GENERATING STORIES FROM LARGE SCALE IMAGE COLLECTIONS

BÜYÜK ÖLÇEKLİ GÖRÜNTÜ DERLEMLERİNDEN ÖYKÜ OLUŞTURMA

İSMAİL BORA ÇELİKKALE

ASSOC. PROF. DR. İBRAHİM AYKUT 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 Doctor of Philosophy

in Computer Engineering

June 2020

To my grandmother. . .

ABSTRACT

GENERATING STORIES FROM LARGE SCALE IMAGE COLLECTIONS

İsmail Bora ÇELİKKALE

Doctor of Philosophy, Computer Engineering Department

Supervisor: Assoc. Prof. Dr. İbrahim Aykut ERDEM

June 2020, 93 pages

Making sense of ever-growing amount of visual data that is available on the web is one of the biggest challenges we face today. As a step towards this goal, this study tackles a relatively less-studied topic in the literature, namely generating structured summaries of large photo collections in a purely unsupervised manner. Our methodology relies on the notion of a story graph which captures the main narratives in the data and their complex relationships by means of a directed graph with a set of (possibly intersecting) paths. Our proposed method identifies coherent visual story lines and exploits submodularity to select a subset of these lines which have the maximum coverage. Various experiments and user studies demonstrate that the approach delivers better performance than the previous methods.

Furthermore, this study explores the role of visual attention and image semantics in understanding image memorability. In particular, we present an attention-driven spatial pooling strategy and show that considering image features from the salient parts of images improves the results of the previous models. We also investigate different semantic properties of images by carrying out an analysis of a diverse set of semantic features which encode meta-level object categories, scene attributes, and invoked feelings. We show that these features which

are automatically extracted from images provide memorability predictions as nearly accurate as those derived from human annotations.

Finally, by incorporating the memorability property together with aesthetics into the story graph generation framework, the effects of intrinsic properties on story graphs are explored. Experiments utilizing these memorable and aesthetic story graphs as a prior knowledge base show further improvements.

Keywords: Visual Storygraph, Structured Summarization, Visual Memorability

ÖZET

BÜYÜK ÖLÇEKLİ GÖRÜNTÜ DERLEMLERİNDEN ÖYKÜ OLUŞTURMA

İsmail Bora ÇELİKKALE

Doktora, Bilgisayar Mühendisliği

Danışman: Doç. Dr. İbrahim Aykut ERDEM

Haziran 2020, 93 sayfa

Web'de mevcut olan ve giderek artan miktarda görsel veriyi anlamak, bugün karşılaştığımız en büyük zorluklardan biridir. Bu hedefe doğru bir adım olarak, bu çalışma literatürde nispeten daha az çalışılmış bir konu olan "tamamen güdümsüz olarak büyük ölçekli fotoğraf kümelerinden yapısal özetler oluşturma" konusunu ele almaktadır. Metodolojimiz, verideki ana anlatıları ve karmaşık ilişkileri yakalayan ve bir dizi (muhtemelen kesişen) öykü yollarından oluşan bir yönlendirilmiş grafik oluşturmaya dayanır. Önerdiğimiz yöntem, veriden tutarlı görsel öykü şeritlerini çıkartır ve bu şeritlerin maksimum kapsama sahip bir alt kümesini seçmek için alt-modülerlikten yararlanır. Çeşitli deneyler ve kullanıcı çalışmaları, yaklaşımın önceki yöntemlerden daha iyi performans sağladığını göstermektedir.

Ayrıca, bu çalışma görsel dikkat ve görüntü semantiğinin görüntü hatırlanabilirliği üzerindeki rolünü araştırmaktadır. Özellikle, dikkate dayalı bir havuzlama stratejisi kullanarak görüntülerin dikkat çekici kısımlarından gelen görüntü özelliklerinin kullanılması, hatırlanabilirlik tahmin sonuçlarını iyileştirdiğini göstermektedir. Ayrıca, meta-düzey nesne kategorilerini, sahne niteliklerini ve duyguları kodlayan özelliklerin bir analizini yaparak görüntülerin farklı semantik özelliklerini araştırmaktadır. Görüntülerden otomatik olarak çıkarılan bu özelliklerin,

neredeyse insanlardan toplanan hatırlanabilirlik tahmin skorlarına yakın hatırlanabilirlik tahminleri sağladığı gösterilmektedir.

Son olarak, hatırlanabilirlik özelliğini estetikle birlikte öykü grafiği oluşturma methodolojisine dahil ederek, içsel özelliklerin öykü grafikleri üzerindeki etkileri araştırılmaktadır. Oluşturulan yeni öykü grafikleri üzerinde gerçekleştirilen deneyler, grafiklerin bir öncül bilgi tabanı olarak kullanıldığında daha iyi sonuçlar verdiğini göstermektedir.

Anahtar Kelimeler: Görsel Öykü Grafiği, Yapısal Özetleme, Görsel Hatırlanabilirlik

ACKNOWLEDGEMENTS

First and foremost, I would like to thank to my supervisor Assos. Prof. Dr. İbrahim Aykut ERDEM and my co-supervisor Assos. Prof. Dr. Erkut ERDEM for their guidance, support, encouragements and toleration throughout the preparation of this thesis.

Furthermore, I would like to thank my thesis committee members, Assos. Prof. Dr. Nazlı İKİZLER CİNBİŞ and Prof. Dr. Pınar KARAGÖZ for reviewing my thesis and their valuable comments.

I would like to thank all of the members of Hacettepe University Computer Vision Laboratory (HUCVL).

I deeply thank, my parents Sema and Şemsettin, and also my brother Arda for supporting and encouraging me throughout my educational life even if there is no such way to thank my family enough. They always trusted and believed in me.

I would also like to thank the Scientific and Technological Research Council of Turkey (TUBITAK) for supporting this thesis with Research Programs, 113E497 and 116E685.

GENİŞLETİLMİŞ ÖZET

Son yıllarda internet üzerinde toplanan ve biriken görsel veri miktarı büyük boyutlara ulaşmaktadır. Özellikle sosyal ağlar ve bulut teknolojilerinin artmasıyla insanların ortak katılımı sağlanarak büyük çapta görüntü veri kümelerinin oluşturulması sağlanmaktadır. Bu verinin büyüklüğüne bağlı olarak sahip olduğu bilgi miktarı da dikkate alınması gereken önem arz etmektedir. Bu bilgiyi yine bu veri yığını içerisinden çıkarmak, bugünün en büyük zorluklardan biridir. Bu çalışma belirtilen problem için bir yaklaşım olarak görsel öykü grafikleri oluşturma yöntemini ele almaktadır.

Çalışmada öncelikle öykü grafiğinin tanımı yapılarak olası kullanım alanları incelenmiş ve bu konuda yapılan çalışmalar araştırılıp, kategoriler altında toplanarak verilmiştir. Daha sonra önerilen öykü grafiği oluşturma metodolojisinin detaylı açıklamasına geçilmiştir. İlk olarak öykü grafiğinin temel özellikleri olan tutarlılık, kapsam ve bağlanılırlık tanımları yapılmış ve her bir özellik için formüller oluşturulmuştur. Öykü grafiğinin oluşturulması da bu özellikleri kullanan bir optimizasyon problemi olarak formülize edilmiştir.

Görüntülerin ifade şekilleri olarak güncel bir çalışma olan "Konvolüsyonların Bölgesel Maksimum Aktivasyonları" (RMAC) derin öğrenme yaklaşımı kullanılmıştır. Buna ek olarak ifade şekillerine fotoğrafların meta verilerinden metin tanımları da dahil edilmiş, ayrıca kısıt olarak zaman damgaları ve coğrafi konum bilgileri de kullanılmıştır.

Oluşturulan ifade şekilleri üzerinde bir tutarlılık grafiği oluşturularak, bu grafikten tutarlı kısa görüntü zincirleri elde edilmiştir. Daha sonra bu zincirlerin üst üste bindirilmesi ile daha uzun öykü şeritleri oluşturulmuştur. Uzun öykü şeritlerinden, fotoğrafların maksimum görsel ve metinsel elemanlarını kapsayan bir alt kümesi belirlenerek ön öykü grafiği oluşturulmuştur. Son olarak yine uzun öykü şeritleri arasından, kapsam miktarında kısıtlı bir azalmaya izin verecek şekilde bağlantı noktalarına sahip şeritler bulunarak değiştirilmiş ve son öykü grafiği elde edilmiştir. Gerçekleştirilen kullanıcı deneyleri sonuçlarına göre önerilen yöntem ile oluşturulan öykü grafikleri, benzer çalışmalardan daha iyi tutarlılık ve kapsam değerlerine sahiptir.

Oluşturulan öykü grafiklerin değerlendirilmesi için bir görsel özetleme deneyi planlanmıştır. Bu deneyde amaç öykü grafiğini öncül bilgi olarak kullanarak nitelikli görsel özetler oluşturabilmektir. Bunun için öncelikle YFCC100M veri kümesi içinden 6 turistik şehir seçilerek bu şehirler için seyahat fotoğrafları taranmıştır. Bu sayede toplamda 25118 fotoğraftan oluşan, her şehir için birer kollektif veri kümesi elde edilmiştir. Öykü grafiklerini kullanarak oluşturulan özetler çeşitli referans yöntemler ile birlikte güncel çalışmalar ile karşılaştırılmış, daha iyi sonuçlar elde edildiği gözlemlenmiştir.

Diğer bir çalışmada görüntülerin dikkat çeken bölgelerinin hatırlanabilirlik üzerindeki etkisi araştırılmıştır. Bunun için görüntülerin sadece dikkat çeken bölgelerindeki özniteliklerini kullanan bir havuzlama yöntemi geliştirilmiş ve bu öznitelikleri kullanan bir ifade şekli oluşturulmuştur. Bu ifade şekli ile, diğer benzer çalışmalardan daha yüksek hatırlanabilirlik tahmin sonuçlarına ulaşılmıştır. Buna ek olarak anlamsal özniteliklerin de eklenmesi sonuçları daha da iyileştirmiş, görüntülerin anlamsan özelliklerinin de hatırlanabilirlik üzerinde etkisi olduğu gösterilmiştir.

Son olarak geliştirilen öykü grafiği oluşturma yaklaşımına, görüntülerin hatırlanabilirlik ve estetik özellikleri de eklenerek bu özelliklerin etkileri araştırılmıştır. Daha önce gerçekleştirilmiş görsel özet oluşturma deneyleri, yeni oluşturulan hatırlanabilir ve estetik öykü grafikleri üzerinde de uygulanarak daha iyi özetleme sonuçlarının elde edildiği gözlemlenmiştir.

