

IMAGE COLLECTION SUMMARIZATION WITH INTRINSIC PROPERTIES

GÖRÜNTÜ KÜMELERİNİN İÇSEL ÖZELLİKLER KULLANILARAK ÖZETLENMESİ

GÖKSU ERDOĞAN

ASSOC. PROF. MEHMET ERKUT ERDEM

Supervisor

Submitted to Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

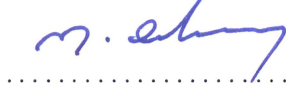
June 2018

This work named “**Image Collection Summarization with Intrinsic Properties**” by GÖKSU ERDOĞAN has been approved as a thesis for the Degree of **MASTER OF SCIENCE IN COMPUTER ENGINEERING** by the below mentioned Examining Committee Members.

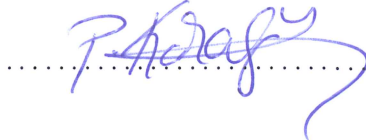
Prof. Dr. Pınar DUYGULU ŞAHİN
Head



Assoc. Prof. Mehmet Erkut ERDEM
Supervisor



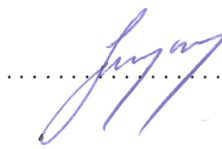
Prof. Dr. Pınar KARAGÖZ
Member



Assoc. Prof. Nazlı İKİZLER CİNBİŞ
Member



Asst. Prof. Hamdi Dibeklioglu
Member



This thesis has been approved as a thesis for the Degree of **MASTER OF SCIENCE IN COMPUTER ENGINEERING** by Board of Directors of the Institute for Graduate Studies in Science and Engineering.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Director of the Institute of
Graduate School of Science and Engineering

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

- ☒ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**

(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etmeniz ve kütüphane bu talebinizi yerine getirirse bile, tezinin arama motorlarının önbelleklerinde kalmaya devam edebilecektir.)

- ☐ **Tezimin/Raporumun tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**

(Bu sürenin sonunda uzatma için başvuruda bulunmadığım takdirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı ve ya tamamının fotokopisi alınabilir)

- ☐ **Tezimin/Raporumun tarihine kadar erişime açılmasını istemiyorum, ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

- ☐ **Serbest Seçenek/Yazarın Seçimi**

17 / 07 / 2018



Gökse ERDOĞAN

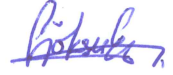
ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate Studies in Science of Hacettepe University,

I declare that

- all the information and documents have been obtained in the base of the academic rules
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- and any part of this thesis has not been presented as another thesis study at this or any other university.

25/06/2018



GÖKSU ERDOĞAN

ABSTRACT

IMAGE COLLECTION SUMMARIZATION WITH INTRINSIC PROPERTIES

Göksu ERDOĞAN

Master of Science, Computer Engineering Department

Supervisor: Asst. Prof. Dr. Mehmet Erkut ERDEM

June 2018, 93 pages

Visual summarization is an inherently complex process as its definition has some subjectivity in itself that there is nothing like a single perfect summary. In general, a good summary consists of two main properties which are (i) coverage and (ii) diversity. A good summary should have a high coverage, i.e it should consist of the key events and the concepts for a given set. At the same time, a good summary should also be diverse, i.e it should not consist of similar events and concepts. In addition to these two main properties, intrinsic image properties such as their aesthetics, their popularity, their sentiment, etc. are assumed to be important especially for the social media applications. In this thesis, we propose an automatic summarization method which considers intrinsic properties of images in addition to coverage and diversity for personal image collection summarization. To evaluate the proposed method, we collected two benchmark datasets where the ground truth summaries are obtained via crowdsourcing. Our experimental analysis reveals that taking intrinsic properties into account improves the summarization performances.

Keywords: image collection summarization, intrinsic properties

GENİŞLETİLMİŞ ÖZET

GÖRÜNTÜ KÜMELERİNİN İÇSEL ÖZELLİKLER KULLANILARAK ÖZETLENMESİ

Göksu ERDOĞAN

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Doç. Dr. Mehmet Erkut ERDEM

Haziran 2018, 93 sayfa

Günümüzde yaşanan teknolojik gelişmeler, akıllı telefonlar ile onlara bağlı olarak sosyal medya uygulamalarını hayatımızın vazgeçilmez bir parçası haline getirmiştir. Sosyal medya uygulamalarının özendirici rolü görüntü verisinin her geçen gün büyük bir hızla artmasına sebep olmaktadır. Görsel veri boyutundaki bu hızlı artış hem büyük hacimli veriyi baş etmek hem de bu veriden faydalanmak açısından zorluk yaratmaktadır. Görsel özetleme yöntemleriyle büyük hacimli veriler temsiliyet kapasitesi yüksek özetlerle ifade edilebilmekte ve hızla artmakta olan verinin sebep olduğu zorluklar aşılabilmektedir.

Görsel özetleme problemleri doğası gereği karmaşıktır. Zira özetin tanımı bir miktar belirsizlik taşımaktadır ve tek bir doğrudan söz edilemez. İyi bir özet, genellikle şu iki temel özelliğe bağlı olarak belirlenir: (i) kapsam ve (ii) çeşitlilik. Bir özetin iyi olması için kapsamının geniş olması gerekir. Başka bir deyişle, özet verilen kümedeki kilit olayları ve kavramları içermelidir. Aynı zamanda iyi bir özetin çeşitliliğinin de iyi olması gerekir yani aynı veya benzer olay ve kavramlar özette tekrarlanmamalıdır.

Çalışmamızda kişisel görüntü kümelerini otomatik özetleme işlemi için, bir özetin sahip olması gereken iki temel özelliği, kapsam ve çeşitlilik, şu yöntemlerle sağlanmıştır:

1. Determinantsal nokta süreçleri (ing. determinantal point processes)
2. Çeşitlilik sıralama (ing. diversity ranking)
3. Farklılık tabanlı seyrek alt küme seçimi (ing. dissimilarity based sparse subset selection)
4. Atlamalı yinelemeli nöral ağ (ing. skipping recurrent neural network)

Sosyal medya uygulamalarının yaygınlaşması ile kapsam ve çeşitliliğe ek olarak görüntülerin kalite, duygu uyandırma, popülerlik gibi içsel özellikleri de önem kazanmıştır. Buna bağlı olarak literatürde görüntülerden farklı içsel özellikler çıkarmaya dayanan yöntemler sıkça çalışılmaya başlamıştır. Çalışmamızda yüksek kaliteli özetler çıkarmak için faydalanılan içsel özellikler şunlardır:

1. Görüntünün insanlarda uyandırdığı duygular
2. Hafızada kalıcılık
3. Estetik
4. Manzaralılık

Özetleme problemlerinin bir diğer zorluğu özetin başarısının nicel olarak ölçülmesinin güç oluşudur. Bu problemin tek bir doğru çözümü olmadığı için uzman biri tarafından çıkarılacak bir özetin dayanak alınarak değerlendirme için kullanılması uygun değildir. Ayrıca özetleme problemi kişisel tercihlere dayandığı için yalnızca bir kişinin özetine göre değerlendirilmesi de doğru bir yaklaşım olmamaktadır. Güvenilir bir değerlendirme için birçok farklı kişiden toplanmış özetlere ihtiyaç duyulmaktadır. Ancak bu özetlerin arasındaki tutarlılıktan faydalanılarak başarılı bir değerlendirme yapmak söz konusudur. Kişisel görüntü kümelerini özetlemek için hali hazırda herkesin erişimine açık ve dayanak özetleri toplanmış bir veri kümesi bulunmadığı için kendi veri kümelerimizi toplamaya yöneldik.

EventSum ve CitySum adlarını verdiğimiz veri kümelerimiz Flickr'dan çekilen görüntü kümelerinden oluşmaktadır. Bu görüntüler İngilizce sorgular kullanılarak getirilmiştir. Her biri 100 görüntüden oluşan toplamda 19 ayrı görüntü kümesi toplanmıştır. Bu görüntü kümelerini özetleme probleminde dayanak olarak kullanılacak insan özetleri, özetleme problemine özel geliştirdiğimiz kendi web uygulamamız aracılığıyla toplanmıştır. Kullanıcılardan gördükleri görüntü kümesini en iyi temsil eden 10 görüntüyü seçmeleri istenmiştir. Kullanıcılara görüntüler zaman etiketlerine göre sıralı bir şekilde gösterilmiştir. Ayrıca iyi bir özetin iki temel özelliği olan kapsam ve çeşitlilik birer cümleyle tanımlanarak kullanıcı bilgilendirilmiştir.

Kısaca, çalışmamızda kişisel görüntü kümelerini otomatik özetleme işlemi için, bir özetin sahip olması gereken iki temel özelliğin yanısıra görüntünün içsel özelliklerini de hesaba katan bir yöntem önerilmiştir. Geliştirilen bu yöntem çeşitli kişisel görüntü kümelerinin kitle kaynaklı toplanan özetleri ile kıyaslanmış ve içsel özelliklerin dikkate alınmasının elde edilen özetleri iyileştirdiği gözlemlenmiştir.

Bu çalışmada bilgisayarlı görü literatürüne yaptığımız katkılar şu şekilde özetlenebilir:

1. Kişisel görüntü kümesi özetleme problemi için kullanıma açık etiketli bir veri kümesi bulunmadığından, veri kümesi toplama yoluna gidilmiştir. Veri kümesindeki her kişisel görüntü kümesi için, dayanak olarak kullanmak üzere kitle kaynaklı (ing. crowdsourcing) olarak kullanıcılardan özetler toplanmıştır.
2. Kişisel görüntü kümesi özetlerken görüntünün içeriğinin yanısıra görüntünün içsel özelliklerinin kullanılmasının çıkarılan özetlerin başarısına etkisi gösterilmiştir.
3. Görüntü kümesi özetleme probleminde insan davranışı analiz edilmiştir.
4. Kullanılan yöntemlerin, özniteliklerden bağımsız olduğu gösterilmiştir.
5. Çıkarılan özetlerin başarısının, farklı değerlendirme yöntemleri karşısında tutarlı olduğu gösterilmiştir.

Anahtar Kelimeler: görüntü kümesi özetleme, içsel özellikler, determinantsal nokta süreçleri, çeşitlilik sıralama, kitle kaynaklı veri kümesi toplama



ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors Assoc. Prof. Mehmet Erkut Erdem and Assoc. Prof. İbrahim Aykut Erdem for their guidance.

Besides, I would like to thank my thesis committee members, Assoc. Prof. Nazlı İkizler Cinbiş, Prof. Dr. Pınar Duygulu Şahin, Prof. Dr. Pınar Karagöz, Asst. Prof. Hamdi Dibeklioglu for reviewing this thesis and giving insightful comments.

I am grateful to Hacettepe University Computer Vision Laboratory (HUCVL) members Çağdaş Baş and Aysun Koçak for their support throughout my thesis work. I am also grateful to all lab members for the nice research environment.

I would like to thank every participant who contributed with their summaries to form the datasets in this thesis and Bora Çelikkale who helped me in every step of the dataset collection process.

I am deeply grateful to Taylan Mor for his support for everything, especially for his countless reviews during the writing of this thesis.

I am deeply grateful to my family, especially to my mother Hatice Erdoğan, for their support throughout my life.

This thesis was partially supported by a grant from The Scientific and Technological Research Council of Turkey (TUBITAK) – Career Development Award 113E497.

CONTENTS

	<u>Page</u>
ABSTRACT	i
ÖZET	ii
ACKNOWLEDGEMENTS	vi
CONTENTS	vii
FIGURES	xii
TABLES	xiii
ABBREVIATIONS.....	xiv
1. INTRODUCTION.....	1
1.1. Scope of the thesis	1
1.2. Contributions	3
1.3. Organization.....	4
2. BACKGROUND	5
2.1. Evaluation Metrics	5
2.1.1. F-Measure	5
2.1.2. V-Rouge	7
2.2. Visual Representations.....	7
2.2.1. Superpixel Representation	8
2.2.2. SIFT Representation	8
2.2.3. CNN Representation	8
2.3. Summarization Methods	10
2.3.1. Determinantal Point Processes.....	10
2.3.2. Submodular Functions	11
2.3.3. Dissimilarity Based Sparse Subset Selection.....	12
2.3.4. Recurrent Neural Networks	12
3. RELATED WORK	14
3.1. Unsupervised Approaches	14
3.2. Supervised Approaches	15

3.3. Intrinsic Image Properties	15
4. OUR APPROACH	18
4.1. Datasets	18
4.1.1. Data Pruning	18
4.1.2. Vacation Photographs	20
4.1.3. EventSum and CitySum.....	20
4.2. Feature Extraction.....	24
4.2.1. Colour Representation	24
4.2.2. Superpixel Representation	25
4.2.3. CNN Representation	26
4.2.4. RMAC Representation.....	26
4.3. Baselines.....	27
4.3.1. Uniform Summarization	28
4.3.2. K-Means Summarization	28
4.3.3. Dissimilarity Based Sparse Subset Selection.....	29
4.3.4. Skipping Recurrent Neural Network	29
4.4. Intrinsic Properties	30
4.4.1. Memorability	30
4.4.2. Sentiments.....	31
4.4.3. Scenicness	34
4.4.4. Aesthetics.....	36
4.5. Proposed Summarization Methods	38
4.5.1. Determinantal Point Processes.....	38
4.5.2. Diversity Ranking	39
5. EXPERIMENTAL ANALYSIS	41
5.1. Quantitative Evaluation	41
5.2. Qualitative Evaluation	50
6. CONCLUSION.....	56
A User Interfaces to Collect Summaries	58
B Image Collections	60

REFERENCES	70
------------------	----



FIGURES

	<u>Page</u>
2.1. Near duplicate images are treated as different for F-Measure	6
2.2. AlexNet [1] architecture	9
4.1. The selected summaries before (left) and after pruning (right) for the Istanbul Tour image collection B3.	20
4.2. Summary distribution of EventSum and CitySum datasets	21
4.3. Example human summary with the maximum F-Measure from an Istanbul Tour image collection B3.	23
4.4. Example human summary with the minimum F-Measure from an Istanbul Tour image collection B3.	24
4.5. Scale invariance example	27
4.6. The most memorable image (left) and the less memorable image (right) in an Istanbul Tour image collection B3.	31
4.7. Distribution of the memorability scores across the images being (red) and not being (blue) parts of the summaries for the EventSum (top) and the CitySum (bottom) datasets	32
4.8. Image with the most positive emotion (left) and image with the most negative emotion (right) in an Istanbul Tour image collection B3.	33
4.9. Distribution of the sentiment scores across the images being (red) and not being (blue) parts of the summaries for the EventSum (top) and the CitySum (bottom) datasets	34
4.10. Image with the maximum scenicness score (left) and image with the mini- mum scenicness score(right) in an Istanbul Tour image collection B3.	35
4.11. Distribution of the scenicness scores across the images being (red) and not being (blue) parts of the summaries for the EventSum (top) and the CitySum (bottom) datasets	36

4.12. Image with the maximum aesthetic score (left) and image with the minimum aesthetic score(right) in an Istanbul Tour image collection B3.....	37
4.13. Distribution of the aesthetic scores across the images being (red) and not being parts (blue) of the summaries for the EventSum (top) and the CitySum (bottom) datasets	38
5.1. Example uniform summary from Istanbul Tour image collection B3.....	51
5.2. Example K-Means summary with maximum F-Measure on an Istanbul Tour image collection B3.	51
5.3. Example K-Means summary with minimum F-Measure on an Istanbul Tour image collection B3.	52
5.4. Example DS3 summary on an Istanbul Tour image collection B3.	52
5.5. Example SRNN summary on an Istanbul Tour image collection B3.	52
5.6. Example DPP summary with the maximum F-Measure on an Istanbul Tour image collection B3.	53
5.7. Example DPP summary with the minimum F-Measure on an Istanbul Tour image collection B3.	53
5.8. Example DR summary on an Istanbul Tour image collection B3.	54
5.9. Example DPP summary with intrinsic property scenicness on an Istanbul Tour image collection B3.	54
5.10. Example DR summary with intrinsic property aesthetics on an Istanbul Tour image collection B3.	55
5.11. Example DPP summary with triplet intrinsic property combination on an Istanbul Tour image collection B3.....	55
5.12. Example DR summary with triplet intrinsic property combination on an Istanbul Tour image collection B3.	55
A1. A snapshot of our summary collection page	58
B1. A Trip to New York image collection from EventSum dataset	60
B2. A Visit to Italy image collection from EventSum dataset	61

B3. An Istanbul Tour image collection from EventSum dataset.....	61
B4. Birthday Party image collection from EventSum dataset	62
B5. Burning Man image collection from EventSum dataset.....	62
B6. London Olympics image collection from EventSum dataset	63
B7. St. Patrick’s Day image collection from EventSum dataset.....	63
B8. World Cup image collection from EventSum dataset	64
B9. Amsterdam Vacation image collection from CitySum dataset	64
B10. Amsterdam Vacation image collection from CitySum dataset	65
B11. Tokyo Vacation image collection from CitySum dataset.....	65
B12. Tokyo Vacation image collection from CitySum dataset.....	66
B13. Venice Vacation image collection from CitySum dataset	66
B14. Venice Vacation image collection from CitySum dataset	67
B15. Paris Vacation image collection from CitySum dataset	67
B16. Paris Vacation image collection from CitySum dataset	68
B17. Paris Vacation image collection from CitySum dataset	68
B18. Paris Vacation image collection from CitySum dataset	69
B19. Paris Vacation image collection from CitySum dataset	69

TABLES

4.1. The consistency of the human summaries based on the V-Rouge and F-Measure metrics on the EventSum dataset.....	22
4.2. The consistency of the human summaries based on the V-Rouge and F-Measure metrics on the CitySum dataset	23
5.1. Comparison of baseline methods on the EventSum dataset with F-Measure metric	41
5.2. Comparison of baseline methods on the CitySum with F-Measure metric.....	42
5.3. Comparison of summarization methods	43
5.4. Effect of features on EventSum with F-Measure metric	44
5.5. Effect of features on CitySum with F-Measure metric	44
5.6. Comparison of methods on Vacation Photographs dataset with F-Measure metric	45
5.7. Effect of intrinsic properties on DPP with F-measure metric	45
5.8. Results with different features EventCitySum with DPP with F-Measure metric	46
5.9. Results with different features EventCitySum with DR with F-Measure metric	47
5.10. Effect of intrinsic properties on EventCitySum with F-Measure metric	47
5.11. Combination of intrinsic properties on EventCitySum with F-Measure metric .	48
5.12. Comparison of summarization methods with intrinsic properties on EventCitySum dataset	48
5.13. Comparison of summarization methods with different set sizes on EventCitySum dataset with F-measure metric	49
5.14. Comparison of summarization methods with intrinsic properties on Vacation Photographs dataset with F-Measure metric	49
5.15. Comparison of summarization methods with intrinsic properties on Vacation Photographs dataset with V-Rouge metric	50

ABBREVIATIONS

BoW	Bag of Words
SIFT	Scale-Invariant Feature Transform
CNN	Convolutional Neural Networks
DPP	Determinantal Point Processes
DR	Diversity Ranking
RNN	Recurrent Neural Network
SRNN	Skipping Recurrent Neural Network
LSTM	Long Short Term Memory
RMAC	Regional Maximum Activation of Convolution
SVR	Support Vector Regression
AVA	Aesthetic Visual Analysis
ADMM	Alternating Direction Method of Multipliers

1. INTRODUCTION

Thanks to the cutting edge technological developments, smartphones have become a crucial part of our lives. The communication functionality, which is the primary function of telephones, has been expanding with these developments. Assembly of the camera to mobile phones can be thought of as a turning point for this technology, this is why we can not imagine a phone without a camera today. This advancement has two impacts on both technology and humans. First, it opens a new gate for future technological developments such as applications on visual data. These applications cause an explosion in the amount of visual data. Understanding and analyzing that large scale visual data become hot topic. Second, it changes human habits through these applications such as social media. Album creation, for example, gains popularity through social media applications. Everybody wants to immortalize his/her memories through photos on special occasions such as a birthday party. For this reason, they create albums from these photos by selecting the good photos with different people at different moments. However, they need to look at all the photos, which is exhausting and time-consuming at the same time, to decide whether or not to add to an album. Today, there are a lot of travelers, who share their experiences during the travel on their blogs. They take a lot of photos about landmark locations of the cities, traditional foods, museums and so on. Then they create albums from these photographs to share in their blogs after the same time-consuming process that is stated above. These reasons show that why automatic summary extraction gains importance.

