view-cells, and use the granularity of view-cells for viewpoint recommendation. In the previous section, we discussed viewpoint modeling which tries to estimate the photo-taking behavior of people for

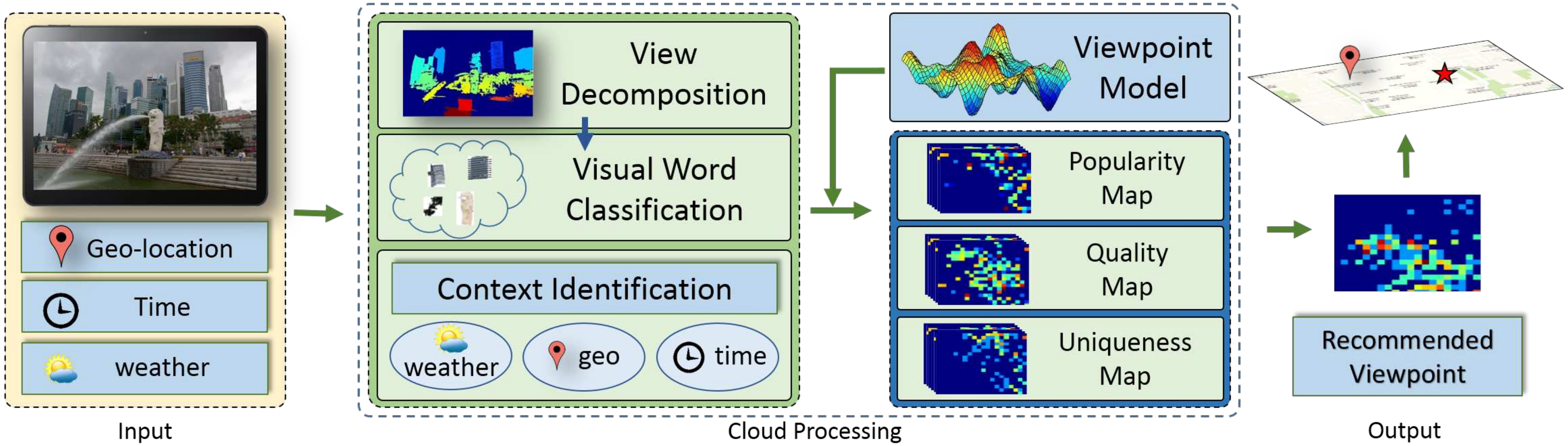


FIGURE 5.4: Overview of *ClickSmart* real-time recommendation

a landmark object at any tourist location. It represents the underlying popularity of a viewpoint for a landmark object based on the number of photographs of that landmark object captured at that location. However, the number of photographs of a landmark object captured at any location does not always indicate the quality of that viewpoint. Therefore we define another metric, *quality* of view-cell, which is an indication of its aesthetic quality and is evaluated based on social media cues.

We first evaluate the aesthetic quality of captured images using social media cues to determine the quality of a viewpoint. The aesthetic quality of an image is computed based on the number of user-views, user-likes and interestingness score assigned by *Flickr* [21]. Since, an old photograph on social media will tend to have more number of user-views and user-likes as compared to a relatively new photograph, the evaluated aesthetic score is adjusted using a time factor.

$$score(i) = \left(1 - \frac{1}{e^{\upsilon v + \beta f + \vartheta I}}\right)\left(\frac{1}{1 + e^{\tau t - \kappa}}\right), \tag{5.8}$$

where, v and f are the number of user-views and likes respectively, I is the interestingness score and t is the number of days between the upload date of the image and the date we crawled the dataset. $\upsilon, \beta, \vartheta, \tau$ and κ are constants whose values are empirically calculated such that photographs with median values of v, f and i get a score of 0.5. Also, equal weightage is given to user views, likes and interestingness in computing the score. The values of θ and κ is set such that the oldest photograph in the dataset has a deteriorating factor of 0.5 and the newest photograph has a deteriorating factor of 1. For our experiments, we set $\upsilon = 0.003, \beta = 0.04, \vartheta = 1, \tau = 0.005$ and $\kappa = 6$.

Based on the aesthetic score of an image, each visual word is also assigned an aesthetic value. This value is propagated from the visual word to the geo-location associated with it. Thus, for each landmark object, we will have aesthetic scores associated with all the locations from where that object was captured. Each geo-location is binned into a view-cell based on its latitude and longitude coordinate values. In this work, we define the size of view-cells as 10×10 square meter blocks. An average score is calculated for each view-cell using the aesthetic value of the locations which belong to this view-cell for a landmark object. This score is termed as the quality of view-cells corresponding to each landmark object.

5.2.5 *Uniqueness* of a view-cell

In any tourist location, there can be some good viewpoints which are not very popular among tourist, but they may be good from a photography perspective. We define such viewpoint locations as rare but interesting for photography and propose a term *uniqueness* to quantify this aspect of viewpoints. The *uniqueness* of a viewpoint depends upon the quality and number of photographs captured at that location. High quality will make a location interesting and therefore *uniqueness* is directly proportional to the quality of a location. A large number of photographs at a location means it is very popular and easily accessible to the users. We define the term *sparseness* to measure the inverse of popularity. Now, the *uniqueness* of a view-cell for a landmark object is computed as the product of quality and sparseness of the view-cell. It is defined as,

$$U_i^l = Q_i^l \times S_i^l, \tag{5.9}$$

where, U_i^l, Q_i^l, and S_i^l represents the *uniqueness*, quality and sparseness of a view-cell i for lmo_l respectively. The quality of a view-cell i for lmo_l is defined as,

$$Q_i^l = \frac{\sum_{j=1}^{N_l} a_j}{N_l}, \tag{5.10}$$

where, a_j is the aesthetic quality score of the j^{th} instance of lmo_l captured in the view-cell i and N_l is the total number of instances of lmo_l captured in the view-cell i. Sparseness of view-cell i for lmo_l is defined as,

$$S_i^l = \begin{cases} \dfrac{1}{1+e^{\zeta N_l - \eta}}, & \text{if } N_l > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{5.11}$$

where, N_l is the total number of photographs with landmark object l captured in the view-cell i. ζ and η are constants which are empirically computed and set as $\zeta = 0.1$ and $\eta = 6$ in the conducted experiments.

5.3 Real-Time Recommendation

The preview on the user's camera and the time and geo-location of the user are taken as input for the recommendation system. An overview of the real-time recommendation is presented in Fig. 5.4. The time and geo-location are used to get the weather condition of the user location from the weather service providers [165, 180]. Image segmentation [2] is performed on the user preview image and the extracted visual words are classified as one of the landmark objects from corresponding tourist location using Nearest Neighbor approach. The input image is represented as a set of landmark objects, $I = \{l_1, l_2, ..., l_N\}$, where N is the number of landmark objects detected.

As described earlier in section (5.2.4), a tourist location is divided into view-cells and a map of view-cells is defined with size $(m \times n)$, where m is the number of view-cells along the latitude and n is the number of view-cells along the longitude. Different tourist locations will have a different number of view-cells which will be based on the total geographical area covered by the captured photographs. A popularity score is estimated for the view-cells corresponding to each detected landmark object using the trained probabilistic model. The geo-location at the center of view-cell is used to estimate the popularity score. If there are faces detected in the preview, the number of faces in the preview is used along with the geo-location to estimate the popularity of the view-cell map for human objects. Finally, we get a popularity map $(\mathbf{PM})$ of size $m \times n$ for each of the detected landmark objects along with the virtual human object.

The quality score of each view-cell in the location is evaluated for the detected landmark objects employing equation 5.10. Thus, we obtain a quality score map $(\mathbf{QM})$ of size $m \times n$ for all the detected landmark objects. Each landmark object is also associated with a popularity score (Pop) which is computed using equation 5.1. We generate a view-cell map $(\mathbf{RM})$ of size $m \times n$ based on the view-cell popularity $(\mathbf{PM})$, view-cell quality $(\mathbf{QM})$ and landmark object popularity (Pop) to make viewpoint recommendation to the user. It is computed as,

$$\mathbf{RM} = \sum_{i=0}^{L} Pop_i \left[\mathbf{PM}_i \circ \mathbf{QM}_i\right] + \mathbf{PM}_h \circ \mathbf{QM}_h, \tag{5.12}$$

where, L is the number of landmark objects detected in the preview, Pop_i, $\mathbf{PM}_i$ and $\mathbf{QM}_i$ are the popularity score, popularity map and the quality map for the i^{th} *lmo* respectively. $(\mathbf{A} \circ \mathbf{B})$ is the Hadamard product of two matrices. $\mathbf{PM}_h$ and $\mathbf{QM}_h$ are the popularity map and quality score for virtual human object respectively (**PM** is set to 0 in case of absence of people in photograph). The view-cells with higher values in the map **RM** are recommended to the user as target viewpoints.

For context-aware recommendation, time and weather conditions are utilized to form feature descriptors as described in section 5.2.3.2. Then a context-aware popularity map $(\mathbf{PM})$ is estimated for the detected landmark objects and virtual human objects using the trained probabilistic model which is utilized in equation 5.12 to generate a context-aware recommendation.

In section 5.2.5, we discussed the idea of *uniqueness* of a view-cell for a landmark object. We integrate this *uniqueness* of view-cells along with the popularity and quality of view-cells to generate rare but interesting viewpoints. The *uniqueness* score of all the view-cells is evaluated for a landmark object using equation 5.9 to generate an *uniqueness* map $(\mathbf{UM})$ for view-cells. The recommendation map is computed as,

$$\mathbf{RM} = \sum_{i=0}^{L} Pop_i \left[\Xi \mathbf{PM}_i \circ \mathbf{QM}_i + \nu \mathbf{UM}_i \right] + \left[\Xi \mathbf{PM}_h \circ \mathbf{QM}_h + \nu \mathbf{UM}_h \right], \tag{5.13}$$

where, $(\mathbf{UM}_i)$ is the *uniqueness* map for i^{th} *lmo* and $\mathbf{UM}_h$ is the *uniqueness* map for virtual human object. Ξ and ν are constants which indicates the preference for quality and *uniqueness*. We set $\Xi = 1$ and $\nu = 1$ in our experiments to assign equal weights to both the factors.

5.4 Experiments and Discussions

5.4.1 Dataset

We used *Flickr* to build our dataset for performing evaluation experiments. To evaluate our proposed method we selected twelve different popular tourist locations and build a dataset of around 67K images (table 5.1). We collected images for *Arc de Triomphe* (Paris, France), *Cologne Cathedral* (Germany), *Eiffel Tower* (Paris, France), *Forbidden City* (Beijing, China), *Gateway of India* (Mumbai, India), *India Gate* (Delhi, India), *Leaning Tower of Pisa* (Italy), *Merlion Park* (Singapore), *Statue of Liberty* (New York, USA), *St. Peter's Basilica* (Vatican City), *Taj Mahal* (Agra, India) and *Tiananmen Square* (Beijing, China). We utilized Flickr's *photos.search* API which allows to search geo-tagged images in order of *interestingness* score and gathered top ranked images for each tourist location.

TABLE 5.1: Details of the dataset and experimental results. Size is the total no. of images and LMO is the no. of landmark objects identified. q-rec is for quality based, t-rec is for time-aware, w-rec is for weather based and uq-rec and ut-rec are for corresponding *uniqueness* based recommendation. P@2 is the average precision score based on top 2 and $nDCG_5$ is the normalized discounted cumulative gain for top 5 recommended viewpoints. (**wtc** - evaluation without time constraint, **tc** - with time constraint).

Location	**Size**	LMO	**q-rec (wtc)**		**t-rec (tc)**		**w-rec (tc)**		**uq-rec (wtc)**		**ut-rec (tc)**	
			P@2	$nDCG_5$	P@2	$nDCG_5$	P@2	$nDCG_5$	P@2	$nDCG_5$	P@2	$nDCG_5$
Arc de Triomphe	5564	334	0.84	0.89	0.82	0.89	0.74	0.81	0.79	0.93	0.74	0.90
Cologne Cathedral	4294	282	0.84	0.95	0.82	0.91	0.72	0.84	0.84	0.93	0.83	0.91
Eiffel Tower	10739	242	0.86	0.87	0.85	0.87	0.68	0.78	0.86	0.89	0.86	0.89
Forbidden City	4985	284	0.77	0.97	0.75	0.95	**0.43**	0.77	0.75	0.74	0.74	0.76
Gateway of India	3521	270	0.74	0.79	0.71	0.82	**0.59**	0.84	0.89	0.88	0.89	0.90
India Gate	3872	392	0.73	0.85	0.64	0.83	**0.47**	0.84	0.82	0.87	0.78	0.87
Leaning Tower of Pisa	5542	279	0.89	0.78	0.86	0.78	**0.51**	0.81	0.89	0.78	0.88	0.81
Merlion Park	4308	232	0.63	0.79	0.59	0.84	**0.56**	0.87	0.73	0.89	0.72	0.89
Statue of Liberty	6358	254	0.91	0.97	0.88	0.95	0.66	0.80	0.84	0.83	0.80	0.82
St. Peter’s Basilica	6722	373	0.85	0.90	0.84	0.90	0.67	0.81	0.86	0.90	0.86	0.91
Taj Mahal	6594	373	0.83	0.91	0.72	0.91	0.72	0.84	0.86	0.93	0.87	0.94
Tiananmen Square	2460	241	**0.45**	0.96	**0.47**	0.94	**0.16**	0.81	0.44	0.70	**0.40**	0.73

TABLE 5.2: Average precision@N and $nDCG_\rho$ score. **S** and **T** are recommendations using [204] corresponding to spatial and temporal methods as defined by the authors. (**wtc** - evaluation without time constraint, all other results are with time constraint.

Method	P@1	P@2	P@5	$nDCG_2$	$nDCG_5$
q-rec (wtc)	0.82	0.78	0.73	0.76	0.89
q-rec	0.53	0.54	0.51	0.58	0.75
t-rec	0.78	0.75	0.70	0.74	0.88
w-rec	**0.57**	**0.57**	**0.56**	0.62	0.83
uq-rec (wtc)	0.81	0.80	0.77	0.70	0.86
ut-rec	0.79	0.78	0.76	0.69	0.86
uw-rec	0.78	0.76	0.73	0.71	0.87
Zhang et al. **S** [204]	0.24	0.31	0.52	0.34	0.66
Zhang et al. **T** [204]	0.41	0.47	0.53	0.44	0.74
Huang et al. [66] (**wtc**)	0.79*	-†	-†	-†	-†

* For *Arc de Triomphe* where we considered only one landmark object.

† [66] recommends one viewpoint, therefore only P@1 is provided.

5.4.2 Evaluation

To evaluate *ClickSmart* we propose a content-based image retrieval (CBIR) technique. Each image is representing using a hybrid set of feature descriptors. First of all, we extract a set of low-level visual features including RGB histogram (768 dimensions), HOG features [36] (64 dimensions) and SURF descriptors [17] (64 dimensions) for each image. Along with these visual features, we also make use of bag-of-visual-words as we have already extracted landmark objects for each location. Based on the presence and absence of a landmark object in any image we obtain a D-dimensional feature descriptor (where D is the number of landmark objects for the corresponding location). Therefore, we extract a (D+896) dimensional feature descriptor for each image.

Now, for each recommendation, the viewpoint from where the input image was captured is defined as source view-cell and the recommended viewpoint is termed target view-cell. The images with a view similar to the input image are retrieved from both source view-cell and target view-cell. View similarity is measured using the extracted (D+896) dimensional feature descriptor.

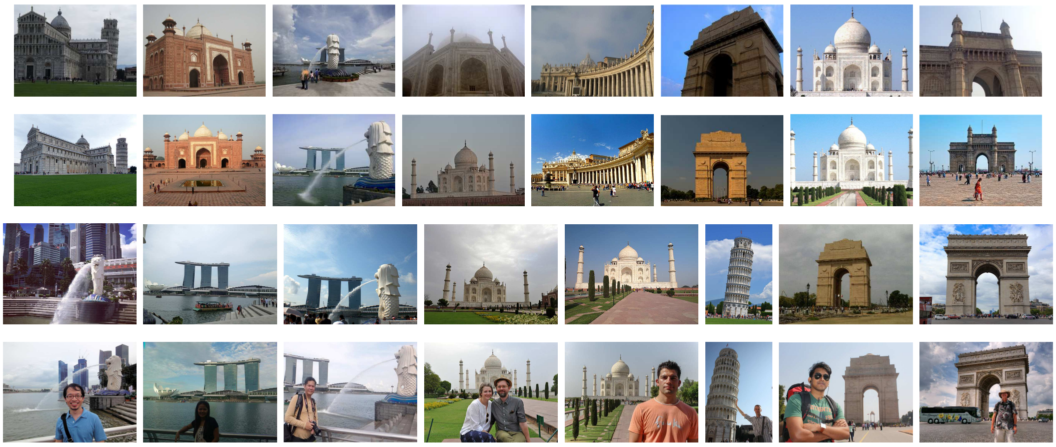


FIGURE 5.5: Sample recommendation results for some of the locations. First and third row are input views and corresponding recommendations are shown in second and fourth row. Fourth row shows recommendations for presence of poeple in the view.