CONTENTS

Page

TABLES

FIGURES

1. Introduction

Traveling and discovering locations contributes a lot to one's life. *"Traveling turns you into a storyteller"* said Ibn Battuta in the middle age. While it is still a valid and effective motive, today we have technology to build those stories for us in terms of visual narratives. When we are planning a trip to a place we have never been before, we usually use a travel app or visit websites such as `tripadvisor.com` or `wikitravel.org` to choose which places to visit and what to do in that destination. City guides which were prepared by professional travelers typically include essential information about the attractions, museums or parks in that city. Hence, each traveler, in a way, joins a collaborative act of living and enjoying the city and its culture. This joint act is clearly visible when we look at related travel photo albums shared on the web. Of course, the individual details can vary across trips, but common elements manifest themselves, providing collaborative stories about a city. Same landmark locations and attractions are visited regularly by tourists, and are being photographed again and again.

Together with shared landmarks and locations on these photo albums, individual photographers have their own taste of aesthetics and preferences while they are taking photos. Inevitably, these personal preferences will take part on their craft. So, when we examine the photo albums we would have different feelings evoked and we personally prefer specific albums to the others. This artistic property of photos adds value to the photo itself and deserves special consideration.

We propose a novel approach to automatically generate an informative visual summary of a specific city directly from a large set of travel photo albums related to that city. We formulate this task as a sub-modular optimization problem in which the structured summary is represented in terms of a *story graph*, providing information about different characteristics of the city. Furthermore, we investigated the effects of intrinsic properties of images on these story graphs. More specifically, we utilized memorability and aesthetics properties of photos on the construction phase of the story graphs.

We put more emphasis on the memorability property due to it's less obvious characteristic and hard to predict nature. We proposed an attention-driven framework to predict memorabilities of photos. Specifically, we utilize attention maps extracted from photos to identify regions to be considered on the prediction scheme. Our intuition is the memorability property of a photo will have extensive effect on the story graphs.

Similarly for the aesthetics property, our intuition is that people generally prefer to see and visit interesting and visually appealing locations during their vacations. So incorporating this property into our visual summary construction solution should yield improved results on user preference. We utilized the state-of-art visual aesthetics prediction method in our framework and analysed the outcomes.

1.1. What is a Storygraph?

In general, a story graph is a representation which allows to illustrate the common relationships between data samples in an informative manner, and has been a topic of interest in the scientific community lately. For instance, story graphs have been used to create summarizes of news articles [9], scientific papers [10], ego-centric videos [11] and the interactions among different characters in a movie or TV series [12].

Story graph representation is useful in terms of exhibiting the associations and interrelations of different aspects over the information overload. Furthermore, it presents the data in a non-linear way that facilitates extracting information and capturing patterns. Consider in Figure 1.1., on left there exist a pile of photos where it consist of a visual information load and it is hard to catch any patterns or extract useful information. On the right a story graph is extracted from the same pile. The graph shows several distinct and coherent paths providing some useful insights about the information the pile is hiding in terms of visual stories or semantic connections between photos.

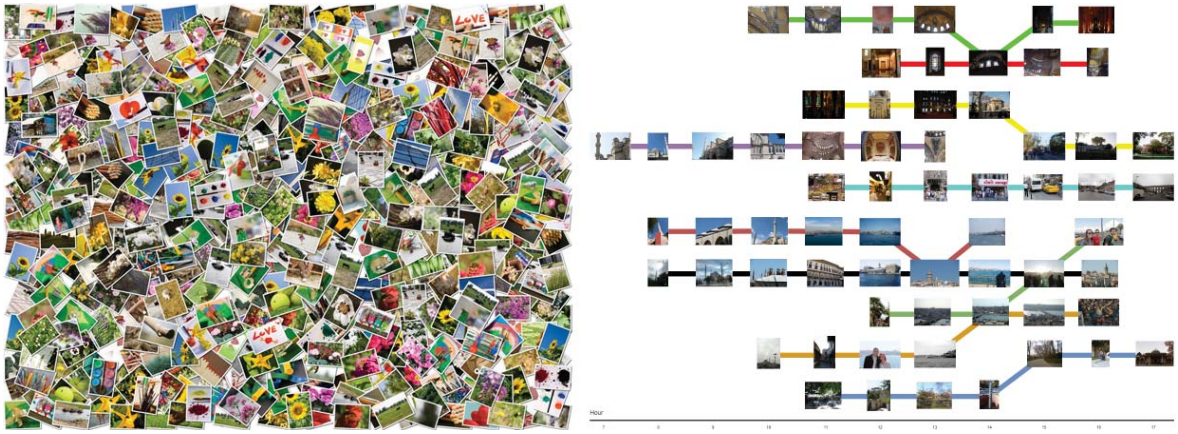


Figure 1.1. Left: A travel photo album consisting of huge amount of photos where it is not practical to extract or acquire information. Right: A story graph constructed from the photo pile. Several distinct story paths shown in different colored lines indicate diverse themes that can be seen during a travel in this location.

1.2. Thesis Statement and Contributions

This thesis research grounds on the following statement:

> "We create visual information maps to enhance user experience over handling collaborative and massive photo collections."

Given tens of thousands of images of a city, in our work, we aim to identify a few story lines that (1) are coherent, i.e. each tells a coherent but different story, (2) cover most of the interesting attractions, i.e. they provide collective information regarding important and salient characteristics of the city, and (3) are connected, i.e. they effectively capture the hidden interconnections. Fig. 1.2. demonstrates an example story graph for the city of Istanbul, reconstructed automatically with our framework by analyzing lots of related travel photo albums. The main contributions of our work are as follows:

- ⋆ We develop a collaborative summarization approach which exploits visual and textual data as well as geospatial and timestamp information to automatically extract a visual story graph for a large collection of photo albums. Our formulation enforces maximum

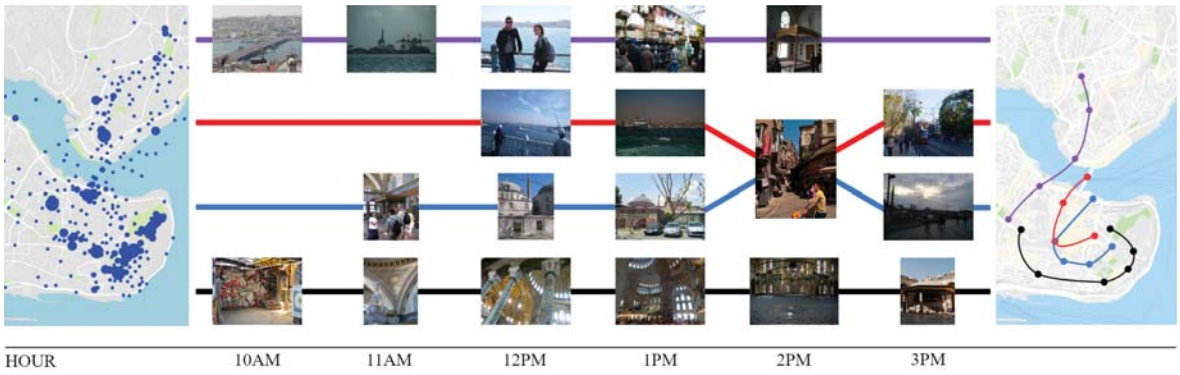


Figure 1.2. A visual story graph generated automatically by our approach for the city of Istanbul. On the left, we show the density map of the geo-tagged images collected from trips to the city of Istanbul. In the middle, we provide some sample story lines which cover coherent and distinct stories. On the right, we draw the story graph on the city map. For illustrative purposes, here we only show four story lines.

degrees of coherency, coverage and connectivity over the extracted story lines, and as it depends on sub-modularity, it is efficient and scalable.

⋆ We introduce YFCC100M-CITIES dataset which includes images of six different cities, annotated with GPS, timestamp tags and textual keywords. It contains in total 132,346 images over 1566 photo albums from 323 users for 6 popular travel destinations in the world.

⋆ We utilize the story graphs generated with our approach as structured abstractions of important concepts, landmarks and events within the photo collections, and demonstrate that they can be employed as a prior in photo album summarization to obtain state-of-the-art results.

⋆ We further demonstrate the effectiveness of our framework with two user studies on next image prediction and tag coverage tasks. Our experimental results show that our model provides better results than the state-of-the-art.

⋆ We analyzed intrinsic properties and their effects on our story graph construction framework. Specifically we showed that utilizing aesthetics and memorability of images further improves the quality of story graphs. Additionally, we proposed an attention-driven framework for memorability prediction of photos.

2. Related Work

We can group the related work on mining large photo collections into three different categories. First group of works focuses on data visualization aspect to allow a user to quickly explore large photo albums [13–19]. Second group addresses summarization of large photo collections by selecting a relatively small set of images based on some desired properties [3, 11, 20–22]. Last group of works, on the other hand, summarizes big visual data in a structured manner by means of story graphs [1, 12]. Instead of selecting a representative set of images, these works aim at reconstructing a narrative where each story line in the graph reflects a major story arc in the image collections.

Additionally, due to the personal nature of story graph characteristics where touristic places and popular locations are tend to take part, intrinsic properties of photos would be a considerable concept. The rest of the works are about this type of intrinsic characteristics. Specifically these works study on the modeling and prediction of memorability scores of photos [4, 6, 23–38] and similarly modeling and prediction of aesthetic scores of photos [39–44].

2.1. Exploratory Data Analysis and Visualization

Recently, there has been much interest in exploratory analysis of big visual data using visualization techniques. Platt *et al.* proposed a method to automatically create an overview of a collection based on clustering and then selecting the representative images from each cluster [13]. Cooper *et al.* suggested a similar framework that depends on clustering of photo collections based on similarities over appearance and temporal characteristics [14]. Kim *et al.* introduced a data-driven method to model and analyze the temporal evolution of the topics of the web images by constructing a large similarity graph of these images through a sequential Monte Carlo based method [16]. Berg and Berg developed an object-centric model to identify canonical images in a set of images collected for a specific object category [15]. Doersch *et al.* proposed a discriminative clustering approach to learn common and distinctive visual elements from large number of photos from a city [17]. Zhu *et al.* introduced

a method which employs average images to let the users browse a large photo collection at ease [18]. More recently, Kleiman *et al*. suggested an approach to search, find and browse similar images on massively large image datasets by projecting their nearest neighbors in a high-dimensional feature space into a 2D layout [19].