1.1. Scope of the thesis

This thesis mainly focuses on automatic summarization of image collections. A summarization method should consider two key properties: (i) coverage and (ii) diversity which are definitive properties of a good summary. We employ the following methods to leverage these two critical properties:

1. Determinantal point processes (DPP) [2]

2. Diversity ranking (DR) [3]
3. Dissimilarity based sparse subset selection [4]
4. Skipping recurrent neural network (SRNN) [5]

Although these properties are critical for a good summary, quality is another important property [6]. A considerable amount of literature has been published on extracting intrinsic properties while it depends on a vivid imagination. For example, people have an enthusiasm for being appreciated by other people on social media. They also want that their photos get many likes on social media or evoke beautiful emotions within them. This requirement can be formulated as a popularity property such as in [7]. On the other hand, another person might want to stick in the mind of his/her friends by his/her image posts. We can associate this need with memorability score of an image similar to [8]. The term intrinsic properties refer to abstract properties of the image and it can be easily extended to any kind of information hidden in the image. By using these intrinsic properties, we aim to include human preferences to album summarization. In this work, we employ the following intrinsic properties to generate high-quality summaries:

1. Sentiments [9]
2. Memorability [8]
3. Scenicness [10]
4. Aesthetics [11]

We investigate the effect of intrinsic image properties on the personal image collection summarization problem. We integrate intrinsic image properties to the summarization problem by the following methods to have two main properties of summaries which are coverage and diversity:

1. Determinantal point processes (DPP) [2]

2. Diversity ranking (DR) [3]

Quantitative evaluation of summaries is very critical in summarization problems because of two reasons. First, there is no unique summary which is extracted by a domain expert that we can assume that as a correct ground truth for the data in consideration. The second reason is that the summarization depends on human preferences, so using only one human summary could reflect only one human preference. We need enough summary to generalize problem for humans. Reliable evaluation can be performed solely by patterns across all summaries. Unfortunately, there is no publicly available visual summarization dataset with a rich set of human reference summaries in the literature, hence we tend toward collecting our own datasets.

Our datasets, referred to as EventSum and CitySum, are collected from the Flickr by queries in English. We implement our website to collect reference human summaries. We ask people to choose ten images that represent the image collection in the best way. These ground truth summaries are necessary for evaluation while evaluation metrics require them. We employ two evaluation metrics to evaluate our summaries which are F-Measure [12] and V-Rouge [13]. F-Measure is a straightforward approach while it measures based on images directly. On the other hand, V-Rouge compares visual words instead of images. These two metrics complete each other for a fair evaluation.

1.2. Contributions

Our main contributions are listed below:

- We collect two photo collection datasets which can be used for benchmarking for visual summarization.

- We perform summarization by using intrinsic properties in addition to image content. These intrinsic properties are integrated into the summarization process to mimic human preferences. We show that taking intrinsic properties into account increases the quality of automatically generated summaries.
- We analyze the image collection summarization behavior of humans by performing a consistency analysis.
- We show that intrinsic properties improve the summarization performance across different evaluation metrics and different image representations.

1.3. Organization

The organization of the thesis is as follows.

Chapter 2. provides the necessary background for the summarization techniques and their evaluation. Chapter 3. gives a brief overview of the recent literature on the image collection summarization problem.

Chapter 4. presents the proposed methods, and gives the details of the collected datasets. Briefly, we present the intrinsic properties used in our experiments and our approach for the summarization frameworks. Chapter 5. presents the analysis of the experiments with the baseline methods and the summarization frameworks by which we integrate intrinsic properties.

Chapter 6. we give our final words on summarization problem and possible future research directions.

We also add the Appendices Section in which we explain the details of this work which are not presented in the main chapters not to destroy the integrity of the thesis. In Appendix A, we introduce our user interfaces which are used for ground truth summary collection. we present the detailed analysis of our datasets. Last of all, in Appendix B, we present the image collections in our datasets.

2. BACKGROUND

In this chapter, we present the key concepts concisely. First and foremost, we introduce the evaluation metrics used for measuring the quality of summaries. Then, we explain visual representations which are used within the employed summarization methods. In the following section, we review common computer vision techniques which are especially related to the methods in our work.

2.1. Evaluation Metrics

One challenge in summarization problems is the quantitative evaluation of summaries with regard to the user preferences. Summary extraction does not have a straightforward solution such as object classification. We can easily distinguish a cat or dog with a few exceptions. On the other hand, in the summarization problem, solution space is not separated by explicit lines as in object classification problem, there is a dependency on human preferences. Because of this reason, evaluation metrics are not straightforward too. In this section, we give two different evaluation metrics to measure how good a summary is. These features are V-Rouge [13] and F-Measure [12] and they are only meaningful when reference ground truth summaries exist.

2.1.1. F-Measure

F-Measure is a commonly used metric in information retrieval literature. This metric is previously used for video summarization problem to analyze human consistency [12]. It is also applicable to measure how much ground truth summaries and generated summaries correlate with each other. F-Measure is defined as the harmonic mean of precision and recall values as follows:

$$\overline{F}_i = \frac{1}{N-1} \sum_{j=1, j \neq i}^N 2 \frac{p_{ij} \cdot r_{ij}}{p_{ij} + r_{ij}} \quad (1)$$

In the equation, N denotes the number of human subjects. p_{ij} is precision and r_{ij} is recall of human selection i using selection j as ground truth.

The F-Measure takes the value of 1 in the best case scenario, whereas in the worst case scenario it gets a value of 0. The advantage of this measure is that no additional data such as features is necessary while it is directly computed over images. On the other hand, the disadvantage of this measure is some images can be very similar to each other and only one of them should be included in the summary. These very similar images are also called near duplicates in literature. Two near duplicate images in Burning Man image collection are shown in Figure 2.1..



FIGURE 2.1.: Near duplicate images are treated as different for F-Measure

F-Measure treats these two images as different images. If the automatic summary selects the left image to the summary and human subject selects the right image to the summary, F-Measure produces low scores. However, an ideal summarization metric should handle that these images are near duplicates and it should produce high scores.

2.1.2. V-Rouge

Rouge metric is a popular metric for text summarization. V-Rouge metric [13] is proposed as an extension of Rouge metric to visual domain. It is computed by the following formula:

$$r_S(A) = \frac{\sum_{w \in W} \sum_{s \in S} \min(c_w(A), c_w(S))}{\sum_{w \in W} \sum_{s \in S} c_w(S)} \quad (2)$$

where S is a set of human-generated reference summaries and W denotes a set of visual words. $c_w(A)$ represents occurrence counts of visual word w in summary A . V-Rouge is proposed for evaluating image collection summarization based on a precomputed set of visual words. Because of this reason, feature extraction is needed as a first step. The intuition behind the measure is that a good summary has similar visual words with the reference ground truth summaries. The main advantage of the V-Rouge metric over the F-Measure metric is that the V-Rouge metric does not suffer from near duplicates while it works at the lower level than F-Measure. In other words, the V-Rouge method accepts two images in the Figure 2.1. as identical based on the visual word similarities between them.

2.2. Visual Representations

In computer vision, images can be represented by the variety of problem-specific features. These features can be interpreted as summaries of images from a specific point of view such as edges, corners, and objects. One common approach in feature extraction process is using Bag of Words (BoW) representation over these features. BoW representation is proposed on textual data in the first place, as the name implies. However, the motivation behind this representation is useful for computer vision field too. In the computer vision field, it basically corresponds to a vector whose elements are occurrence counts of visual words in the images. For example, if one dimension of our feature vector corresponds to wheels and we have four of them, a classifier might classify that image as a car.

Here, we present two classical hand-crafted feature representations before the rise of Convolutional Neural Network (CNN) features which are superpixel representation and Scale-Invariant Feature Transform (SIFT) representation. Then, we briefly describe the CNN based representation while it produces state-of-the-art results in many computer vision problem.

2.2.1. Superpixel Representation

Images have millions of pixels thanks to high-resolution cameras in use today. However, using each pixel as the primitive feature is not feasible for many methods. Superpixel is a group of pixels with a similar color or brightness values. The aim of superpixel representation is reducing the complexity by grouping pixels and dealing with less visual units. It is more useful than using a fixed sized square or rectangular patches because of the similarity between pixels. We employ the quick shift clustering algorithm [14] to find segments in the image.

2.2.2. SIFT Representation

One of the most popular feature representation is the SIFT [15]. SIFT features are extracted from several key points in the image. These key points correspond to the area around a specific x,y position in the image. SIFT feature is basically a histogram of the gradients of the specific key point. SIFT owes its popularity to a rotation and scale invariance.

2.2.3. CNN Representation

Thus far, we explain the so-called shallow conventional feature extraction methods. We now turn to the deep convolutional neural networks. Deep CNNs have revolutionized the computer vision field in recent years, obtaining state-of-the-art results in many different problem domains. CNN is one specific method that employs a deep learning approach.

Despite the long history of artificial neural networks in the literature, they recently become a hot topic due to the technological developments. In the first place, this approach is not feasible because of power and data limitations. Nowadays, by virtue of GPUs and large-scale datasets available for public access on the Internet, deep learning becomes reachable for everyone. The deep term refers to the number of layers which is used to model solution between input and output. This depth provides high representation power to model any complex relationship between an input and an output.

CNN features are used to encode high-level information that corresponds to objects and scene information in the image. AlexNet [1] is one of the architecture that highlights deep CNNs. Alexnet architecture is shown in Figure 2.2.

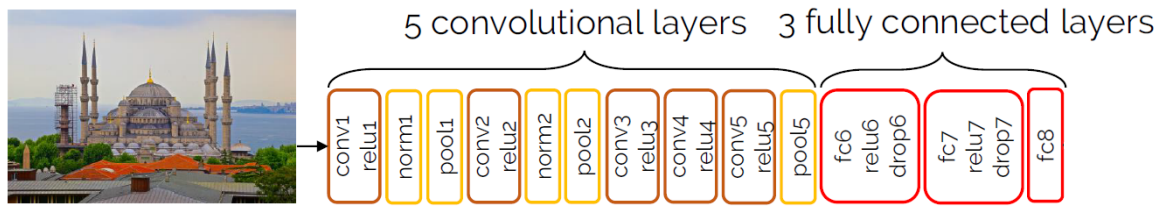


FIGURE 2.2.: AlexNet [1] architecture

Some of these layers are common among various CNN methods. These common layers are input layer, convolutional layer, pooling layer, fully connected layer and output layer. In the computer vision field naturally, inputs are images and outputs differ based on the problem. For instance in the AlexNet input is an image and output is a probability distribution of object classes. Convolutional layers are the main strength of CNNs. The responsibility of convolutional layers changes with depth due to the hierarchical structure. For example, the first convolutional layers become special for low-level features of images such as edges, while the high-level convolutional layers become special for high-level features such as object and scenes. Pooling layer is responsible to provide dimension reduction and invariance from position and rotation. Last important layer is a fully connected layer that combines responses in the previous layer. The output of the fully connected layers replaced the conventional image representations depending on high representation power.

2.3. Summarization Methods

In this section, we explain the general summarization frameworks which are employed throughout this thesis work. First, we discuss DPP which is directly used as a summarization framework. Then, we give submodular functions which are necessary to understand the DR method. In the next section, we explain DS3 subset selection method. Lastly, we summarize SRNN, whose skipping connections provide powerful summarization framework.

2.3.1. Determinantal Point Processes

DPP is a method for diverse subset selection which makes it a qualifiable method for summarization problem. DPP is a probabilistic model which takes advantage of determinant computation, as its name implies. It uses negative correlations to sample diverse examples. These negative correlations are represented on similarity matrices.

Assume that, we have a set $\mathcal{Y}$ with N elements: $\mathcal{Y} = \{1, 2, 3, \dots, N\}$. $2^{\mathcal{Y}}$ denotes subsets of $\mathcal{Y}$. DPP on $\mathcal{Y}$ corresponds to probability measure for any $Y \subseteq \mathcal{Y}$ on $\mathcal{Y}$ denoted by $\mathcal{P}$ [2]. L matrix denotes DPP kernel on L_Y matrix denotes submatrix of L for indexes Y . L is a real symmetric matrix. Probability of Y is :

$$\mathcal{P}(Y \subseteq \mathcal{Y}) = \det(L_Y) \quad (3)$$

Another representation of DPP is marginal kernel K . Transformation between K and L matrix is pretty easy with $K = (L + I)^{-1}L$ formulation. To understand how DPP models similarities, K kernel is more appropriate because this kernel carries all necessary information to compute the probability of any subset Y .

$$\begin{aligned}
\mathcal{P}(i, j \in \mathcal{Y}) &= \begin{vmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{vmatrix} \\
&= K_{ii}K_{jj} - K_{ij}K_{ji} \\
&= \mathcal{P}(i \in \mathcal{Y})\mathcal{P}(j \in \mathcal{Y}) - K_{ij}^2
\end{aligned} \tag{4}$$

For example, if $Y = \{i\}$ corresponds to $\mathcal{P}(i \in \mathcal{Y}) = K_{ii}$. In this case, K_{ii} represents the probability of individual inclusion. When this value gets closer to 1, its probability to be chosen gets closer to a hundred percent. Let's extend the example with a set whose size is 2 e.g $Y = \{i, j\}$. Diagonal represents individual probabilities as in the previous example. On the other hand, off-diagonal elements represent the negative relationship between pairs. In other words, increase in K_{ij} decrease the coexistence of them. This property ensures the main diversity condition on DPP. If $K_{ij} = \sqrt{K_{ii} \cdot K_{jj}}$, then we infer that element i and element j is very similar, so we can make a strong assumption that they are not going to be included to the same summary.

Until this point, we discussed how DPP provides diversity by using a similarity kernel. This kernel can be denoted by a Gram matrix: $L = B^T B$. As the last step, they find the solution with a polynomial algorithm which has the exponential complexity.

2.3.2. Submodular Functions

Submodular functions have a wide area of applications for various problems such as game theory, active learning, sensor placement, document summarization and image collection summarization. These problems are addressed by submodular function while submodular functions provide diminishing returns property. Based on this property, set with $A \subseteq B \subseteq V - v$ provides $f(A \cup v) - f(A) \geq f(B \cup v) - f(B)$ relationship. This property guarantees that adding a new element to a small set will produce more gain than adding a new element to

a big set. This property provides diversity intrinsically. The main advantage of these methods is finding an optimum solution with a greedy algorithm.

2.3.3. Dissimilarity Based Sparse Subset Selection

As the name implies, dissimilarity based sparse subset selection uses pairwise dissimilarities to choose the subset of the set under interest. These pairwise dissimilarities come from two different sets which are source set and target set. Although it is proposed to perform subset selection for two different sets at the beginning, it is applicable when these two sets are identical. They formulate the problem as row-sparsity regularized trace minimization problem which is NP-hard which NP-hard. They employ the Alternating Direction Method of Multipliers (ADMM) framework which decreases the complexity to quadratic form. This method is referred to the DS3 term in this work. This method is reliable against outliers in both source and target sets.

2.3.4. Recurrent Neural Networks

Recurrent neural networks are a special kind of neural networks which focus on modeling time series data. This kind of modeling is useful regarding on prevalence of sequential data over nature. It is also densely used in vision and language domains problems like image captioning, sentiment classification and machine translation. Sequential models need to take historical relationship into consideration in the data. It can be thought of as a memory which is the main strength of Recurrent Neural Networks (RNN).

SRNN [5] introduces the skipping connections to train the RNN for visual summarization. As it is mentioned, one important point is eliminating consecutive images in the summarization problem. Instead of learning each connection between consecutive images which are not needed, they skip these consecutive images and model connection between remote images.

They formulate the problem as a maximization over loglikelihood on the future images. However, including all possible subsets to computation is not feasible. This limitation forces them to use the Expectation Maximization algorithm to solve the problem. Unsupervised learning is the main advantage of SRNN for visual summarization, while it eliminates ground truth data collection which is a great challenge for summarization.



3. RELATED WORK

In this chapter, we provide a brief overview of the literature that is related to visual summarization problem. We discuss unsupervised approaches in section 3.1.. Then, we present supervised approaches for summarization problem in section 3.2.. Lastly, in section 3.3., we explain the various intrinsic properties studied in the literature.

3.1. Unsupervised Approaches

As the name implies, unsupervised learning is a machine learning approach without supervision which is ground truth labeled data. This section attempts to provide a brief summary of the literature relating to unsupervised methods that are employed throughout this thesis work.

Summarization can be formulated as a subset selection problem under coverage and diversity constraints. Although there are many subset selection methods, only a few of them have our problem-specific constraints or support for the integration of them. Determinantal Point Processes (DPP) [16] is one of these subset selection methods whose main focus is choosing diverse elements of the set. K-DPP [2] is a variant of DPP with a fixed subset size that leverages Newton identities. Elhamifar et al. [4] proposed a method for subset selection problem based on pairwise dissimilarities between the source and target sets by using the trace minimization with a regularized row-sparsity term.

Submodular functions are set functions which have a diminishing returns property which is pretty common through summarization problems while it provides both diversity and coverage. Diversity ranking method [3] optimizes a submodular function. This property guarantees that adding a new element to a small set will produce more gain than adding a new element to a big set. The main advantage of these methods is finding the optimum solution with a greedy algorithm. Singla et al. [17] take advantage of diminishing returns property with adaptive sampling strategy to select images that are marginally significant.

Sequential modeling is important while we come across frequently in the real world this is why deep learning framework also applied to sequential problems. SRNN [5] focuses on extracting long-term correlations by skipping images to select ordered images from the sequence. Unlike RNN [18] and Long Short Term Memory (LSTM) [19], SRNN does not model consecutive relationships, models transitions between images which have skipping connections.

3.2. Supervised Approaches

One common approach before deep learning dominates the literature was the Support Vector Machines (SVMs). Structured SVM was employed to order and select images from albums by using three kinds of features which are face features, global scene features, and image quality [20]. They also collected a dataset which is comprised of image collections.

As it is stated in the previous chapter, submodular functions are very common for summarization problem and it is also used with the supervised approach. Tschitschek et al. [13] propose a method for image collection summarization problem based on submodular functions. Their method learns a set of submodular functions by large-margin structured prediction. They also provide an evaluation metric V-Rouge which is also used in this work.

3.3. Intrinsic Image Properties

Extracting high-level properties from images is the main focus of computer vision. These high-level properties can be concrete concepts such as objects, faces or abstract concepts such as memorability, interestingness, scenicness. Our focus in this work is abstract concepts that are related to the meaning in the image which is called intrinsic properties in the literature. These intrinsic properties are taken into account to import human preferences into the summaries.

Borth et al. [9] propose a method to measure sentiments in the image. They construct an ontology, has 3000 concepts, on top of a Youtube and Flickr images and labels. These concepts cause strong feeling and they also relate to emotions. They use adjective-noun pairs instead of only adjective or noun. They aim to eliminate neutral concepts by adjective-noun pairs. For instance, 'butterfly' might not cause strong emotion. On the other hand, when we think 'colorful butterfly', we have a strong positive emotion. They train a concept receptors to detect 1200 adjective-noun pairs in the image. Then each adjective and noun separately is associated with a value between -1 and 1. As it is expected, negative values correspond to negative emotions conversely positive emotions correspond to positive emotions. By doing so, they compute the sentiments and emotions of images.

In addition to sentiments, there is a lot of intrinsic property in the literature. For example, Gygli et al. [21] investigate interestingness of images for humans. Khosla et al. [7] propose a method to measure the popularity of an image which is represented by various levels of image features. Another interesting intrinsic property is memorability which corresponds to a measure of human memory. In addition to popularity, Khosla et al. investigate memorability [8]. As a learning framework, they use Support Vector Regression (SVR) framework using a linear kernel to predict image popularity. Wang et al. [22] propose a new ranking loss in siamese CNN architecture to assign event-specific importance scores to images. Scenicness [10] property rates images according to their scene existence. They present a large dataset with ground truth scenicness scores and use deep convolutional neural networks with different loss functions.

Aesthetics is another important property which is considered for images. Kong et al. [23] collect a benchmark dataset for the aesthetics problem. Talebi and Milanfar also measure aesthetics in [11]. They train CNN models for two different kinds of qualities which are technical and aesthetics. In the technical quality, they measure effects which decrease the image quality because of the noise, blur, compression artifacts. On the other hand in the evaluation of aesthetic, they measure high-level properties related to sentiments and beauty in images.

Intrinsic properties that are searched in the literature increases every day. In this work, we use four kind of intrinsic properties which are sentiments [9], memorability [8], aesthetics [11] and scenicness [10]. First, the contribution of each intrinsic property to summarization problem is evaluated separately. Then their combinations are evaluated to find the best combination which improves the summarization process.