Each retrieved image has an associated aesthetic score which is evaluated using equation 5.8. For context-aware recommendation (time and weather based), images captured within one hour window from input image time-stamp are retrieved. The average aesthetic score is evaluated for source view-cell and target view-cell based on the retrieved images. The recommendation is considered positive if the average aesthetic score of the target view-cell is greater than the source view-cell. Based on this, an average precision@N score is computed to evaluate the recommendation. We also use Normalized Discounted Cumulative Gain $(nDCG)$ [20] to evaluate the quality of recommendation results. Discounted cumulative gain for top ρ recommendations (DCG_ρ) is computed as,

$$DCG_\rho = \sum_{i=1}^{\rho} \frac{2^{rel_i} - 1}{log(i+1)}, \tag{5.14}$$

where, $rel_i \in \{0,1\}$ is the relevance of i^{th} recommendation which is the computed average aesthetic score for recommended view-cell. The DCG value is normalized for each recommendation result by the maximum possible DCG, also known as Ideal $DCG(IDCG)$. The $nDCG$ values are averaged for all recommendation queries to obtain a measure of the average performance.

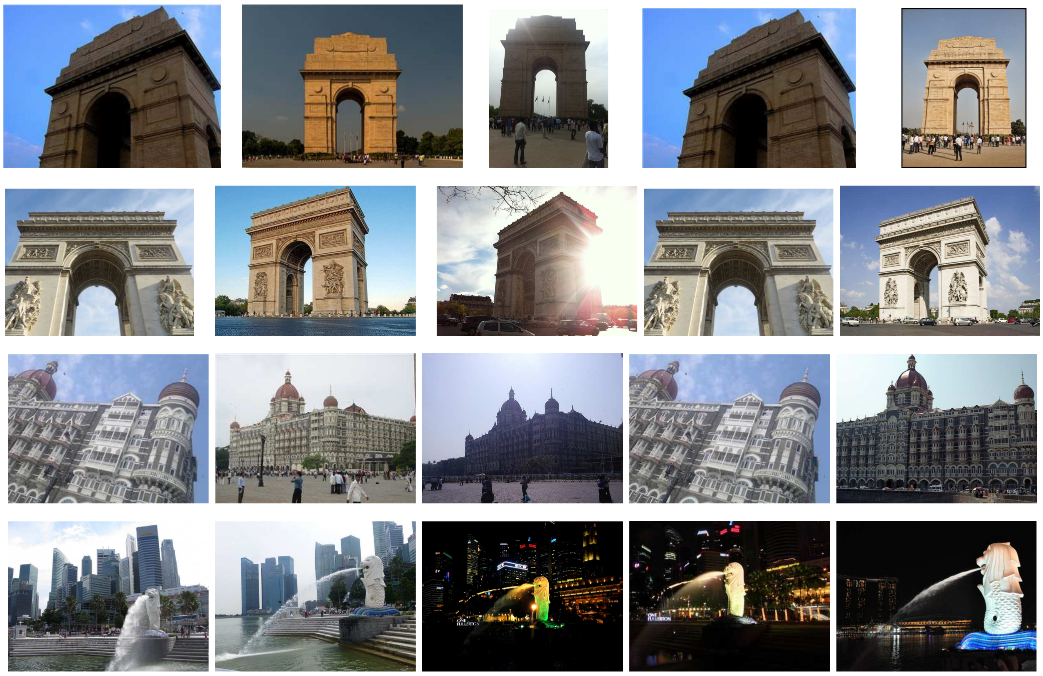


FIGURE 5.6: Time-based recommendation. First and the fourth column shows the input view and corresponding recommendations are shown in second and fifth column. The third column shows image from viewpoint similar to second column but at a different time.

For each tourist location, 10% of the images are kept for testing the proposed method. As described in section 5.3, six different type of recommendations are generated for each image. We call these, quality based (q-rec), time-aware (t-rec), weather-aware (w-rec), *uniqueness*-quality based (uq-rec), *uniqueness*-time based (ut-rec) and *uniqueness*-weather based (uw-rec) recommendation. For quality based recommendation, popularity model for landmark objects based on geographical coordinates is used with quality score of view-cells in equation 5.12. In time-aware recommendation, popularity model of landmark objects based on both geo-location and time information is used in equation 5.12. Similarly, to generate weather-aware recommendation, popularity model based on all the three factors, location, time and weather conditions, is employed. *Uniqueness* based recommendation is generated with equation 5.13 using time-based popularity model, quality score and *uniqueness* score of view-cells.

5.4.3 Aesthetic Score Validation

We performed a quantitative comparison of the aesthetic scores with Datta et al. [39]. We used the method proposed by [39] to compute an aesthetic score for all the images in our dataset. Considering this as ground truth we get a Mean Squared Error (MSE) of 0.092 with our aesthetic scores. We further considered images with aesthetic score >0.75 as good and score <0.25 as bad. We considered this as ground truth and get a precision of 0.76 and a recall of 0.52 for aesthetic scores predicted using our approach. The low recall indicates that there are images with good aesthetics but less social media popularity. It is also important to note that the aesthetic evaluation of images using social media cues can be augmented with any other state of the art content-based aesthetic evaluation model such as [157] and [160] to get a better estimate of the aesthetic quality of images.

We also conducted a short user study to validate the aesthetic score computed using social media cues in which 5 skilled photographers participated. We randomly selected 10 images, 5 with aesthetic score >0.9 and 5 with aesthetic score <0.1. In the user study, we asked the participants to rate the images from 1-5 (1-poor and 5-Excellent). The images were presented in a random order to the participants. The average rating score for images with aesthetic score >0.9 was 4.56 and images with aesthetic score <0.1 was 2.4.

5.4.4 Results and Analysis

The experimental results are presented in table 5.1 and 5.2. In table 5.2 we have shown average precision@N (for N = 1, 2 and 5) and $nDCG_{\rho}$ (for ρ = 2 and 5) scores for the complete dataset. We observed that precision and $nDCG$ score for weather-based recommendation is lower as compared to quality and time-based recommendations. Table 5.1 shows average precision@2 and $nDCG_5$ scores for each of the location in the dataset independently. We further observed

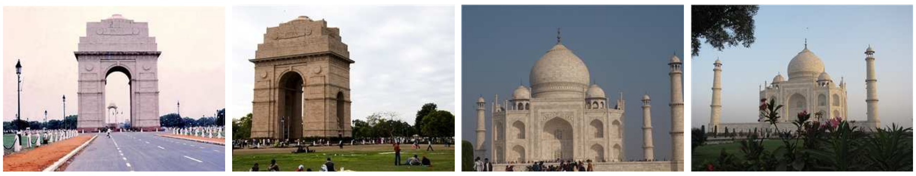


FIGURE 5.7: Sample images from user survey for *uniqueness* based recommendation

that low precision score was mostly observed for locations with smaller dataset size. One of the reasons for this low precision can be the insufficient number of images which may lead to over-fitting of GMM in high dimensional space. Another reason can be inaccuracy in weather forecast services. We came across some images in the dataset with a clear sky when the weather reports have indicated cloudy conditions. An improvement in precision and $nDCG$ score was observed after integrating the *uniqueness* factor with weather based recommendation. This is mainly because it was integrated as an additive factor (equation 5.13) rather than multiplicative, which minimizes the error propagation to the generated recommendation.

Figure 5.5 presents some viewpoint recommendation results generated by *ClickSmart*. The images in the first and third row are input images and the images in the second and fourth row are images captured from the viewpoint recommended by *ClickSmart*. It can be observed that the images from recommended viewpoints are better as compared to images from the original viewpoint. The fourth row shows the recommendation corresponding to input views in third rows for the presence of people.

To investigate the role of context, we evaluate the quality-based recommendation similar to context-based by retrieving images within 1-hour window of capture time-stamp of the input image. The average precision@1 was 0.53, which is quite low as compared to time-based recommendation (0.78). Figure 5.6 shows some time-based recommendation results using *ClickSmart*.

The first and the fourth column are the input views and the second and the fifth column are images captured from corresponding recommended viewpoints. The third column shows images from viewpoints same as the second column but at a different time of the day. We can observe that how the viewpoint recommendation changes for the same view in the fifth column. In the first and second row, the recommended viewpoints in the second and fifth column are at the opposite side of the monument. Also, if we generate a recommendation for overcast weather conditions for the first two rows, the viewpoint in the second column is still in the recommended viewpoints list.

To evaluate the image retrieval step we manually annotated around 1% (535 images) of complete dataset. First we randomly sampled 10 set of images from *Merlion* dataset and use *ClickSmart* to generate viewpoint recommendation. Then the images from the source and the recommended view-cell are retrieved and annotated manually for relevance. The Mean Average Precision (MAP) score for the 20 queries (2 queries for each image) was found to be 0.80 when we retrieve top N results (N is different and known for each of the images). When we set a threshold value of similarity for retrieval we get an average precision of 0.77 and an average recall of 0.78. The use of context (geo-location) can be attributed to this high precision and recall.

We also evaluated the retrieval step with Holidays [76] and ZuBuD [147] dataset which are publicly available. With Holidays dataset we get a MAP score of 0.60 and on ZuBuD dataset a MAP score of 0.49. ZuBud is a dataset of different kind of buildings whereas Holidays dataset consists of images with natural sceneries. In these two datasets, the images are mainly annotated for content similarity rather than view similarity.

5.4.5 User Study

To further evaluate the recommendation of *ClickSmart*, we conducted a user study in which 26 users participated including 21 amateur users and 5 skilled photographers with at least 5 years of experience in single-lens reflex camera. In the survey, there were four set of questions to evaluate different types of recommendation generated by *ClickSmart*. We evaluated our model for quality based (q-rec), time-aware (t-rec), *uniqueness*-based (ut-rec) and p-rec method (where people are present in the photograph). For each question, a pair of images with similar view captured from different viewpoints were shown to the user and asked the following question, '*Which of the two images is captured from a better viewpoint?*'. In the pair of images, one is the input image and the other is an image captured from the view-cell recommended by *ClickSmart*. For each type 10 set of questions were presented to the user.

Figure 5.8 shows the average percentage of users who finds the recommended viewpoint better than the original viewpoint. Overall 75% of participants liked the viewpoint recommended by our system. The agreement percentage is higher among skilled photographers (76.5%) as compared to amateur users (73.5%). It can be noticed that for *uniqueness* based recommendation the agreement score is low for amateur users (56%). The *uniqueness* based recommendation suggest non-popular viewpoints, however, amateur users gave preference to popular views for images with comparable quality. For example, in figure 5.7, we have images for two different locations (first and third image) and corresponding images from viewpoints recommended by *ClickSmart* (second and fourth image). Almost 52% of the amateur users preferred the original views, which are popular. However, the recommended viewpoints were preferred by almost 70% of the skilled users.

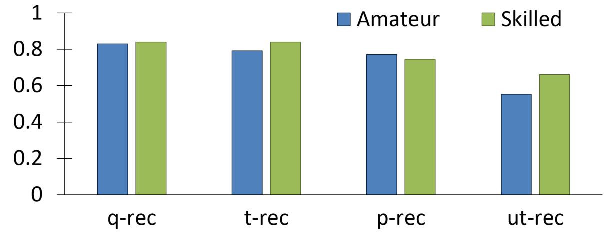


FIGURE 5.8: Average agreement values from user study. q-rec is quality based, t-rec is time-aware, p-rec is for photographs with people and ut-rec is for *uniqueness* based.

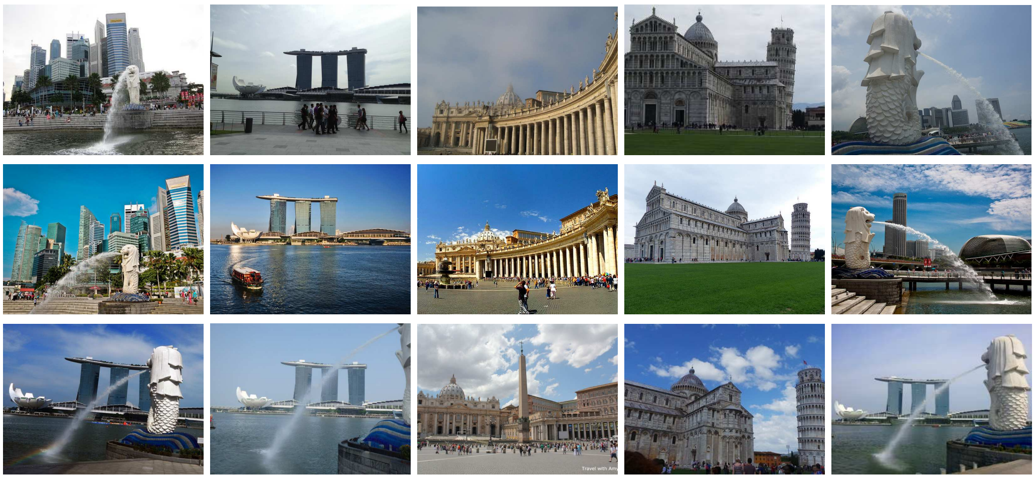


FIGURE 5.9: Comparison of recommendations with [204]. The first row is the input view, second row shows results with *ClickSmart* and third row using [204].

5.4.6 Comparison

We first compared the proposed method with the state of the art methods in photography assistance based on their approach and problem formulation. A detailed analysis of the comparison is presented in table 5.3. We can observe that the state of the art methods which provide location recommendation [66, 132, 149, 204] are not interactive and ignore the presence of people and user context for making the recommendation. It is important to note that photography assistance methods such as [138, 190] are focused on improving image composition and not viewpoint. Therefore, these methods are complementary to *ClickSmart* and can be combined with it to provide a better photography experience to a user.

We further performed a quantitative and qualitative comparison of our proposed method with

[204] and [66]. A quantitative comparison is given in table 5.2. For [204], a ranked list of viewpoints is generated as proposed by the authors for both spatial and temporal recommendations. We can observe that both the precision and *nDCG* score are lower as compared to our method. The main reason for this low score is irrelevance of the recommended viewpoint for the input user view. Fig 5.9 shows a comparison of recommendation results of [204] with our method. The first row is the input view, the second row shows recommendations using our method and the third row shows the results of [204]. It can be observed that the recommended viewpoint using [204] are popular viewpoints but not good for capturing the given user input views.

Huang et al. [66] proposed a method which identifies popular viewing directions for a landmark object and then recommends a viewpoint based on the quality of images in the popular viewing direction. We used this method to generate a viewpoint recommendation for the given user views and compared it with our approach. First, the viewing direction is identified using the geo-location of the user and then the geo-location of the best-captured image in that viewing direction is recommended as a target viewpoint. This method works for locations where we have only one landmark object present and it performs similarly to our approach when the time factor is ignored (table 5.2). We performed quantitative experiments for one of the location (*Arc de Triomphe*, Paris, France), which has only one landmark object, and employed [37] for computing the aesthetic quality of images. Figure 5.10 presents the recommendation results obtained using [66] for some other locations. First and the second column shows the case when there are more than one landmark objects and in the third and fourth column, we can observe the bad recommendations as time factor is not considered. The recommended viewpoints are in fact good viewpoints for capturing other views (first and second) and at a different time during the day (third and fourth). Corresponding recommendations using our method are shown in figure 5.5 and 5.6.

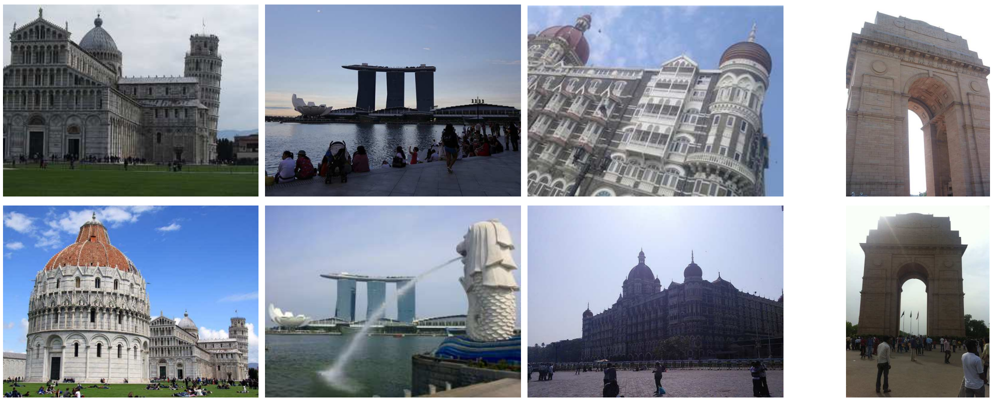


FIGURE 5.10: Recommendations using [66]. First row is the input view and second row is the corresponding recommendation.

TABLE 5.3: Comparison with State of the Art methods. **C**(composition), **L**(location), **I**(interactive), **V**(user view), **T**(time), **W**(weather)

The Work	Recommendation Type			User Context		
	C	**L**	**I**	**V**	**T**	**W**
[32, 111, 128, 156, 181]	✓		✓	✓		
[139, 190]	✓		✓	✓	✓	✓
[132, 149]		✓				
[66, 204]		✓			✓	
Proposed Method		✓	✓	✓	✓	✓

5.4.7 Running-time Analysis

In the real-time recommendation, processing is required for image segmentation, object classification and generating the recommendation on the cloud. For this research, we conducted our experiments on a 8 core processor running at 3.40 GHz with 8 GB of RAM for cloud processing. The average time for complete recommendation phase on this machine for a 640 × 480 pixel image is around 800 milliseconds. Therefore, *ClickSmart* can be used to develop a cloud-based service which can provide a real-time viewpoint recommendation to the user.

5.5 Summary

In this work, we propose *ClickSmart*, a method of viewpoint recommendation which can guide users to capture high-quality images in popular tourist locations. The proposed method leverage on publicly available images and social media cues to learn the photo-taking behavior of people. We presented the idea of view-cells and defined their *popularity*, *quality* and *uniqueness* which are further utilized for viewpoint recommendation. We also investigated the role of context, such as time and weather conditions, for viewpoint recommendation in photography. The experimental results and user study shows that the proposed method can make effective viewpoint recommendation to the user. *ClickSmart* can be extended to a system which can provide viewpoint recommendation for user defined compositions and it will be the focus of our future work. We also plan to exploit the photo-taking behavior of people for making a personalized recommendation which will also take into account the photography taste of a person. Also, the idea of view-cells can be further utilized for assisting a user in capturing a high-quality videos and, other location based recommendations.