2.2. Summarization of Visual Data

A large body of works aims at analyzing big visual data by selecting the most representative images among a given set of images by eliminating the redundant ones. The selection process amounts to capturing the most salient or interesting visual information depending on the task or motivation at hand. For instance, Simon *et al*. developed a photo collection summarization technique which extract the most interesting images over the collection by using a SIFT co-occurrences based clustering framework with a RANSAC loop [20]. Obrador *et al*. approached the summarization process from a supervised learning perspective in which the information from users' online social networks are used as cues [21]. Lu and Grauman proposed a summarization method for egocentric video which relies on segmenting the video into shots and identifying important objects in each shot and then extracts the summary by enforcing coherency based on common objects shared in consecutive shots [11]. Sadeghi *et al*. suggested a method for automatically creating a photo album from a large, unordered collection images, which can be also regarded as an unstructured summarization [22]. The authors, in particular, employ a discriminative structured model to capture compelling visual narratives through features encoding faces, scene context and certain visual attributes. More recently, Sigurdsson *et al*. have used recurrent neural networks to model long-term temporal relations among photo albums to extract visual story lines and summaries [3].

2.3. Visual Story Graphs

Compared to the previous line of works, visual story graphs have been one of the least investigated topics in the computer vision literature. They serve as means for discovering hidden patterns and structures in large sets of images or videos while summarizing events

and activities in the visual data. In their pioneering work [1], Kim and Xing formulated generating visual story graphs as inferring a sparse time-varying directed graph from multiple photo albums which are collected on a single topic. Tapaswi *et al*. developed a similar graph based summary of videos over the interactions among different characters [12]. Like the aforementioned studies, our approach also differs from the conventional summarization techniques in the sense that it outputs a structured summary depicting different aspects of the photo collections in the form of a story graph. In that regard, the most similar related work to our approach is the method of Kim and Xing [1]. However, our method is fundamentally different from this work in several aspects. Most notably that the approach in [1] does not explicitly try to maximize coverage and connectivity of the story graphs, whereas the approach presented here actually enforces this together with a coherency measure. Here, while the coverage leads to diversity of the images in the story graph, connectivity allows to extract the common aspects which are essential for photo album summarization. Moreover, while the story graphs in [1] are constructed with the nodes as the visual elements, the nodes of our story graphs correspond to individual images.

The story graphs generated from large photo collections can be also interpreted as a prior graph collaboratively constructed for a particular interest. This property makes the proposed approach a convenient tool for photo album summarization since the generated story graphs both provide diverse information regarding the image collections but also encode particular aspects of the visual data that are shared among many users.

2.4. Intrinsic Properties

There had been high interest on predicting intrinsic properties of images such as popularity, aesthetics and memorability. [23] predicted the popularity of an image in terms of how many views it can take on a social site by using both hand crafted features and social cues of the owner of the photo. Similarly by using user information from a social site as features, on [24] authors predicted the popularity based on the relations of other photos in their temporal

neighborhood. Same authors proposed a deep network architecture [25] incorporating both temporal and attention mechanisms to predict popularity.

[26] worked on predicting the popularity of a photo from social websites with a cold start scenario where there is no or limited metadata. They predicted the popularity in terms of number of views and comments. Another work [27] similarly studied on social media photo popularity prediction, utilizing visual sentiment features. They analysed which sentiments has effect on the popularity of a photo. More recently deep neural networks are being proposed for popularity prediction. [28] proposed a multi-modal deep network model with an attention mechanism by utilizing both textual and visual features.

When we look at memorability works, before today's deep learning paradigm dominance which shifted out hand-crafted representation extraction, all the existing image memorability models in the literature followed the general framework of [4]. In the training step, some low and high-level visual features are extracted from the images and they are used together with the corresponding ground-truth memorability scores to train a support vector regression (SVR) machine, which can then be used to predict the memorability score of a given test image. In [4], the authors suggested representing images by means of some low-level image features such as SIFT [45], HOG [46], SSIM [47], GIST [29] and color histograms, and/or some semantic features which can be extracted from object and scene annotations. Their proposed model predicts image memorability significantly better than chance, illustrating that such image memorability models can be developed. Then, a number of models [6, 30–32] have been proposed to improve the results of [4]. In general, these studies examine the prediction problem by investigating new hand-crafted features that the authors consider to be relevant to intrinsic memorability of images.

More recently image memorability works are generally focused on training complex deep learning architectures on large scale data sets. Khosla *et al.* [33] collected a large scale data set for image memorability and greatly improved memorability prediction by fine-tuning a Convolutional Neural Network (CNNs). Similarly Baveye *et al.* [34] fine-tuned a different CNN model and achieved better performance on memorability prediction. They showed that

the architecture of the CNN model is also important on the task. Following the same direction, Zarezadeh *et al*. [35] utilized 3 different CNN models as feature extractors and used their fully connected layers as features. They showed that different layers of the CNN networks have distinct performances on memorability prediction. In [36], Fajtl *et al*. created a 4 layer network combining a convolution layer with an attention mechanism and then a recurrent layer. They improved memorability prediction accuracy with their mixed architecture on two major memorability photo sets. Sidorov O. [37] used pre-calculated memorability scores of photos on a Generative Adversarial Network (GAN) to generate human faces and analysed what characteristics of faces are changing with respect to memorability. [38] carried out an analysis of current works on memorability and aggregated which properties of photos are effective on memorability prediction.

Similarly when we look at the works on visual aesthetics prediction, formerly hand-crafted features were being used. [39] used basic light, colorfulness, hue, saturation and object information to predict aesthetics of the photo. From a different perspective, [40] made use of edge and color distribution together with hue and blurriness for the same task. More recently, after deep learning paradigm became popular, visual aesthetic research shifted to that direction. [41] proposed a deep network utilizing a multi-patch aggregation method and by adding some novel layers they increased the effectiveness. Similarly [42] used a convolutional neural network model extended with a multi-scale adaptive spatial pooling layer and achieved improved visual asthetic prediction accuracy. Another multi-patch based novel convolutional neural network model [43] combined with another layout-aware network for to form hybrid presentation for aesthetic prediction. Finally [44] used again a multi-patch aggregation approach on an end-to-end deep model together with an attention mechanism, achieving state-of-art performance for visual aesthetic prediction. They showed attention is also correlated with aesthetics.

One of our goals in this work is to explore the function of visual attention in predicting intrinsic memorability of images. In that respect, our work shares some motivating factors with the models suggested in [30, 32]. In [30], the authors presented a probabilistic model to measure memorability of image regions, which can be used to predict image memorability

as well as the regions that are more likely to be remembered. Within their framework, they suggested to use saliency maps of images as features along with some other visual features. In [32], the authors performed an eye-tracking experiment on a subset of the images in order to observe which parts of those images attract subjects' attention. They have observed that there is a strong correlation between fixation durations and the memorability scores. In addition, the authors proposed two attention-guided (saliency-oriented) features which are shown to be useful in predicting image memorability.

Beyond visual attention-based features, we specifically aim to investigate the use of attentional mechanisms for selecting relevant features to image memorability. Previous models [30, 32] employ saliency maps or saliency-oriented features as additional images features. In contrast, our key insight is that the visual content in the regions that attract attention is as important as or even more valuable than the whole image content in predicting intrinsic memorability of an image. Thus, whereas prior work [4, 30, 32] employs a fixed pooling layout for feature pooling, we propose to consider a pooling scheme that focuses on salient regions within images. We expect this additional feature selection mechanism will allow us to capture characteristics of images relevant to memorability and accordingly improve the prediction performances of dense image features. The details of our feature pooling scheme will be given in Section 6.1.1..

In this work, we also consider ways to boost the success of memorability predictions by employing high-level descriptors that encode the semantic content of images. Similar to [4, 6, 30, 31], we make use of information regarding to objects in images, scene knowledge and/or attributes. In [6], Isola et al. investigated the use of annotated visual attributes to estimate memorability of images. Their study revealed that exploiting available human-describable attributes greatly increases the quality of the predictions. To deeply understand which attribute is a better indicator of memorability, they investigated a greedy feature selection approach to select the best set of relevant attributes. In another study [31], the authors proposed two novel spatial features which can be extracted from the object annotations exist in the dataset. While the first feature measures the importance of the object in terms of how close it is to the image center and how large it is, the second feature is related to how much unusual the

coverage of the object is among all other objects from the same visual category. Their results show that both of these features improve the memorability prediction accuracy.

As compared to [6, 31], however, the semantic features that we employ, which encode meta-level object categories [48], scene attributes [7], and invoked feelings [8], have quite a number of distinct benefits. While the semantic features used in the previous models are based on manual image annotations collected from human subjects, these features can be automatically extracted from the images. This allows us to develop a prediction model which can work in the absence of this sort of high-level annotations. Our approach, thus, requires no supervision and has dramatically less complexity in the training and testing. Notably, among the previous studies, only [30] employed such an automatically extracted semantic feature which is composed of the responses of many pre-trained generic object detectors from ObjectBank [49]. However, these ObjectBank features can be considered as limited as compared to our features, specifically the meta-level object categories [48] which represent an image by means of abstract classes of objects in a hierarchical structure obtained by grouping similar object classes and putting forward higher level common features. The details of our semantic features will be given in Section 6.1.2..

To our knowledge, no previous work attempted to improve image memorability prediction based on an attention-guided feature selection mechanism. In this chapter, we will give details of our proposed attention-driven pooling strategy on visual memorability prediction. We experimented on the MIT Memorability dataset, and perform a thorough experimental analysis to validate that selecting features from the salient image regions via our proposed attention-driven pooling strategy can indeed make more accurate predictions of memorability scores. In addition, we study a group of semantic features related to meta-level object categories [48], scene attributes [7], and invoked feelings [8] that can automatically extracted from images (Section 6.1.2.), and analyze their roles in predicting memorability of images. Thus, we provide additional discussion of the results and related work, and include new quantitative comparisons of our combined framework against the state-of-the-art.

3. Background

In this chapter, we will give brief summaries and descriptions of the works that we benefited from in our framework and experimental analysis. These works are:

- ⋆ *Particular Object Retrieval With Integral Max-Pooling of CNN Activations:* We utilized the RMAC representation of Tolieas etal. [2] where it extracts visual patches from the convolution output of the Convolutional Neural Network and form the representation vectors based on these patches.
- ⋆ *Dissimilarity-based Sparse Subset Selection:* We used the D3S method [50] in our visual summarization experiments where we selected the summary subset of photos using our story graphs as a prior knowledge.
- ⋆ *AMNet: Memorability Estimation with Attention:* While we extract our memorable story graphs, we make use of the state-of-art visual memorability prediction framework AMNet [36] in order to extract memorability prediction scores of the photos.
- ⋆ *Attention-based Multi-patch Aggregation for Image Aesthetic Assessment:* Similar to the memorable ones, to construct our aesthetic story graphs we utilized the current state-of-art multi-patch aggregation framework for visual aesthetic prediction of Sheng etal. [44].