4. OUR APPROACH

This chapter can best be treated under five headings: datasets, feature extraction, baselines, intrinsic properties, proposed summarization methods. In the section 4.1., we introduce our datasets which are collected for image collection summarization problem. In the subsequent section, we explain how we represent images via different visual features. In the third section, we discuss baseline methods especially their strengths and weaknesses that we pay attention during selection process. Then, we revisit the intrinsic properties in this main section to explain our tricks to integrate them with summarization methods. After that we explain these preliminary steps, we delve into details of summarization methods.

4.1. Datasets

First we explain the pruning strategies that are used to clean data from noisy samples. In the next section, we describe the only publicly available dataset which can be used for summarization problem. Then, we present our two datasets in the same section while they are collected in the same way. Main difference between the two datasets are the themes that image collections have.

4.1.1. Data Pruning

As we stated in the previous sections, the summarization problem has some ambiguities. This ambiguity causes a variance between human reference summaries. High variance means less patterns in the data and it is a real problem for learning summarization. Even if we ignore ambiguity, crowd-sourcing has some disadvantages. Problems like summarization, there is no expert knowledge, the problem is naturally based on human preferences and crowd-sourcing is the only option. For example, it is not guaranteed that every user pays the minimum required attention to summarization of image collection that directly affects the quality of summaries. Pruning methods, in the case of summarization, basically aim to find

most consistent summaries across all collected summaries. Although there is a variety of pruning strategies, we choose simple and efficient pruning strategy by inspring the method in [13]. The pseudocode of our pruning algorithm can be seen in Algorithm 1.

Algorithm 1 Prune summaries

Require: summary size N , user summaries S

```

while  $N < S.size$  do
   $y \leftarrow 1$ 
  for  $y \leq S.size$  do
     $X = S - S_y$ 
     $Y_y = \text{computeScore}(S_y, X)$  {compute score of  $S_y$  based on  $X$ }
  end for
   $S_{lowest} = \min(Y)$ 
   $S = S - S_{lowest}$ 
end while
return  $S$ 

```

This algorithm is an iterative algorithm that prunes the summary, that has the lowest score at each iteration. We associate the degree of quality of summary with a score. Low scores correspond to low quality summaries and high scores corresponds to high quality summaries depending on our assumption. This algorithm filters low quality summaries until reaching the requested size N . In our experiments, we use V-Rouge and F-Measure as the scoring functions.

The effect of our pruning algorithm is demonstrated in Figure 4.1. The left plot shows the distribution the selected images before pruning and the right plot illustrates the distribution after pruning. Here, the pruning process is performed by considering V-Rouge metric. In this Figure, each vertical line shows the images included in one human summary which consists of 10 different images. As a result, horizontal lines correspond to the patterns in the summaries. In other words, if an image is selected to be a part of the summary by most of the users, small squares form a horizontal line in that position. By performing pruning, the groups in the summary become more clear.

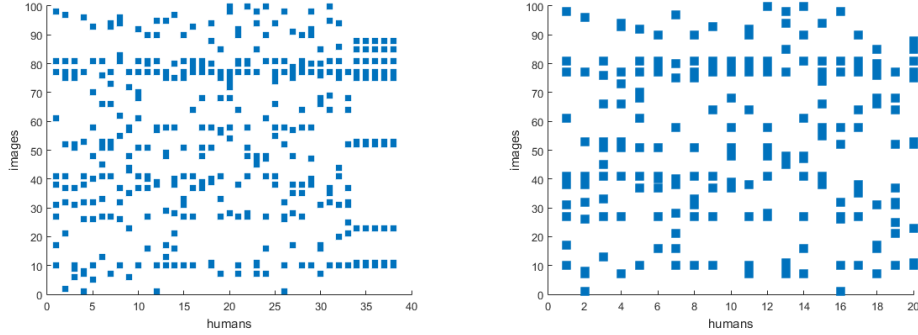


FIGURE 4.1.: The selected summaries before (left) and after pruning (right) for the Istanbul Tour image collection B3.

4.1.2. Vacation Photographs

Vacation Photographs dataset [20] consists of 63 image collections on 5 vacations topics which are Disney Trips, London, Paris, Washington, Beach Vacation. Their image collections have varying size from 44 to 1353. However, this dataset is not rich with the ground truth human summaries as much as our datasets. They state that they collect four annotations for each image collection. One kind of annotation is a shot annotation in which they aim to generate shots of near duplicate images to use in evaluation metric. Although they released album annotation with varying size such as 5, 10, 15, only summaries with size 10 are employed to keep summary size consistent with our datasets. They collect these summaries by asking annotators to select and order photographs from an album in a way that “tells a story”. They do not directly ask the annotators the summarize the image collection.

4.1.3. EventSum and CitySum

Our two datasets, EventSum and CitySum, are gathered for image collection summarization problem. The both are downloaded from Flickr ¹. Only difference between these datasets is the query concept. In the EventSum dataset, as the name implies, images are queried by event-based keywords. EventSum dataset keywords are: A Trip to New York [B1.], A

¹<https://www.flickr.com/>

Visit to Italy [B2.], An Istanbul Tour [B3.], Birthday Party [B4.], Burning Man [B5.], London Olympics [B6.], St. Patrick’s Day [B7.], World Cup [B8.]. In the CitySum dataset, we narrow our focus from all events to only one specific event of vacation. CitySum dataset keywords are: Amsterdam Vacation [B9., B10.], Paris Vacation [B15., B16., B17., B18., B19.], Tokyo Vacation [B11., B12.], Venice Vacation [B13., B14.]. We choose popular keywords to increase our chance to find appropriate image collections.

We take following questions in consideration when including an image collection to the dataset:

1. Are the images relevant with the query?
2. Do the images have time-tags?
3. Do the images represents an event for EvenSum? Do the images represents a vacation for CitySum?

“Does image collection have 100 images?” is not a question that we consider because it is very hard to ensure. For this reason, we filter some images to make all collection sizes equal. Uniform filtering help us to filter images without ruining the natural flow in the image collection.

Number of human summaries for EventSum and CitySum datasets are shown in Figure 4.2..

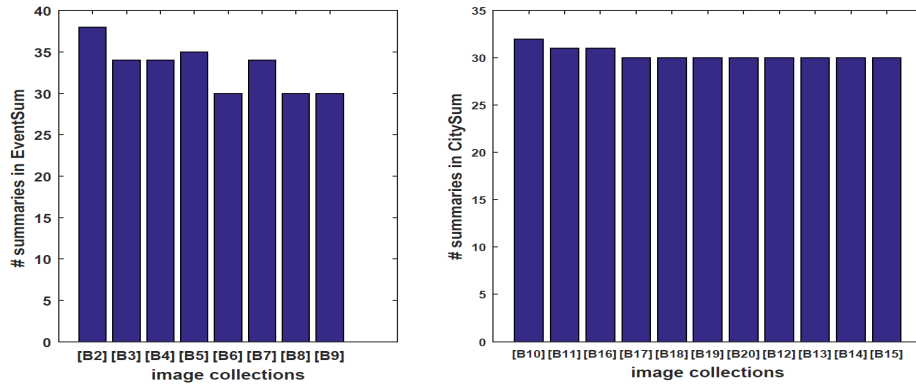


FIGURE 4.2.: Summary distribution of EventSum and CitySum datasets

As it is mentioned, we measure the quality of summaries through V-Rouge and F-Measure. We prune human ground truth summaries until we have 20 consistent summaries. Original human summary consistencies and pruned human summary consistencies based on these two metrics for EventSum and CitySum are shown in Table [4.1.] and Table [4.2.] respectively.

TABLE 4.1.: The consistency of the human summaries based on the V-Rouge and F-Measure metrics on the EventSum dataset

Category	V-Rouge		F-Measure	
	Before Pruning	After Pruning	Before Pruning	After Pruning
A Trip to New York	0.387	0.484	0.205	0.309
A Visit to Italy	0.406	0.462	0.215	0.244
An Istanbul Tour	0.473	0.538	0.279	0.323
Birthday Party	0.419	0.470	0.219	0.287
Burning Man	0.425	0.473	0.232	0.269
London Olympics	0.329	0.371	0.182	0.208
St. Patrick's Day	0.449	0.508	0.208	0.259
World Cup	0.370	0.412	0.169	0.179
Mean	0.407	0.465	0.213	0.260

TABLE 4.2.: The consistency of the human summaries based on the V-Rouge and F-Measure metrics on the CitySum dataset

Category	V-Rouge		F-Measure	
	Before Pruning	After Pruning	Before Pruning	After Pruning
Amsterdam V. [B9.]	0.505	0.587	0.269	0.372
Amsterdam Vacation [B10.]	0.522	0.603	0.240	0.321
Paris V. [B15.]	0.532	0.607	0.330	0.441
Paris V. [B16.]	0.438	0.488	0.206	0.255
Paris V. [B17.]	0.485	0.554	0.255	0.346
Paris V. [B18.]	0.456	0.526	0.232	0.287
Paris V. [B19.]	0.448	0.496	0.209	0.254
Tokyo V. [B11.]	0.392	0.439	0.240	0.293
Tokyo V. [B12.]	0.387	0.434	0.197	0.240
Venice V. [B13.]	0.470	0.547	0.274	0.364
Venice V. [B14.]	0.542	0.597	0.298	0.379
Mean	0.471	0.534	0.250	0.323



FIGURE 4.3.: Example human summary with the maximum F-Measure from an Istanbul Tour image collection B3.

As it is expected, in both metrics pruning process increase the patterns in the summaries. These results are important to compare the automatic summaries against the human summaries. These low consistency values show that the problem is challenging for even humans.

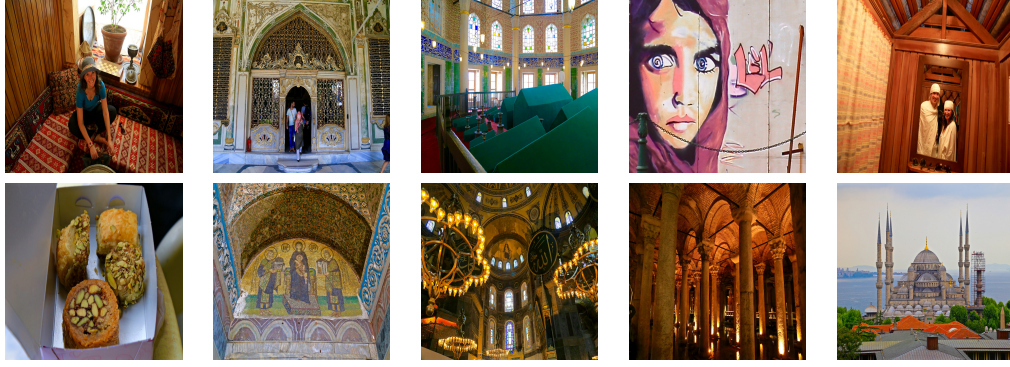


FIGURE 4.4.: Example human summary with the minimum F-Measure from an Istanbul Tour image collection B3.

These results can be thought of an upper bound for success of automatic summarization methods.

4.2. Feature Extraction

We extract four kinds of features from each image that encode different type of information from high level to low level in the image. First three features are combined to represent an image which are colour representation, superpixel representation and CNN representation. After that we extract each individual feature, we simply concatenate these three features while they complement each other. By doing this, our representation for each image becomes one dimensional feature vector with the size 511. However, these features are not scale invariant. As a last feature we employ Regional Maximum Activation of Convolution (RMAC) method [24] to reliably embrace the objects in images in varying scales.

4.2.1. Colour Representation

Colour is one of the significant features of images. In addition to be a one of the important characteristics of the images, colour eases the perception of objects and makes borders more visible. Although the methods mostly use simple BoW representation of values in the RGB and HSV colour spaces, there are more complex representation of colors regarding to human

perception. For example, Van De Weijer et al. [25] provides a method to represent the colours in the image by 11 frequent colours in the nature such as black, blue, brown, grey, green, orange, pink, purple, red, white, yellow. Instead of using simple histograms in a colour space such as RGB and HSV, we prefer to use BoW representation of more prominent colours in the human perception. For this reason, we extract the patches in size 10x10 pixel from the image. Then we associate each patch with most dominant colour in that patch. By doing this, we build BoW representation of the image in terms of 11 colours. We normalize the resulting feature vector with L1 norm as a last step in feature extraction process.

4.2.2. Superpixel Representation

By using superpixel representations, we compute features that encode relationship between more meaningful patches instead of random fixed size patches. This method includes two-stage dictionary learning. This feature extraction process starts with computation of dense SIFT features of images, however resulting huge feature size is not feasible for dictionary learning process. For this reason, we employ 10 pixel step size to make the features sparse. First dictionary, with 200 visual word, is learnt on dense SIFT features with k-means clustering algorithm. Then, we generate BoW representations of these visual words for each superpixel region. In other words each superpixel region is represented by its SIFT visual words. Although the resulting feature is sufficient to encode points in the image under scale, rotation, illumination and viewpoint variance, it does not encode colour information. For this reason, we concatenate them with simple 16-bin RGB colour histogram which is extracted from the corresponding superpixel. Then we again build second dictionary with 200 visual word on these SIFT and colour features and encode the final features by that dictionary. This feature results in a 200 dimensional vector depending on the dictionary size. We normalize the resulting feature vector with L1 norm as a last step in feature extraction process similar to previous color representation.

Our SIFT features are 128 sized vectors similar to common usage of SIFT. Size comes from quantizations of 8 bins and 4x4 squares in the key point.

Supapixel extraction algorithm converts image from RGB space to LAB space as a first step. One critical point in the superpixel extraction process is to select parameters similar to various computer vision problems. There are three parameters which effect the resulting superpixels which are kernel size, maximum distance and ratio. We choose kernel size as 2 which specifies the area in which density estimation takes place. Second parameter is ratio which weights the colour and spatial information according to each other. In our setting, we choose the ratio as 0.5 to have equal importance. As the name implies, maximum distance specifies the maximum distance in feature space that might be connected depending on increase in density which is 20 in this setting.

4.2.3. CNN Representation

Similar to colour feature extraction process, we extract patches from images. However, instead of keeping size of them fixed as in colour feature extraction process, we divide image directly 10x10 patches with 0.5 overlap, so the patch sizes varied by image size. As a CNN features, we employ VGG16 'fc7' features [26]. After that we cluster them to 300 centers by k-means clustering algorithm, we encode them by kernel codebook encoding [27]. This feature results in a 300 dimensional vector depending on the dictionary size. Similar to other feature extraction methods, we compute BoW representations over encoding results. We normalize the resulting feature vector with L1 norm as a last step in feature extraction process similar to previous two representations.

4.2.4. RMAC Representation

Thus far, we introduce four kind of features however none of these features handle scale changes of objects which is an important property for summarization. Similar to previous examples, images share similar visual features should not be found together in a summary. Two images of the birthday cake from Birthday Party [B4.] image collection can be seen in

Figure 4.5.. Summarization methods should extract summaries by considering these commonalities across images event if they are not at the same scale. RMAC method [24] extracts r number of regions and s number of scales from the CNN response map. In this method, regions overlap with each other similar to the CNN representation 2.2.3.. After these steps, each image is represented by m dimensional regional vectors which represents a matrix whose size is $r \times m$. As a last step, they sum the regional feature vectors to have one feature vector for the image and normalize with L2 norm.



FIGURE 4.5.: Scale invariance example

For this purpose, we employ a feature extraction method which extracts r number of regions and s number of scales from the CNN response map.

4.3. Baselines

In this work, we compare the summarization methods with two basic baseline methods which are called uniform summarization and k-means summarization. In addition to these two simple baselines, we also choose two strong baseline methods which are DS3 and SRNN.

4.3.1. Uniform Summarization

Uniform summarization is the first thing that comes to mind. Initially, we order images according to their time-tags. Then, we sample with constant step size by starting with the first image. In our case the step size is selected as 10 regarding to the collection size which is 100, hence the summary includes 10 images.

Although this method seems simple, it is pretty reasonable baseline because time is one of key factors which is important in summarization process. By jumping in time, we almost guarantee filtering of near duplicates while they are mostly at consecutive order. Different events also occur at different times with a high degree of probability, so it provides coverage too.

4.3.2. K-Means Summarization

Our second baseline is k-means summarization which generates a summary with k-means clustering algorithm. In the k-means sampling, first we cluster the images to 10 centers as it is expected, then images closest to the cluster centers are used to form a summary. It is shown in equation 5:

$$v = [v_{c1}v_{c2}v_{c3}v_{c4}v_{c5}v_{c6}v_{c7}v_{c8}v_{c9}v_{c10}] \quad (5)$$

In the case of clusters are well formed, k-means provides diverse images with high coverage at the same time, so it is a very significant baseline from summarization point of view. We run k-means clustering algorithm 100 times to find an average performance.

An advantage of using uniform summarization is taking time information into account while various summarization techniques benefit from time information because it is significant for summarization problem. It also produces same summaries in each run. On the other hand, it does not consider image content. For example, if we have a very similar 50 images in the

beginning of the image collection, half of our selections will be very similar to each other which results in poor summary. K-Means overcome this weakness by using image representations. However, K-Means generates varying results across different runs depending on initialization conditions.

4.3.3. Dissimilarity Based Sparse Subset Selection

As stated in the Section 3.1., DS3 algorithm allows us to summarize a target set depending on a source set. In our case, we have only one set which corresponds to both source and target set. When this is the case, only pairwise differences in the set are taken into account for subset selection. Dissimilarity is measured by cosine distance. DS3 method provides a clustering framework such as K-Means, however it does not suffer from initialization conditions as the K-Means summarization.

4.3.4. Skipping Recurrent Neural Network

SRNN method is based on a training a neural network as it is mentioned in the Section 2.3.4.. Training neural networks requires a lot of data and training a model over only one image collection is not reasonable. For this reason, we employ the dataset in [28] for six cities. However, we have still image collections which are not covered by these six cities dataset. Hence, we extend six cities dataset to cover all of our image collections. While SRNN method takes the time ordering into account, first we sort images by their time-tags. Then, we train the SRNN which models the skipping connections between images. SRNN results are also averaged over 100 runs similar to K-Means baseline because SRNN generates different summaries in each run.

4.4. Intrinsic Properties

In this section, we focus on analyzing which intrinsic properties play a crucial role in summarization process. As we mentioned, we use four intrinsic properties in our experiments that are sentiments [9], memorability [8], scenicness [10] and aesthetics [11]. Each property assigns a score to an image. To be clear, image features were not replaced by intrinsic properties, both of them integrated in the methods.

4.4.1. Memorability

Khosla et al. [8] collect the LaMem dataset that includes 600,000 images with its ground truth memorability scores. They follow the literature and use transfer learning with CNN. For this purpose, they finetune hybrid CNN [29]. We employ this intrinsic property in our work, because we believe that memorability can improve our summaries in the case of choosing images which stick to people’s mind. The most memorable and the less memorable images for an Istanbul Tour image collection B3. are shown in Figure 4.6..

We expect that memorability correlates with quality, aesthetics and positive sentiments in the images. Contrary to expectations, memorability correlates with human existence in the images. On the other hand, scene images are less memorable images. Minimum and maximum scored images are relevant with the memorability work, however they are not relevant with our insight when including this intrinsic property in this work.