Chapter 6

Optimal Foraging Theory for Photography and Exploration

Animals search for food in their environment with a decision strategy which keeps them fit. Optimal Foraging Theory (OFT) models this foraging behavior to determine the optimal decision strategy followed by animals. This theory has been successfully applied to humans as they search for information and is termed as Information Foraging. When people visit a tourist location, they follow a similar strategy to move from one spot to another and collect information by capturing photographs. This behavior has similarities with the foraging behavior of animals which has been widely studied by researchers. In this work, we propose to employ OFT to help tourists explore a location and capture photographs in an optimal way. We use this theory to determine a decision strategy for tourists which provides a list of micro-pois to visit and the corresponding stay time at each of the micro-poi. Finally, we solve an optimization problem to find an optimal path which can be followed to explore a tourist location.

6.1 Introduction

There are usually multiple hot-spots in any tourist location and people follow some trajectory which passes through these hot-spot locations. In this work, we termed these hot-spot locations as micro-locations or micro-pois (micro point of interest). People also capture photographs at micro-locations which are good from the photography perspective. If people are not familiar with the tourist location, then usually they follow their intuitions or other tourists to explore hot-spots of the location. This strategy is not always successful for tourists, and people usually spend a lot of time exploring the location rather than enjoying the hot-spots.

A very similar kind of problem is faced by animals while foraging where they need to search for their food and they move from one food patch to another in search of food. Researchers have well studied the problem of animal foraging behavior and observed that an optimal foraging behavior is followed by the animals for their survival. OFT [155] is one such study which tries to model the animal behavior for foraging which ensures their survival. This theory has also been adapted successfully to model human behavior as he searches for information [133].

Inspired by this analogy, we propose a novel problem in which we attempt to identify the optimal tourist behavior at tourist locations. More specifically, we want to find an optimal path to follow and the amount of time to spend at each micro-location for a given user context at any given tourist location. We leverage on social media images captured at any tourist location and associated metadata to understand the past tourist behavior and the location environment. Thereafter, we employ concepts from OFT to find optimal paths to follow between the micro-pois and the amount of time to spend at each of the visited micro-locations.

The availability of a large number of geo-tagged photographs shared by users on social media platform has motivated the research in location recommendation. This available source of information has been widely utilized by researchers to identify Points of Interests [97, 107, 150, 185] and recommend tourist locations to users [94, 103, 110, 193, 198, 199]. These methods are focused on automatically detecting the points-of-interest and recommending them to the users based on their previous travel history. The works in [88, 132, 204] focused on photography hotspots and recommend points-of-interests which are good from photography perspective. However, the existing methods do not provide any particular order or strategy in which these locations should be visited.

To overcome this limitation, the authors in [30, 31, 53, 104, 114] proposed methods to recommend travel routes which guide users to follow a path as they visit different attractions. The most popular traveled paths are determined based on trajectory clustering techniques and recommended to the user. The existing methods of route recommendation generate a path from one attraction to another and do not provide any guidance on how each particular attraction should be explored.

In this work, we focus on route recommendation within an attraction where the routes are dynamically created based on the user context and provide details such as which micro-pois to visit and corresponding stay time for taking photographs. We observe that each tourist attraction has multiple hots-spots (micro-pois) which are visited by the tourists. There can be multiple ways to visit these micro-pois and searching for an optimal path is an NP-hard problem. We make use of social media images to learn previous patterns in the environment and employ OFT to determine an optimal path for exploring the attraction and capturing photographs. The recommended path not only provides a route but also includes a list of micro-pois in the attraction where the user should visit. In addition, the amount of time to spend at each of the micro-poi location is also recommended.

The rest of the chapter is organized as follows. In section 6.2 we present an overview of the proposed method. In section 6.3 we will introduce the OFT and discuss how we adapt it for exploring tourist attractions. Section 6.4 and 6.5 will present the proposed method in detail and the experimental results will be discussed in section 6.6. Finally, we will conclude this chapter along with future research direction in section 6.7.

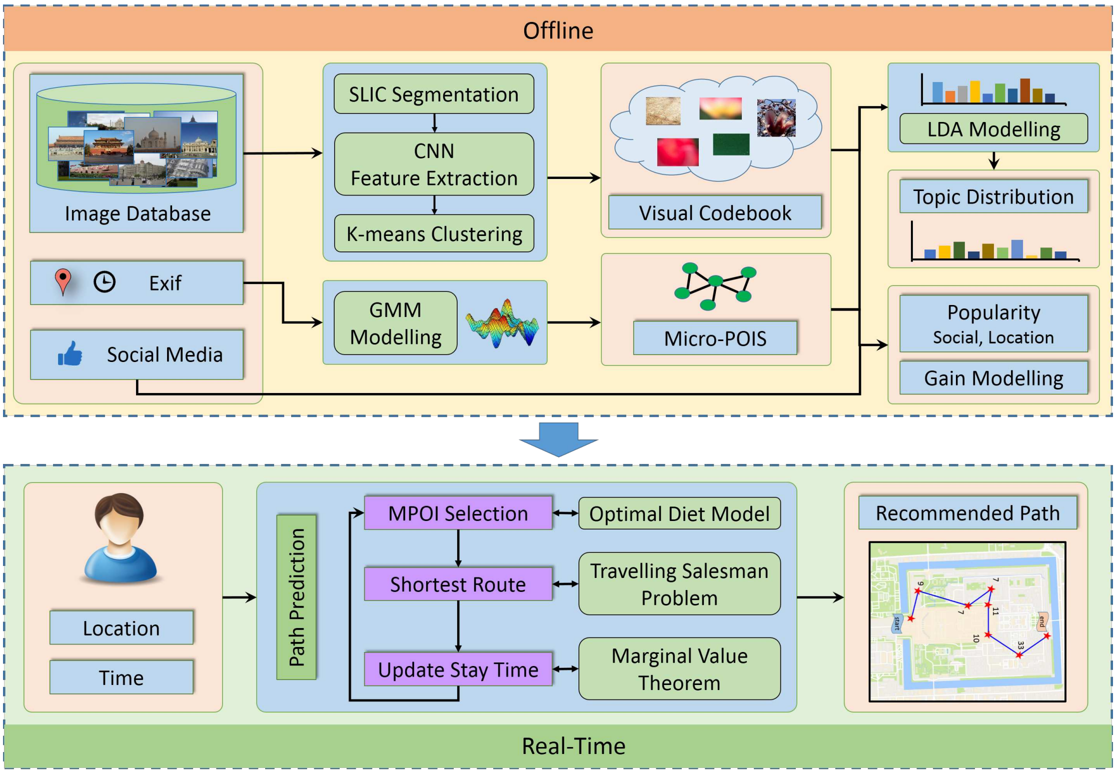


FIGURE 6.1: Overview of the proposed method.

6.2 Overview

The proposed method is composed of two phases. The overview of the proposed method is outlined in Figure 6.1. An offline phase where the social media images and associated metadata is utilized to understand the environment of a tourist location. In this phase, the micro-pois present within a tourist location and their corresponding parameters related to photography are determined. A three-dimensional network graph is developed for each location in which the node represents micro-pois and the edge connection determines the connectivity between these micro-pois. The three dimensions in the network represent latitude, longitude and the visit time for each of the micro-pois. The past photography behavior of tourists and the response from social media on their shared photographs is utilized to determine the various parameters such as quality of micro-pois.

In the online phase, OFT is employed to predict a suitable path based on the user-context. User-context indicates the time of arrival and the total time for which the tourist intend to spend at that location. The predicted path includes a list of micro-pois along the path and a stay time at each of the micro-pois which is computed based on OFT. More specifically, we employ the Marginal Value Theorem for predicting optimal stay time at each of the micro-pois and the Optimal Diet Selection Algorithm to find out a list of micro-pois to visit. Later, we pose path prediction as Traveling Salesman Problem and employ Simulated Annealing to find an optimal path through these micro-pois.

6.3 Optimal Foraging Theory for Photography

Optimal Foraging Theory is a model which is used to predict the foraging behavior of animals as they search for their food [155]. The energy gain from the food depends not only on the acquired food item but also on the foraging behavior as searching the food also require energy and time. Therefore, animal wants to maximize the energy gain as they forage in their environment to remain fit. OFT aims at predicting the best foraging strategy to achieve this goal.

OFT has also been successfully applied to develop Information Foraging Theory [133] which models human behavior as they search for information. Information Foraging Theory is based on the assumption that humans use an inbuilt foraging mechanism that evolved from animal foraging behavior as they search for information. This theory models the human behavior where they are in search of information and have to decide whether stay at the same location and try to find additional information or move on to another site and which path to follow. We observe that capturing photographs at tourist locations and moving from one spot to another is analogous to gathering information for capturing the experience. Inspired by this analogy, we propose to use OFT to understand human foraging behavior as they capture photographs and most importantly make a recommendation to users so that they follow an optimal way of capturing photographs.

Modeling of animal foraging behavior requires a currency variable, such as energy gain per unit time, which the animals are trying to maximize under the constraints, such as travel time, in the environment. OFT aims at predicting the best foraging strategy or an optimal decision rule for a given currency and environmental constraints. The average rate of gain (R) is the key factor that characterizes the efficiency of a forager. It is defined as a ratio of the net gain accumulated, G, divided by the total time spent between and within patches,

$$R = \frac{G}{T_B + T_W}, \tag{6.1}$$

here, T_B is the total between patch time and T_W is the total within patch time spent during foraging. The average rate of patch encounter is defined as,

$$\lambda_i = \frac{1}{t_{Bi}}, \tag{6.2}$$

where, t_{Bi} is average time for finding patch of type i. Now, if we have P different types of patches, the total gain can be represented as,

$$G = \sum_{i=1}^{P} \lambda_i T_B g_i(t_{Wi}), \tag{6.3}$$

where, g_i is the expected gain function from a patch i in terms of stay time t_{Wi}. Similarly, the total amount of time spent within patches is represented as,

$$T_W = \sum_{i=1}^{P} \lambda_i T_B t_{Wi}. \tag{6.4}$$

Now, after substituting equation 6.3 and 6.4 in equation 6.1, we get the overall average rate of gain as,

$$R = \frac{\Sigma_{i=1}^{P} \lambda_i g_i(t_{Wi})}{1 + \Sigma_{i=1}^{P} \lambda_i t_{Wi}}. \tag{6.5}$$

This is known as Holling's Disk Equation [63] which serves as the basis for deriving several optimal foraging models. In this work, we consider micro-pois as patches and the net gain is determined in terms of visual information in the captured photographs as a function of time spent by the users in a tourist location. We employ Optimal Diet Selection [155] and Marginal Value Theorem [27] from OFT to find the best strategy for taking photographs at a tourist location. We will present these two models in the following section and discuss how they can be used to solve the proposed problem.

6.3.1 Optimal Diet Model

Optimal Diet Model, also known as contingency model, helps in deciding whether a predator should consume the prey at hand or search for a more profitable prey item. This model predicts that the predator should ignore low-profit prey items when high-profit prey items are present in abundant. The profitability of a prey item is defined as the rate of energy gain as a function of time. If a prey item can provide a total energy gain g with a handling time of t_W, then the profitability is defined as,

$$\pi = \frac{g}{t_W}. \tag{6.6}$$

Based on the Optimal Diet Model, a predator should consume a prey item only if its profitability is greater than the overall profitability during foraging. We use this model to select micro-pois in a tourist location. The act of photo capture is associated with energy gain and the goal is to predict a strategy to maximize this gain in an optimal amount of time. We utilize the shared social media images to determine the profitability of micro-pois and selection of micro-pois is predicted using Optimal Diet Model.

6.3.2 Marginal Value Theorem

Marginal Value Theorem [27] is used to determine whether an organism searching for food should stay in the current patch or search for a new patch. The model helps in predicting when it is economically favorable to leave the current food patch to maximize the overall energy gain during foraging. When the animal forages within a patch, finding food becomes more difficult and it experiences the law of diminishing returns. This may happen because of the depletion in current food patch. Finding new patch also involves cost as the animal loses foraging time as well as energy while searching.

Marginal Value Theorem optimizes the net energy gain per unit time (Equation 6.3) in the foraging strategy. Figure 6.2 shows a plot of diminishing returns in terms of experience gain as a user capture photographs. If net experience gain is the currency then it can be represented as the slope of the line which starts at the search start time and intersects the gain curve. Marginal Value Theorem states that in order to maximize the net energy gain, one should leave the patch when this line touches the diminishing curve. In order to determine the optimal stay time at a micro-poi, we utilize shared social media images to compute the diminishing gain curves. The act of photo capture measures the energy gain and if a user continues to capture photographs at the same location, then the gain from each successive photograph will diminish due to redundancy. Based

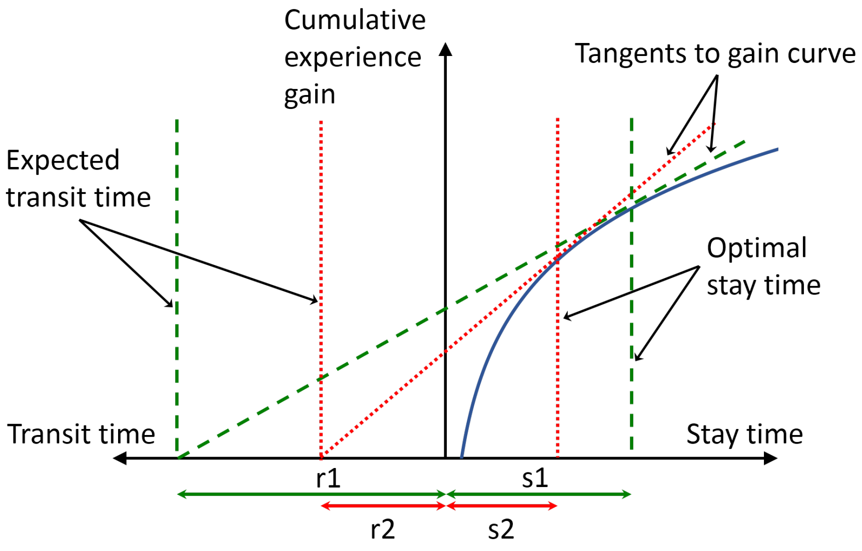


FIGURE 6.2: Marginal Value Theorem. The y-axis represents cumulative experience gain in terms of captured visual concepts in the photograph and x-axis represents time. The green and red lines corresponds to two different transit times (r1 and r2) and, s1 and s2 are the predicted optimal stay time for r1 and r2 respectively.

on this assumption we model the diminishing gain curve for each of the micros-poi and utilize it to determine the optimal stay time.

6.4 Graph-Based Micro-POI Modeling

There are usually multiple hot-spots for photography at any tourist location. In this work, we term these locations as micro-pois. Tourists explore a location by visiting these micro-pois in some order as they capture photographs along their way. Therefore to generate a path for a recommendation we first need to identify these micro-pois.

6.4.1 Micro-poi Identification

We utilize the social media images shared by users to identify these micro-pois. We observe that each micro-poi may not be suitable for photography throughout the day because of the changing lighting conditions. Therefore we also incorporate the time factor as we identify these micro-pois. The *Exif* meta-data associated with the shared photographs can be used to determine the location as well as the time of image capture. We use the geo-coordinates and the time-stamp to develop a generative model to determine the micro-pois at a tourist location. The spatial distribution of location and time pair is assumed to be a Gaussian Mixture Model (GMM). For each photograph i we define $\mathbf{x}(i) = (latitude, longitude, time)^T$, where $(latitude, longitude)$ and $time$ represents the geo-location and time of capture respectively. The probabilistic distribution of location and time pair at an attraction can be expressed as,

$$P(\mathbf{x}) = \sum_{i=1}^{N} w_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i), \tag{6.7}$$

where, $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ denotes a Gaussian component, N is the no. of Gaussian components and w_i indicates the prior for each component. We make use of Bayesian information criterion (BIC) [145] to estimate the number of Gaussian components and the parameters $(\mu^k, \Sigma^k$ and $w^k)$ of Gaussian mixture model are estimated using expectation-maximization (EM) algorithm [40].

The components of obtained generative model represent the identified micro-pois. Each micro-poi has a geo-location and a time-stamp associated with it. We associate each of the captured photographs at the corresponding attraction to one of the micro-poi.

6.4.2 Micro-poi Profiling

We compute a set of properties for each of the identified micro-poi which we will use later for recommendation. The total number of photographs captured at any micro-poi indicates its popularity among the visitors. We denote this as location-popularity ($LPop$) and it is computed as,

$$LPop(i) = \frac{N_i}{N_{max}}, \tag{6.8}$$

where, N_i is the total number of photos captured at i^{th} micro-poi and N_{max} is the maximum number of photographs captured at any micro-poi.

The social media images also have associated social media cues such as, user likes and user views, along with them which indicate their popularity among social media users. We utilize these cues to compute the popularity of each of the image as well as the popularity of each micro-poi. The image popularity (q^i) for image i is computed as proposed by [189] which assigns a score between 0-1 to each photograph,

$$q^i = 1 - \frac{1}{exp(\upsilon * v + \beta * f + \gamma * c)}, \tag{6.9}$$

where, v is the number of user-views, f is the number of user-favorites, c is the number of user comments and υ, β and γ are constants empirically set to 0.003, 0.1 and 0.1 respectively. The social media popularity ($SPpop$) for a micro-poi i is computed as an average of the quality score assigned to the photographs captured at that micro-poi,

$$SPop(i) = \frac{1}{N_i} \sum_{j=1}^{N_i} q^j, \tag{6.10}$$

where, q^j is the popularity of a photograph captured at i^{th} micro-poi computed using equation 6.9 and N_i is the total number of photographs captured at i^{th} micro-poi.