For the rest of this chapter, we will explain the key points of these works.

3.1. Regional Maximum Activations of Convolutions

The *Regional Maximum Activations of Convolutions* (RMAC) [2] is an compact image representation based on the convolution layer output which are the activation maps from a convolutional neural network (CNN).

The convolution layer output from a trained convolutional neural network has the dimensions of $WxHxK$ where W and H represents the width and height of the activation maps of the last convolution layer. K is the number of filters or feature channels. From the WxH sized 2D maps, 40% overlapping multi-scaled square regions are extracted. The sizes of the squares are formed from L different scales where each scale dimension is calculated by $min(W, H)/(l_i + 1)$ where $(l_0..l_L) = 1..L$ shows the scale dimension indexes. Totally R regions are obtained.

For each region r_i from $R = r_0..r_R$ the max-pool is performed and the output of each region from the corresponding filter map is combined, resulting with a K dimensional representation vector for that region r_i. Finally L_2 normalization followed by PCA-whitening and again L_2 normalization is applied to the vector as the post-processing. For the image representation, each K dimensional region vector is summed and L_2 normalized to achieve the final feature vector.

The authors showed by experiments on Oxford Buildings dataset [51] and Paris dataset [52] that the RMAC representation improves the performance on image retrieval and image re-ranking tasks and provides efficient object localization.

3.2. Dissimilarity-based Sparse Subset Selection

Dissimilarity-based Sparse Subset Selection [50] method is a subset selection algorithm that finds the representative photos from a large image collection using dissimilarities of the images.

The algorithm takes two sets of images: the source set $\mathbb{X}$ and the target set $\mathbb{Y}$. The idea is to find a representative set from $\mathbb{X}$ which encodes the elements of $\mathbb{Y}$ based on their dissimilarities. Here constructing the dissimilarity matrix is the crucial step that the representatives will be determined based on it. The dissimilarity matrix is given in Equation 1.

$$\mathbf{D} \triangleq \begin{bmatrix} \mathbf{d}_1^T \\ \vdots \\ \mathbf{d}_M^T \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & ... & d_{1N} \\ \vdots & \vdots & & \vdots \\ d_{M1} & d_{M2} & ... & d_{MN} \end{bmatrix} \in \mathbb{R}^{MxN} \tag{1}$$

Here $d_i \in \mathbb{R}^N$ corresponds to the i^{th} photo's dissimilarity vector with respect to other photos. The goal is to find a small subset that will represent the elements of $\mathbb{Y}$. The method allows to use any dissimilarity metric such as KL divergence, Hamming, Euclidean etc. As dissimilarity is the opposite of similarity, both can be utilized with this algorithm.

After the construction step of dissimilarity matrix D, next step is to find the subset which corresponds to the representative photos of $\mathbb{X}$. In order to achieve this, the algorithm uses an optimization program to find some binary variables z_{ij}. These variables are associated with d_{ij} and are indicator for x_i representing y_i. Equation for binary z_{ij} matrix is shown in Equation 2.

$$\mathbf{Z} \triangleq \begin{bmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_M^T \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & ... & z_{1N} \\ \vdots & \vdots & & \vdots \\ z_{M1} & z_{M2} & ... & z_{MN} \end{bmatrix} \in \mathbb{R}^{MxN} \tag{2}$$

On this equation $\sum_{i=1}^{N} z_{ij} = 1$ should be satisfied to ensure that each $y_i \in \mathbb{Y}$ is represented. The optimization problem to select the subset from $\mathbb{X}$ is defined as a row-sparsity regularized trace minimization as formalized in Eq 3.

$$\begin{aligned} \min_{\{z_{ij}\}} \quad & \lambda \sum_{i=1}^{M} I(\| \ z_i \ \|_p) + \sum_{j=1}^{N} \sum_{i=1}^{M} d_{ij} z_{ij} \\ s.t. \quad & \sum_{i=1}^{M} z_{ij} = 1, \forall j; z_{ij} \in \{0,1\}, \forall i,j \end{aligned} \tag{3}$$

Here I is the indicator function where We want to select as few representative photos as possible, so it is 0 if the l_p norm of z_i is zero, 1 otherwise. The second term corresponds to the cost of encoding $\mathbb{Y}$ with $\mathbb{X}$ which is $\sum_{j=1}^{N}\sum_{i=1}^{M} d_{ij}z_{ij}$. Finally the regularization parameter λ adjusts the number of representative photos that we want to select. Because the formalization is NP-hard, the authors used convex programming and provided an efficient and parallelizable implementation. Furthermore, they showed the method can be used for scene categorization and deals effectively with outliers.

3.3. AMNet: Memorability Estimation with Attention

AMNet: Memorability Estimation with Attention [36] proposes an end-to-end trainable deep neural network together with a visual attention mechanism for visual memorability estimation. The network consists of 4 layers:

1. Convolutional Network layer
2. Soft Attention Network layer
3. Long-Short Term Memory (Recurrent Network) layer
4. Fully Connected layer

For the first convolutional network layer, authors preferred to use transfer learning on ResNet50 [53] trained on ImageNet which achieves high prediction accuracy on image classification tasks. So, given a single image $\mathcal{X}$ the first CNN layer outputs a tensor having dimensions (W, H, D). Here W and H correspond to output resolution of feature maps and D shows the number of filters or length of the feature vectors. For AMNet, these dimensions are $(14, 14, 1024)$.

For the second layer, they used a soft attention mechanism which consists of a network to learn probabilities of discrete elements on the image and a gating function to weight the data based on those probabilities.

The third layer is a $L = 3$ step LSTM network where at each step the corresponding state is calculated with the Equation 4.

$$h_t = \phi(h_{t-1}, z_t) \qquad t = [0, T), h \in \mathbb{R}^B \tag{4}$$

Here each state h_t is calculated with a function ϕ which uses the previous state h_{t-1} and the transition image features z_t calculated as a weighted sum using the image itself $\mathcal{X}$ and α_t which is the probability of calculated attention weights given the image $\mathcal{X}$ and the previous LSTM state h_{t-1}, as shown in Equation 5.

$$\begin{aligned} z_t &= \sum_{i=1}^{L} \alpha_{t,i} x_i \qquad & z_t \in \mathbb{R}^D \\ \alpha_t &\sim p(\alpha_t | x, h_{t-1}) \qquad & \alpha_t \in \mathbb{R}^L \end{aligned} \tag{5}$$

At the end of each step of LSTM, the produced output h_t is converted to a memorability score m_t with the Equation 6.

$$m_t = f_m(h_t) \tag{6}$$

Here f_m is the function that converts the output of corresponding LSTM state h_t to the memorability score m_t. Basically it is a two-layered network with a single output.

Finally the memorability score y of image $\mathcal{X}$ is simply the sum of these memorability score at each step, as shown in Equation 7.

$$y = \sum_{t}^{T} m_t \tag{7}$$

AMNet achives average Spearman's rank correlation score of 0.649 and mean squared error (MSE) of 0.011 on SUN Memorability dataset [54]. Similarly 0.677 average Spearman's rank correlation and 0.0082 MSE scores on LaMem dataset [33]. These scores show the state-of-art performance on image memorability prediction in the literature. It should be noted that the human performance from memorability prediction experiments has 0.68 average Spearman's rank correlation score. AMNet closes the gap between human and machine predictions.

3.4. Attention-based Multi-patch Aggregation for Image Aesthetic Assessment

The visual aesthetic prediction framework of Sheng etal. [44] is a multi-patch aggregation method for image aesthetic prediction. The framework basically consists of a convolutional neural network and then an attention mechanism. Here the attention mechanism plays an important role that after the prediction, during the back-propagation phase it increases the weights of the object patches that are predicted incorrectly resulting in a boosting in accuracy.

The method bases on an energy maximization function as shown in Equation 8.

$$\underset{x}{\operatorname{argmin}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} Pr(\tilde{y} = \hat{y}|p, \theta) \tag{8}$$

Here $\mathcal{P}$ is the set of patches extracted from photos of the dataset, $\tilde{y}$ shows the predicted aesthetic score and θ is the network parameters that maximizes the probability of correct label prediction. Based on this optimization function, authors devised 3 multi-patch weight method for testing where each method uses different weighting schemes for the patches. MP_{avg} which corresponds to average weighting takes all patches into the consideration and optimizes based on the average prediction scores. MP_{min} takes only patches with lowest prediction confidence into consideration from an image and ignores others. MP_{ada} which corresponds to adaptive-weighting scheme gives adaptively changing weights to incorrectly predicted patches to increase their prediction accuracies.

For training, authors preferred tuning the 18-layer ResNet architecture trained on the ImageNet ILSVRC2012 dataset. They used AVA dataset which is the largest open photo set for visual aesthetic assessment consisting of 250.000 photos together with their aesthetic scores. During training they did not change the aspect ratio of photos where for aesthetic prediction it plays an important role. Instead they resized photos by keeping aspect ratio fixed and cropped 224 x 224 sized patches from them. They also applied random horizontal flipping for data augmentation.

Their proposed MP_{ada} weighting scheme achieves the state-of-art prediction accuracy of 83.03 for visual aesthetic prediction. Other two schemes which are MP_{avg} and MP_{min} similarly achieve 81.76 and 80.50 scores respectively.

Additionally, authors emphasized the effectiveness of their proposed method by further making experiments using different network architectures (VGG16), measuring correlations between pre-trained (ImageNet) and fine-tuned networks and investigating the effects of image resizing.

4. Visual Storygraph Generation

In this section we give the formal definitions of story graphs and introduce our framework with details of the steps to extract visual story graphs from large image collections. We start with constructing dictionaries for visual and textual elements from the given sets of images. These elements serve as fundamental building blocks in finding coherent and intersecting story lines. Then we explain the construction steps of the story graphs in detail.

4.1. Definition of a Story Graph

A story graph is a pair $\mathcal{S} = (G, \mathcal{P})$ where $G = (V, E)$ represents a directed graph, $\mathcal{P}$ denotes a set of chains (paths) which includes the story lines in G, the nodes of G correspond to the representative images from a large photo collection and its edges symbolize the connections among them. In an ideal case, a story graph, as a whole, should provide a visual collaborative summary of the photo collection from which it is extracted. This goal can be achieved by constructing it by considering three key properties, namely *coherence*, *coverage* and *connectivity* [55].