To gain a deeper understanding of the relationship between memorability score and human summarization, we analyse the ground truth human summaries. For this reason, we observe the distribution of memorability scores in the dataset by dividing normalized memorability scores to equivalent bins. Then we plot two distribution graphics. The first plot shows the distribution of intrinsic image properties for the images that are selected by humans at least ten times. The second plot shows the distribution of intrinsic image properties for the images that are not selected by the users for the summary. Image count in each bin is normalized

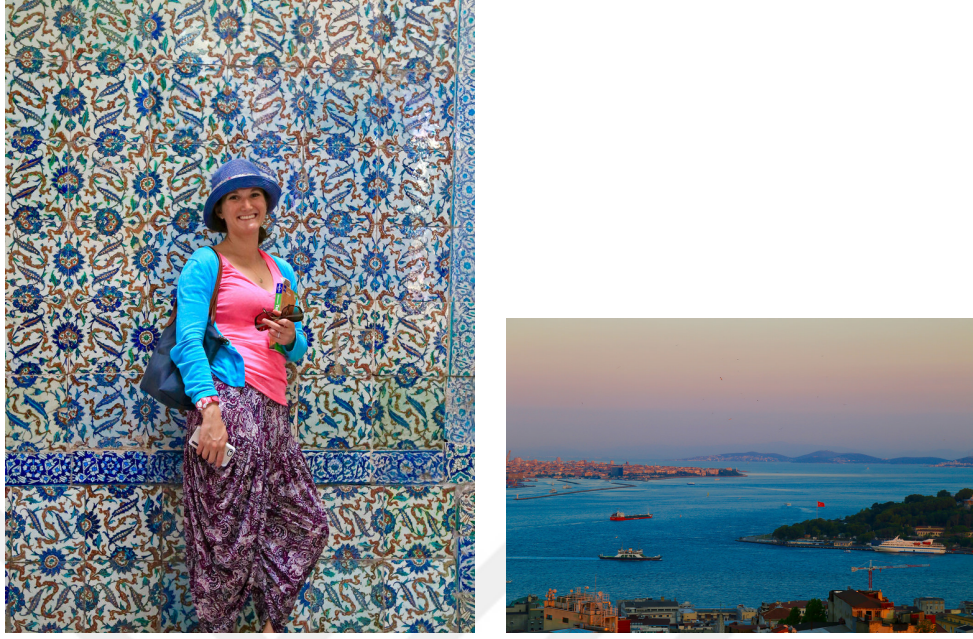


FIGURE 4.6.: The most memorable image (left) and the less memorable image (right) in an Istanbul Tour image collection B3.

by the total number of images in all bins to convert image counts to probabilities for easy interpretation. Results are shown in Figure 4.7. for memorability score.

According to the Figure 4.7., two datasets show different distribution characteristics. However, it is notable that images in the bins with low memorability scores have higher inclusion ratio than the remaining bins.

4.4.2. Sentiments

We benefitted from work of Borth et al.[9] to measure sentiments in the image. We assume that images cause to positive sentiments are more likely to be selected to summaries. First we need to assign a score to each adjective noun pair. For this reason, we simply compute sentiment score of each pair by the following equation:

$$s(\text{pair}) = s(\text{adjective}) + s(\text{noun}) \quad (6)$$

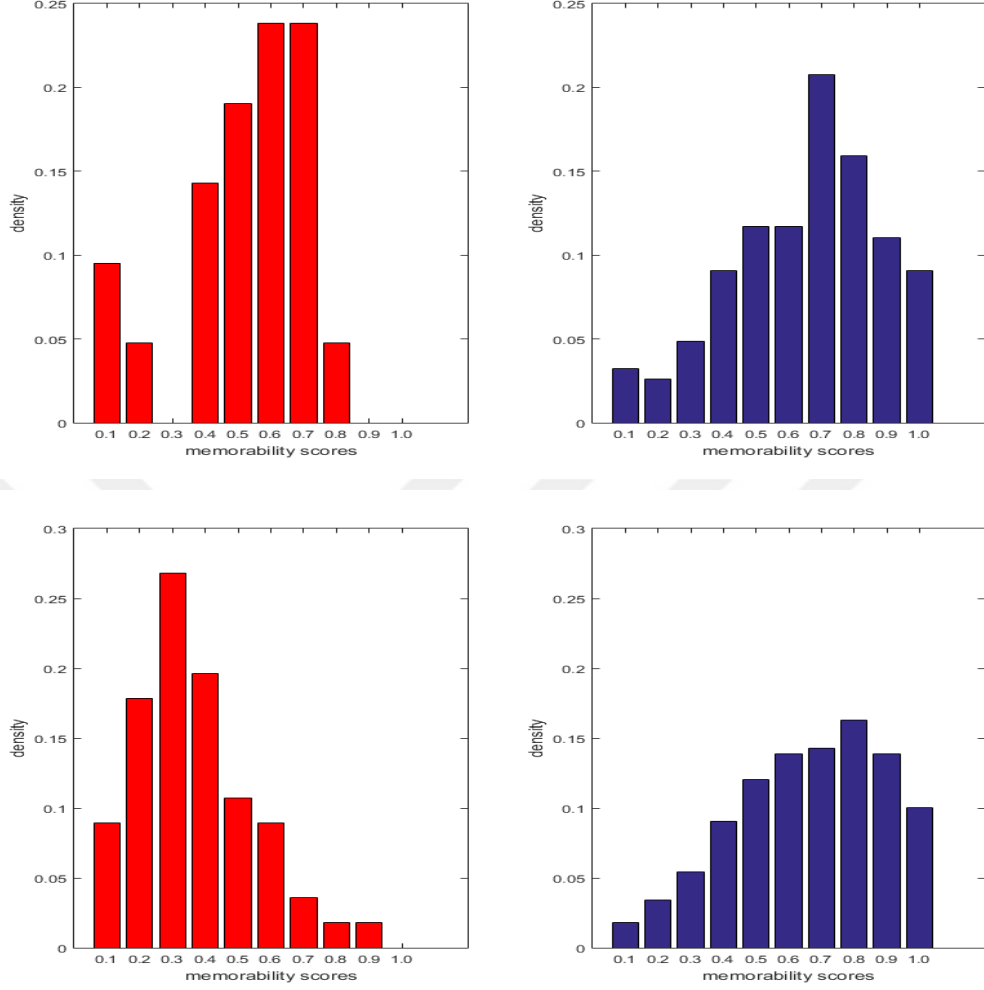


FIGURE 4.7.: Distribution of the memorability scores across the images being (red) and not being (blue) parts of the summaries for the EventSum (top) and the CitySum (bottom) datasets

Then, we extract 1200-length detector response vectors. We did not use all the responses while they are not included in the image at the same degree. We use only dominant responses which are higher than a specific threshold which is 0.85 in our experiments. However we need one real valued score for each image instead of a vector. Hence, we perform vector to real value conversion by Equation 7. σ denotes vector of dominant receptor responses, s denotes the output of Equation 6. This score guarantees that the images that cause negative feelings have smaller score than the images that cause positive feelings which is our motivation behind choosing this intrinsic property for summarization.

$$q = \sum_{i=1}^{1200} \sigma_i \times s_i \quad (7)$$

Images with the most positive emotion and the most negative emotion are shown in Figure 4.8. for an Istanbul Tour image collection B3..



FIGURE 4.8.: Image with the most positive emotion (left) and image with the most negative emotion (right) in an Istanbul Tour image collection B3.

In the case of sentiments, these two images are a little bit confusing because there is a girl with a big smile in both of them. If we ask people to choose most positive and most negative images from the emotion point of view, they do not choose these two images with high probability. The cause of this situation might be dense pattern in the right image which might dominate the other adjective noun pairs and smile might be eliminated while it is non-dominant sentiment.

Similar analysis which is done for memorability score is repeated for sentiment score. The corresponding analysis is shown in Figure 4.9.. In this case, distribution characteristics for the both datasets are pretty similar to each other and most of the images that are selected to

the human ground truth summaries are above the mean sentiment score. We can infer from the results that humans select images with high sentiment scores.

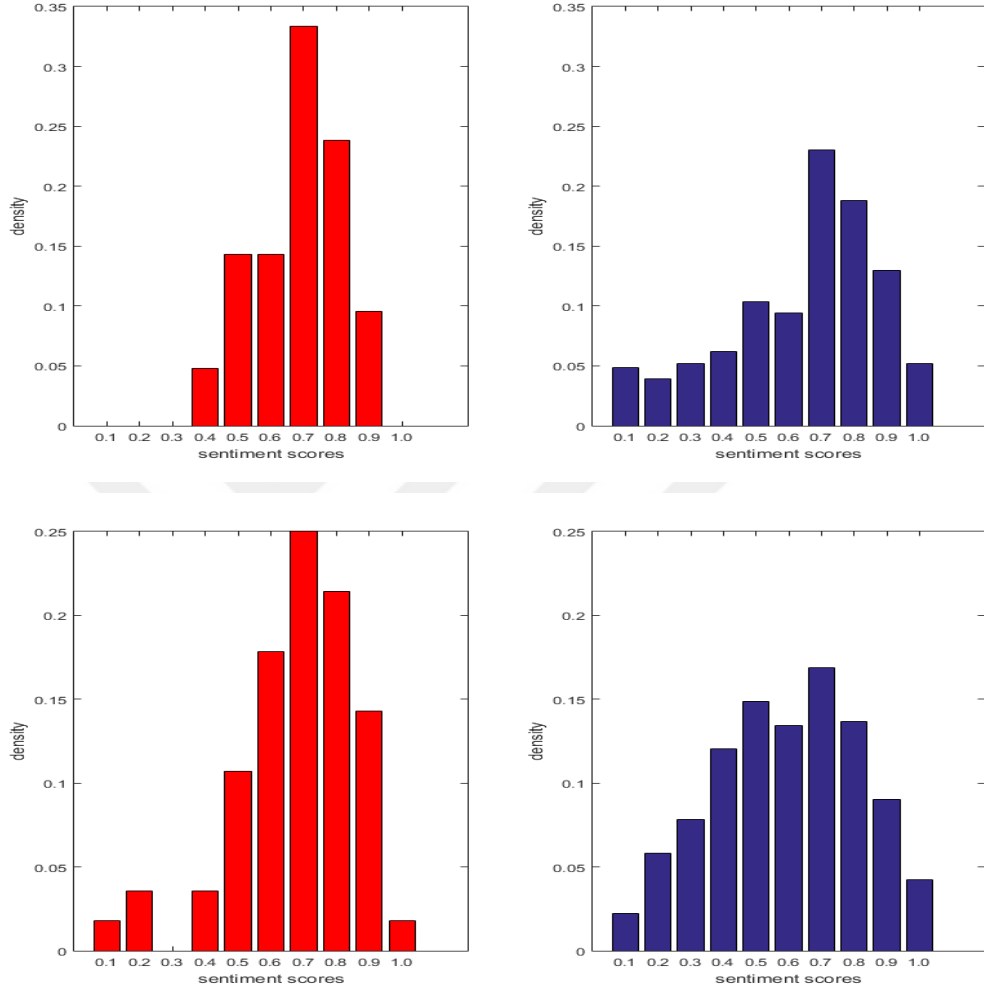


FIGURE 4.9.: Distribution of the sentiment scores across the images being (red) and not being (blue) parts of the summaries for the EventSum (top) and the CitySum (bottom) datasets

4.4.3. Scenicness

Scenicness is proposed as a measure of natural beauty. Workman et. al [10] build a game on their database with 185,584 images. They want from users to rate images based on scenicness, or natural beauty. By doing this, they investigate aesthetic judgement of users. Similar to memorability, they finetune CNN for scenicness prediction. Although they experiment

with three variant of CNN architecture, they report the best result on test set in a multinomial distribution setting. In this setting they use each image and its all possible ratings to finetune the network.

We leverage their multinomial network to predict scenicness scores of images in our datasets. Images with maximum and minimum scenicness scores in an Istanbul Tour image collection B3. are shown in Figure 4.10..



FIGURE 4.10.: Image with the maximum scenicness score (left) and image with the minimum scenicness score(right) in an Istanbul Tour image collection B3.

Scenicness scores reflect the scenicness properly. Image with the maximum scenicness is a great scene of Bosphorus without a doubt. The minimum scenicness score is also coherent while it is not a scene which is just a door. This score carries our intuition to choose it for summarization in the first place.

Similar distribution analysis which is done for previous intrinsic scores is repeated for the scenicness score. The resulting distributions can be seen in Figure 4.11.. Blue distributions of the EventSum and CitySum datasets share similar characteristics. Images that are not selected for the summaries have low scenicness scores with a high probability on both datasets.

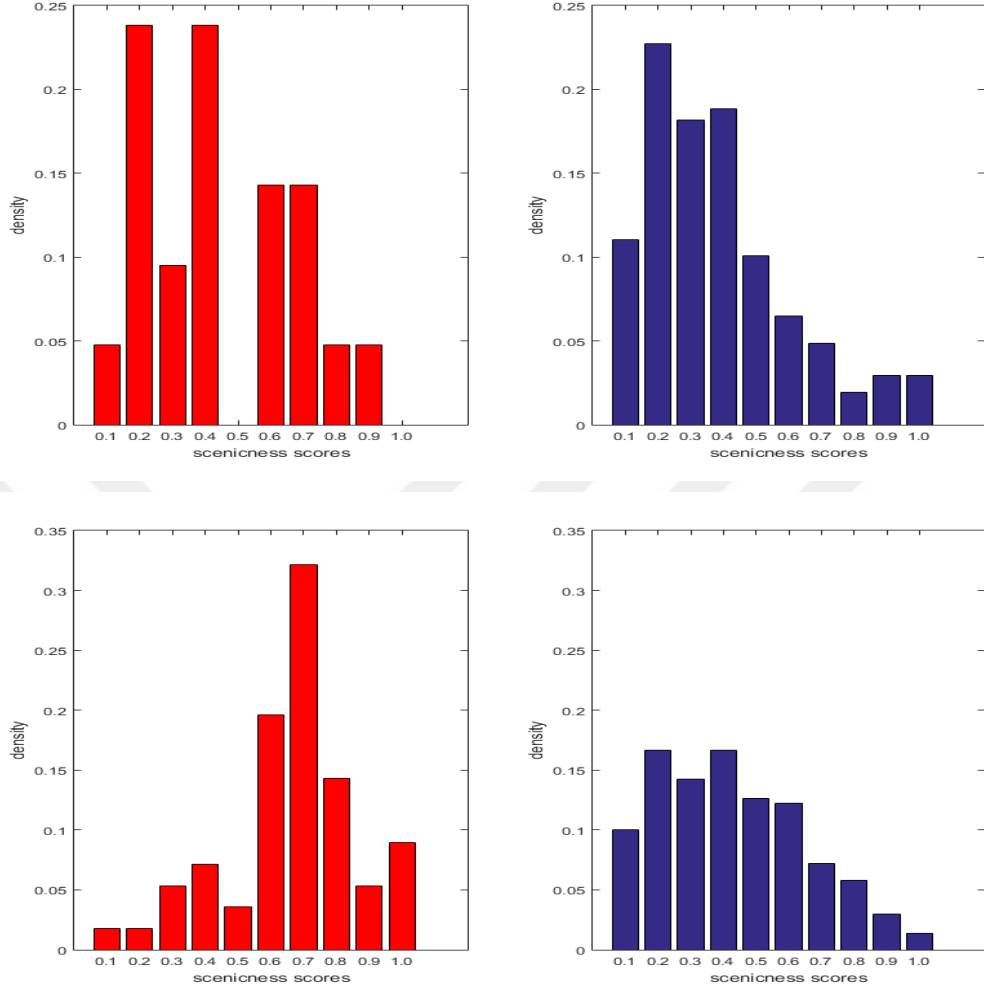


FIGURE 4.11.: Distribution of the sceniness scores across the images being (red) and not being (blue) parts of the summaries for the EventSum (top) and the CitySum (bottom) datasets

4.4.4. Aesthetics

Aesthetics corresponds to beauty in the image which is one of the strongest intrinsic property which we suppose that it plays an substantial role in human summarization process. We leverage from [11] to compute aesthetic scores because it presents good results with state-of-the-art methods. In this work, we compute aesthetics quality of image by using Inception-v2 baseline [30] while they report their best results in that setting. Inception-v2 network is a

improved version of Inception [31] with parallelization on convolution and pooling computations. They train the network on Aesthetic Visual Analysis (AVA) database [32] which is a large scale database for aesthetic visual analysis. Although it has a size smaller than LaMem [8] and scenicness [10] databases in the case of images, it abounds in annotations. They collected 200 annotations for each image in the database. Images with the maximum and the minimum aesthetics scores in an Istanbul Tour image collection B3. are shown in Figure 4.12..



FIGURE 4.12.: Image with the maximum aesthetic score (left) and image with the minimum aesthetic score(right) in an Istanbul Tour image collection B3.

These results are coherent with the idea of aesthetics. It assigns the highest score to a impressive scene of a mosque. On the other hand, it assigns the smallest value to dark scene which seems logical.

Distribution analysis for aesthetics score is shown in Figure 4.13.. Similar to sentiment distribution, aesthetic scores of summary images are mostly located above 0.5 which shows that the humans select the aesthetic images for the summary.

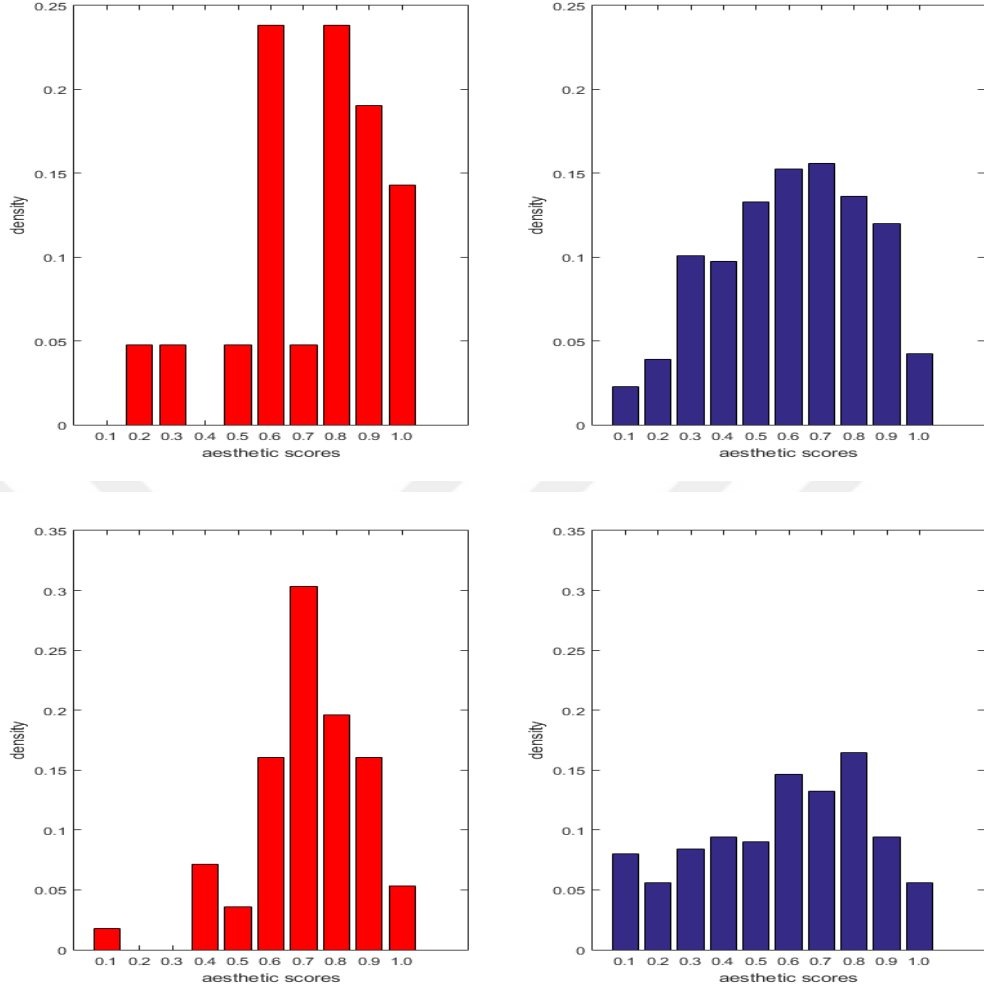


FIGURE 4.13.: Distribution of the aesthetic scores across the images being (red) and not being parts (blue) of the summaries for the EventSum (top) and the CitySum (bottom) datasets

4.5. Proposed Summarization Methods

This section presents the details of how we integrate intrinsic properties to the DPP and the DR methods.

4.5.1. Determinantal Point Processes

DPP method uses affinity matrices which encode pairwise similarities based on a specific metric. This representation only takes image features, diversity features, into account. One

critical parameter of affinity matrices is scale parameter. We choose mean of matrix to have a shared parameter between DPP and DR methods and EventSum and CitySum datasets.

We integrate intrinsic properties and image features by Equation 8 in which q corresponds to intrinsic properties and A_{Dis} corresponds affinity matrix based on dissimilarity features.

$$\begin{aligned} Q &= qq^T \\ A &= QA_{Dis} \end{aligned} \tag{8}$$

We employed K-DPP method to generate fixed size summaries. DPP method does not generate same summary in each run. This instability is handled by averaging multiple runs. We run DPP method 100 times to have more stable solution for healthy comparison with other methods.