To determine a visual representation of micro-pois, we utilize the pixel information from the images captured at that micro-poi. First, we build a dictionary of visual words for a tourist location based on the captured images. We perform segmentation [2] on images and collect all the visual patches. Each patch is represented using a visual feature extracted using Convolutional Neural Network (CNN) [91]. A network trained on the ImageNet dataset is used and visual features are extracted from the fully-connected layer (fc7), prior to prediction layer, in the network. Then, we employ clustering on these patches to build a dictionary of visual words. Each photograph can be represented as a feature vector using this dictionary to indicate the presence of visual words.

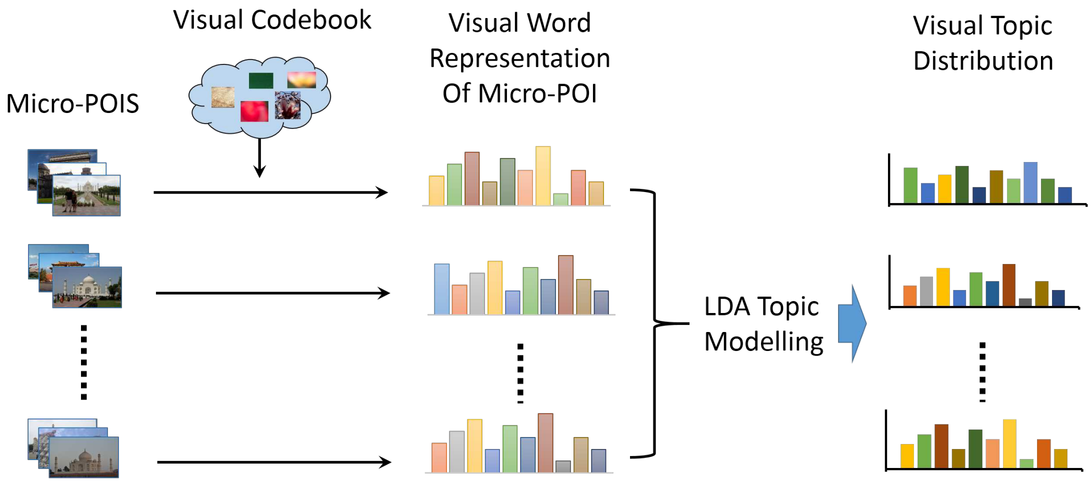

Visual Codebook
Visual Word Representation Of Micro-POI
Visual Topic Distribution
Micro-POIS
LDA Topic Modelling

FIGURE 6.3: Overview of the LDA topic modeling performed for micro-poi profiling.

The micro-poi locations can be semantically categorized into different groups based on the presence of visual words in the photographs captured at any micro-poi. We perform topic modeling with Latent Dirichlet Allocation (LDA) [61] to determine the latent categories of the micro-poi (Figure 6.3). Each micro-poi is represented as a document, where the captured photographs are considered as sentences and the visual words present in the photograph corresponds to the words in the sentence. The topic model determines a set of latent topics for the tourist location along with the association of each topic with the identified micro-pois.

6.4.3 Modeling Information Gain

Modeling and automatic quantification of the information gain as a tourist move from one micro-poi to another is a very difficult task. However, taking photographs as we explore a tourist location is a common practice followed by most of us. Therefore, we associate this information gain with the photographs captured by the users along the exploration. As discussed in section 6.4.2, each photograph can be represented as a set of visual words. A micro-poi location will be associated with a subset of these visual words which is based on the photographs which can be captured from this micro-poi.

Now, as a user takes a photo, there will be a gain associated with it which will depend on the visual words present in the photograph. With each consecutive photograph captured at any micro-poi, this gain will follow a diminishing curve as some of the visual words might already be captured in previous photographs. Finally, the gain will saturate at a certain level when all the visual words have been captured by the user. The cumulative information gain as a user captured i^{th} photograph is computed as,

$$G^i = \sum_{j=1}^{T} max(v^i_j, g^{i-1}_j), \tag{6.11}$$

where, T is the total number of visual words in the dictionary, $v^i_j \in (0,1)$ which indicates the presence of the visual word j in the i^{th} photograph and g^{i-1}_j is gain from the visual word j in the previous photograph which is computed as $max(v^{i-1}_j, g^{i-2}_j)$. The gain corresponding to each of the visual words before capturing any photograph is initialized with 0.

This information gain will be different for different users based on their photo-taking behavior. To determine the gain pattern for each micro-poi we perform regression analysis on the information gain observe for the previously captured photographs at a micro-poi. We utilize a logarithmic diminishing gain function as proposed by [155] for modeling information gain,

$$G(t) = \upsilon ln(t+\Gamma)+\varepsilon, \tag{6.12}$$

where, $G(t)$ is the information gain after time t, υ, and ε are constants which are determined using regression analysis and Γ indicates the amount of time before capturing the first photograph. We set this to 60 seconds, however, this constant will not have any effect on computation of υ and Γ. The constants for equation 6.12 can be determined using least-square linear regression analysis. We also compute average gain (G_a) and average stay time (T_a) for each of the micro-poi which will be used to compute profitability.

6.4.4 User Profiling

The previous photographs captured by a user can be used to determine the preference corresponding to the semantic visual categories. To quantify user interest we represented each user as a document and the personal image collection corresponds to sentences with the detected visual patches as words in each sentence. The trained LDA model as described in section 6.4.2 is used to determine the preference of a user for the identified categories.

6.4.5 Graph Modeling

To determine the optimal path for a user we first represent a tourist location as a graph (V, E) in 3-dimensional space. Here, V represents a set of nodes in the graph which corresponds to the identified micro-poi in the location, and E is the set of edges corresponding to the path connecting these micro-pois. The 3 dimensions refer to the latitude, longitude and time.

The photographs captured at a location are first utilized to identify the tours which people have followed in the past. A tour is defined as a set of photographs which are captured in a sequence within a day. Each photograph in the sequence has associated geo-location and time-stamp. Each tour will pass through a set of micro-pois and we can determine the stay time as well as transit time between different micro-pois from each tour. Stay time at each micro-poi is computed as a difference between the time-stamp of the first captured photograph and the last captured photograph in that micro-poi. And, the transit time is computed as the difference between the time-stamp of first captured photograph at a micro-poi and the last captured photograph at the previous micro-poi in the sequence. Finally, an average stay time for each micro-poi and an average transit time between two micro-pois is computed using all the tours traveled in the past.

6.5 Path Prediction

The 3-dimensional network graph of the tourist location is used to find an optimal path for a given user context. The user context indicates the current user location, visit time, trip duration and final location, and are used to determine the start and last node in the path from the graph. The location and time information is used to identify the graph node closest to the user. The trip duration is added to the current time to identify the last node in the path. If destination location is not provided by the user, a round trip is computed.

We employ Optimal Diet Algorithm to determine the micro-pois which should be included in the path. The location popularity ($LPop$), social media popularity ($SPop$), average gain (G_a) and average stay time T_a is utilized to compute profitability for Optimal Diet Algorithm. The profitability of a micro-poi i (Π_i) is computed as,

$$\Pi_i = \Delta * \frac{G_a^i}{T_a^i} + \Theta * SPop_i + \theta * LPop_i, \tag{6.13}$$

where, G_a^i is the average information gain, T_a^i is the average stay time, $SPop_i$ is the social media popularity, $LPop_i$ is the location popularity and Δ, Θ and θ are constants to assign weights to these parameters.

ALGORITHM 4: *OPT_PATH*

```
Input: Graph G(V, E), start node ($m_s$), end node ($m_e$), trip duration ($td$)
Output: Recommended path {$m_s, m_1, ..., m_e$}, stay time {$s_s, s_1, ..., s_e$}
$tt := 0.0$                                                    // current trip time
$P := \{m_s, m_e\}$                                              // initialize path
$\Pi := sort(\Pi)$                          // sort the profitability in decreasing order
for $\pi_i$ in $\Pi$ do
    $P.append(m_i)$                          // add the node corresponding to $\pi_i$ to path
    $P := TSP(P)$                          // find shortest path through these micro-pois
    $S := \{\}$                                                // initialize stay time
    for $m_i$ in $P$ do
        $s_i := MVT(m_i, P)$                                       // update stay time
        $S.append(s_i)$                                            // maintain a list
    end
    $tt := update\_trip\_time(P, S)$                              // update trip time
    if $tt > td$ then
        $break$
    end
end
```

The stay time at each micro-pois in the path is predicted using Marginal Value Theorem. The following function for net gain at a micro-poi is optimized to determine the optimal stay time.

$$t_{opt} = \underset{t_s}{\operatorname{argmax}} \frac{G(t_s)}{t_r + t_s}, \tag{6.14}$$

where, t_{opt} is the predicted stay time, t_s is the stay time to be optimized, $G(t_s)$ is the estimated gain at time t_s and t_r is the estimated reach time. Finally, an optimal path is constructed through the selected micro-pois by solving a Traveling Salesman Problem. A path through the micro-pois is constructed to minimize the total travel time,

$$t_{total} = \min \sum_{i=1}^{N_{mpoi}} (t^i_{opt} + t^i_r), \tag{6.15}$$

where, N_{mpoi} is the total number of micro-pois in the path, t^i_{opt} is the predicted stay time for i^{th} micro-poi and t^i_r is the estimated reach time for i^{th} micro-poi. We employ Simulated Annealing to determine the path in real-time. The complete path prediction process is presented in Algorithm 4.

6.5.1 Personalization

Personalization in route recommendation can be incorporated by taking into account the user preference for the visual content of each micro-poi. As discussed in section 6.4.4, we determine the user preference based on the past captured photographs using the trained LDA model for visual topics. We find the preference of a user for each micro-poi by computing a cosine similarity measure between the topics present in the user preference with the topics present at micro-poi. The similarity between user preference (TD_u) and i^{th} micro-poi's topic distribution (TD_i) is computed as,

$$Sim(u,i) = \frac{\sum_{j=1}^{K} TD_u^j * TD_i^j}{||TD_u|| \cdot ||TD_i||}, \tag{6.16}$$

where, TD_u is the topic distribution for user, TD_i is the topic distribution for i^{th} micro-poi and K is the total number of topics present in the LDA model. Now, for personalized recommendation, the profitability equation is updated as follows,

$$\Pi_i = \Delta * \frac{G_a^i}{T_a^i} + \Theta * SPop_i + \theta * LPop_i + \varepsilon * Sim(u,i), \tag{6.17}$$

where, ε is the weight given to personal preference in computing profitability.

6.6 Experiments and Results

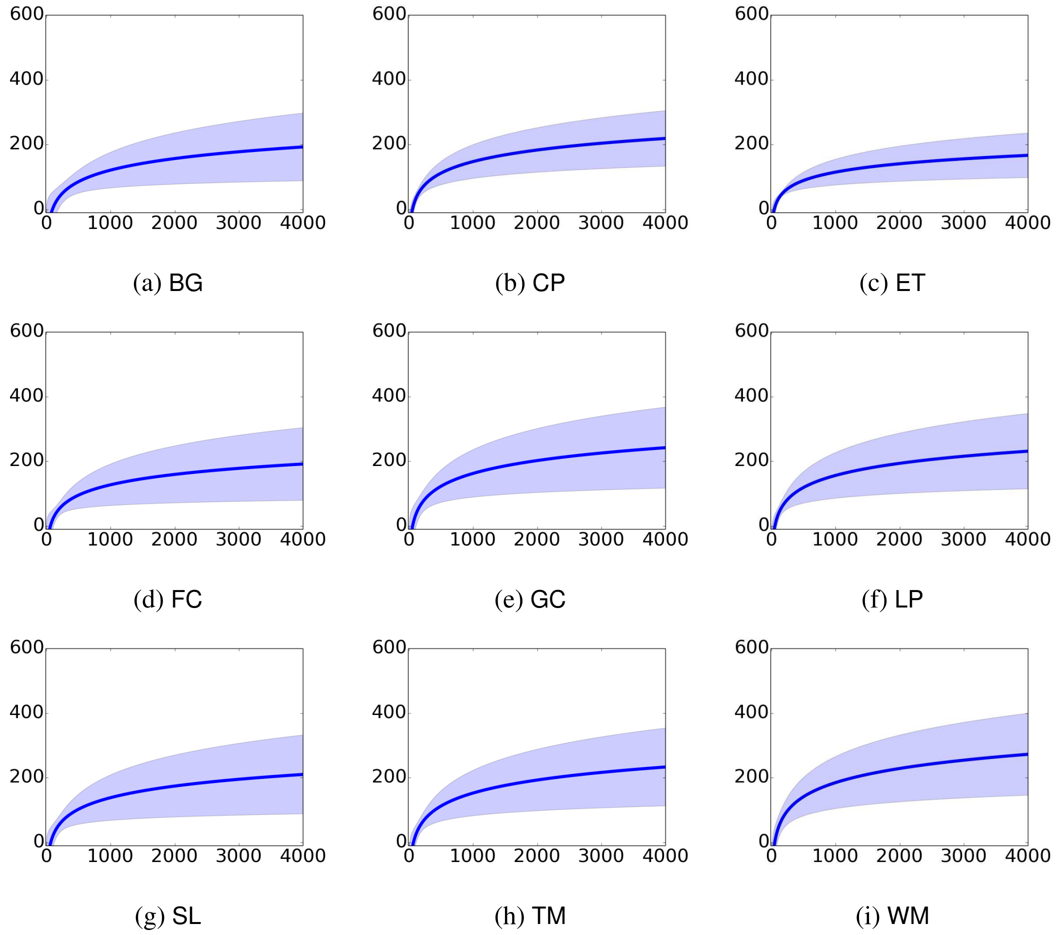
In this section, we will discuss the evaluation of the proposed method in terms of route recommendation.

6.6.1 Dataset

We use Flickr YFCC100M dataset [163] to create a dataset of around 330K images from 9 tourist locations around the world. The details of the dataset are provided in table 6.1.

TABLE 6.1: Details of the dataset along with the identified micro-pois, average trip durations and average R2 scores for the linear regression modeling of gain curves at each of the tourist location.(* trip time in seconds.)

Location	Total images	Unique users	Avg. photos per user	Total trips	Identified micro-pois	Avg. trip time*	R2 score
Botanical Gardens, Singapore (BG)	3914	229	17	305	69	3291	0.64
Central Park, New York, USA (CP)	127858	6269	20	10631	265	2267	0.59
Eiffel Tower, Paris, France (ET)	41303	4716	8	4678	90	2172	0.57
Forbidden City, Beijing, China (FC)	3481	317	10	278	82	2384	0.68
Grand Canyon, Arizona, USA (GC)	20310	1155	17	1491	160	3085	0.61
Leaning Tower of Pisa, Pisa, Italy (LP)	6854	681	10	594	128	2416	0.65
Statue of Liberty, New York, USA (SL)	6974	1222	5	663	116	1897	0.66
Taj Mahal, Agra, India (TM)	6152	487	12	406	87	3649	0.64
Washington Monument, DC, USA (WM)	113931	3917	29	7746	271	2716	0.55



(a) BG (b) CP (c) ET (d) FC (e) GC (f) LP (g) SL (h) TM (i) WM

FIGURE 6.4: Plots of average gain curves along with standard deviation learned employing regression analysis for each of the location. The average gain pattern and standard deviation is varying among different locations.

6.6.2 Micro_poi Identification

We employ generative model (GMM) to identify the micro-pois present in each location. The BIC score was measured to determine the number of micro-pois and we tested it for a range of 10-400 components. Table 6.1 presents the number of micro-poi identified at each of the location in the dataset. The dictionary of visual words was created using k-means clustering algorithm where we set the dictionary size to 1000. A topic modeling using LDA was performed to determine the visual content distribution of the micro-pois. We set the number of topics to 50, prior of document topic distribution to 0.02 and prior of word topic distribution to 0.02 for topic distribution learning.

Table 6.1 also shows the total number of trips for each tourist location. A trajectory is considered a trip only if it passes through at least 2 micro-pois.

6.6.3 Modeling Experience Gain

The information gain curve for each of the micro-poi is determined using equation 6.11. Equation 6.11 is converted to a linear equation by setting the value of Γ and taking log of the time (t) dimension. We employ Linear Regression to identify the parameters of the gain curve. Table 6.1 shows the average Coefficient of Determination (R2 score) for each of the micro-poi for this regression analysis. The trained gain curves are further utilized to determine the optimal stay time at each of micro-poi as we predict a tour for exploring a tourist location.