4.1.1. Coherence

Intuitively, we want our story graphs to tell coherent stories. Hence, we need a mechanism to measure the consistency across each story line of our story graph. We employ visual and textual elements as means for forming coherent visual stories through these story lines. Specifically, we connect the images with the visual and textual elements shared among them. We define the overall coherence gained by a story line $\mathcal{P}_i = (p_1, ..., p_n) \in \mathcal{P}$ by the following equation:

$$Coherence(\mathcal{P}_i) = \min_{k=1..n-1} \sum_{e} \mathbb{1}(e \text{ is active in } p_k \text{ and } p_{k+1}) \tag{9}$$

where $\{e\}$ denotes the set of elements, p_k represents the kth photo in the story line. We consider an element e as active if its importance is above a certain pre-defined threshold for both p_k and p_{k+1}.

In particular, here, we ensure that all the consecutive pairs of photos on the story line share at least an element e which can either be a visual or a textual element. The function $\mathbb{1}$ is an indicator function which enforces that the element should be active among the photos p_k and p_{k+1}. The final coherence value is then determined by the weakest pair among the whole story line. Hence, for a coherent chain, the behavior of all of the elements should provide a transition as smooth as possible throughout the story line. Refer to 4.2.1. and Section 4.2.2. for the details of how we construct the visual and textual elements and decide whether an element is active or not for an image.

In Fig. 4.1., we show some sample coherent and incoherent chains based on visual and textual elements shared among the images in the chain and plots. As can be seen, the characteristics of the images change rapidly in an incoherent chain without producing consistent stories, which is valid for both visual and textual domains.

4.1.2. Coverage

Coverage property ensures that the photos among the story line cover a diverse set of elements. That is, if a story line sufficiently covers an element, there is no need to add it to the story graph. This brings the so-called diminishing return property that tells as new story lines are added to the graph, if the new story line covers an element that has already been covered, it should contribute very little to the total coverage. With this property in mind, each element's coverage through a story graph $\mathcal{S}$ is given by the following equation:

$$Coverage_{\mathcal{S}}(e) = 1 - \prod_{p \in photos(\mathcal{S})} (1 - Coverage_p(e)) \tag{10}$$

where $Coverage_p(e) \in [0, 1]$ denotes how important that element is for describing the photo p, defined differently for visual and textual elements. If the story graph $\mathcal{S}$ has photos covering

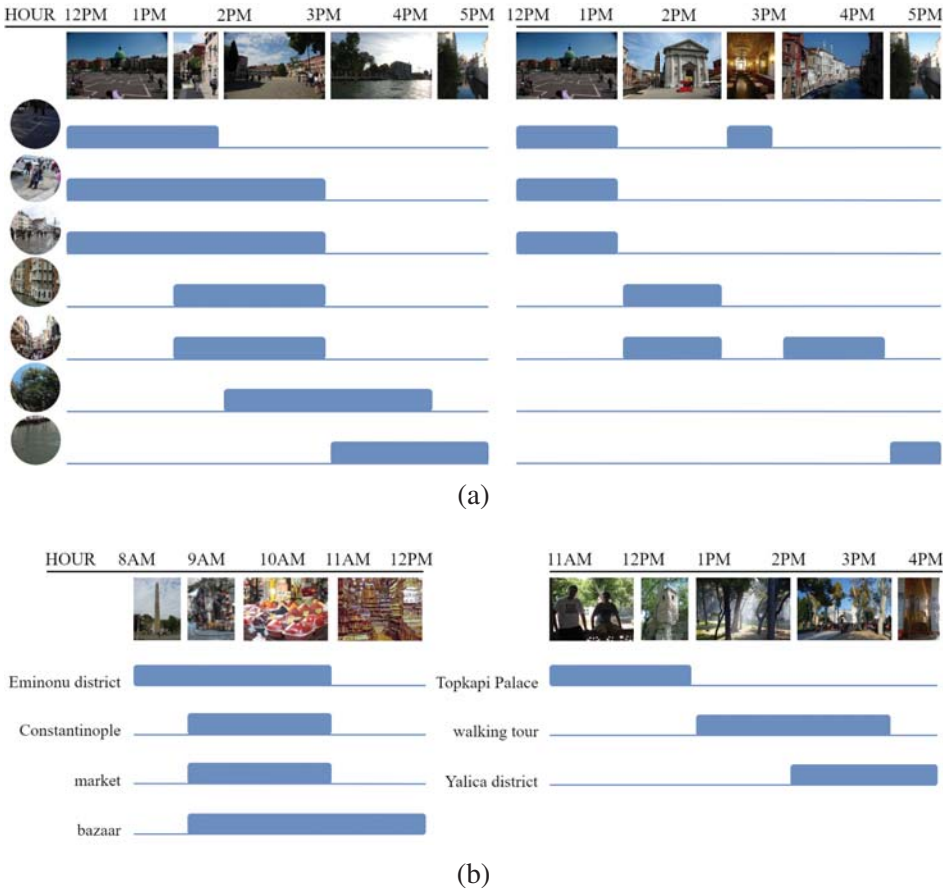


Figure 4.1. Coherent and incoherent chain examples in terms of (a) visual elements and (b) textual elements. For each case, we show a number of images composing a story. The bars indicate the elements that are active on the images. The coherent chain given on the left tells a consistent story through smooth transitions over the active elements. On the other hand, within the incoherent chain shown on the right, the active elements change very rapidly over the images, which result in inconsistencies in the story told.

the element e well, the coverage of the whole map on element e, $Coverage_S(e)$, will be close to 1 which means there is no need to select any other photos covering the same element e. If a new story line has been added to the graph, it should cover different elements, resulting in a more diverse chains of photos. In our framework, visual elements connects photos via visual patches whereas textual ones creates semantic connections through textual keywords. However, of course, not all elements are equally important. Some visual elements such as the sky regions are so common among the images that it is not feasible to use them to form story lines. Similarly, specific textual keywords such as *White House* which shows a singular location should have higher importance than generic location names like *garden* or *museum*.

Total coverage of a story graph is then computed as the summation of the coverage of both visual and textual elements as given below:

$$Coverage(\mathcal{S}) = \alpha \sum_{v \in \mathcal{V}} Coverage_{\mathcal{S}}(v) + (1 - \alpha) \sum_{t \in \mathcal{T}} Coverage_{\mathcal{S}}(t) \tag{11}$$

where $v \in \mathcal{V}$ denotes a visual element, $t \in \mathcal{T}$ represents a textual element, and $\alpha \in (0, 1)$ is a scalar representing relative significance of textual and visual elements. In our experiments, we empirically set the value of α to 0.1.

4.1.3. Connectivity

Connectivity enforces that the story lines should share some photos which amounts to the crossing points between the chains. This is a unique property that gives a story graph a nonlinear story structure as compared to the simple linear story model. The story graph is more informative when it shows hidden connections between diverse paths. In other words, without connectivity, the output will be linear summaries of individual photo collections. Although it seems contradicting with the coverage property, we look for minor connections between story lines after selecting a diverse set, preserving diversity together with a few individual photo similarities. Formally, the connectivity of a graph can be defined in terms of a value denoting the sum of the number of lines that intersect in story graph $\mathcal{S}$:

$$Connectivity(\mathcal{S}) = \sum_{i<j} \mathbb{1}(\mathcal{P}_i \cap \mathcal{P}_j \neq \emptyset) \tag{12}$$

with $\mathcal{P}_i$ and $\mathcal{P}_j$ denoting the ith and jth story lines in the story graph $\mathcal{S}$.

4.2. Constructing the Story Graph

We cast the story graph construction as an optimization task defined over extracted coherent story lines $\mathcal{S} = (\mathcal{P}_i, .., \mathcal{P}_n)$. That is, we compute the optimal story graph $\mathcal{S}^*$ by first extracting most coherent story lines and then selecting a diverse set of important ones which intersect

Algorithm 1 Steps of finding the optimal story graph $\mathcal{S}^*$ from a large collection of images denoted by $\mathcal{I}$

1: **for** each image p_i in the input photo collection $\mathcal{I}$ **do**
2: Estimate the importance weights for the visual elements (Section 4.2.1.)
3: Estimate the importance weights for the textual elements (Section 4.2.2.)
4: Compute the coherence graph G based on the transitions over elements (Section 4.2.3.)
5: Extract a set of high coverage chains from G (Section 4.2.4.)
6: Perform a local search to improve the connectivity (Section 4.2.5.)

with each other to a certain extent by considering the following equation that is built upon *Coverage*, and *Connectivity* characteristics:

$$\begin{aligned} \mathcal{S}^* &= \text{argmax}_{\mathcal{S}}\, Connectivity(\mathcal{S}) \\ &\quad \text{s.t. } Coverage(\mathcal{S}) \geq C \end{aligned} \tag{13}$$

where C denotes a coverage score that is smaller than the highest coverage score that can be obtained without considering the connectivity property.

An optimum approach to find $\mathcal{S}^*$ is not trivial, hence, instead, we use a greedy approach by exploiting the sub-modular structure that exist in our problem. That is, we first maximize coverage and then try to maximize connectivity over story lines by allowing some decrease in the maximum possible coverage score (please refer to Section 4.2.4. for the details about how the maximal coverage can be defined. The whole algorithm is summarized in Algorithm 1.

4.2.1. Visual representation

Our visual representations are based on bag of visual elements. In particular, we approach the extraction of the visual elements from a dictionary learning perspective. In particular, we employ a recently proposed deep feature called Regional Maximum Activation of Convolutions (RMAC) [2] which achieves the state-of-the-art performance for the image retrieval task. Specifically, the RMAC representation that we use in our work depends on the VGG16 [56] model pre-trained on ImageNet. It is extracted from the last pooling layer, resulting in a 3D tensor having $W \times H \times K$ dimensions where K denotes the number of filters. Then, for these

Figure 4.2. Sample visual elements from the visual dictionary constructed from the Paris vacation photo albums. These elements are visualized by finding the image patches having the closest RMAC representations [2]. While some of them captures the details from touristic attractions (left), some correspond to very ordinary regions such as trees, clouds, and sky (right).

$W \times H$ response maps they sample R uniform square regions at L different scales with 40% overlap. For each region $r \in R$ max-pooling is performed on each channel and obtained a feature vector of K dimensions as shown in Eq. 14. The last step is the L_2-normalization to get a single region vector.

$$f_{r \in R} = [f_{r,1} \dots f_{r,i} \dots f_{r,K}]^T \tag{14}$$

In our work, we cluster these region features with K-Means clustering algorithm and form the visual dictionary for a city accordingly. We set the size of this dictionary as 1024. This approach captures various structures that persistently exist in the image collections, reflecting the visual characteristics of a city and the popular landmarks within. In Fig. 4.2., we demonstrate sample image regions which are close to some of the visual elements from the Paris dataset. As can be seen, some of these regions correspond to the details from the touristic attractions such as *Eiffel Tower, Arc de Triomphe, Notre Dame and Louvre Museum* as given on the left. However, since our dictionary learning procedure does not use any prior knowledge about the cities, some of the extracted visual elements might correspond to very common image regions such as sky, trees, etc. as shown on the right. Hence, for each image p_i we assign a certain importance weight to each visual element v, which is defined inversely proportional to the number of occurrences of this visual element in the whole image collection.