4.5.2. Diversity Ranking

Diversity ranking method [3] optimizes a submodular function. Diversity ranking decreases the information that repeated in the images by choosing diverse images. In the best case, it finds 10 data center which is in the same size and equal distance. By doing this, it ensures centrality that is important to gather similar images into close to each other. However, these centers should have distance to provide diversity. It represents relationship between feature vectors by Gaussian similarity.

$$d_{xy} = \begin{cases} \exp(-\beta \|g(x) - g(y)\|^2), & \text{if } (x, y) \in \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Selection process optimizes the formula below:

$$\begin{aligned}
& \max \sum_{x \in V} u(x) \\
& a_x = \sum_{(x,y) \in \epsilon} d_{yx} + z_x \text{ if } u(x) = \frac{1}{a_x} \sum_{(x,y) \in \epsilon} d_{yx} u(y) \\
& s \in S \subset V, |S| \leq K \text{ for } u(g) = 0, u(s) = 1
\end{aligned} \tag{10}$$

$u(g)$ results in zero when no element is selected. $u(s)$ results in 1 when that element is selected. K corresponds to set size in our experiment. While the equation seems complex, it simply takes weighted sum of its neighbours. Here, z parameter controls trade-off between diversity and centrality that is a application specific parameter. Using a large value of z decreases the neighbourhood window. In this work, we choose z according to mean.

Intrinsic properties are taken into account by using mass points in diversity ranking method. Hence images with higher intrinsic scores become more likely to be chosen to a summary, while higher scores weighted more than the scores with lower scores.

In this method, we also use an affinity matrix exactly same with DPP for fair comparison.

5. EXPERIMENTAL ANALYSIS

This section presents the performance of the proposed summarization methods. We aim to generate summaries with high diversity, high coverage, and high quality by using these methods. We analyse our experiments with the qualitative and the quantitative evaluations.

5.1. Quantitative Evaluation

First, we compare the baseline methods and the proposed summarization methods without integrating intrinsic properties in Table 5.1. and in Table 5.2. for EventSum and CitySum datasets. These proposed summarization methods can be thought of as a baseline, while we aim to pass their results by using intrinsic image properties. The most successful method for each image collection is highlighted in the corresponding tables. The most successful method shows great variability across the image collections. Hence, we can not report the best method or at least dominant method among baselines. This situation shows the strong dependency on image collection under interest.

TABLE 5.1.: Comparison of baseline methods on the EventSum dataset with F-Measure metric

Category	Uniform	K-means	DS3	SRNN	DPP	DR
A Trip to New York	0.068	0.108	0.158	0.124	0.106	0.132
A Visit to Italy	0.042	0.097	0.090	0.075	0.106	0.200
An Istanbul Tour	0.163	0.085	0.158	0.105	0.098	0.068
Birthday Party	0.116	0.082	0.190	0.090	0.110	0.058
Burning Man	0.090	0.082	0.053	0.123	0.110	0.095
London Olympics	0.163	0.100	0.068	0.080	0.108	0.047
St. Patrick’s Day	0.084	0.092	0.042	0.128	0.106	0.147
World Cup	0.163	0.105	0.090	0.115	0.107	0.179
Mean	0.112	0.094	0.106	0.105	0.106	0.116

TABLE 5.2.: Comparison of baseline methods on the CitySum with F-Measure metric

Category	Uniform	K-means	DS3	SRNN	DPP	DR
Amsterdam [B9.]	0.105	0.158	0.132	0.117	0.103	0.158
Amsterdam [B10.]	0.084	0.097	0.063	0.147	0.104	0.063
Paris [B15.]	0.032	0.099	0.019	0.136	0.096	0.032
Paris [B16.]	0.111	0.099	0.090	0.085	0.106	0.105
Paris [B17.]	0.132	0.099	0.079	0.0120	0.104	0.163
Paris [B18.]	0.137	0.096	0.195	0.082	0.111	0.068
Paris [B19.]	0.042	0.086	0.074	0.104	0.105	0.100
Tokyo [B11.]	0.121	0.108	0.116	0.119	0.104	0.111
Tokyo [B12.]	0.216	0.096	0.174	0.067	0.107	0.153
Venice [B13.]	0.090	0.101	0.126	0.131	0.111	0.174
Venice [B14.]	0.079	0.093	0.084	0.124	0.100	0.095
Mean	0.104	0.103	0.105	0.112	0.105	0.111

In the EventSum dataset, DR baseline outperforms all baselines in the F-Measure evaluation metric. On the other hand, SRNN performs the best summarization based on the F-Measure metric on the CitySum dataset. Although the train set and the test set come from different datasets that are collected for different reasons, we still have good results which prove the strength of the learning methods.

Evaluation of the uniform summaries varies between 0.042 and 0.163 in Table 5.2.. This shows how powerful is the time ordering in the some of the sets and how useless in the remaining. By using DPP, we have a more stable solution regarding quality, unlike the uniform baseline.

Thus far, we present the results with detailed collection-based comparisons. Now, we continue with the comparison of summarization methods on average across all image collections.

First, we compare summarization methods DPP, DR with F-Measure and V-Rouge evaluation metric to show diversity and coverage capacities of these methods. A comparison of the methods in the EventSum and the CitySum datasets can be seen in Table 5.3..

TABLE 5.3.: Comparison of summarization methods

	Dataset	Uniform	K-Means	DS3	SRNN	DPP	DR
F-Measure	EventSum	0.112	0.094	0.106	0.105	0.106	0.116
	CitySum	0.104	0.103	0.105	0.112	0.105	0.111
	Mean	0.108	0.099	0.106	0.109	0.106	0.114
V-Rouge (rmac)	EventSum	0.316	0.330	0.356	0.348	0.339	0.363
	CitySum	0.321	0.316	0.332	0.337	0.319	0.305
	Mean	0.319	0.323	0.344	0.343	0.329	0.334

For both datasets, results show that the DR method is the best based on the F-Measure metric. DR method passes DPP method with huge gap based on the F-Measure metric. DS3 method passes SRNN algorithm in V-Rouge metric with a very small negligible difference which is 0.01. However, this small gap is not enough to say that the DS3 method is better than the SRNN method. We can not easily make an inference based on only one metric in the problems like summarization which involves great ambiguity and human preferences. This is the reason why we present the results of two different metrics for each experiment. We evaluate the performance of the methods as a good if they take the first rank at least one of the metrics. According to our second metric, V-Rouge, DS3 generates the best summaries. In light of these two metrics, we can say that DR and DS3 methods generate better summaries than the other methods.

To extend analysis, we also show the effect of features on the results. Until here, we present only experiments with RMAC features while it is pretty straightforward to extract from each image. Results show summarization methods improve the quality of summaries, but these results might depend on the RMAC features. We repeat the experiments with another feature which we call it combined features to show that success of summarization methods are

independent of used features if these features represent the images appropriately. Combined feature corresponds to the combination of different features as is evident from its name. In this work, it is the concatenation of three features which encode different level of image information from low to high. Colour features, superpixel features, and CNN features are combined to form combined features. Results for each dataset is shown in 5.4. and 5.5..

TABLE 5.4.: Effect of features on EventSum with F-Measure metric

Category	Uniform	K-Means	DPP	DR
Combined features	0.112	0.099	0.103	0.077
RMAC	0.112	0.094	0.106	0.116

On the EventSum dataset, DPP and DR methods whose main interest is providing a diverse set of elements fail to pass uniform baseline. However, the uniform baseline on EventSum image collection is very successful which is only beaten by the DR method with RMAC features. This result shows how important the time ordering in summarization problem.

TABLE 5.5.: Effect of features on CitySum with F-Measure metric

Category	Uniform	K-Means	DPP	DR
Combined features	0.104	0.107	0.111	0.100
RMAC	0.104	0.103	0.105	0.111

Results on CitySum dataset shows that baseline methods are beaten by the DPP method. Ranking between the summarization methods are not preserved, but methods, whose aim to provide diversity which is one of the main features of the summary, pass the baselines.

Until this point, we only present experiments on our datasets. Vacation Photographs dataset is only applicable dataset for our work even if it has the very small number of ground truth human summaries. Results on this dataset can be seen in Figure 5.6..

TABLE 5.6.: Comparison of methods on Vacation Photographs dataset with F-Measure metric

Category	Uniform	K-Means	DS3	DPP	DR
Beach Vacation	0.223	0.184	0.196	0.191	0.146
Disney Trips	0.086	0.135	0.229	0.152	0.121
London	0.300	0.221	0.150	0.202	0.167
Paris	0.236	0.158	0.207	0.161	0.214
Washington	0.288	0.279	0.288	0.278	0.313
Mean	0.227	0.195	0.214	0.197	0.192

We compare the methods based on different metrics and different visual features until this point. By doing this, we analyse summarization methods whose success in providing diversity and coverage. We can not report one best summarization method based on the previous results, while best methods regarding to average can fail badly in some of the image collections. This means that summarization methods might generate poor summaries. We propose to use intrinsic properties to increase the qualities of these poor summaries. We assume that the integration of intrinsic properties should increase the quality of summaries. However, all summarization methods are not suitable for integration of intrinsic properties, this integration only applicable to DPP and DR methods. In this section, we demonstrate each method with four different intrinsic properties which are explained in the Section 4.4. in detail. Then, we show comparisons of methods with intrinsic properties. Results of the DPP method after integration of intrinsic properties are shown in Table 5.7..

TABLE 5.7.: Effect of intrinsic properties on DPP with F-measure metric

Intrinsic properties	EventSum	CitySum
No property	0.106	0.105
Memorability	0.090	0.068
Sentiment	0.107	0.114
Scenicness	0.113	0.147
Aesthetics	0.122	0.124

Sentiment, scenicness and aesthetic scores improve the summarization quality for both datasets. Surprisingly, memorability reduces the results in both cases. In the EventSum dataset, the focus of the dataset is humans, on the other hand, in the CitySum dataset focus of the dataset is cities. Regard to this difference, we had an assumption that the effect of intrinsic properties might change. For this reason, until here we present results in two datasets EventSum and CitySum. However, results show that the effect of intrinsic properties is consistent across our two datasets. As we explained that our datasets EventSum and CitySum, these datasets are pretty similar to each other. Their collection process makes them combinable to form a bigger image collection which is named EventCitySum. Based on our findings about the effect of intrinsic properties and similarity between the two datasets, we merge these two datasets to use in remaining experiments. For this reason, we also apply the summarization methods on the entire image collections in the EventCitySum dataset.

We measure the effect of features on summarization with intrinsic properties similar to our previous reasoning. Results of the DPP method are shown in Figure 5.8..

TABLE 5.8.: Results with different features EventCitySum with DPP with F-Measure metric

Intrinsic properties	Combined Features	RMAC
No property	0.107	0.106
Memorability	0.084	0.079
Sentiment	0.154	0.111
Scenicness	0.132	0.130
Aesthetics	0.123	0.123

In the pure summarization experiments, we face with small changes across the usage of different features which affect the ranking between methods. Although shifts in the values still exist, the direction of effect stays same. In other words, sentiment, scenicness and aesthetics have a positive effect on the summarization process. This consistency is very important to conclude that summarization process has a minimum dependency to the visual features when they have enough expressive capacity.

The contribution of each individual intrinsic property to the DR method for two visual features is shown in Table 5.9..

TABLE 5.9.: Results with different features EventCitySum with DR with F-Measure metric

Intrinsic properties	Combined Features	RMAC
No property	0.089	0.114
Memorability	0.064	0.048
Sentiment	0.089	0.094
Scenicness	0.124	0.136
Aesthetics	0.115	0.119

In the DR methods, we find that scenicness and aesthetic scores improve the summaries. We could not observe the same positive effect in the DR method. However, memorability consistently worsens the results in both methods DPP and DR.

After an individual evaluation of each method, we compare the summarization methods. Resulting summaries were evaluated by F-Measure evaluation metric as shown in Table 5.10..

TABLE 5.10.: Effect of intrinsic properties on EventCitySum with F-Measure metric

Intrinsic properties	DPP	DR
No property	0.106	0.114
Memorability	0.079	0.048
Sentiment	0.111	0.094
Scenicness	0.130	0.136
Aesthetics	0.123	0.119

Memorability score degrades the summarization process clearly by decreasing the evaluation metric for both methods. Scenicness and aesthetics improve the summarization process consistently. However, it is hard to make these kinds of evaluations for sentiment intrinsic property, because it's effect is not stable across summarization methods. For example, the DPP method improves the quality of the summary from 0.106 to 0.111 while it decreases the quality of the summary from 0.114 to 0.094. We also can not surely say that DPP is better than DR or vice versa while the amount of improvement varies between methods.

Then, we extracted summaries with the combination of intrinsic properties to see maximum benefit might be gained from the inclusion of intrinsic properties. Memorability is excluded while it does not provide any improvement on the summaries. In the pairwise combinations, we give equal weight to each intrinsic property which 0.5. Like pairwise combinations in the

triple combination, we assign equal weight which is 0.33 for each intrinsic property. Results can be seen in Table 5.11..

TABLE 5.11.: Combination of intrinsic properties on EventCitySum with F-Measure metric

Intrinsic properties	DPP	DR
No property	0.106	0.114
Senti - Scenic	0.125	0.151
Senti - Aesth	0.120	0.134
Scenic - Aesth	0.129	0.166
Senti - Scenic - Aesth	0.130	0.175

Triple combination is our best combination according to Table 5.11.. This result can be interpreted such that each intrinsic property is valuable for the summarization process.

Lastly, we present the effect of intrinsic properties against baseline methods in Table 5.12.. Similar to previous experiments, to increase the reliability of results, we extend our analysis with V-Rouge to show coherency across different evaluation metrics.

TABLE 5.12.: Comparison of summarization methods with intrinsic properties on EventCitySum dataset

	Intrinsic properties	Uniform	K-Means	DS3	SRNN	DPP	DR
F-Measure	No property	0.108	0.099	0.113	0.109	0.106	0.114
	Best Combination	-	-	-	-	0.130	0.175
V-Rouge	No property	0.319	0.323	0.344	0.343	0.329	0.334
	Best Combination	-	-	-	-	0.366	0.394

Our results clearly show that the integration of intrinsic properties to summarization methods improves the summaries. People choose more aesthetic, more scenic and more positive emotional images to a summary.

We also investigate the size of the image collection on the summarization process. For this reason, we decrease the size of our original sets by sampling. First, we sample the images that occur in the ground truth human summaries. Then, we sample randomly to reach to corresponding set size. This approach works when unique images in the ground truth human

summaries do not pass the sampling size 75 and 50. Unique image count in ground truth data is lower than 75 in our dataset EventCitySum. Hence, we can employ all the image collections in the experiments when we fixed the set size to 75. On the other hand, when we experiment with set size 50, we can employ only four image collections which fulfil the condition. Results are shown in Table 5.13..

TABLE 5.13.: Comparison of summarization methods with different set sizes on EventCitySum dataset with F-measure metric

Set Size	Uniform	K-Means	DPP	DR
100	0.108	0.099	0.130	0.175
75	0.151	0.139	0.172	0.207
50	0.229	0.200	0.242	0.254

Results show that the even if we change the size of the image collection that is summarized, the positive effect of intrinsic properties is preserved. Proposed summarization methods also outperform the simple baselines.

We also show the effects of intrinsic image properties on the Vacation Photographs dataset in Table 5.14. based on the F-Measure metric and in Table 5.15. based on the V-Rouge metric.

TABLE 5.14.: Comparison of summarization methods with intrinsic properties on Vacation Photographs dataset with F-Measure metric

Intrinsic properties	Uniform	K-Means	DS3	DPP	DR
No property	0.227	0.195	0.214	0.197	0.192
Memorability	-	-	-	0.203	0.208
Sentiment	-	-	-	0.225	0.266
Scenicness	-	-	-	0.246	0.270
Aesthetics	-	-	-	0.233	0.253

TABLE 5.15.: Comparison of summarization methods with intrinsic properties on Vacation Photographs dataset with V-Rouge metric

Intrinsic properties	Uniform	K-Means	DS3	DPP	DR
No property	0.367	0.318	0.404	0.336	0.284
Memorability	-	-	-	0.311	0.238
Sentiment	-	-	-	0.363	0.356
Scenicness	-	-	-	0.376	0.338
Aesthetics	-	-	-	0.358	0.312

In Table 5.14., all intrinsic properties increase the quality of summaries. However, individual use of scenicness and aesthetics outperform the simple baseline methods in both summarization methods.

5.2. Qualitative Evaluation

In this section, we visualize the summaries of the baseline methods and the proposed methods. As it is discussed in previous sections, the summarization problem has no straightforward quantitative evaluation. This is why we enrich the results through qualitative evaluations. Visualization is performed on only one selected image collection which is Istanbul Tour B3. for fair comparison among results. These example summaries help us to visualize what these numbers correspond to.

Example uniform summary for Istanbul Tour image collection B3. is shown in Figure 5.1.. This summary can be treated as a good summary because it consists of diverse images and it provides good coverage. Although uniform summarization seems simple, its success is proven by both quantitative and qualitative evaluations.

Our second simple baseline based on K-Means generates different summaries in each run. As we stated, we run K-Means 100 times. For this reason, we present two summaries which have the minimum and the maximum F-Measure scores to show the upper bound and lower

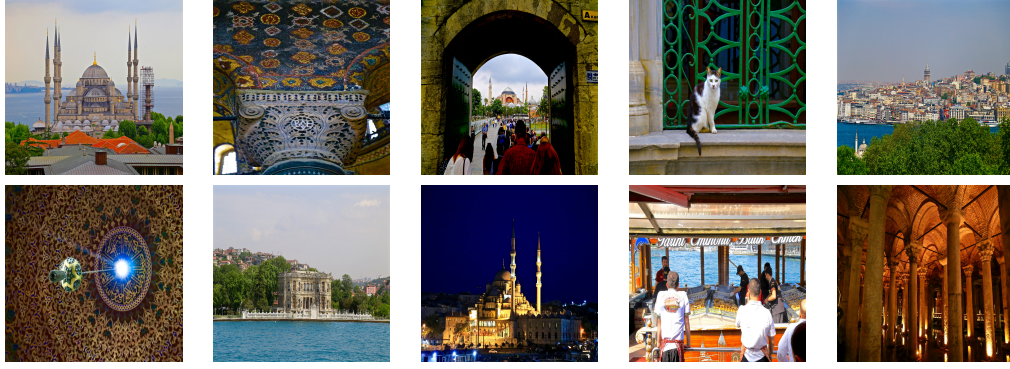


FIGURE 5.1.: Example uniform summary from Istanbul Tour image collection B3.

bound of the K-Means method. Example k-means summaries for B3. image collection is shown in Figure 5.2. and 5.3..



FIGURE 5.2.: Example K-Means summary with maximum F-Measure on an Istanbul Tour image collection B3.

Summary with the maximum F-Measure in Figure 5.2. carries the main summarization properties such as diversity and coverage. It consists of images different from each other and key points of the Istanbul. On the other hand, the summary with minimum F-Measure in Figure 5.3. contains six very similar images which share the same texture. Under these observations, we can say that our quantitative evaluations are consistent with qualitative evaluations.

Example summary that is generated by the DS3 method can be seen in Figure 5.4.. This summary can be interpreted as good summary based on the previous reasoning that holds for uniform and K-Means baselines.

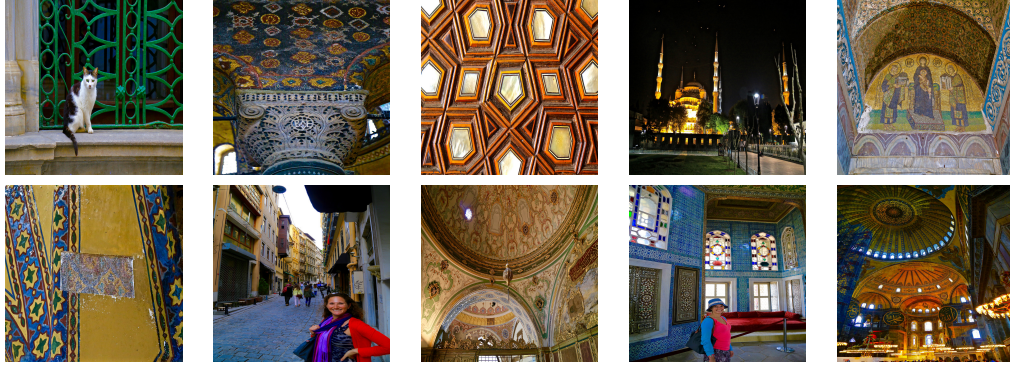


FIGURE 5.3.: Example K-Means summary with minimum F-Measure on an Istanbul Tour image collection B3.



FIGURE 5.4.: Example DS3 summary on an Istanbul Tour image collection B3.

Only learning based method SRNN generates the summary as in Figure 5.5. which consists of images from varying key points of Istanbul such as big mosques, nice views, and towers.