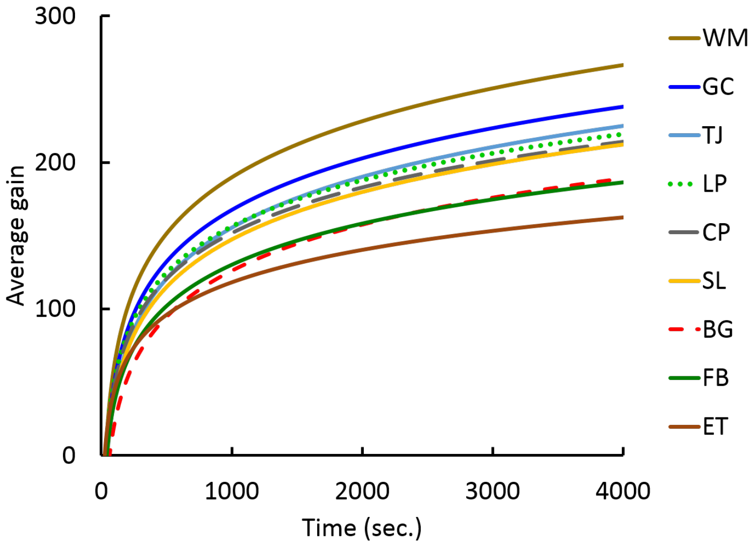
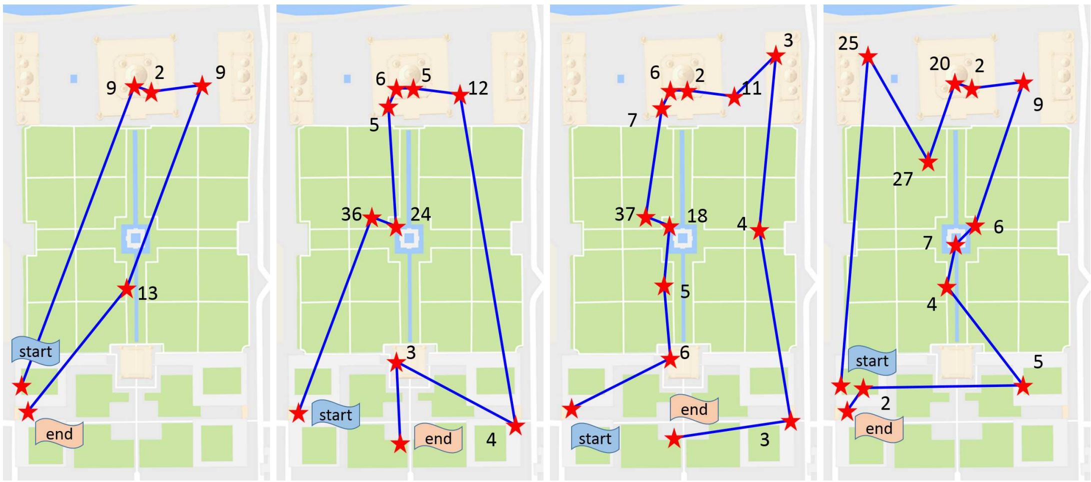


FIGURE 6.5: Average gain curve for locations in the dataset.

In figure 6.5 we have shown the average of all the gain curves corresponding to different micro-pois at each location for all tourist locations in our dataset. The variation in average gain curves corresponding to different tourist location shows different photography behavior of people at these locations. Apart from varying average gain pattern, we also observe different variation in gain among micro-pois from similar tourist location. Figure 6.4 shows the average gain curves along with standard deviation for each of the tourist locations. We can observe that each micro-poi has a different gain pattern. In addition, the variation in this gain pattern is also different for different tourist locations. In section 6.6.5.4, we will discuss how gain and visual diversity are correlated.

6.6.4 Path Recommendation

A tour recommendation is generated based on user visit time and trip duration. The tour includes a list of micro-pois, which should be visited in an order, and corresponding stay time at each micro-poi included in the path. Equation 6.13 is utilized to determine the list of micro-pois to include in the tour and corresponding stay time for each micro-poi is computed using Marginal Value Theorem (section 6.3.2).



(a) 1 PM, 1.5 hour (b) 10 AM, 3 hour (c) 10 AM, 4 hour (d) 1 PM, 4 hour

FIGURE 6.6: Sample tour recommendations at Taj Mahal for different visit times and varying trip durations. The star marks are micro-pois in the predicted tour and the number indicates a recommended stay time in minutes.

Figure 6.6 shows the recommended tours for different visit time and varying trip duration at Taj Mahal. We can observe how the trip path changes with a change in visit time and also a larger trip with more number of micro-pois is recommended for longer trip durations. In Figure 6.7, we have shown the recommended tour along with sample images captured at each of the micro-poi present in the tour (Figure 6.6b) for Taj Mahal location.

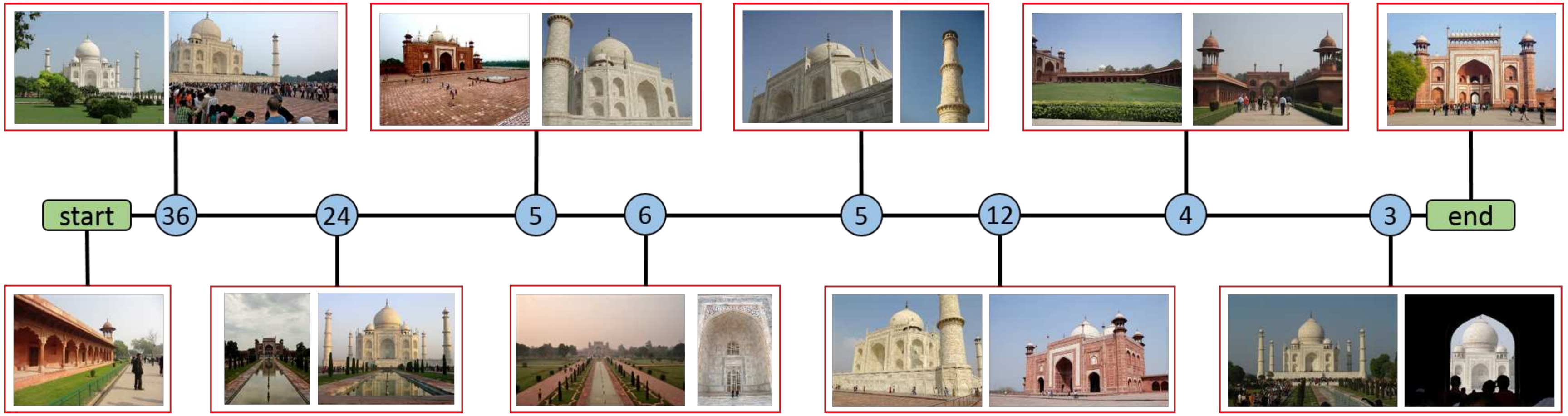
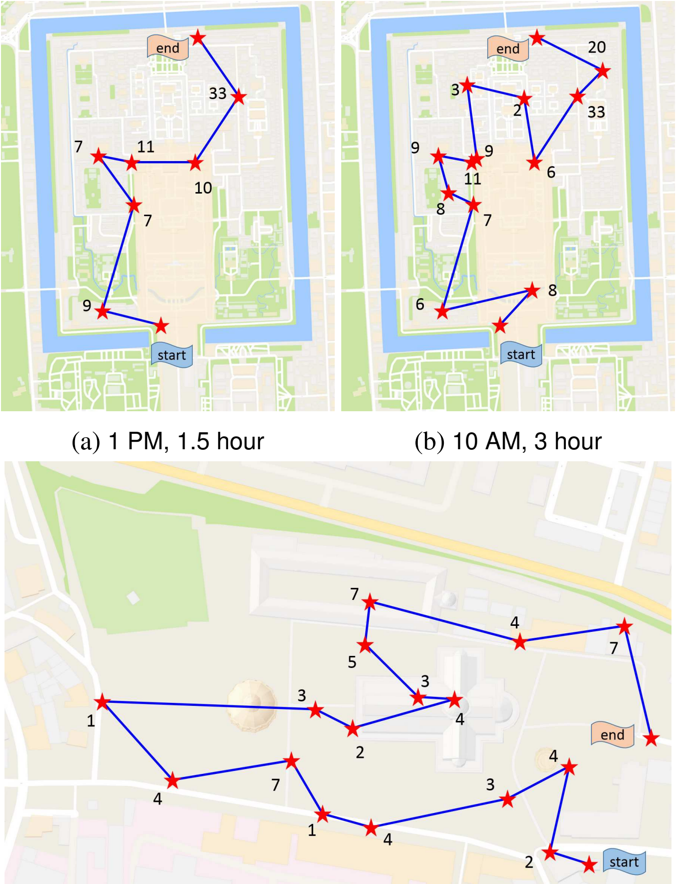


FIGURE 6.7: Visualization of recommended tour showing sample images captured at each of the micro-poi locations in the path for Taj Mahal location.

In Figure 6.8, we have shown recommended tours for Forbidden City and Leaning Tower of Pisa. We observe that the average stay time for micro-pois in the predicted tour at Leaning Tower of Pisa is around 4 minutes which is relatively lower as compare to Forbidden City (13 minutes) and Taj Mahal (12 minutes) locations. This information can be useful for tourists in making their selection for visiting tourist locations.



(a) 1 PM, 1.5 hour

(b) 10 AM, 3 hour

(c) 10 AM, 4 hour

FIGURE 6.8: Sample tour recommendations at Forbidden City and Leaning Tower of Pisa.

6.6.5 Evaluation

Quantitative evaluation of the recommended tours is a challenging task due to unavailability of ground truth. In addition, obtaining ground truth for varying user context (visit time and trip duration) is a non-trivial task. To overcome this difficulty, we make use of social media cues to determine the tours which are most popular among social media users.

We extract user trips at a tourist location which are popular on social media and meet certain criteria to establish ground truth trips. The criteria include minimum trip duration, which was set to 1 hour, and the minimum number of micro-pois in the tour, which was set to 8. To determine the popularity of a trip, quality score is computed for each of the photographs in the trip using equation 6.9. Then, an average score is computed for the complete trip using corresponding photographs and trips with a score >0.6 are considered popular. The photographs in these ground truth trips are kept for testing and excluded from the training dataset.

The proposed method is used to predict tour recommendations corresponding to the extracted ground truth trips. The user context (visit time and trip duration) of the ground truth trip is utilized to generate the recommended tour. The generated tour is evaluated based on its similarity to the ground truth trip. We propose three different metrics, micro-poi similarity, edge similarity and path similarity, for the evaluation.

The micro-poi similarity is measured based on the number of overlapping micro-pois in the ground truth trip and the recommended trip. It is computed as,

$$mpoi\_sim = \frac{n_{common}}{N_{mpoi}}, \tag{6.18}$$

where n_{common} is the number of common micro-pois in ground truth and recommended trips and N_{mpoi} is the total number of micro-pois in the recommended trip. Edge similarity between the two trips is computed as,

$$edge\_sim = \frac{e_{common}}{E_{mpoi}}, \tag{6.19}$$

where e_{common} is the number of common edges in ground truth and recommended trips and E_{mpoi} is the total number of edges in the recommended trip. An edge is defined as the path from one micro-poi to another in the trip. We compute the coefficient of determination (R2 score) to measure path similarity. For each micro-poi (poi_i) in the recommended trip, its closest micro-poi (poi_i^g) from the ground truth trip is determined. Then, the path similarity is computed as,

$$path\_sim = 1 - \frac{\sum_{i}^{N_{mpoi}} (poi_i^g - poi_i)^2}{\sum_{i}^{N_{mpoi}} (poi_i^g - \overline{poi}^g)^2}, \tag{6.20}$$

where $(poi_i^g - poi_i)$ represents the distance between corresponding micro-pois in the 3-dimensional space of latitude, longitude and time and $(\overline{poi}^g)$ is the mean position of the identified micro-pois in the ground truth trip. Finally, the average of these three similarity measures is computed for the evaluation.

6.6.5.1 Baseline

We propose three baseline methods to compare the generated recommendation results. The recommendation is generated using algorithm 4 for all the baselines with variation in the selection of micro-poi and prediction of stay time at each micro-poi. The first method (BL1) performs a random selection of micro-pois for path generation. In the second baseline (BL2), the social media popularity score $SPop$ is used for micro-poi selection and finally, in the third baseline (BL3), the micro-pois are selected based on the location-popularity score $(LPop)$. The parameter configuration in equation 6.13 will be ($\Delta = 0, \Theta = 1$ and $\theta = 0$) for BL2 and ($\Delta = 0, \Theta = 0$ and $\theta = 1$) for BL3. For all the baselines, average stay time of each micro-pois is considered instead of predicting the stay time using the proposed method which is based on Marginal Value Theorem.

TABLE 6.2: Quantitative comparison of the results for proposed and baseline methods. BL1, BL2, and BL3 are described in section 6.6.5.1. PR1 is the proposed method without using social and location popularity, PR2 is the proposed method which also makes use of social and location popularity and PR3 is the proposed personalized recommendation. (ST ratio is the stay time and travel time ratio for a trip.)

Method	Similarity score	net-gain	ST ratio
BL1	0.13	0.03	0.44
BL2	0.25	0.04	0.52
BL3	0.26	0.05	**2.94**
PR1	0.21	0.12	0.64
PR2	0.26	0.12	0.69
PR3	**0.27**	**0.13**	0.70

6.6.5.2 Comparison

The comparison results for the proposed and baseline methods is shown in table 6.2. We generate two different types of recommendation for the evaluation. The first method (PR1) is based only on the gain of micro-pois and uses the parameter settings of ($\Delta = 1, \Theta = 0$ and $\theta = 0$) for selecting micro-pois in equation 6.13. The second method (PR2) is based on gain, social media popularity and location popularity with a parameter setting of ($\Delta = 1, \Theta = 1$ and $\theta = 1$). We observe that the method based on social and location similarity performs better than PR1 in terms of path similarity. As we make use of social media popularity in selecting our ground truth, it may have some influence on the similarity measure. However, after integrating the social and location popularity in the proposed recommendation (PR2) we observe a higher similarity score.

To further investigate the quality of recommended path, we compute net-gain for each of the predicted trips and compare with the baseline methods. We observe that the proposed methods (PR1 and PR2) outperform the other baselines in terms of net-gain in the trip. In addition, we also measure the ratio of stay time and travel time. Although this ratio will be location dependent, a more favorable tour should have a balanced travel and stay time for a user to better enjoy the trip. The results are shown in column 3 and 4 of Table 6.2. In addition, we also observe that the method based on location popularity has a slightly higher stay/travel ratio as compared to other methods. The reason is that the location-popularity is computed based on the number of photographs captured at any micro-poi and hence the corresponding micro-poi may have a larger spatial area leading to a higher stay time. To validate this further we compute Spearman's rank correlation coefficient between stay time and location popularity (SPop). We found a weak positive correlation of 0.37 between the stay time and location popularity.

6.6.5.3 Personalization

We generate personalized recommended trips for each of the ground truth trips to evaluate personalized recommendation which employs equation 6.17 for selecting micro-poi locations. The personal preference of a user is determined by considering the photographs captured by the user as discussed in section 6.4.4 and section 6.5.1. The evaluation results are shown in table 6.2 (PR3). We can observe that adding personalization improves the performance in terms of path similarity with the ground truth trips while maintaining a higher net-gain and stay/travel ratio.

6.6.5.4 Gain and Stay Time Analysis

To validate the experience gain modeling at each of the micro-poi, we compare the actual gain observed in ground truth trips with the gain predicted using the models learned from social media images. We use the trained model to predict estimated gain at each micro-poi location in a ground truth trip based on the observed stay time. The quality of prediction is validated by computing a Mean Squared Error (MSE) using the actual net-gain (total-gain/trip-time) and predicted net-gain. We observe an average MSE score of 0.002 for the predicted net-gain as compared to the actual net-gain in the ground truth trips of all the locations.