Each image is decomposed into a set of local image regions, each encoded via a RMAC feature. Then, Locality-constrained Linear Coding (LLC) [57] is applied over these regions

to obtain the final representation by max pooling of each region's code vector over the extracted visual elements. LLC encoding yields a sparse representation where only the most prominent visual elements are considered in the final representation. Importances of visual elements are then defined in terms of this LLC encoding scheme.

In Eqn. 9, the coherence score is estimated through the active visual elements over a story line. The decision about whether a visual element is active or not is made by inspecting the weights of this visual element within the LLC encodings of the image pairs. If they are above a certain threshold, we assume the element is shared between the images and considered as active.

4.2.2. Textual Representation

In our work, we represent the images in the photo-collections in a multi-modal manner. As we mentioned earlier, representing images visually is carried out by first learning a visual dictionary from the training images and then by extracting visual elements from each image. Apart from this, we also consider a semantic representation of images that depends on textual information. In particular, each image can be tagged by a list of words by employing a pre-trained set of image classifiers that identify the visual characteristics of the image. In our work, we alternatively assume that each image has been already associated with a set of keywords. By this way, we can utilize a dictionary of words extracted from all of the images in the collections and then represent each image in terms of these keywords. To determine the importance of textual elements, we employ a tf-idf weighting scheme.

Similar to the visual elements, the coherence score due from the textual elements is computed by taking into account the textual elements that are shared over a story line. While deciding these shared textual elements, we utilize their importance scores of indicated by their tf-idf weights. We assume that a textual element is active if its score is above a pre-defined threshold value.

4.2.3. Finding Coherent Story Lines

We start with modeling story lines by the transitions of the extracted visual and textual elements. The brute-force solution to optimize the energy function in Eqn. 13 inspects every pair of images for the occurrence of all elements, and thus it takes time proportional to $N^2 \times D$ where N is the number of images and D is the total number of elements. Since this is intractable for large image collections, we use a divide-and-conquer approach to build story lines. First, we extract short chains of images with smooth transitions being observed over some visual and/or textual elements. Then, we combine these short chains which overlap through some common images to obtain longer story lines that constitute our coherence graph G.

Our algorithm starts with a RANSAC [58] loop where at each iteration we randomly choose two images from the collection, which share at least a visual or textual element to satisfy the coherence property and which correspond to the end points of a short chain. Hence, to determine the images in between these two, for each shared element we search for images that also share the same element. Specifically, we enforce a smooth transition across the story line as in [1]. For each shared visual element of the end point images, we fit a line over the activation scores coming from the LLC encoding [57] and validate the consistency of a candidate image by analyzing how well it fits to this linear activation transition function [55] by its corresponding element. For each shared textual element of the end point images, we check whether the element is active in the candidate image or not.

In our framework, we also utilize additional meta-data about the photos, namely the time-stamps and GPS location information to enforce additional constraints to improve the quality of the transitions. First, each image over a story line should be captured after the time the photo preceding it is taken. This eliminates the possibility of ambiguous ordering of images such that a night time image follows a day time. Second, an image should be close to its previous image in geospatial terms. This enhances both the structure and the overall visual appearance of the story line in that nearby locations are more likely to share similar visual

structures. In our experiments, we empirically set the length of the short chains as 3. Larger values, in general, fail to find sufficient number of high quality chains.

Once we extract the coherent short chains, the next step is to construct a coherence graph G. We combine the shorter chains by the common images that they share and accordingly obtain longer chains, each of which denotes a coherent story line.

4.2.4. Finding Story Lines with High Coverage

In the previous subsection, we show how to extract all coherent story lines on a coherence graph G we build based on short chains. Finding story lines with high coverage corresponds to selecting a subset of those from G that maximize the coverage as Eqn. 10 indicates. This can be formulated as an *orienteering problem*, aka *prize-collecting TSP* [55, 59], in which the goal is to maximize rewards collected while walking on the graph subject to a budget on the tour length and given two endpoints. The reward function is given by $f : 2^V \rightarrow \mathbb{R}^+$, which returns a non-negative value to every subset of nodes. Exhaustively searching for an optimum solution is infeasible but we can exploit the submodularity of our coverage function (Eqn. **??**) where greedy algorithms with good approximation guarantees exist in the literature [59].

A set function $f : 2^V \rightarrow \mathcal{R}$ is *submodular* if $f(\mathcal{A} \cup a) - f(\mathcal{A}) \geq f(\mathcal{B} \cup a) - f(\mathcal{B})$ and for all $\mathcal{A} \subseteq \mathcal{B} \subseteq V$. This property is referred to as the *diminishing returns*, meaning that adding a new item to a smaller set provides a larger gain than adding it to a larger set.

After we extract our coherent story lines, we define the following incremental coverage notion to measure the gain in the coverage score when we add the story line to our story graph S for each story line $\mathcal{P}_i$ as follows:

$$IncCoverage(\mathcal{P}_i|\mathcal{S}) = Coverage(\mathcal{P}_i \cup \mathcal{S}) - Coverage(\mathcal{S}) \tag{15}$$

To sum up, in order to find the set of story lines that have the highest coverage over the visual elements, we follow an incremental search strategy. Starting with the story line having the

highest coverage value, we gradually enlarge the story graph by analyze each not included story line by its contribution to the current coverage (Eqn. 15) and add the one that contributes the most. This procedure is repeated until there is no additional gain.

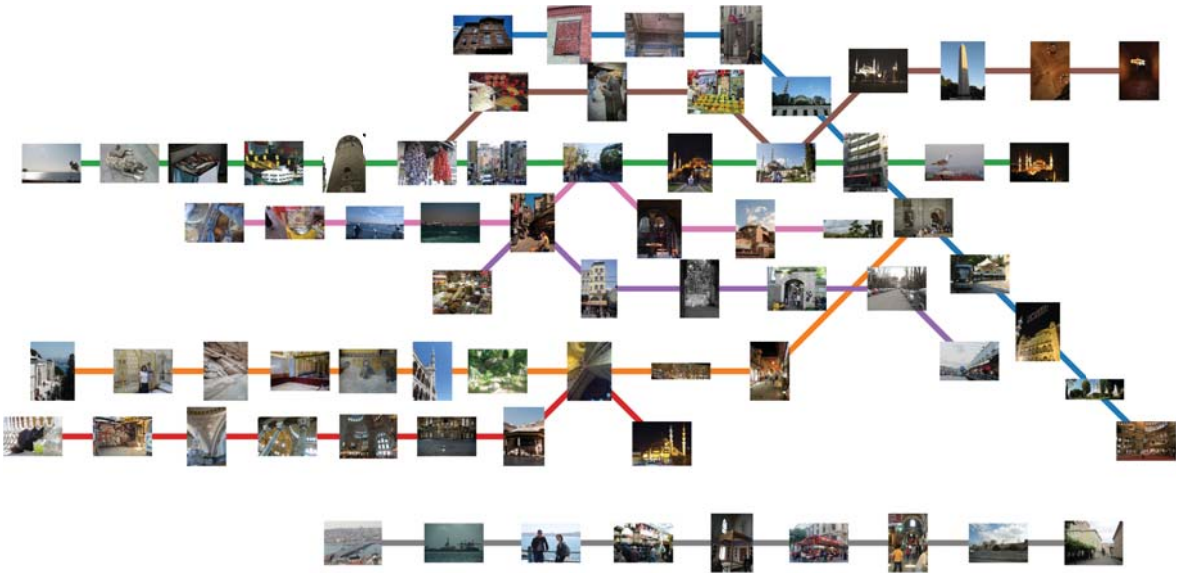
4.2.5. Increasing Connectivity

Increasing the connectivity is important to discover nonlinear story structures. We perform a local search operation on the extracted coherence graph G by using the story graph $\mathcal{S}^+$ with the highest coverage as an initial point. In particular, we fix the story line having the highest individual coverage and perform a search among all of the other story lines forming the coherence graph G. Our aim is to find story lines alternative to the ones in $\mathcal{S}^+$, which increases the connectivity by allowing a reasonable amount of degradation in the total coverage value. Of course, the key question here is how much coverage drop can be tolerated. Allowing too much drop in the coverage results in story graphs with low coverage whereas limiting it to a low value prevents finding an appropriate chains for the replacements. In our work, we empirically observe that a 7% drop in the total coverage score generally gives satisfactory results. In Figures. 4.3., 4.4. and 4.5., we provide the story graphs for the cities of Amsterdam, Istanbul, New York, Paris, Tokyo and Venice which are automatically constructed by our approach from large sets of travel photo albums collected from the web.

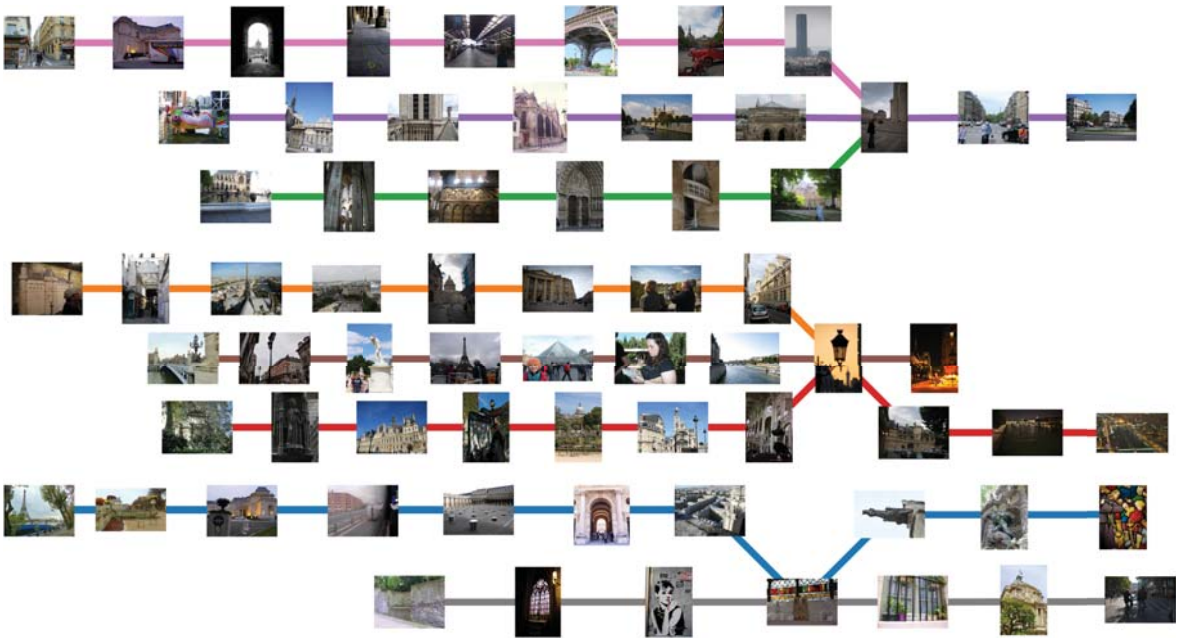
In the story graph figures, each story line is shown in a different colored line. Each story line consists of photos with similar but as many diverse visual characteristics as possible. Together with this property and multiple distinct story lines, a story graph forms an excellent prior information about the specific city, which we will make use of in a visual summarization application in the next chapter.