FIGURE 5.5.: Example SRNN summary on an Istanbul Tour image collection B3.

We also use the summarization methods without intrinsic properties to generate summaries. These methods can be handled as a baseline method because our aim is to pass these results

by taking intrinsic properties into account. On discussed setting, example DPP summaries and DR summaries are shown in Figure 5.6., 5.7. and 5.8.. Even in the summary with smallest F-Measure, diversity and coverage are provided.

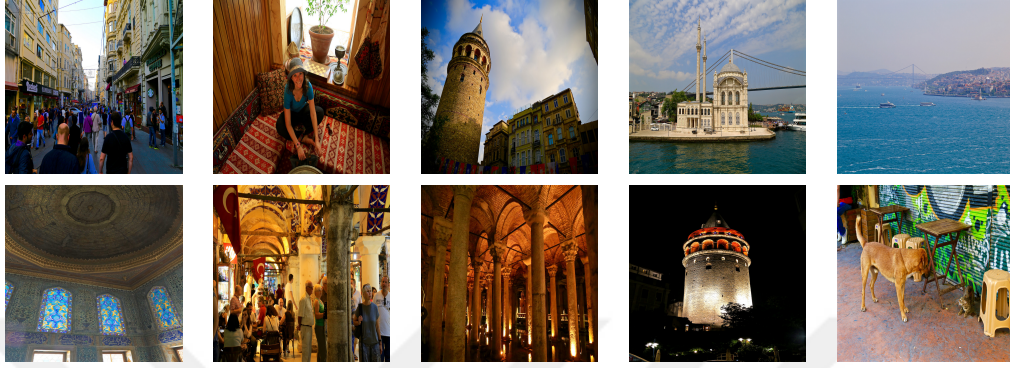


FIGURE 5.6.: Example DPP summary with the maximum F-Measure on an Istanbul Tour image collection B3.

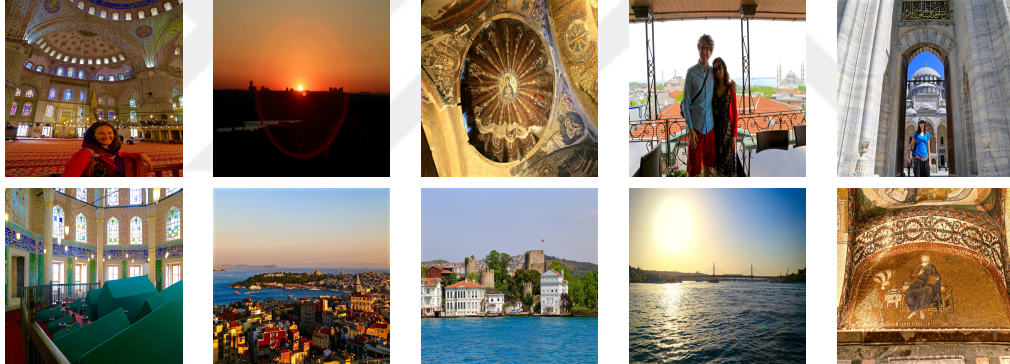


FIGURE 5.7.: Example DPP summary with the minimum F-Measure on an Istanbul Tour image collection B3.

For qualitative evaluation, we show the most successful intrinsic properties for each method. In the DPP method, we get the best result with the scenicness property. In the DR method, we get the best result with aesthetics property. Effect of intrinsic properties varied across methods. In Figure 5.9., the effect of scenicness very obvious. Almost all images contain sea views. Scenicness property suppresses the diversity principle which is not exactly we look for. In Figure 5.10., the aesthetic score does not violate diversity in the DR method. The contribution of intrinsic properties can be observed by the increase in the quality of images.

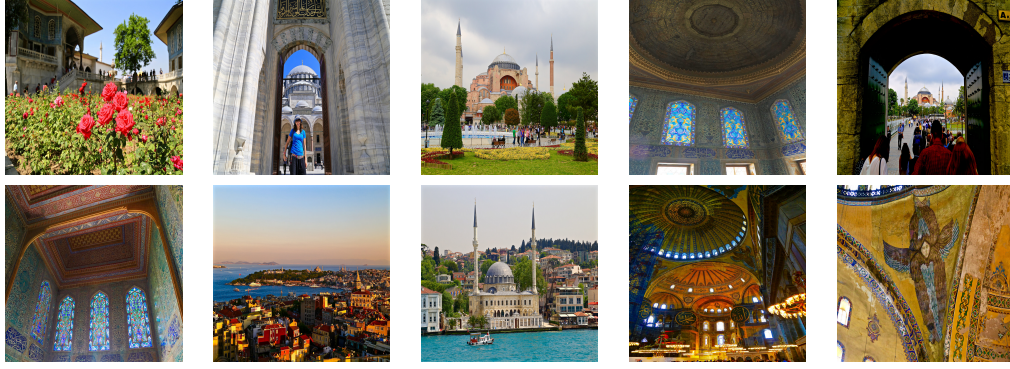


FIGURE 5.8.: Example DR summary on an Istanbul Tour image collection B3.



FIGURE 5.9.: Example DPP summary with intrinsic property scenicness on an Istanbul Tour image collection B3.

As the last example, we present our best results according to quantitative evaluation in which three intrinsic properties are taken into account by equal weights. These intrinsic properties are sentiment, scenicness and aesthetics. Example summaries for each summarization methods are shown in Figure 5.11. and 5.12..

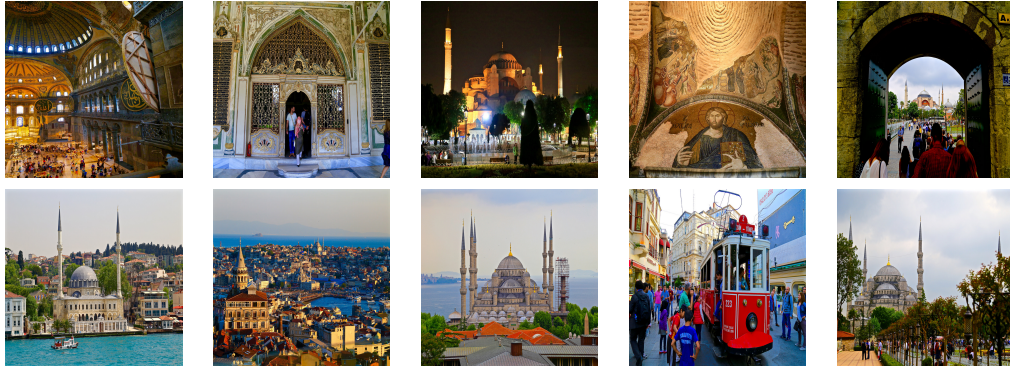


FIGURE 5.10.: Example DR summary with intrinsic property aesthetics on an Istanbul Tour image collection B3.



FIGURE 5.11.: Example DPP summary with triplet intrinsic property combination on an Istanbul Tour image collection B3.

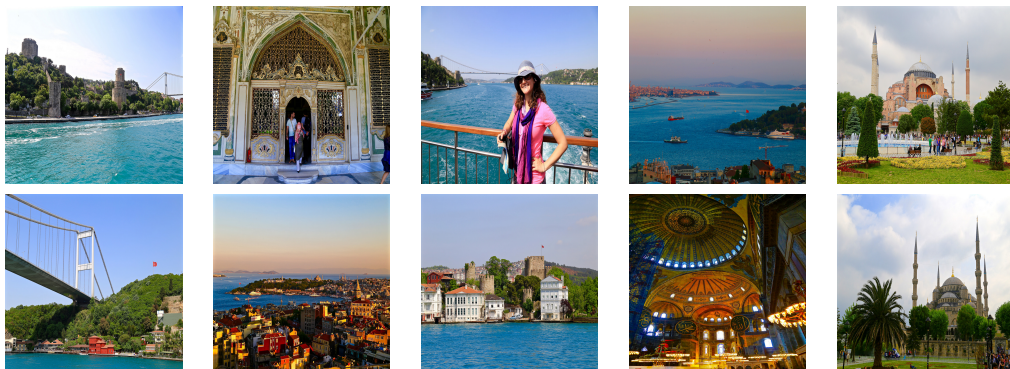


FIGURE 5.12.: Example DR summary with triplet intrinsic property combination on an Istanbul Tour image collection B3.

6. CONCLUSION

In this thesis work, we investigate the problem of visual summarization by using image collections. The main challenges of the summarization task are near duplicate images and ambiguity of the summarization process. We try to overcome these difficulties with various summarization methods and intrinsic properties.

Although near-duplicate detection is a very straightforward process for humans, it is not the case for computers. For filtering near duplicates, image representations play an important role. If our representations are not reliable, we can not expect to get good results from the methods. The first step in the near-duplicate detection, representations of near-duplicates should be close to each other in the feature space. Then, the method should select only one of these similar features. However same objects might be included in the images at various scales which should be also handled with the image representation. These representations should be robust against scale changes to detect near duplicates, this is why we choose RMAC image representation. After that we decide our feature representation, we choose methods which selects images by considering diversity. By choosing diverse images, we prevent the near-duplicate problem. DPP, DR DS3, and SRNN methods focus on the selection of diverse images as the summarization requires.

The second problem for summarization is ambiguity which is proven by low human consistency scores on the image collections. Human consistency scores correspond to upper bound for the problems as in our case. These low human consistency scores also give clues of the hardness of the summarization problem. After the analysis of these consistency values, our first action is employing pruning strategies to find the more consistent subset of summaries which is also necessary to model summarization.

Another critical problem for image collection summarization is the lack of a benchmark dataset. Aesthetics and scenicness works are investigated on huge datasets with thousands of images with their ground truths. Although summarization problem is one of the active research areas, datasets in literature are not published or only partially published. There

is no available dataset which is rich with the ground truth summaries for image collection summarization such as EventSum and CitySum.

In addition to summarization methods, we employ intrinsic properties to improve image collection summarization. These intrinsic properties could be thought of as a quality measure. We experiment with intrinsic properties that we believe that will increase the quality of resulting summaries. However, memorability does not meet with our expectations. It accurately degrades the quality of summaries across all over the experiments with different summarization methods, different features, and different evaluation metrics. On the other hand, remaining intrinsic properties such as sentiments, aesthetics, and scenicness helps us to generate better summaries.

We show that summarization has the minimal dependency on evaluation metrics and image representation by additional experiments. Two evaluation metrics are applied which are V-Rouge and F-Measure each has weaknesses and strengths according to each other. We also use two kinds of features which are combined features and RMAC. Results show that the effect of intrinsic properties preserves similar behaviour across these features and evaluation metrics.

To sum up, we experiment with several summarization methods and intrinsic properties to understand the underlying causes behind human summarization process in this thesis work. For this purpose, we collected two datasets that are EventSum and CitySum. We show that the success of summarization methods are limited. By using intrinsic properties, we can go beyond these limitations.

A User Interfaces to Collect Summaries

We collected human summaries through a website ¹. While this web page is designed specifically for human summary collection and offers functionality which is relevant to summarization. Main summarization page is shown in Figure A1..

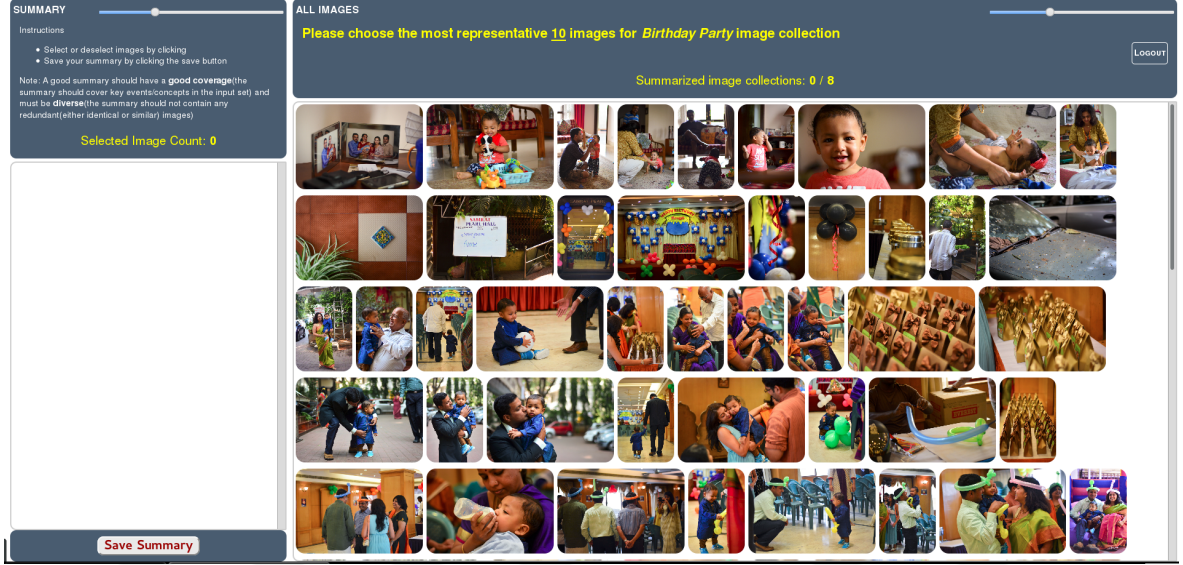


FIGURE A1.: A snapshot of our summary collection page

For people that are not familiar with computer vision field, summarization is a concept on textual data. Hence, they can get confused easily in front of a bunch of images. This situation is called visual crowding in the literature. Visual crowding affects the quality of summaries in a negative way. To avoid visual crowding and to provide clear experiment environment, we kept the following points in mind during the design and implementation of the web page:

1. User specific accounts are preferred to provide users to give a break when they get tired. Our aim is to reduce the effects of distraction on the summaries.
2. Image sizes are critical for summarizing image collections because it specifies the amount of detail in the image that can be perceived by humans. For example, in the

¹<http://vissum.cs.hacettepe.edu.tr/project>

case of eliminating near duplicates, we need more details to see similarities and dissimilarities. On the other hand, when we inspect image collection from the coverage point of view, we need to see the entire collection at a glance. These requirements make zoom in and zoom out property crucial for healthy summary collection process.

3. Displaying image collection and summary at the same time is another important property that directly affects the quality of summaries. For this reason, we distinguish summary images from entire image collection by using two different panels instead of only marking or highlighting the selected images on the same panel. If we only highlight or mark the summary images, diversity and coverage properties cannot be evaluated reliably because of the visual crowd.
4. We sort images by time because the time ordered image collection seems more fluent to a user than the randomly ordered collection. Time ordering is a very powerful method to avoid the visual crowd. It also makes easier near-duplicate detection.

We inform the users only for the following points:

1. We add the name of image collections to our question that is used to download images from Flickr. By doing this, we aim to give general insight to users about the image collection.
2. We remind users that image collection summarization shares similar theoretical baseline with textual summarization by adding a brief explanation :
"A good summary should have a good coverage (the summary should cover key events/-concepts) in the input set) and must be diverse (the summary should not contain any redundant (either identical or similar) images)"
3. Our question is short and clear: "Please choose the most representative 10 images for X image collection" Representativeness corresponds to summaries while each summary element is added to the summary to increase the representativeness of summary.

B Image Collections

Each image collection in our dataset is shown in this section. In our experimental analysis, summarization methods and effect of intrinsic properties on summarization methods are visualized on An Istanbul Tour image collection B3.. Other image collections are also presented in this section, while the summarization problem is investigated through these image collections.