We further analyze the recommended stay time at each of the micro-poi location included in the predicted path. An optimal stay time is predicted for each of the micro-poi in the ground truth

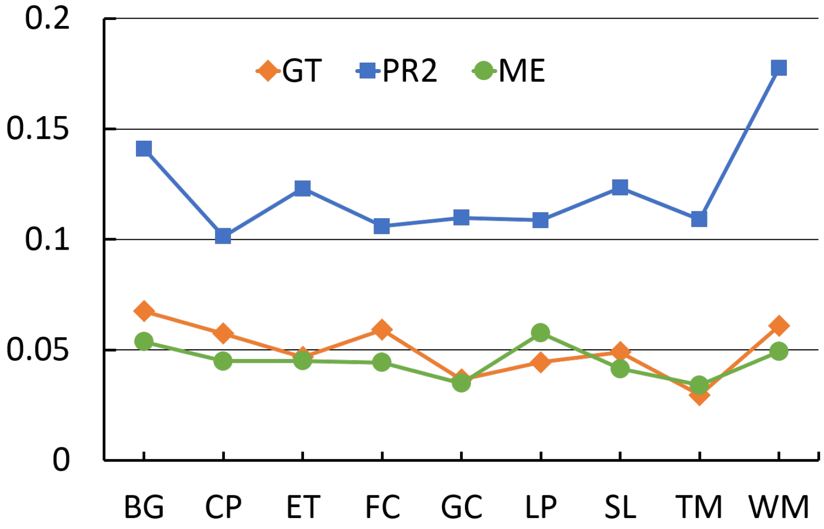


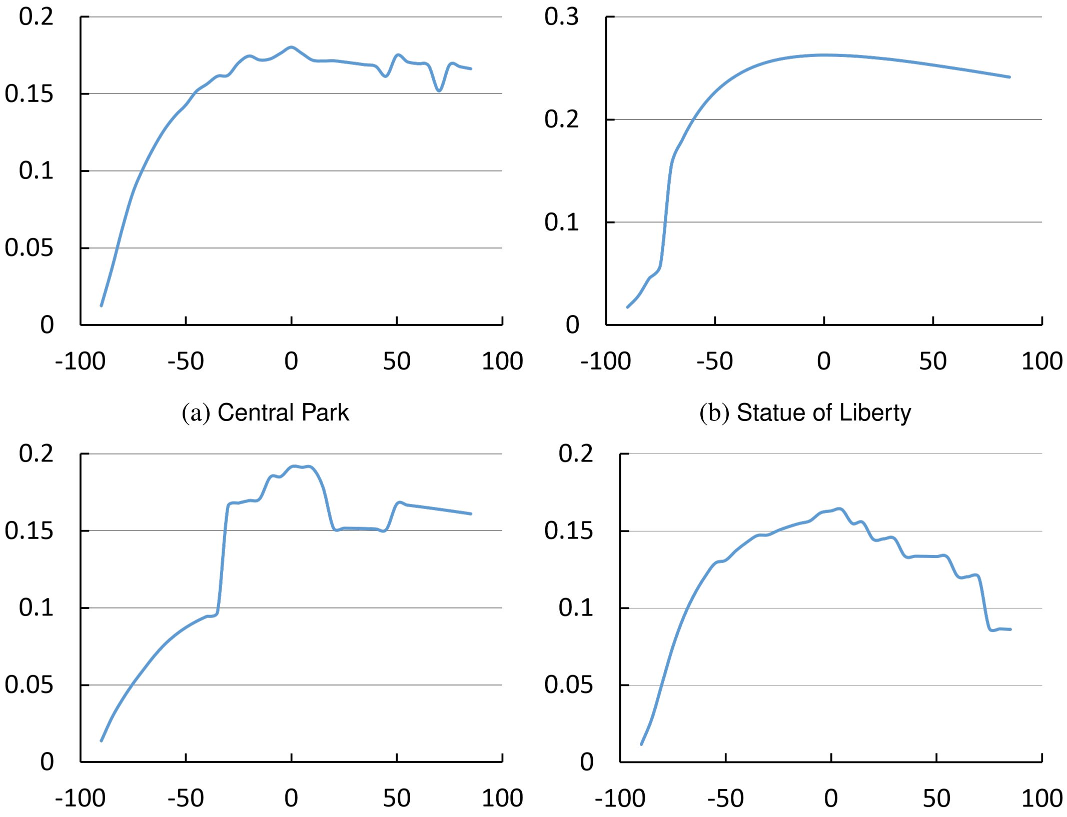
FIGURE 6.9: Comparison of net-gain observed using proposed method (PR2) with ground truth (GT) and employing mean values (ME) for different locations.

trips using the proposed method (PR2). Then, a net-gain is computed for the trip based on predicted stay time at each micro-poi location. We compare this net-gain with the actual net-gain in the ground truth trip and observed that the predicted stay time leads to a better net-gain. We also computed net-gain estimated when average stay time is used and found that it performs somewhat similar and sometimes worse than the ground truth gain. The comparison is shown in Figure 6.9 for all the nine locations in our dataset.

Figure 6.10 shows the variation of net-gain estimated for some predicted tours corresponding to a ground truth trips from different locations as we vary the stay time at each of the micro-poi location in the trip. The stay time estimated using MVT for each of the micro-poi in the predicted tour is varied as follows,

$$st^* = \frac{st_0 * \Lambda}{100}, \tag{6.21}$$

where, st_0 is the estimated stay time using MVT, st^* is the updated stay time and Λ is varied from -90 to +90. We can observe that the net-gain reduces as we move away from the optimal predicted stay time using MVT.



(a) Central Park

(b) Statue of Liberty

(c) Taj Mahal

(d) Washington Monument

FIGURE 6.10: Variation of net-gain from the recommended trip, as we change the stay time at micro-pois in the trip, corresponding to sample ground truth trips for different locations.

We observed varying gain patterns at different micro-pois. Since each micro-poi has different visual topic distribution, it can be one of the reasons for this variation in gain pattern. Therefore, we investigate the relation between gain and diversity of topic distribution at a micro-poi to understand the variation in gain patterns across different micro-pois in a location. To quantify the diversity of topic distribution at a micro-poi we employ Shanon's diversity index which is computed as,

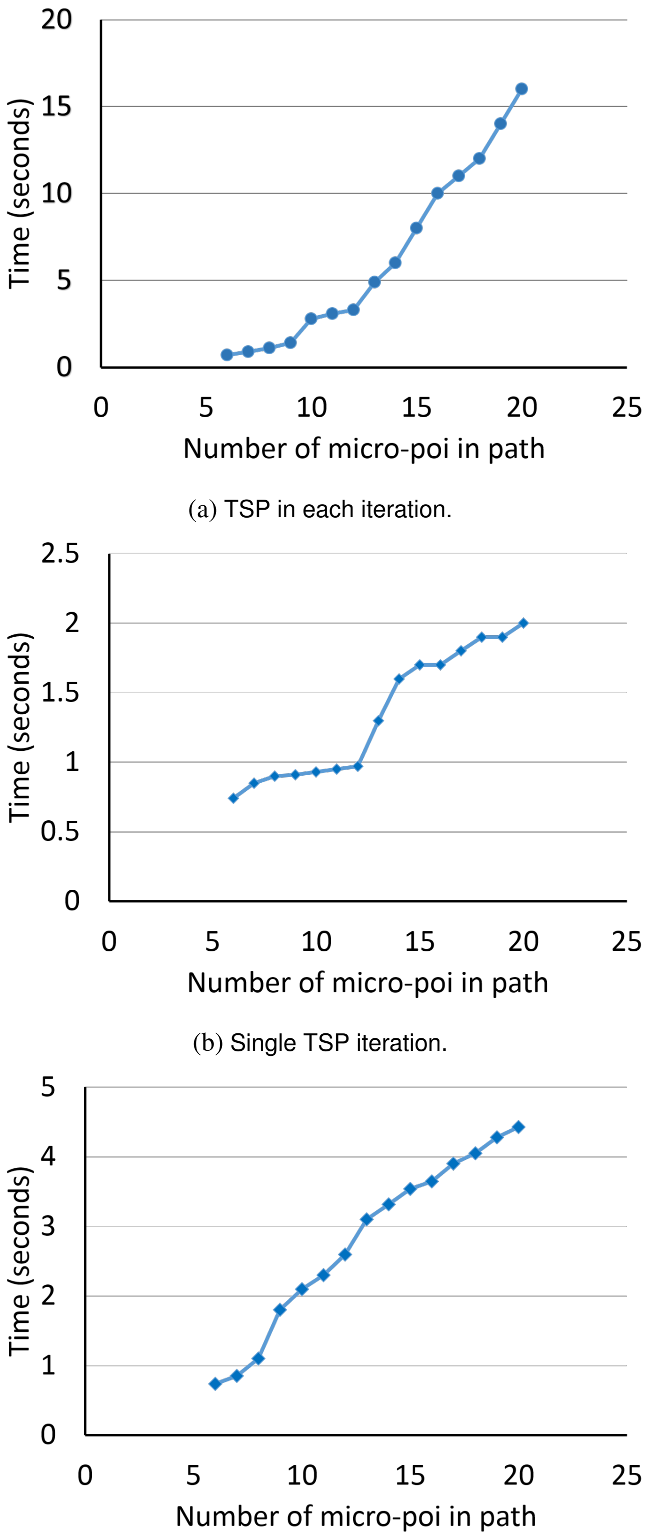
$$H_s = -\sum_{i}^{N_t} p_i ln(p_i), \tag{6.22}$$

where N_i is the total number topics present in the model (50 for our experiments) and p_i is the distribution of i^{th} topic at a micro-poi. We compute Spearman's rank correlation coefficient to find the correlation between gain and diversity of a micro-poi. We observe a positive weak correlation of 0.37 between the gain and diversity which indicates that visual diversity of a location has some impact on the observed gain. We also observe a moderate positive correlation of 0.56 between the stay time and observed gain which was expected as the gain at any location increases as we increase the stay time.

6.6.6 Running-time Analysis

The experiments for the proposed system were performed on a 8 core Intel processor running at 3.40 GHz and 8 GB of RAM using unoptimized python code. Solving Traveling Salesman Problem is the most time-consuming step in the recommendation process. The time required to determine an optimal path through a set of micro-pois also depend on the total number of micro-pois in the path. The average running-time to generate a path recommendation for a 2-hour tour for Taj Mahal location takes around 1.5 seconds. The running-time for path generation also varies for different locations as different locations will have a different number of micro-pois in the recommended path for a similar trip time.

We have shown the variation of running-time to generate a path with varying number of micro-pois in the tour in Figure 6.11a. We can observe that the running time increases exponentially with increase in the number of micro-pois in the path. This long running-time was mainly observed for Leaning Tower of Pisa location where the stay time at each micro-poi is smaller as compared to other locations. This leads to a relatively larger number of micro-pois in the recommended path. To overcome this problem, we analyze the running time of TSP algorithm for single iteration as we vary the number of micro-pois in the path (Figure 6.11b). A running time of 2 seconds was observed for a path with around 20 micro-pois.



(a) TSP in each iteration.

(b) Single TSP iteration.

(c) Minimizing TSP invocation.

FIGURE 6.11: Running time analysis.

We modify Algorithm 4 to reduce the running-time for path recommendation. The TSP invocation was not performed during each iteration of the algorithm and the average reach time of each micro-poi was utilized to update the trip time. We invoke TSP only when the total trip time is closer to the required trip time. This brings down the total running time for path recommendation to around 5 seconds even for a tour with 20 micro-pois.

6.7 Summary

In this work, we propose a trip recommendation method for photography and exploration of tourist locations based on OFT. The recommended trip includes a list of micro-poi locations a user should visit and corresponding stay time to spend at each micro-poi locations for capturing photographs. The recommendation can also be personalized based on the past photography behavior of a user. We evaluated the proposed method on a dataset drawn from YFCC100M [163] for 9 different tourist locations. The experimental results demonstrated the effectiveness of proposed method. The current work focuses on providing a recommendation based on optimal foraging behavior. However, different users may have different foraging behavior for photography and exploration and understanding individual user behavior is also important. Therefore, understanding the photography and exploration foraging behavior of users and employing it for personalized recommendations can be a future direction in this research.

Chapter 7

Conclusion and Future Work

In this thesis, we looked at the problem of providing real-time photography assistance to a user. We mainly focused on camera guidance for improving image quality and location recommendation for improving photography experience of users at tourist locations. We leveraged on social media content for generating relevant feedback for the user.

In chapter 3, we presented a context based photography learning method which can provide composition and camera parameter guidance to the user based on context. It can also provide human position and camera motion guidance to improve the image composition. We also presented the idea of photographic composition basis, *eigenrules* and *baserules* to substantiate the proposed composition learning. The idea of *eigenrules* and *baserules* can be further exploited to better understand photographic composition.

In chapter 4, we focused on obtaining a visual balance in an image frame and providing real-time assistance to users for capturing high-quality group photographs. We extended the idea of spring-electric graph model and augmented it with the concept of color energy to obtain visual balance in a system with elements of art. The proposed model for visual balance can have a wide range of applications in visual arts.

In chapter 4, we proposed *ClickSmart*, a method of viewpoint recommendation which leverage on publicly available images and social media cues to learn the photo-taking behavior of people.

We presented the idea of geo-pixels and defined their *popularity*, *quality* and *uniqueness* which are further utilized for viewpoint recommendation. We also investigated the role of contexts, such as time and weather conditions, for viewpoint recommendation in photography.

In chapter 6, we propose a trip recommendation method for photography and exploration of tourist locations based on OFT. The recommended trip includes a list of micro-poi locations a user should visit and corresponding stay time to spend at each micro-poi locations for capturing photographs. This work is focused on providing a recommendation based on optimal foraging behavior. However, different users may have different foraging behavior for photography and exploration. Therefore, understanding this behavior of users is also important which can be utilized further for personalized recommendations.

7.1 Future Work

In our future work, we plan to extend the current research to real-time videography assistance. To this end, we have identified the following set of problems for the future research.

7.1.1 Videography Assistance

The additional dimension of time and involvement of camera and user motion makes the problem of videography assistance more challenging as compared to photography assistance. We have factored videography assistance problem into three subproblems, which includes, understanding video aesthetics, computational characterizing of various cinematography shots used by professionals and further applying this knowledge to provide real-time assistance to a user.

7.1.1.1 Video Aesthetics

Understanding what makes a video professional and aesthetically pleasing is important for guiding a user to capture high-quality videos. Therefore, we first want to explore the factors which are responsible for the aesthetic quality of videos. We plan to employ Deep Learning framework along with the art of cinematography to better understand video aesthetics.

7.1.1.2 Cinematic Shot Characterization

The video shots in professional videos and movies are carefully captured by cinematographers which follow some rules and guidelines. Professional cinematographers make use of their experience and knowledge in cinematography to capture such cinematic shots. In this research problem, we aim to automatically identify and characterize various cinematographic shots based on some visual categories. This will allow us to understand video aesthetics on a higher level as compared to low-level features.

7.1.1.3 Real-time Assistance

The understanding of video aesthetics and computational knowledge of cinematic shots can be used to provide videography assistance to a user for taking high-quality videos. One of the challenges towards this goal is to find out which cinematography rules should be applied to user video. Application of video aesthetics to provide guidance in real-time is another issue as the too much computational processing in real-time will not be feasible. We plan to solve this problem by abstract classification of different video shots which will help us to infer the target cinematographic rule to be applied in a given user context.

References

[1] R. Abdullah, M. Christie, G. Schofield, C. Lino, and P. Olivier. Advanced composition in virtual camera control. In *International Symposium on Smart Graphics*, pages 13–24, 2011.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[3] R. Achanta and S. Süsstrunk. Saliency detection using maximum symmetric surround. In *IEEE International Conference on Image Processing*, pages 2653–2656, 2010.

[4] B. Adams, S. Venkatesh, and R. Jain. IMCE: Integrated media creation environment. *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(3):211–247, 2005.

[5] S. Ahern, M. Davis, D. Eckles, S. King, M. Naaman, R. Nair, M. Spasojevic, and J. Yang. Zonetag: Designing context-aware mobile media capture to increase participation. In *Proceedings of the Pervasive Image Capture and Sharing, 8th International Conference on Ubiquitous Computing*, 2006.

[6] R. Arnheim. *Art and visual perception : A psychology of the creative eye*. University of California Press, Berkeley, 2004.

[7] R. Arnheim. *The power of the center : A study of composition in the visual arts*. University of California Press, Berkeley, 2009.

[8] T. O. Aydın, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):31–42, 2015.

[9] S. Bae, A. Agarwala, and F. Durand. Computational rephotography. *ACM Transactions on Graphics*, 29(3):24:1–24:15, 2010.

[10] J. Baek, D. Pajak, K. Kim, K. Pulli, and M. Levoy. WYSIWYG computational photography via viewfinder editing. *ACM Transactions on Graphics*, 32(6):198:1–198:10, 2013.

[11] W. A. Bainbridge, P. Isola, I. Blank, and A. Oliva. Establishing a database for studying human face photograph memory. In *Annual Conference of the Cognitive Science Society*, page 1302–1307, 2012.

[12] S. Banerjee and B. Evans. In-camera automation of photographic composition rules. *IEEE Transactions on Image Processing*, 16(7):1807–1820, 2007.

[13] S. Banerjee and B. L. Evans. Unsupervised automation of photographic composition rules in digital still cameras. In *SPIE Conference on Sensors, Color, Cameras, and Systems for Digital Photography*, pages 364–373, 2004.

[14] W. Bares. A photographic composition assistant for intelligent virtual 3d camera systems. In *Smart Graphics*, volume 4073, pages 172–183. Springer, 2006.

[15] O. Barinova, V. Lempitsky, E. Tretiak, and P. Kohli. Geometric image parsing in man-made environments. In *Proceedings of the 11th European Conference on Computer Vision: Part II*, pages 57–70, 2010.

[16] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. Algorithms for drawing graphs: An annotated bibliography. *In Computational Geometry: Theory and Applications*, pages 235–282, 1994.

[17] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006.

[18] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the International Conference on Multimedia*, page 271–280, 2010.

[19] S. Bourke, K. McCarthy, and B. Smyth. The social camera: A case-study in contextual image recommendation. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, page 13–22, 2011.

[20] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[21] D. Butterfield, C. Fake, C. Henderson-Begg, and S. Mourachov. Interestingness ranking of media objects, 2006. US Patent US20060242139.

[22] Z. Byers, M. Dixion, W. D. Smart, and C. M. Grimm. Say cheese! experiences with a robot photographer. *AI Magazine*, 25(3):37–46, 2004.

[23] G. Cai, K. Lee, and I. Lee. Discovering common semantic trajectories from geo-tagged social media. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 320–332, 2016.

[24] B. Celikkale, A. Erdem, and E. Erdem. Visual attention-driven spatial pooling for image memorability. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 976–983, 2013.

[25] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[26] Y.-Y. Chang and H.-T. Chen. Finding good composition in panoramic scenes. In *IEEE 12th International Conference on Computer Vision*, pages 2225–2231, 2009.

[27] E. L. Charnov. Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9(2):129–136, 1976.

[28] D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 737–744, 2011.

[29] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou. A visual attention model for adapting images on small displays. *Multimedia Systems*, 9(4):353–364, 2003.

[30] Y.-Y. Chen, A.-J. Cheng, and W. H. Hsu. Travel recommendation by mining people attributes and travel group types from community-contributed photos. *IEEE Transactions on Multimedia*, 15(6):1283–1295, 2013.

[31] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu, and H.-Y. M. Liao. Personalized travel recommendation by mining people attributes from community-contributed photos. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 83–92, 2011.

[32] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. In *Proceedings of the International Conference on Multimedia*, page 291–300, 2010.

[33] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu. Color harmonization. In *ACM Transactions on Graphics*, volume 25, pages 624–630, 2006.

[34] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*, pages 761–770, 2009.

[35] M. Cristani, A. Perina, U. Castellani, and V. Murino. Content visualization and management of geo-located image databases. In *Extended Abstracts on Human Factors in Computing Systems*, page 2823–2828, 2008.