4.3. Summary

In this chapter, we proposed an approach to automatically extract story graphs from large collections of photo albums, which serves as a collaborative and structured summary of these



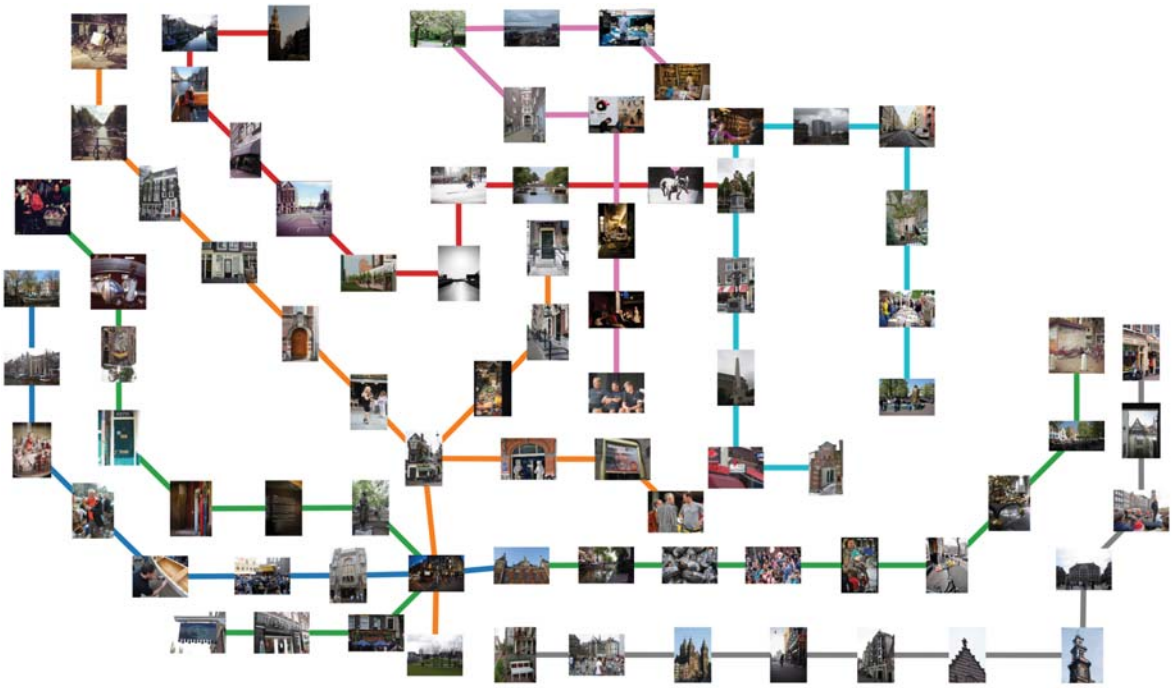
(a)



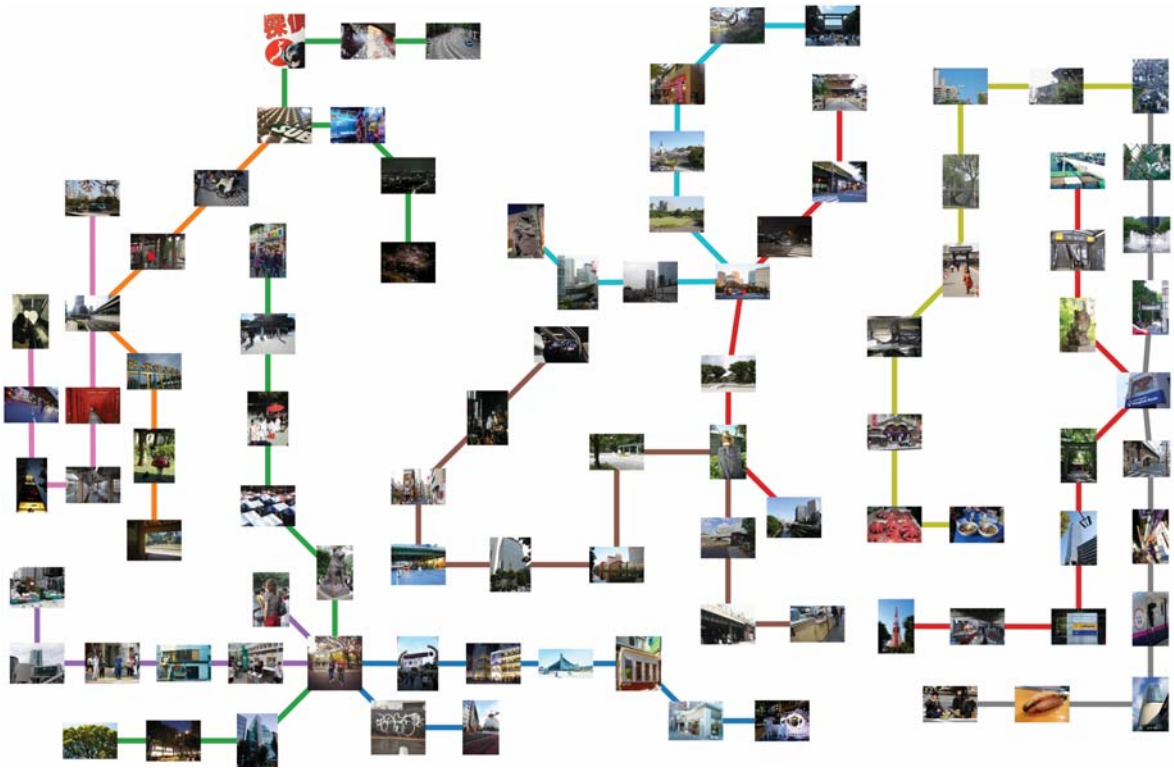
(b)

Figure 4.3. The story graphs of (a) Istanbul and (b) Paris, which are based on travel photo albums collected from the web. The nodes (images) of the graphs are arranged based on the available timestamp information.

albums. We treated this task as a sub-modular optimization problem and formulated a greedy approach to find a graph that maximizes the degrees of coherence, coverage and connectivity of the story lines. Next we will describe an application where the generated story graphs can

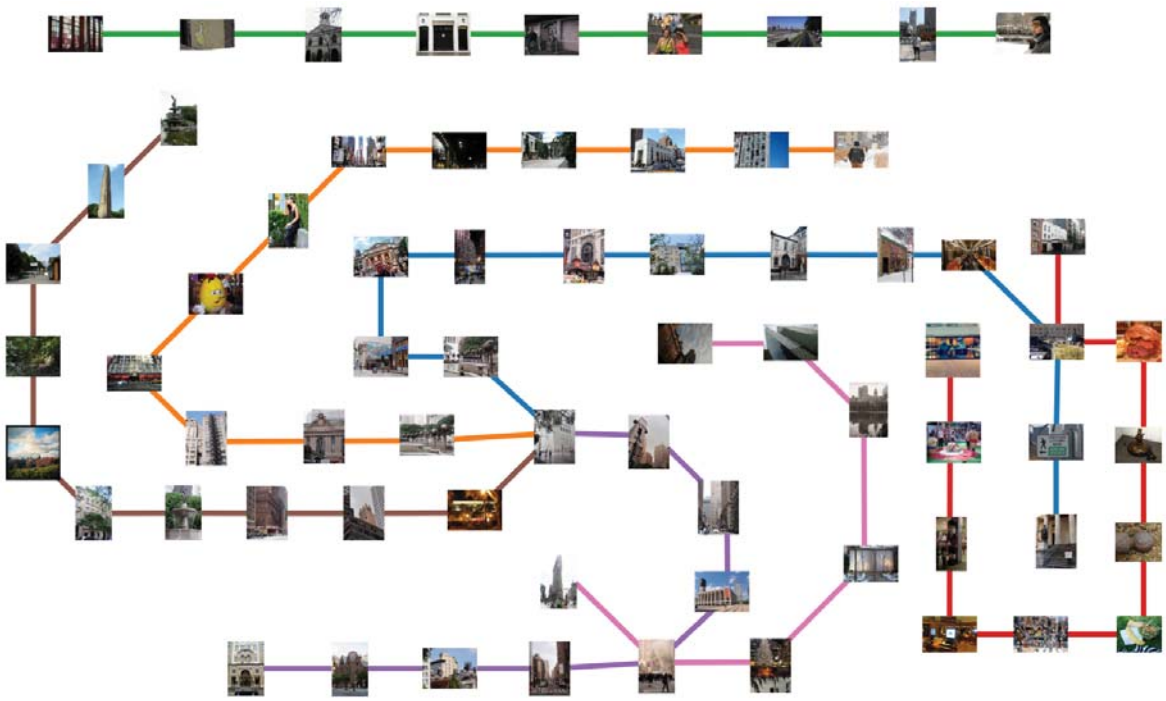


(a)