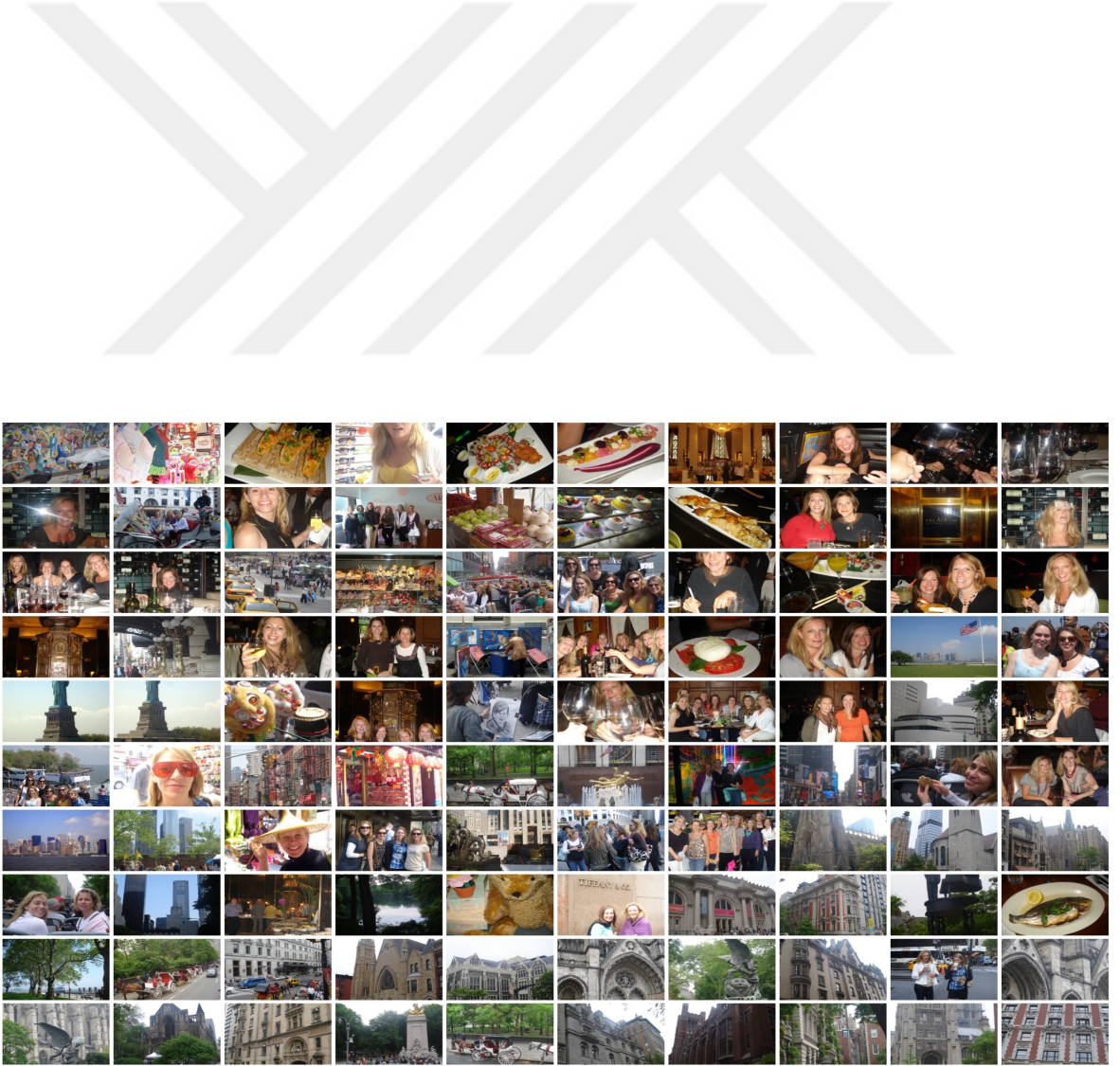


FIGURE B1.: A Trip to New York image collection from EventSum dataset



FIGURE B2.: A Visit to Italy image collection from EventSum dataset

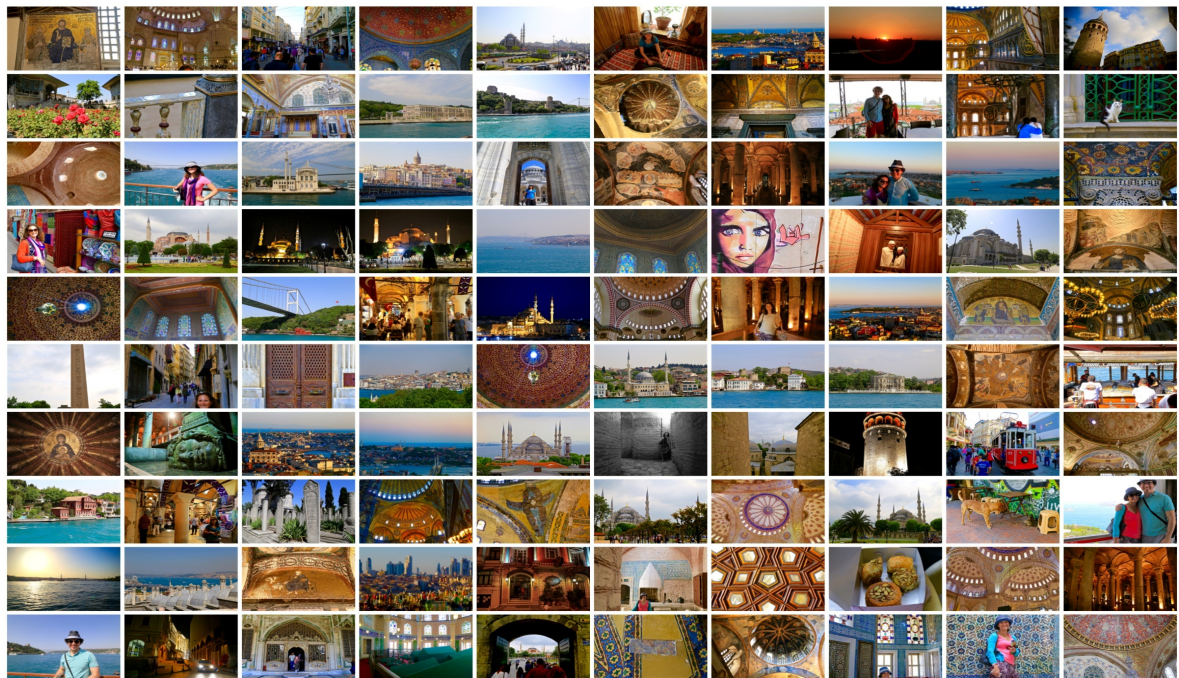


FIGURE B3.: An Istanbul Tour image collection from EventSum dataset



FIGURE B4.: Birthday Party image collection from EventSum dataset

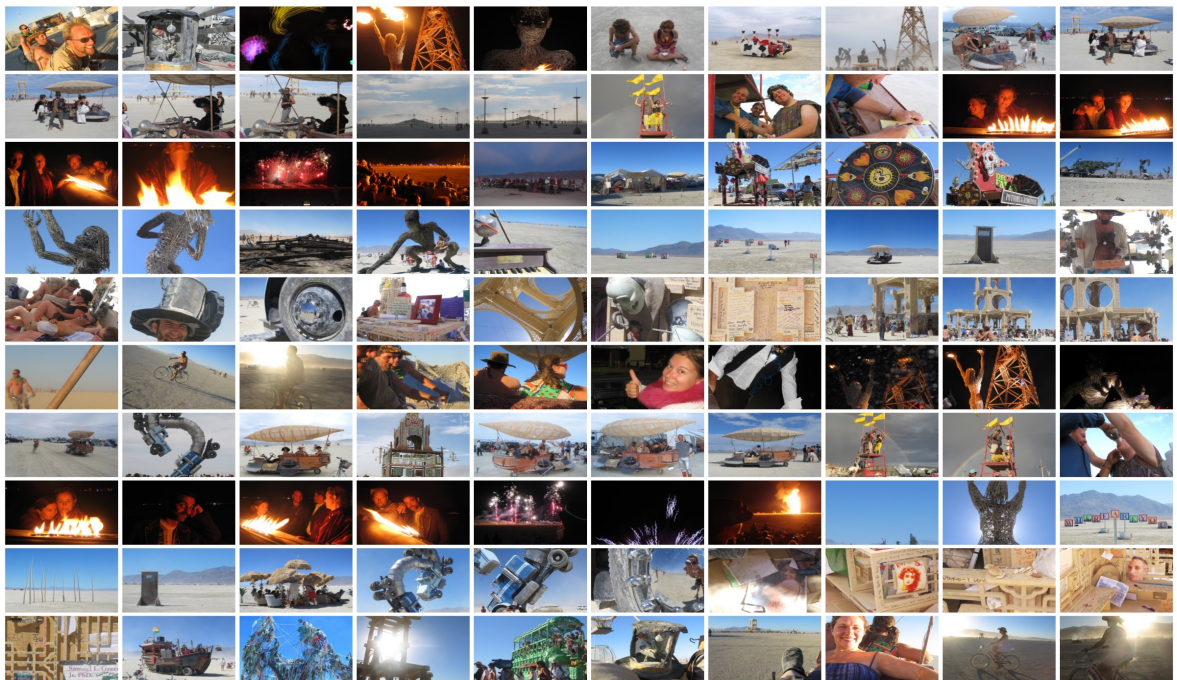


FIGURE B5.: Burning Man image collection from EventSum dataset

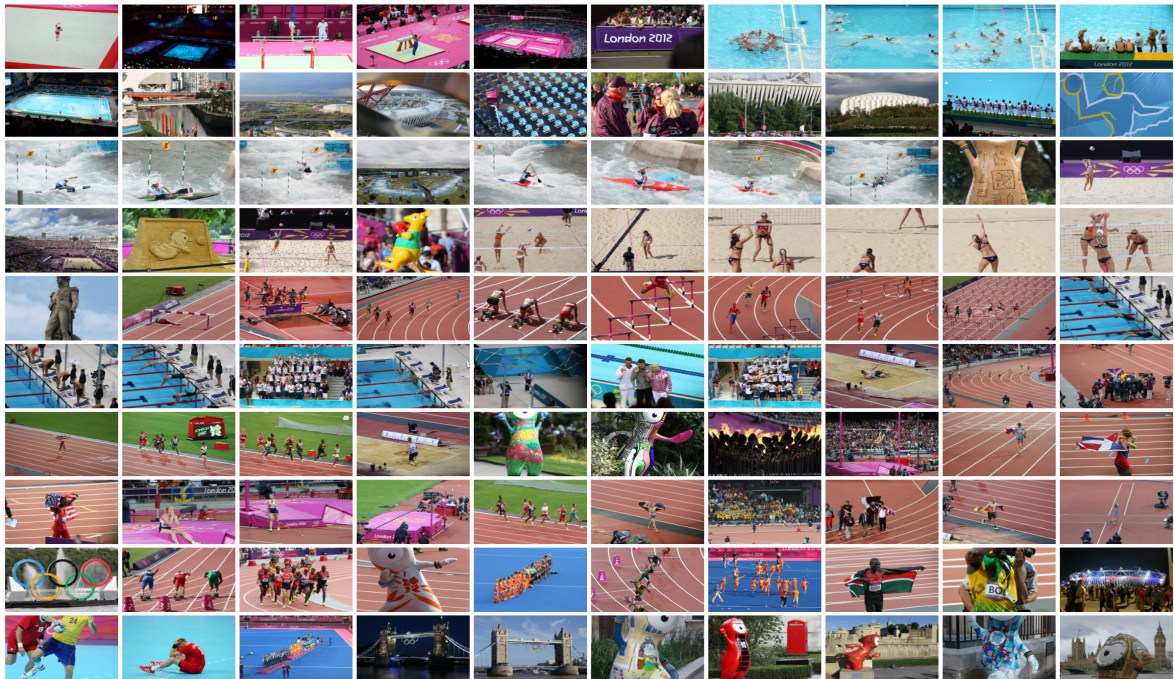


FIGURE B6.: London Olympics image collection from EventSum dataset



FIGURE B7.: St. Patrick's Day image collection from EventSum dataset

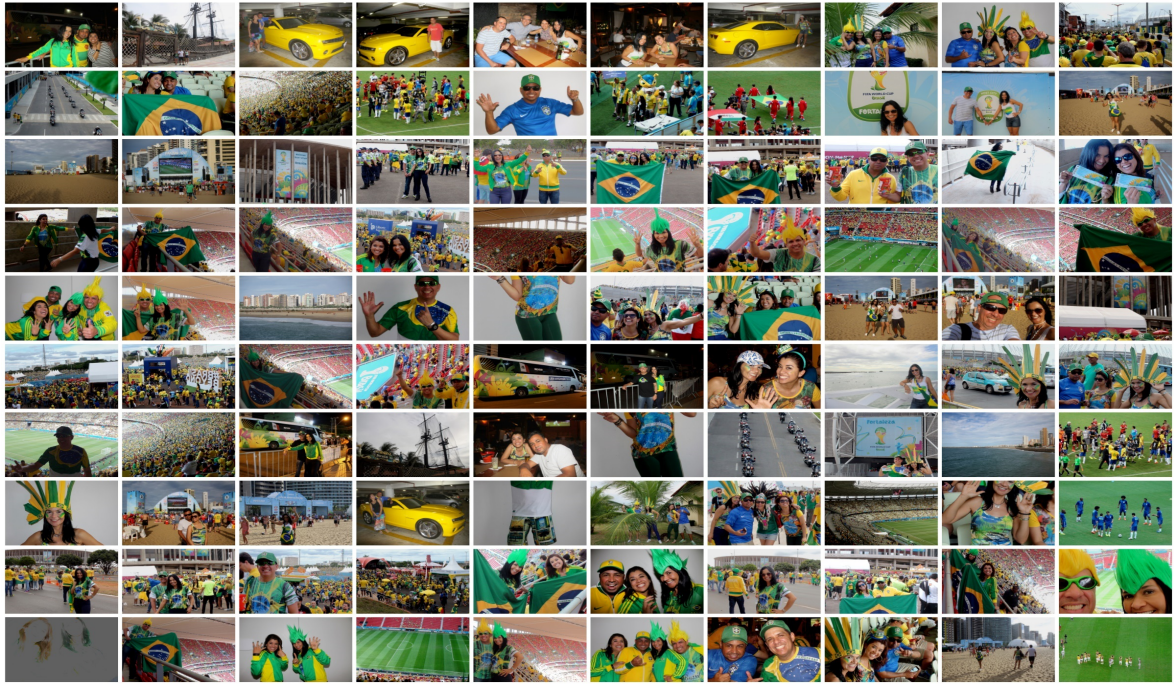


FIGURE B8.: World Cup image collection from EventSum dataset

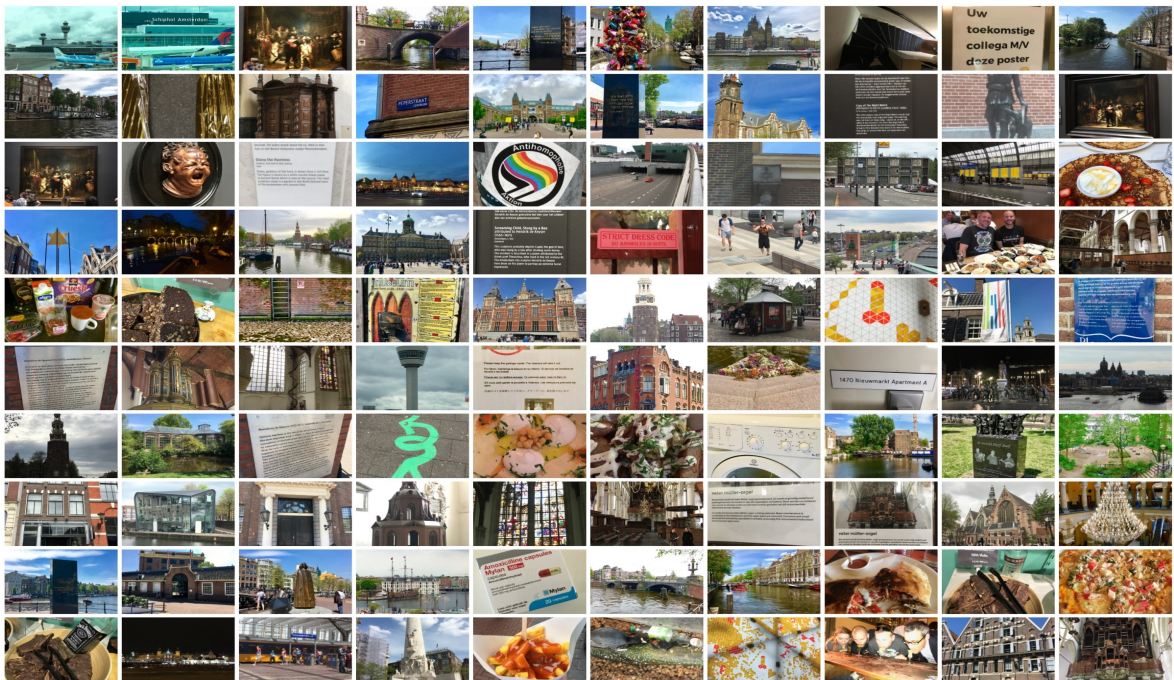


FIGURE B9.: Amsterdam Vacation image collection from CitySum dataset

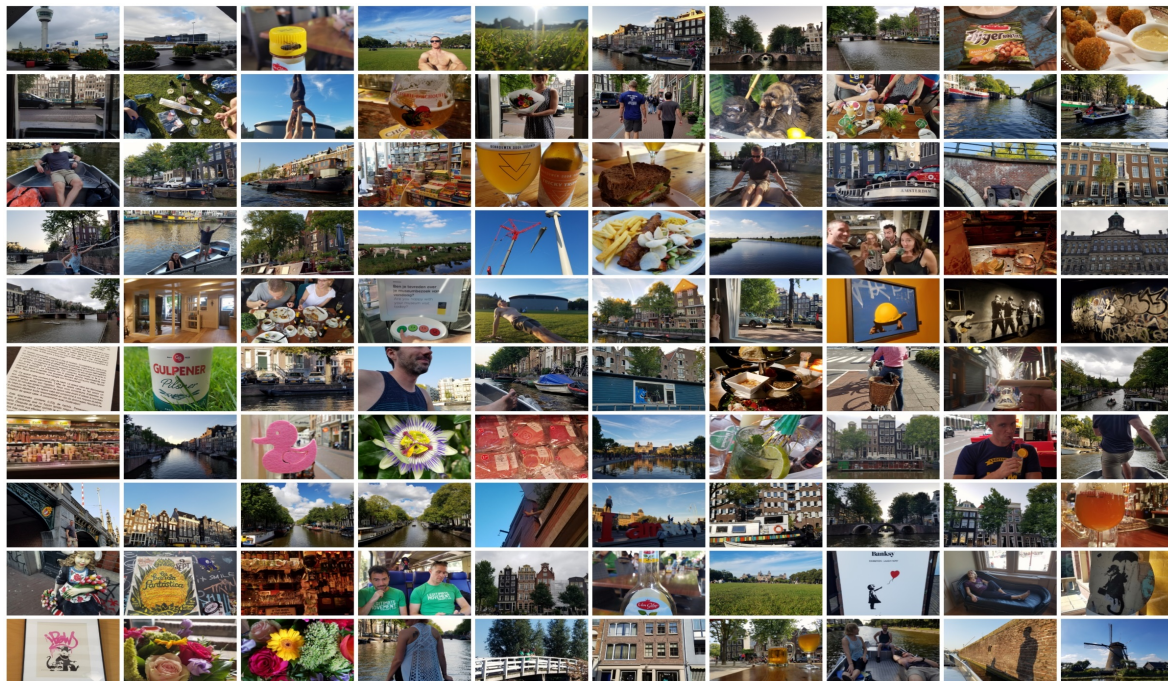


FIGURE B10.: Amsterdam Vacation image collection from CitySum dataset

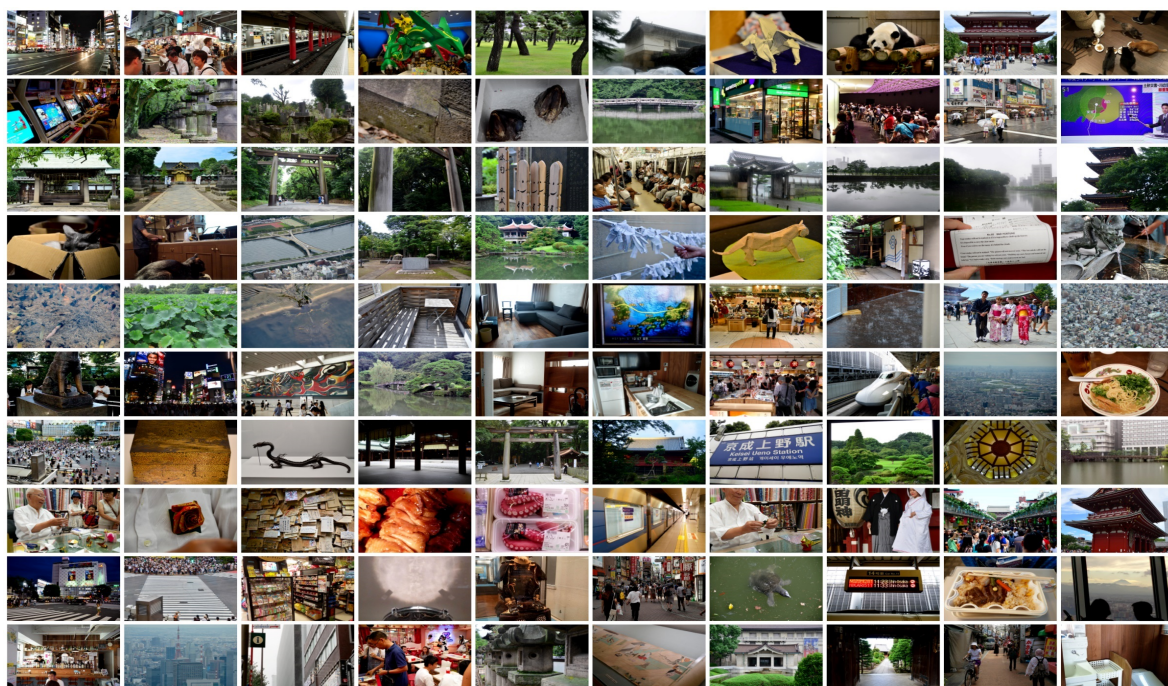


FIGURE B11.: Tokyo Vacation image collection from CitySum dataset

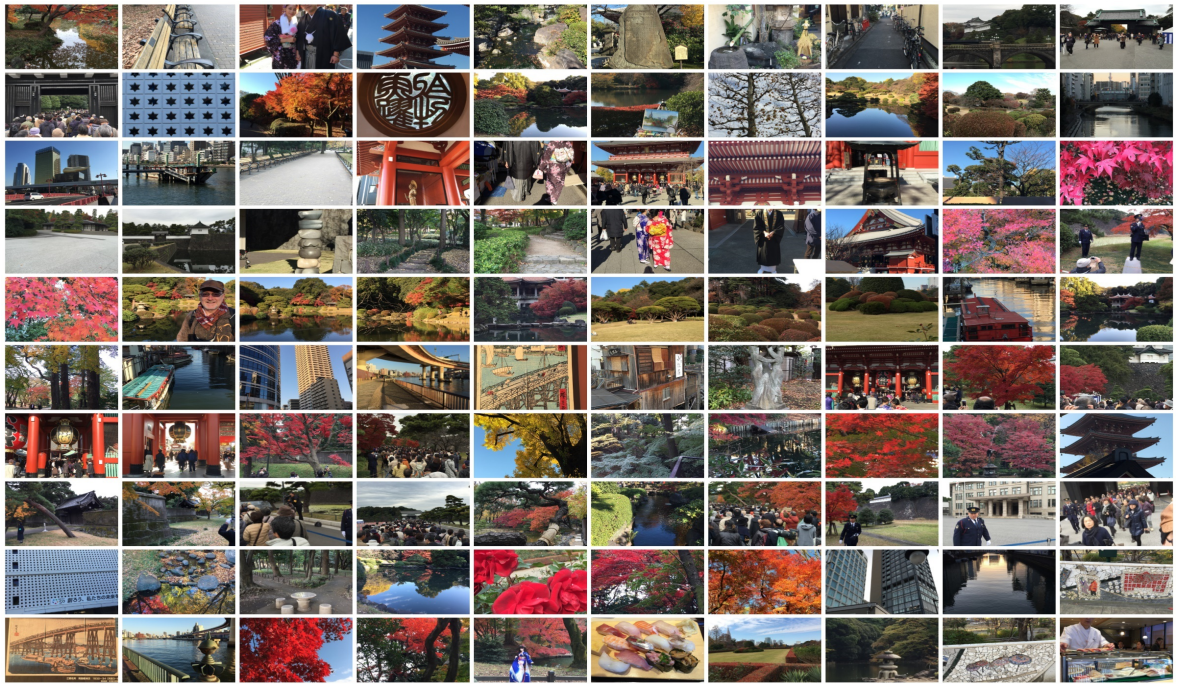


FIGURE B12.: Tokyo Vacation image collection from CitySum dataset

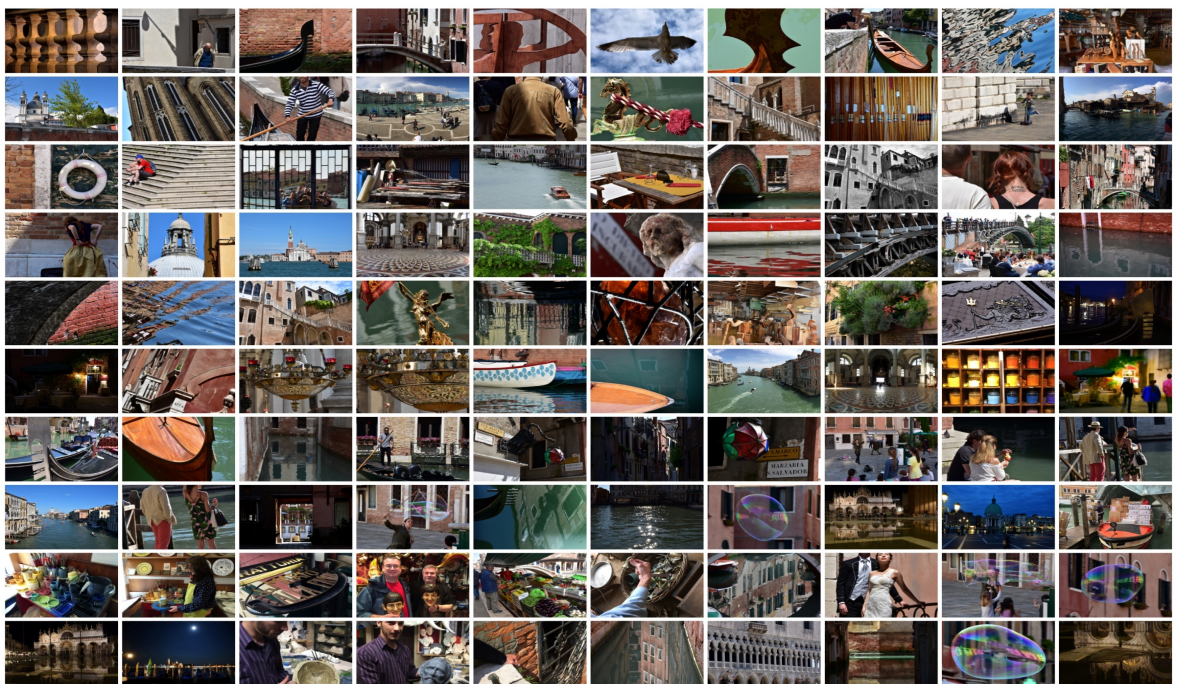


FIGURE B13.: Venice Vacation image collection from CitySum dataset



FIGURE B14.: Venice Vacation image collection from CitySum dataset

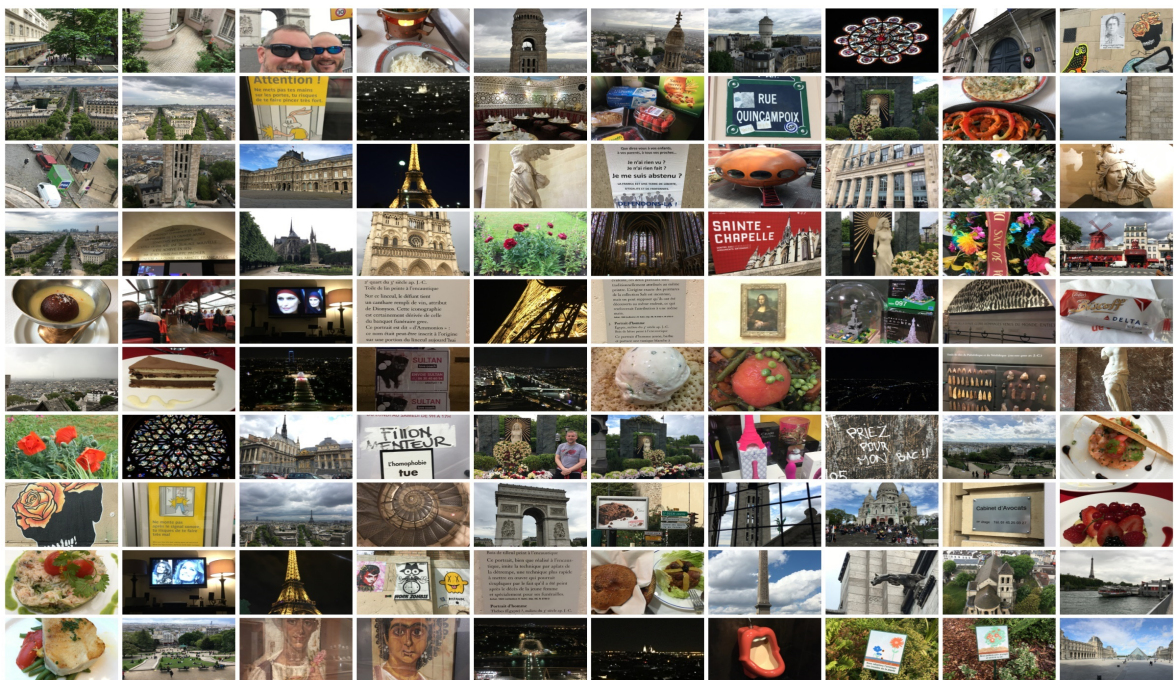


FIGURE B15.: Paris Vacation image collection from CitySum dataset

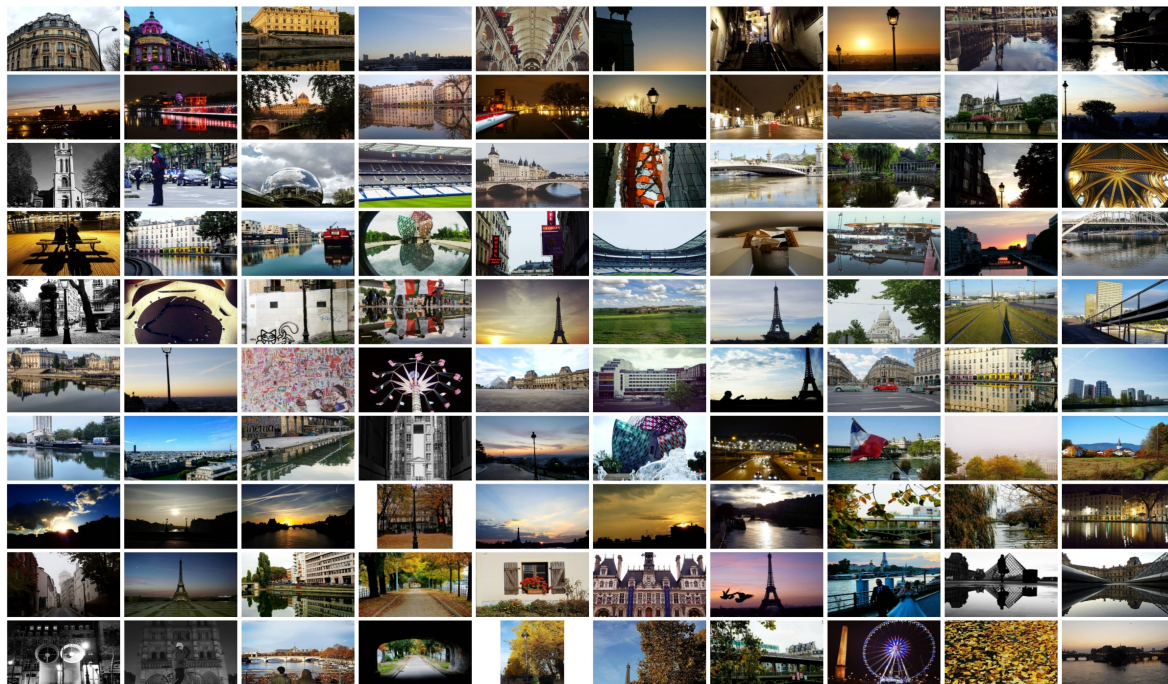


FIGURE B16.: Paris Vacation image collection from CitySum dataset

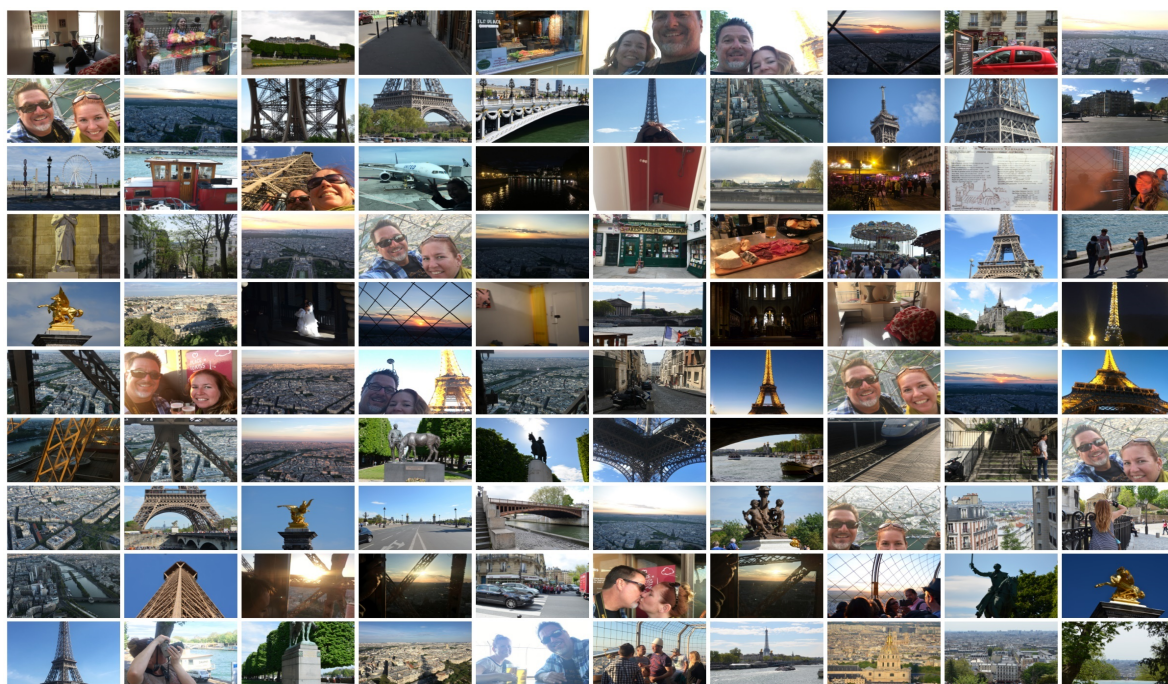


FIGURE B17.: Paris Vacation image collection from CitySum dataset

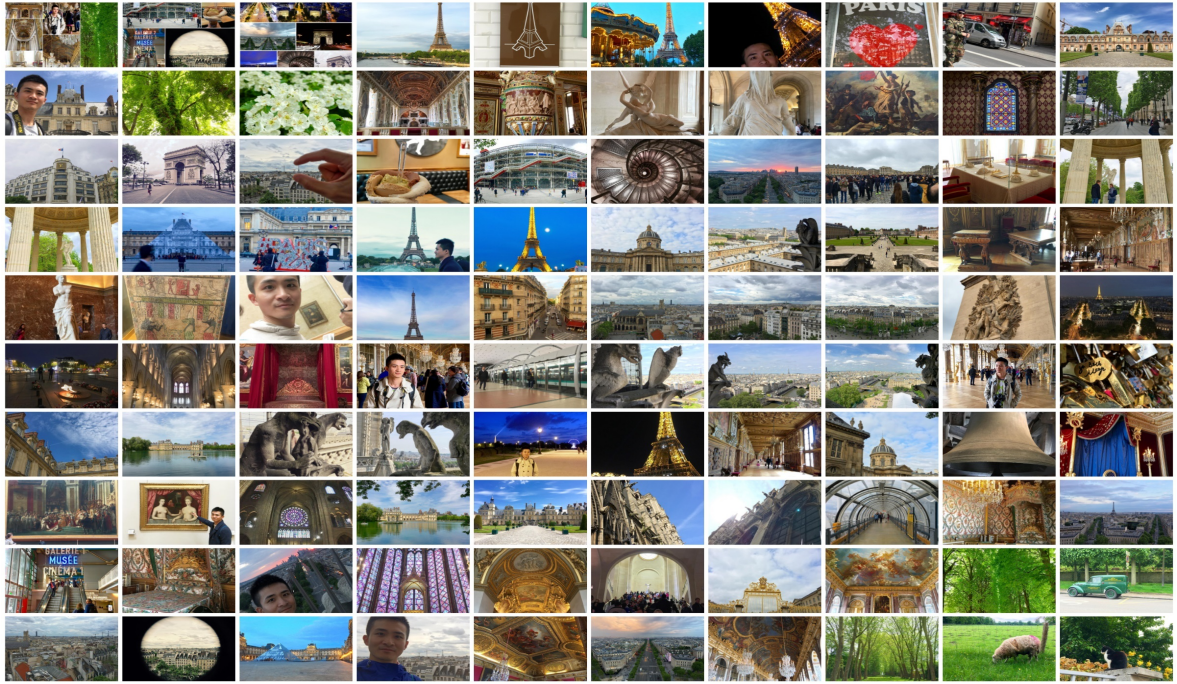


FIGURE B18.: Paris Vacation image collection from CitySum dataset

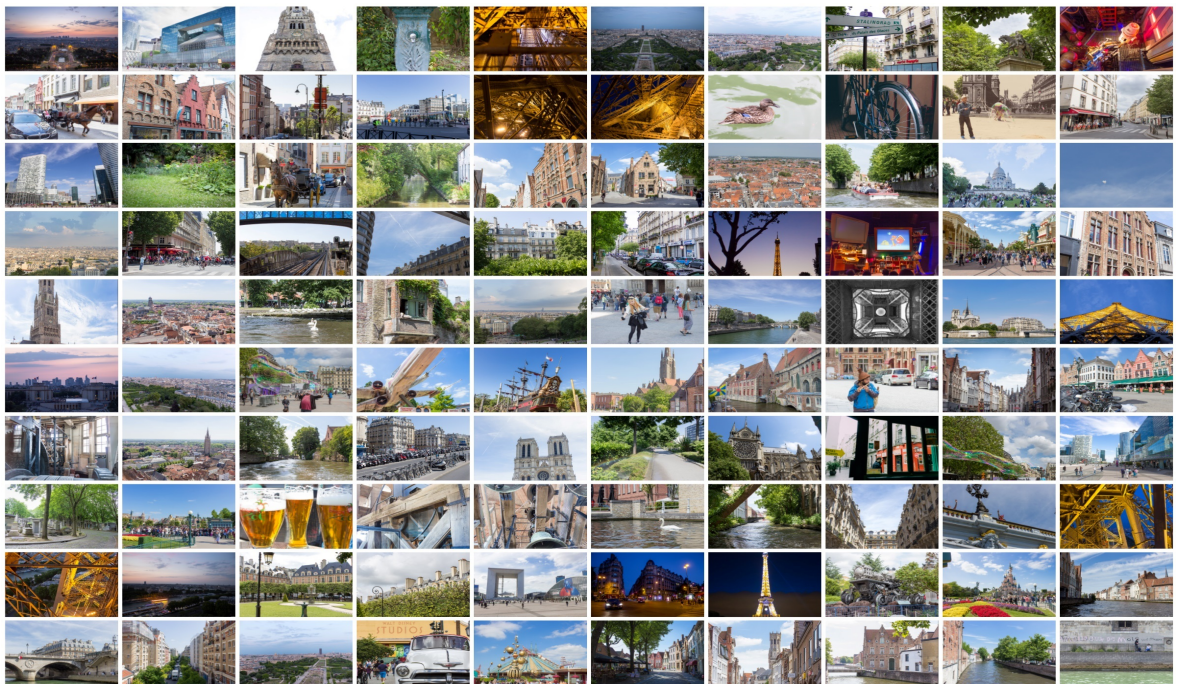


FIGURE B19.: Paris Vacation image collection from CitySum dataset

REFERENCES

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. **2012**.
- [2] Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML-11*, pages 1193–1200. **2011**.
- [3] Gunhee Kim, Eric P Xing, Li Fei-Fei, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, pages 169–176. IEEE, **2011**.
- [4] Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. Dissimilarity-based sparse subset selection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2182–2197, **2016**.
- [5] Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. Learning visual storylines with skipping recurrent neural networks. In *ECCV*, pages 71–88. Springer, **2016**.
- [6] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 4. ACM, **2011**.
- [7] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. ACM, **2014**.
- [8] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *ICCV*, pages 2390–2398. **2015**.
- [9] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs.

- In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, **2013**.
- [10] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and Mapping Natural Beauty. In *ICCV*. **2017**.
 - [11] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *arXiv preprint arXiv:1709.05424*, **2017**.
 - [12] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*. **2014**.
 - [13] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summarization. page 1413–1421. Curran Associates, Inc., **2014**.
 - [14] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*, pages 705–718. Springer, **2008**.
 - [15] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157. Ieee, **1999**.
 - [16] Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, **1975**.
 - [17] Adish Singla, Sebastian Tschiatschek, and Andreas Krause. Noisy submodular maximization via adaptive sampling with applications to crowdsourced image collection summarization. In *AAAI*, pages 2037–2043. **2016**.
 - [18] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, **1990**.
 - [19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, **1997**.