[36] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893 vol. 1, 2005.

[37] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III*, page 288–301, 2006.

[38] R. Datta, J. Li, and J. Z. Wang. Learning the consensus on visual quality for next-generation image management. In *Proceedings of the 15th International Conference on Multimedia*, pages 533–536, 2007.

[39] R. Datta and J. Z. Wang. ACQUINE: Aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *Proceedings of the International Conference on Multimedia Information Retrieval*, page 421–424, 2010.

[40] A. P. Dempster et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 1977.

[41] P. Denis, J. H. Elder, and F. J. Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, pages 197–210, 2008.

[42] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1664, 2011.

[43] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1338–1345, 2012.

[44] B. Dunstan. *Composing Your Paintings*. London, Studio Vista, 1979.

[45] P. Eades. A heuristics for graph drawing. *Congressus Numerantium*, 42:146–160, 1984.

[46] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal Computer Vision*, 59(2):167–181, 2004.

[47] M. Freeman. *The Photographer's Eye: Composition and Design for Better Digital Photos*. Focal Press, May 2007.

[48] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[49] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[50] H. Fu, X. Han, and Q. H. Phan. Data-driven suggestions for portrait posing. In *SIGGRAPH Asia 2013 Emerging Technologies*, pages 7:1–7:3, 2013.

[51] R. Gadde and K. Karlapalem. Aesthetic guideline driven photography by robots. In *Proceedings of the International Joint Conference on Artificial Intelligence - Volume III*, pages 2060–2065, 2011.

[52] K. D. Gavric, D. R. Culibrk, P. I. Lugonja, M. R. Mirkovic, and V. S. Crnojevic. Detecting attractive locations and tourists' dynamics using geo-referenced images. In *International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services*, volume 1, pages 208–211, 2011.

[53] A. Gionis, T. Lappas, K. Pelechrinis, and E. Terzi. Customized tour recommendations in urban areas. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 313–322, 2014.

[54] B. Gooch, E. Reinhard, C. Moulding, and P. Shirley. Artistic composition for image creation. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 83–88, 2001.

[55] T. Grill and M. Scanlon. *Photographic composition*. Amphoto, 1990.

[56] L. Guo, J. Shao, K. L. Tan, and Y. Yang. Wheretogo: Personalized travel recommendation for individuals and groups. In *2014 IEEE 15th International Conference on Mobile Data Management*, volume 1, pages 49–58, 2014.

[57] N. Hariri, B. Mobasher, and R. Burke. Query-driven context aware recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 9–16, 2013.

[58] C. Hauff. A study on the accuracy of flickr's geotag data. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1037–1040, 2013.

[59] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[60] J. Hedgecoe. *The Photographer's Handbook*. Knopf, New York, 1992.

[61] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864, 2010.

[62] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[63] C. S. Holling. Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist*, 91(07):385–398, 1959.

[64] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[65] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang. Integrating visual saliency and consistency for re-ranking image search results. *IEEE Transactions on Multimedia*, pages 653 – 661, 2011.

[66] Y.-T. Huang, K.-T. Chen, L.-C. Hsieh, W. Hsu, and Y.-F. Su. Detecting the directions of viewing landmarks for recommendation by large-scale user-contributed photos. In *Proceedings of the 20th ACM International Conference on Multimedia*, page 997–1000, 2012.

[67] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2011.

[68] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1469–1482, 2013.

[69] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–152, 2011.

[70] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[71] S. Iwabuchi, M. Kumano, M. Koseki, K. Ono, and M. Kimura. Visualizing attractive periods of popular photo spots using flickr data. In *ACM SIGGRAPH 2013 Posters*, page 112:1–112:1, 2013.

[72] R. Jacobson. *The Manual of Photography: Photographic and Digital Imaging.* Focal Press, 2000.

[73] A. Jahanian, J. Liu, Q. Lin, D. Tretter, E. O'Brien-Strain, S. C. Lee, N. Lyons, and J. Allebach. Recommendation system for automatic design of magazine covers. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, pages 95–106, 2013.

[74] S. Jain, S. Seufert, and S. Bedathur. Antourage: mining distance-constrained trips from flickr. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1121–1122, 2010.

[75] C. Jayant, H. Ji, S. White, and J. P. Bigham. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, page 203–210, 2011.

[76] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In A. Z. David Forsyth, Philip Torr, editor, *European Conference on Computer Vision*, volume I of *LNCS*, pages 304–317, 2008.

[77] S. Jiang, X. Qian, J. Shen, and T. Mei. Travel recommendation via author topic model based collaborative filtering. In *International Conference on Multimedia Modeling*, pages 392–402, 2015.

[78] D. Joshi, A. Gallagher, J. Yu, and J. Luo. Exploring user image tags for geo-location inference. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 5598–5601, 2010.

[79] D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, page 37–46, 2008.

[80] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7 – 15, 1989.

[81] Y. Kao, C. Wang, and K. Huang. Visual aesthetic quality assessment with a regression model. In *2015 IEEE International Conference on Image Processing*, pages 1583–1587, 2015.

[82] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, page 419–426, 2006.

[83] S. Kelby. *The digital photography book: the step-by-step secrets for how to make your photos look like the pros'!* Peachpit Press, [Berkeley, CA], 2006.

[84] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*, pages 35:1–35:4, 2012.

[85] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 305–313. 2012.

[86] J. Kim, S. Yoon, and V. Pavlovic. Relative spatial features for image memorability. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 761–764, 2013.

[87] J.-G. Kim, H. S. Chang, J. Kim, and H.-M. Kim. Efficient camera motion characterization for mpeg video indexing. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 1171–1174 vol.2, 2000.

[88] K. Kimura, H.-H. Huang, and K. Kawagoe. Photo-taking point recommendation with nested clustering. In *IEEE International Symposium on Multimedia*, pages 65–68, 2012.

[89] S. Kobayashi. *Color image scale*. Kosdansha International Distributed in the U.S. by Kodansha America, Tokyo New York New York, NY, 1991.

[90] C. Kofler, L. Caballero, M. Menendez, V. Occhialini, and M. Larson. Near2me: An authentic and personalized social media-based recommender for travel destinations. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 47–52, 2011.

[91] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[92] D. Laptev, A. Tikhonov, P. Serdyukov, and G. Gusev. Parameter-free discovery and recommendation of areas-of-interest. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2014.

[93] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Proceedings of the conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[94] T. V. Le, S. Liu, H. C. Lau, and R. Krishnan. Predicting bundles of spatial locations from learning revealed preference data. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1121–1129, 2015.

[95] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[96] H. Lee, E. Shechtman, J. Wang, and S. Lee. Automatic upright adjustment of photographs with robust camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2013.

[97] I. Lee, G. Cai, and K. Lee. Points-of-interest mining from people's photo-taking behavior. In *Hawaii International Conference on System Sciences*, pages 3129–3136, 2013.

[98] K. W.-T. Leung, D. L. Lee, and W.-C. Lee. Clr: a collaborative location recommendation framework based on co-clustering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 305–314, 2011.

[99] C. Li, A. C. Loui, and T. Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the International Conference on Multimedia*, page 827–830, 2010.

[100] H. Li, L. Yi, J. Tang, and X. Wang. Capturing a great photo via learning from community-contributed photo collections. In *Proceedings of the 19th ACM International Conference on Multimedia*, page 809–810, 2011.

[101] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In D. Forsyth, P. Torr, and A. Zisserman, editors, *European Conference on Computer Vision*, volume 5302 of *Lecture Notes in Computer Science*, pages 427–440. 2008.

[102] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classification in large-scale image collections. In *IEEE 12th International Conference on Computer Vision*, pages 1957–1964, 2009.

[103] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 831–840, 2014.

[104] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera. Personalized tour recommendation based on user interests and points of interest visit durations. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1778–1784, 2015.

[105] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2368–2382, 2011.

[106] H. Liu, T. Mei, J. Luo, H. Li, and S. Li. Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing. In *Proceedings of the 20th ACM International Conference on Multimedia*, page 9–18, 2012.

[107] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan. Discovering areas of interest with geo-tagged images and check-ins. In *Proceedings of the 20th ACM International Conference on Multimedia*, page 589–598, 2012.

[108] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. In *Computer Graphics Forum*, volume 29, page 469–478, 2010.

[109] Q. Liu, Y. Ge, Z. Li, E. Chen, and H. Xiong. Personalized travel package recommendation. In *2011 IEEE 11th International Conference on Data Mining*, pages 407–416, 2011.

[110] X. Liu, Y. Liu, K. Aberer, and C. Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 733–738, 2013.

[111] K.-Y. Lo, K.-H. Liu, and C.-S. Chen. Intelligent photographing interface with on-device aesthetic quality assessment. In J.-I. Park and J. Kim, editors, *Asian Conference on Computer Vision*, volume 7729 of *Lecture Notes in Computer Science*, pages 533–544. 2013.

[112] S. Lok, S. Feiner, and G. Ngai. Evaluation of visual balance for automated layout. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pages 101–108, 2004.

[113] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.

[114] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang. Photo2trip: generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 143–152, 2010.

[115] Z. Lu, X. Yang, W. Lin, H. Zha, and X. Chen. Inferring user image-search goals under the implicit guidance of users. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(3):394–406, 2014.

[116] C. Lujun, Y. Hongxun, S. Xiaoshuai, and Z. Hongming. Real-time viewfinder composition assessment and recommendation to mobile photographing. In *Advances in Multimedia Information Processing*, number 7674, pages 707–714. 2012.

[117] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In D. Forsyth, P. Torr, and A. Zisserman, editors, *European Conference on Computer Vision*, number 5304, pages 386–399. 2008.

[118] S. Ma, Y. Fan, and C. W. Chen. Finding your spot: A photography suggestion system for placing human in the scene. In *IEEE International Conference on Image Processing*, pages 556–560, 2014.

[119] S. Ma, Y. Fan, and C. W. Chen. Pose maker: A pose recommendation system for person in the landscape photographing. In *Proceedings of the International Conference on Multimedia*, pages 1053–1056, 2014.

[120] L. Mai, H. Le, Y. Niu, and F. Liu. Rule of thirds detection from photograph. In *IEEE International Symposium on Multimedia*, pages 91–96, 2011.

[121] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen. A system for mining interesting tourist locations and travel sequences from public geo-tagged photos. *Data & Knowledge Engineering*, 95:66–86, 2015.

[122] M. Mancas and O. Le Meur. Memorability of natural scenes: the role of attention. In *2013 20th IEEE International Conference on Image Processing*, 2013.

[123] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *IEEE 12th International Conference on Computer Vision*, pages 2232–2239, 2009.

[124] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE International Conference on Computer Vision*, pages 1784–1791, 2011.

[125] A. A. Michelson. *Studies in Opticss*. University of Chicago Press, 1927.

[126] H. Mitarai, Y. Itamiya, and A. Yoshitaka. Interactive photographic shooting assistance based on composition and saliency. In *Computational Science and Its Applications – ICCSA 2013*, volume 7975 of *Lecture Notes in Computer Science*, pages 348–363. 2013.

[127] S. Musse and D. Thalmann. A model of human crowd behavior : Group inter-relationship and collision detection analysis. In *Computer Animation and Simulation*, Eurographics, pages 39–51. 1997.

[128] B. Ni, M. Xu, B. Cheng, M. Wang, S. Yan, and Q. Tian. Learning to photograph: A compositional perspective. *IEEE Transactions on Multimedia*, 15(5):1138–1151, 2013.

[129] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *IEEE International Conference on Image Processing*, pages 3185–3188, 2010.

[130] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[131] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Modeling photo composition and its application to photo rearrangement. In *2012 19th IEEE International Conference on Image Processing*, pages 2741–2744, 2012.

[132] T. Phan, J. Zhou, S. Chang, J. Hu, and J. Lee. Collaborative recommendation of photo-taking geolocations. In *ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, 2014.

[133] P. Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2007.

[134] A. Popescu, G. Grefenstette, and P.-A. Moëllic. Mining tourist information from user-supplied collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 1713–1716, 2009.

[135] H. C. Purchase, J. Hamer, A. Jamieson, and O. Ryan. Investigating objective measures of web page aesthetics and usability. In *Proceedings of the Australasian User Interface Conference*, pages 19–28, 2011.

[136] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 103–110, 2007.

[137] Y. S. Rawat. Real-time assistance in multimedia capture using social media. In *Proceedings of the ACM Conference on Multimedia Conference*, pages 641–644, 2015.

[138] Y. S. Rawat and M. S. Kankanhalli. Context-based photography learning using crowdsourced images and social media. In *Proceedings of the ACM International Conference on Multimedia, Grand Challenge*, 2014.

[139] Y. S. Rawat and M. S. Kankanhalli. Context-aware photography learning for smart mobile devices. In *ACM Transactions on Multimedia Computing, Communications, and Applications*, pages 19:1–19:24, 2015.

[140] Y. S. Rawat and M. S. Kankanhalli. Clicksmart: A context-aware viewpoint recommendation system for mobile photography. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[141] Y. S. Rawat and M. S. Kankanhalli. Optimal foraging theory for photography and exploration. In *to be submitted*, 2016.

[142] Y. S. Rawat and M. S. Kankanhalli. A spring-electric graph model for socialized group photography. In *IEEE Transactions on Multimedia, (under review)*, 2016.

[143] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Transactions on Graphics*, 27(3):16:1–16:9, 2008.

[144] J. San Pedro. Synesthetic enrichment of mobile photography. In *Proceedings of the 2013 ACM International Workshop on Immersive Media Experiences*, page 41–44, 2013.

[145] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):pp. 461–464, 1978.

[146] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch. Automatic image retargeting. In *Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia*, pages 59–68, 2005.

[147] H. Shao, T. Svoboda, and L. Van Gool. Zubud-zurich buildings database for image based recognition. *Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Tech. Rep*, 260:20, 2003.

[148] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[149] M. Shirai, M. Hirota, H. Ishikawa, and S. Yokoyama. A method of area of interest and shooting spot detection using geo-tagged photographs. In *ACM SIGSPATIAL*, 2013.

[150] M. Shirai, M. Hirota, S. Yokoyama, N. Fukuta, and H. Ishikawa. Discovering multiple HotSpots using geo-tagged photographs. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, page 490–493, 2012.

[151] F. Simond, N. Arvanitopoulos Darginis, and S. Süsstrunk. Image Aesthetics Depends on Context. In *IEEE Proceedings of the International Conference on Image Processing*, 2015.

[152] J. Smolak. Candid town photography : “Exposure triangle – How to achieve correct exposure" (www.candidtown.com). Accessed: 2014-03-03.

[153] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM Transactions on Graphics*, volume 25, page 835–846, 2006.

[154] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal Computer Vision*, 80(2):189–210, 2008.

[155] D. W. Stephens and J. R. Krebs. *Foraging theory*. Princeton University Press, 1986.

[156] H.-H. Su, T.-W. Chen, C.-C. Kao, W. Hsu, and S.-Y. Chien. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, 14(3):833–843, 2012.

[157] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien. Scenic photo quality assessment with bag of aesthetics-preserving features. In *Proceedings of the 19th ACM International Conference on Multimedia*, page 1213–1216, 2011.

[158] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, pages 95–104, 2003.

[159] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computational visual attention model. In *Proceedings of the 17th ACM International Conference on Multimedia*, page 541–544, 2009.

[160] X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013.

[161] B. Thomee. Localization of points of interest from georeferenced and oriented photographs. In *Proceedings of the 2nd ACM international workshop on Geotagging and its applications in multimedia*, pages 19–24, 2013.

[162] B. Thomee, I. Arapakis, and D. A. Shamma. Finding social points of interest from georeferenced and oriented online photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 12(2):36, 2016.

[163] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: The new data in multimedia research. *ACM Communications*, 59(2):64–73, 2016.

[164] X. Tian, Z. Dong, K. Yang, and T. Mei. Query-dependent aesthetic model with deep learning for photo quality assessment. *IEEE Transactions on Multimedia*, 17(11):2035–2048, 2015.

[165] Timeanddate.com. Sunrise and sunset calculator. http://www.timeanddate.com/worldclock/sunrise.html, 2015. Last visited on March 3, 2014.

[166] H. Tong, M. Li, H.-J. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In K. Aizawa, Y. Nakamura, and S. Satoh, editors, *Advances in Multimedia Information Processing - PCM 2004*, volume 3331 of *Lecture Notes in Computer Science*, pages 198–205. 2005.

[167] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[168] B. Tversky and D. Baratz. Memory for faces: Are caricatures better than photographs? *Memory & Cognition*, 13(1):45–49, 1985.

[169] M. Vazquez and A. Steinfeld. An assisted photography method for street scenes. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, page 89–94, 2011.

[170] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511, 2001.

[171] P. Wang, W. Zhang, J. Li, and Y. Zhang. Online photography assistance by exploring geo-referenced photos on MID / UMPC. In *IEEE 10th Workshop on Multimedia Signal Processing*, pages 6–10, 2008.

[172] W. Wang, W. Lin, Y. Chen, J. Wu, J. Wang, and B. Sheng. Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach. In *European Conference on Computer Vision*, pages 756–771, 2014.

[173] X. Wang, Y.-L. Zhao, L. Nie, Y. Gao, W. Nie, Z.-J. Zha, and T.-S. Chua. Semantic-based location recommendation with multimodal venue semantics. *IEEE Transactions on Multimedia*, 17(3):409–419, 2015.

[174] Y. Wang et al. Where2stand: A human position recommendation system for souvenir photography. In *ACM Transactions on Intelligent Systems and Technology*, volume 7, page 9, 2015.

[175] L.-Y. Wei, Y. Zheng, and W.-C. Peng. Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*, pages 195–203, 2012.

[176] S. White, H. Ji, and J. P. Bigham. EasySnap: real-time audio feedback for blind photography. In *Adjunct Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, page 409–410, 2010.

[177] L.-K. Wong and K.-L. Low. Saliency-enhanced image aesthetics class prediction. In *IEEE International Conference on Image Processing (ICIP)*, pages 997–1000. IEEE, 2009.

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

[179] L.-K. Wong and K.-L. Wong. Enhancing visual dominance by semantics-preserving image recomposition. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 845–848. ACM, 2012.

[180] Wunderground.com. Weather forecast & reports. "http://www.wunderground.com/history/", 2015. Last visited on March 3, 2014.

[181] P. Xu, H. Yao, R. Ji, X.-M. Liu, and X. Sun. Where should i stand? learning based human position recommendation for mobile photographing. *Multimedia Tools and Applications*, 69(1):3–29, 2014.

[182] Z. Xu, L. Chen, and G. Chen. Topic based context-aware travel recommendation method exploiting geotagged photos. *Neurocomputing*, 155:99–107, 2015.

[183] T. Yamasaki, A. Gallagher, and T. Chen. Personalized intra-and inter-city travel recommendation using large-scale geotags. In *Proceedings of the 2nd ACM International Workshop on Geotagging and its Applications in Multimedia*, pages 25–30, 2013.

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

[185] Y. Yang, Z. Gong, and L. H. U. Identifying points of interest by self-tuning clustering. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 883–892, 2011.

[186] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li. OSCAR: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision*, 96(3):353–383, 2012.

[187] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *Proceedings of the International Conference on Multimedia*, page 211–220, 2010.

[188] W. Yin, T. Mei, and C. W. Chen. Assessing photo quality with geo-context and crowdsourced photos. In *IEEE Visual Communications and Image Processing*, pages 1–6, 2012.

[189] W. Yin, T. Mei, and C. W. Chen. Crowdsourced learning to photograph via mobile devices. In *IEEE International Conference on Multimedia and Expo*, pages 812–817, 2012.

[190] W. Yin, T. Mei, C. W. Chen, and S. Li. Socialized mobile photography: Learning to photograph with social context via mobile devices. *IEEE Transactions on Multimedia*, 16(1):184–200, 2014.

[191] H. Yoon, Y. Zheng, X. Xie, and W. Woo. Social itinerary recommendation from user-generated digital trails. *Personal and Ubiquitous Computing*, 16(5):469–484, 2012.

[192] F. X. Yu, R. Ji, and S.-F. Chang. Active query sensing for mobile location search. In *Proceedings of the 19th ACM international conference on Multimedia*, page 3–12, 2011.

[193] Q. Yuan, G. Cong, and A. Sun. Graph-based point-of-interest recommendation with geographical and temporal influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 659–668, 2014.

[194] J. Zahálka, S. Rudinac, and M. Worring. New yorker melange: Interactive brew of personalized venue recommendations. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 205–208, 2014.

[195] R. D. Zakia and D. A. Page. Photographic composition. pages i – ii. Focal Press, Oxford, 2011.

[196] H. Zettl. *Sight, Sound, Motion: Applied Media Aesthetics*. Wadsworth Pub. Co., 2010.

[197] C. Zhang, J. Gao, O. Wang, P. Georgel, R. Yang, J. Davis, J.-M. Frahm, and M. Pollefeys. Personal photograph enhancement using internet photo collections. *IEEE Transactions on Visualization and Computer Graphics*, 20(2):262–275, 2014.

[198] J.-D. Zhang, C.-Y. Chow, and Y. Li. Lore: Exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 103–112, 2014.

[199] J.-D. Zhang, C.-Y. Chow, and Y. Li. igeorec: A personalized and efficient geographical location recommendation framework. *IEEE Transactions on Services Computing*, 8(5):701–714, 2015.

[200] L. Zhang, Y. Gao, C. Zhang, H. Zhang, Q. Tian, and R. Zimmermann. Perception-guided multimodal feature fusion for photo aesthetics assessment. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, pages 237–246, 2014.

[201] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma. Auto cropping for digital photographs. In *IEEE International Conference on Multimedia and Expo*, pages 4–pp, 2005.

[202] W. Zhang and J. Kosecka. Image based localization in urban environments. In *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 33–40, 2006.

[203] Y. Zhang, X. Sun, H. Yao, L. Qin, and Q. Huang. Aesthetic composition represetation for portrait photographing recommendation. In *IEEE International Conference on Image Processing*, pages 2753–2756, 2012.

[204] Y. Zhang and R. Zimmermann. Camera shooting location recommendations for landmarks in geo-space. In *IEEE 21st International Symposium on Modeling, Analysis Simulation of Computer and Telecommunication Systems*, pages 172–181, 2013.

[205] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. *ACM Transactions on Intelligent Systems and Technology*, 5(3):50, 2014.

[206] Y.-T. Zheng, Y. Li, Z.-J. Zha, and T.-S. Chua. Mining travel patterns from gps-tagged photos. In *Advances in Multimedia Modeling*, volume 6523 of *Lecture Notes in Computer Science*, pages 262–272. 2011.

[207] Y.-T. Zheng, S. Yan, Z.-J. Zha, Y. Li, X. Zhou, T.-S. Chua, and R. Jain. GPSView: A scenic driving route planner. *ACM Transactions on Multimedia Computing, Communications and Applications*, 9(1):3:1–3:18, 2013.

[208] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua. Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology*, 3(3):56:1–56:18, 2012.

[209] C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa. Anaba: An obscure sightseeing spots discovering system. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2014.

List of Publications

Journal

- Y. S. Rawat and M. S. Kankanhalli. Context-aware photography learning for smart mobile devices. In *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2015.
- Y. S. Rawat and M. S. Kankanhalli. Clicksmart: A context-aware viewpoint recommendation system for mobile photography. In *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2016.
- Y. S. Rawat, M. Song, and M. S. Kankanhalli. A spring-electric graph model for socialized group photography. In *IEEE Transactions on Multimedia (TMM)*, **under review**.

Conference

- Y. S. Rawat. Real-time assistance in multimedia capture using social media. In *Proceedings of the ACM International Conference on Multimedia (MM), Doctoral Symposium*, 2015.
- Y. S. Rawat and M. S. Kankanhalli. Context-based photography learning using crowdsourced images and social media. In *Proceedings of the ACM International Conference on Multimedia (MM), Grand Challenge*, 2014.
- Y. S. Rawat and M. S. Kankanhalli. Optimal foraging theory for photography and exploration. **to be submitted**.

Appendix A

Rules of Photography

There are no fixed rules in photography, but there are guidelines which can often help us to enhance the impact of photographs. There are number of established composition guidelines which can be applied in almost any situation, to enhance the impact of a scene [47], [60]. Here we list some of the important composition rules of photography which are generally used as a guideline in capturing photographs.

The Rule of Thirds According to this rule an image should be imagined as divided into nine equal parts by two equally-spaced horizontal lines and two equally spaced vertical lines. With this grid in mind the *Rule of Thirds* now identifies four important parts of the image that we should consider placing points of interest as we frame the image. It also gives us four 'lines' that are also useful positions for elements in the photograph. Aligning a subject with these points creates more tension, energy and interest in the composition than simply centering the subject. If we place points of interest in the intersections or along the lines then the photo becomes more balanced and will enable a viewer of the image to interact with it more naturally. In figure A.1, the house is placed at one of the destined points and the lighthouse is aligned with one of the line.

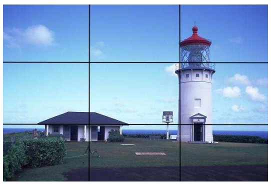
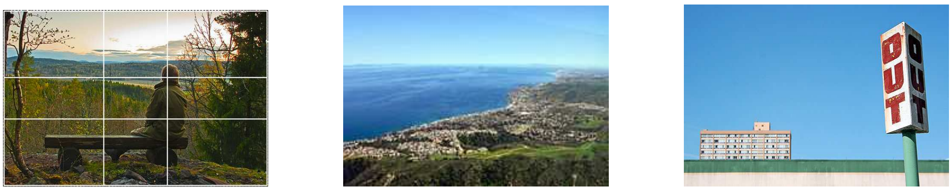


FIGURE A.1: *Rule of Thirds* [1]

The Golden Ratio Rule Similar to the *Rule of Thirds*, it is a way of dividing the image frame into rectangular segments. These 'golden rectangles' have proportions that the ancient Greeks thought to be especially harmonious and pleasing to the eye [47]. Placing compositional elements of importance either inside of or at the intersection of these rectangles can give them greater prominence and create a well-balanced image. This rule requires the ratio between areas of rectangles formed because of the horizon line be equal to the golden mean, 1.618, to

[1] Image source: www.photographymad.com

be more pleasing to the eye. In figure A.2a, the head of the person is placed on one of the intersection point of lines following golden ratio.



(a) *Golden Ratio Rule* [2] (b) *Rule of Diagonal* [3] (c) *Balancing Elements* [4]

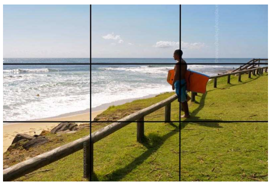
FIGURE A.2

Diagonal Rule One side of the picture is divided into two, and then each half is divided into three parts. The adjacent side is divided so that the lines connecting the resulting points form a diagonal frame. According to the Diagonal Rule, important elements of the picture should be placed along these diagonals. Linear elements, such as roads, waterways, and fences placed diagonally, are generally perceived as more dynamic than horizontally placed ones. In figure A.2b, the beach line is placed diagonally to make the photograph more pleasing.

Balancing Elements When we place the main subject off-centre, following the rule of thirds, it will create a more interesting photo, but it can leave a void in the scene which can make it feel empty. The scene should be balanced by the “weight" of the subject by including another object of lesser importance to fill the space. As we can see in figure A.2c, the photographer has placed the bigger object according to Rule of Thirds and then tried to balance it using a smaller object on the other side of the image.

Leading Lines When we look at a photo our eye is naturally drawn along lines. By thinking about how we place lines in your composition, we can affect the way we view the image, pulling us into the picture, towards the subject, or on a journey “through" the scene. There are many different types of line - straight, diagonal, curvy, zigzag, radial etc - and each can be used to enhance our photo's composition. In figure

FIGURE A.3: *Rule of Leading Lines* [5]



[2] Image source : http://dasawhartonphotography.wordpress.com

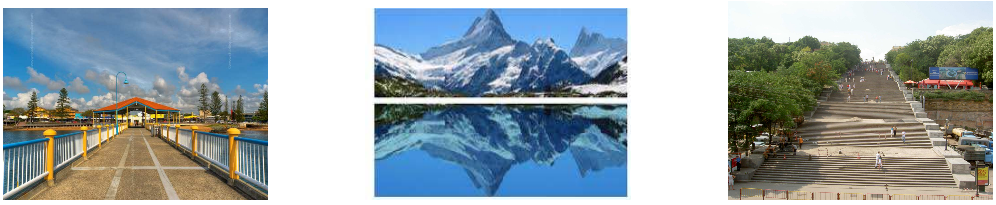
[3] Image source : www.wikipedia.org

[4] Image source: www.photographymad.com

[5] Image source : www.charlesphotoplace.com

A.3, the photographer has used *Rule of thirds, Diagonal* and *leading lines* to make the photograph more attractive.

Symmetry and Patterns We are surrounded by symmetry and patterns, both natural and man-made. They can make for very eye-catching compositions, particularly in situations where they are not expected. Another great way to use them is to break the symmetry or pattern in some way, introducing tension and a focal point to the scene. In figure A.4a and A.4b, we can see how the photographers have used symmetry in their captured images.



(a) Symmetry [5] (b) Symmetry [5] (c) View Point Selection[6]

FIGURE A.4: Rule of Symmetry and View Point Selection

Viewpoint Viewpoint has a massive impact on the composition of our photo, and as a result it can greatly affect the message that the shot conveys. Rather than just shooting from eye level, we can consider photographing from high above, down at ground level, from the side, from the back, from a long way away, from very close up, and so on. In figure A.4c, the photographer has captured the image from near ground level to produce a special effect.

Depth of Field Photography is a two-dimensional medium and we have to choose our composition carefully to convey the sense of depth which is present in the actual scene. We can create depth in a photo by including objects in the foreground, middle ground and background. Another useful composition technique is overlapping, where we deliberately partially obscure one object with another. The human eye naturally recognizes these layers and mentally separates them out, creating an image with more depth. In figure A.5, the photographer has used rock, waves, clouds and ocean, which are at different depth in the scene.

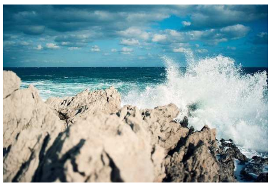


FIGURE A.5: Photograph with proper depth of field [7]

[6] Image source : www.wikipedia.org

[7] Image source : www.dpshots.com

Framing While focusing on the main subject of the scene we can also utilize surrounding objects to form a kind of frame for the main subject. By placing these objects around the edge of the composition we can help to isolate the main subject from the rest of the scene. The result is a more focused image which draws our eye naturally to the main point of interest. In figure A.6, the photographer has used the surrounding trees to create a frame for the house which is the main object in the scene.

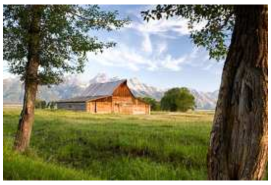


FIGURE A.6: Photograph with frame boundary [8]

Cropping Focus Most often a photo lacks impact because the main subject is so small it becomes lost among the clutter of its surroundings. By cropping tight around the subject we eliminate the background “noise", ensuring the subject gets the viewer’s undivided attention. This can be done using smaller field of view and focusing only on the main subject of the scene. In figure A.7, the photographer has used a smaller field of view and focused only on the main subject to make it distinctly visible to the viewer.

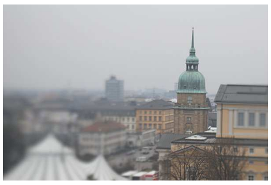


FIGURE A.7: Photograph with cropped focus [9]

[8] Image source: digital-photography-school.com

[9] Image source: jr-worldwi.de

Appendix B

Camera Controls

In this section we present some basic photographic terms which we used in this report. They are useful for adjusting various available controls in modern cameras for capturing good quality photographs [60], [83].

Aperture Aperture is the size of the opening in the lens when a picture is captures. It is measured in f-stops, for example f2.8, f8, f22.

TABLE B.1: Common Full Stop Aperture sizes

1	1.4	2	2.8	4	5.6	8	11	16	22

Since f-stops are actually fractions, the smaller the number the bigger the opening. 1/2 is greater than 1/2.8. Aperture f5.6 requires 16 times more light to expose correctly than f1.4. Aperture controls depth of field which is the area in the image which remain in focus, while the rest of the image gets blurry. A large aperture opening will produce a very shallow depth of field only keeping the subject in sharp focus, while blurring everything else. A small aperture opening(f22) will produce a very long depth of field showing most of the image in sharp focus.

Shutter speed Shutter speed is the amount of time the sensor is exposed to the light coming through the aperture.

TABLE B.2: Common Full Stop Shutter Speed

2	4	8	15	30	60	125	250	500	1000

Just like aperture, the values are fractions. 1/2 is greater than 1/4 which. Shutter speed 1/2 will let 16 times more light through, than 1/32. Shutter speed controls motion. When we set a very fast shutter speed, we will freeze that moment, but when we choose a much slower shutter speed and allow the subject to get a little bit blurry, even in a single frame we can simulate motion. We can show object being in motion when we choose the right shutter speed.

ISO ISO defines the sensor's light sensitivity rating. ISO-100 is the less sensitive and requires the most amount of light to expose correctly. ISO-1600 requires 16 times less light to expose correctly.

TABLE B.3: Common Full Stop ISO settings

100	200	400	800	1600	3200	6400

ISO controls the sensitivity rating of the sensor. When we increase the ISO from 200 to 400 we double the sensitivity of the sensor, which means that at 400 we only require half the light to properly expose as we did at 200. When we go from 200 to 800, we only need a quarter of the light to still achieve correct exposure.

F-stop F-Stop is the unit of measure for the aperture size. Increasing the aperture by 1 stop means opening it wider to allow twice as much light in. Decreasing the f-stop by 1 means making the aperture opening smaller to half the amount of light getting through.

Depth Of Field Depth of field is the area of the frame that is in focus. More precisely, its a the distance between the nearest and farthest objects in a scene that appear acceptably sharp in an image. In portrait photography people would be photographed with their faces being sharp, but the background blurry, which leads to shallow focus. In landscape on the other hand, most of the frame will be sharp causing deep focus.

Exposure Triangle Exposure triangle is a term used to describe a relationship between Aperture, Shutter Speed and ISO. These three factors are adjusted when composing a scene in different light conditions. Each factor is directly related to the other two, so changes to one of those three will have to be compensated with a change to one of the other two if we were to maintain the same exposure level.

White Balance White balance controls the colors in the images as accurate as possible. Sometimes images can come out with an orange, blue, yellow, etc, despite the fact that to the naked eye the scene looked quite normal. The reason for this is that, different sources of light have a different *color* for images. Fluorescent lighting adds a bluish cast to photos whereas tungsten (incandescent/bulbs) lights add a yellowish tinge to photos.

The way a digital camera produces image is that it reads raw data from the sensor, applies the setting of the camera to the raw data and produces the final image. Among other things, the digital camera needs to know the color of light before it can produce the final image. The WB (White Balance) setting on the digital camera is used to convey the color of light.