- [20] Fereshteh Sadeghi, J Rafael Tena, Ali Farhadi, and Leonid Sigal. Learning to select and order vacation photographs. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 510–517. IEEE, **2015**.
- [21] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *ICCV*, pages 1633–1640. IEEE, **2013**.
- [22] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. Event-specific image importance. In *CVPR*, pages 4810–4819. **2016**.
- [23] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*. **2016**.
- [24] Giorgos Tolias, Ronan Sifre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, **2015**.
- [25] Joost Van De Weijer, Cordelia Schmid, and Jakob Verbeek. Learning color names from real-world images. In *CVPR*, pages 1–8. IEEE, **2007**.
- [26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, **2014**.
- [27] Ken Chatfield, Victor S Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, volume 2, page 8. **2011**.
- [28] Bora Çelikkale, Göksu Erdoğan, Aykut Erdem, and Erkut Erdem. Submodular visual story graph construction from large scale image collections. In *submitted*. **2018**.

- [29] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495. **2014**.
- [30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826. **2016**.
- [31] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, **2015**.
- [32] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415. IEEE, **2012**.

CURRICULUM VITAE

Credentials

Name,Surname : Göksu ERDOĞAN
Place of Birth : Ankara
Marital Status : Single
E-mail : goksuerdogan@cs.hacettepe.edu.tr
Address : Computer Engineering Department of Hacettepe University
Beytepe,ANKARA

Education

High School : Milli Piyango High School (2006-2010)
BSc : Computer Engineering, Hacettepe University, Ankara, Turkey (2010-2015)
Exchange Semester : Albert-Ludwigs University, Freiburg, Germany (Spring 2014)

Foreign Languages

English : Upper Intermediate
German : Basic Knowledge

Work Experience

Research Assistant at Hacettepe University, Ankara (September 2015- February 2017)
Summer Intern at Aselsan, Ankara (June 2015- August 2015)
Summer Intern at Arcelik, Ankara (June 2013- August 2013)

Areas of Experiences

My research interests are in Computer Vision and Machine Learning.
I have been actively working on these fields since last year of my undergraduate study.

Projects and Budgets

This thesis was partially supported by a grant from The Scientific and Technological Research Council of Turkey (TUBITAK) – Career Development Award 113E497.

Publications

“Summarizing personal image collections with intrinsic properties”, 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, Turkey, May 2016, Göksu Erdoğan, Bora Çelikkale, Aykut Erdem, Erkut Erdem.

“HUCVL at MediaEval 2016: Predicting Interesting Key Frames with Deep Models”, Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 2016, Göksu Erdoğan, Erkut Erdem, Aykut Erdem.

Oral and Poster Presentations



HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
TO THE DEPARTMENT OF COMPUTER ENGINEERING

Date: 17/07/2018

Thesis Title / Topic: Image Collection Summarization with Intrinsic Properties

According to the originality report obtained by my thesis advisor by using the *Turnitin* plagiarism detection software and by applying the filtering options stated below on 17/07/2018 for the total of 58 pages including the a) Title Page, b) Introduction, c) Main Chapters, d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 8 %.

Filtering options applied:

1. Bibliography/Works Cited excluded
2. Quotes excluded
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Science and Engineering Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

17.07.2018

Date and Signature

Name Surname: Gökse Erdoğan

Student No: N14321215

Department: Computer Engineering

Program: Computer Engineering

Status: ☒ Masters ☐ Ph.D. ☐ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED.

Assoc. Prof. Mehmet Erkut Erdem