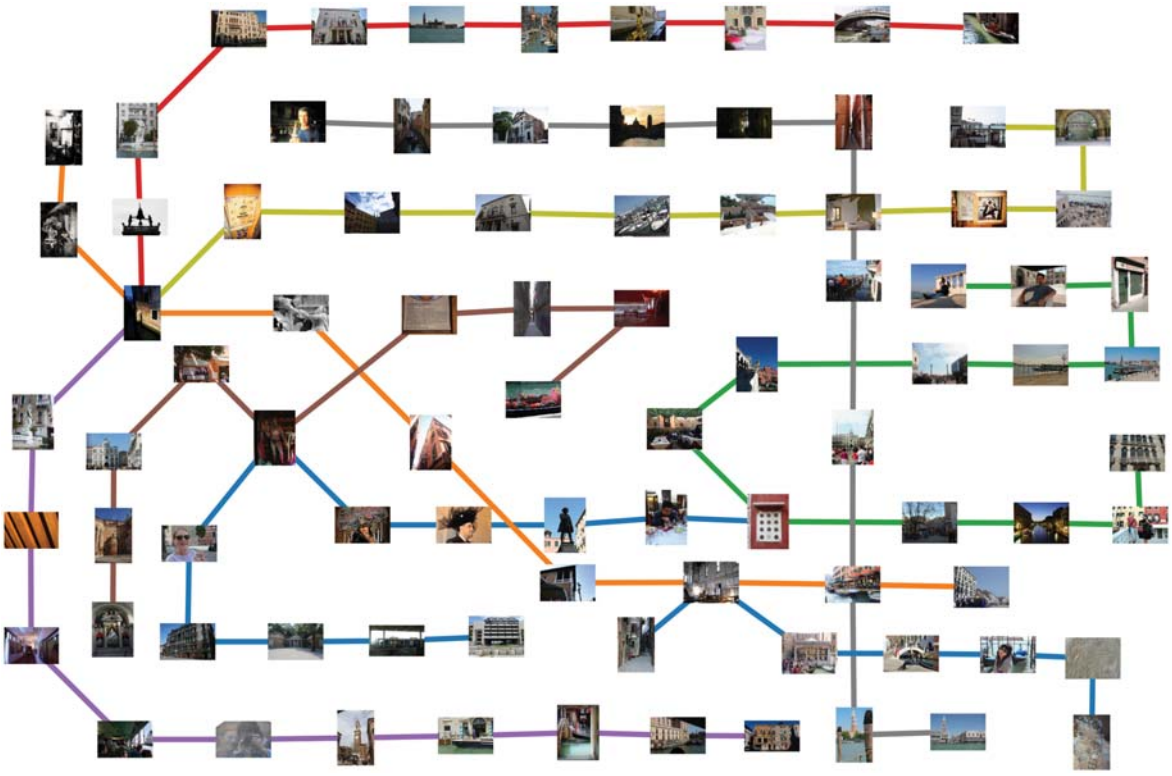


(b)

Figure 4.4. The story graphs of (a) Amsterdam and (b) Tokyo, which are based on travel photo albums collected from the web. The nodes (images) of the graphs are arranged based on the available timestamp information.



(a)



(b)

Figure 4.5. The story graphs of (a) New York and (b) Venice, which are based on travel photo albums collected from the web. The nodes (images) of the graphs are arranged based on the available timestamp information.

be utilized.

5. Application: Visual Summarization Using Story Graphs

In the previous section, we develop a method to generate visual story graphs from a large collection for photo albums. These story graphs are collaborative structured summaries containing coherent visual story lines and providing a comprehensive overview of specific topics of interest. With these characteristics, story graphs can be interpreted as prior graphs representing important concepts, landmarks and events within the photo collections.

An extensive application that make use of a story graph is photo album summarization task. In that regard, in this section, we demonstrate a way to obtain more effective summaries of photo albums and albums that cover the topics encoded in the story graphs generated by our approach.

5.1. Story-Graph Guided Photo Album Summarization

Given a photo album $\mathbb{X}$, our goal is to extract a small number of images from $\mathbb{X}$ that represents the whole set. We additionally assume that another set of images are given in the form of a story graph $\mathbb{Y}$. Here, we formulate the summarization task as a subset selection task. For this purpose, we particularly employ the D3S algorithm [50] which formulates subset selection as a row-sparsity regularized trace minimization problem which can be easily solved via convex optimization.

In short, the D3S algorithm solves a special subset selection problem when side information is available in the form of dissimilarities between the source set $\mathbb{X}$ and a target set $\mathbb{Y}$, defined as:

$$\begin{aligned} \min_{\{z_{ij}\}} \quad & \lambda \sum_{i=1}^{M} \|\mathbf{z}_i\|_p + \sum_{j=1}^{N} \sum_{i=1}^{M} d_{ij} z_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{M} z_{ij} = 1,\ \forall j;\ z_{ij} \geq 0,\ \forall i, j \end{aligned} \tag{16}$$

where z_{ij} is the indicator of the source item $x_i \in \mathbb{X}$ representing the target item $y_j \in \mathbb{Y}$ and d_{ij} denotes the dissimilarity between x_i and y_j. In our experiments, we use the KL-divergence as our dissimilarity measure. The parameter λ provides a trade-off between the number of representative samples and the encoding quality with smaller values of λ causing more number of samples selected as representative. Here, the first term penalizes the size of the representative subset and the second term is the encoding cost. In [50], the authors show that an optimal solution can be found using an Alternating Direction Method of Multipliers (ADMM) approach in an effective manner.

Notice that here we suggest to let $\mathbb{Y}$ denote the set of images available in the input story graph. Hence, while extracting a summary from the given photo album denoted by $\mathbb{X}$, the representative samples of $\mathbb{X}$ in the generated summary cover the themes available in $\mathbb{Y}$. Alternatively, we can let $\mathbb{Y} = \mathbb{X}$ by selecting the target set same as the source set. If this is the case, it becomes a self-summarization problem [50].

5.2. YFCC100M-CITIES Dataset

To evaluate our proposed story graph generation approach, we curated a new dataset by selecting and annotating images from the publicly available YFCC100M dataset [60]. In short, YFCC100M dataset [60] which contains 99.2M photos and associated metadata such as time stamps, geolocation information and keywords from Flickr. However, most of the time, the user generated keywords are noisy, since the users are from different countries, they use different languages while providing them.

In our work, we particularly collected vacation photographs from 6 different cities, namely Amsterdam, Istanbul, New York, Paris, Tokyo, Venice which are among the most visited cities around the world. We eliminated the photo albums that consist of only close-up pictures of humans or cover just one topic such as flowers in a garden. For user generated keywords, we filtered out highly generic words or words that are unrelated to the topic of interest. We then grouped similar and synonym words into common concepts by taking into account non-English words as well. In total, we have collected 132K geotagged images from 323 users

Amsterdam Istanbul New York Paris Tokyo Venice

Figure 5.1. The distribution of photos in our YFCC100M-CITIES dataset. The area of a circle is proportional to the density of the photos in that location.

Table 5.1. Statistics of YFCC100M-CITIES.

City Albums	Number of Users	Number of Photo Sets	Number of Photos	Number of Unique Words
Amsterdam	39	100	9,923	1,460
Istanbul	58	167	13,645	979
New York	54	428	30,443	18,538
Paris	39	178	21,819	1,521
Tokyo	71	514	36,787	4,007
Venice	62	179	19,729	2,032
Total	323	1,566	132,346	25,118

and 1.5K photo albums. Fig. 5.1. and Table 5.1. show the basic statistics of our dataset, which we named as YFCC100M-CITIES dataset.

5.3. Experiments

We performed an extrinsic evaluation of our story graphs in which we leverage them as a prior to guide photo album summarization (Section 5.3.2.). Another common approach to the visual evaluation task is performing user studies. Based on our formalism, a good story graph must first meet two criteria. It must be composed of *coherent* chains and these chains should all together should *cover* most of the important aspects. However, it is difficult to quantitatively evaluate these two notions so we decided to perform controlled user studies, on which we compare against the previous work by Kim and Xing [1]. To assess coherence, we employ the next image prediction task proposed in [1] (Section 5.3.3.), but to evaluate coverage we devised a new experiment (Section 5.3.4.) since there has been no particular attention to this essential property.

Table 5.2. Statistics of additional photo set for summarization experiments.

City	Number of Albums	Number of Photos
Amsterdam Trip	2	200
Istanbul Trip	1	100
New York Trip	1	100
Paris Trip	5	500
Tokyo Trip	2	200
Venice Trip	2	200
Total	13	1300

5.3.1. Evaluation Dataset

For summarization experiments, additional to the YFCC100M-CITIES dataset we collected, we need another unseen photos to perform a fair comparison. Thus, we use another photo set of our work [61] where we collected distinct photo collections having different photos than YFCC100M-CITIES. Both sets consist of user vacation photos and are collected from the same touristic cities which are Amsterdam, Istanbul, Paris, New York, Tokyo and Venice. The dataset consists of 13 albums each having 100 photos, totally 1300 photos as we give the statistics on Table 5.2.

5.3.2. Photo Album Summarization

As we have mentioned earlier, story graphs provide a collaborative summary of photo albums on specific themes, which can be used as priors. As for our first experiment, we conduct an extrinsic evaluation of our proposed summarization framework by utilizing story graphs as priors in photo album summarization. For this task together with YFCC100M-CITIES, we used the additional evaluation dataset as we mentioned in Section 5.3.1. We collected 20 human generated summaries for each city for comparison of automatic and human summaries.

For comparison, we test two simple baselines, which are uniform sampling (Uniform) and K-Means clustering (K-Means), the skipping recurrent neural network model (S-RNN) by

Sigurdsson *et al*. [3], two subset selection based summarization methods by Iyer *et al*. [62], which respectively employ simple color histograms of hue and saturation channels (DSS-S), and deep features from the last fully connected layer of the VGG network (DSS-D)[1] and the DS3 model performing self-summarization with $\mathbb{Y} = \mathbb{X}$. In addition to those, we constructed three story graphs using our framework by taking into account (1) only visual features ($\mathbb{S}^V$), (2) visual features along with GPS information ($\mathbb{S}^{VG}$), and (3) both visual, GPS and textual information ($\mathbb{S}^{VGT}$).

We quantitatively evaluate the results using V-ROUGE [63] which is an extension of the ROUGE metric used for document summarization and F-measure [64] metric. V-ROUGE simply measures how the automatic summaries correlate with the human generated ones based on occurrence-counts of visual elements. F-measure similarly measures the accuracy of automatic summaries considering both precision and recall with respect to human generated summaries.

In Table 5.3. and Table 5.4. we report the V-ROUGE and F-measure scores, respectively. As can be seen, the quality of summaries obtained with the simple baselines, Uniform and K-means, is lower than other approaches. S-RNN also gives unsatisfactory results although its formulation is based on modeling how a story evolves within a photo album. DSS method with simple features (DSS-S) produces slightly better summaries than S-RNN, but it is beaten by DSS-D, which is somewhat expected as deep features provide better semantic representations. The summaries obtained by different versions of our proposed framework, $\mathbb{S}^V$, $\mathbb{S}^{VG}$ and $\mathbb{S}^{VGT}$, are far better than the competing methods, including the deep approaches deep learning based models S-RNN [3] and DSS-D [62]. Moreover, we observe that our fully featured story graph $\mathbb{S}^{VGT}$ which employs both visual, GPS and textual information, in general, achieves the best summarization performance. For some cities our story graphs without additional meta-data which are $\mathbb{S}^V$ and $\mathbb{S}^{VG}$ give better results which also show that generally our

[1]Here, we intentionally use VGG model for a fair comparison with our approach, which employs features from the same base network.

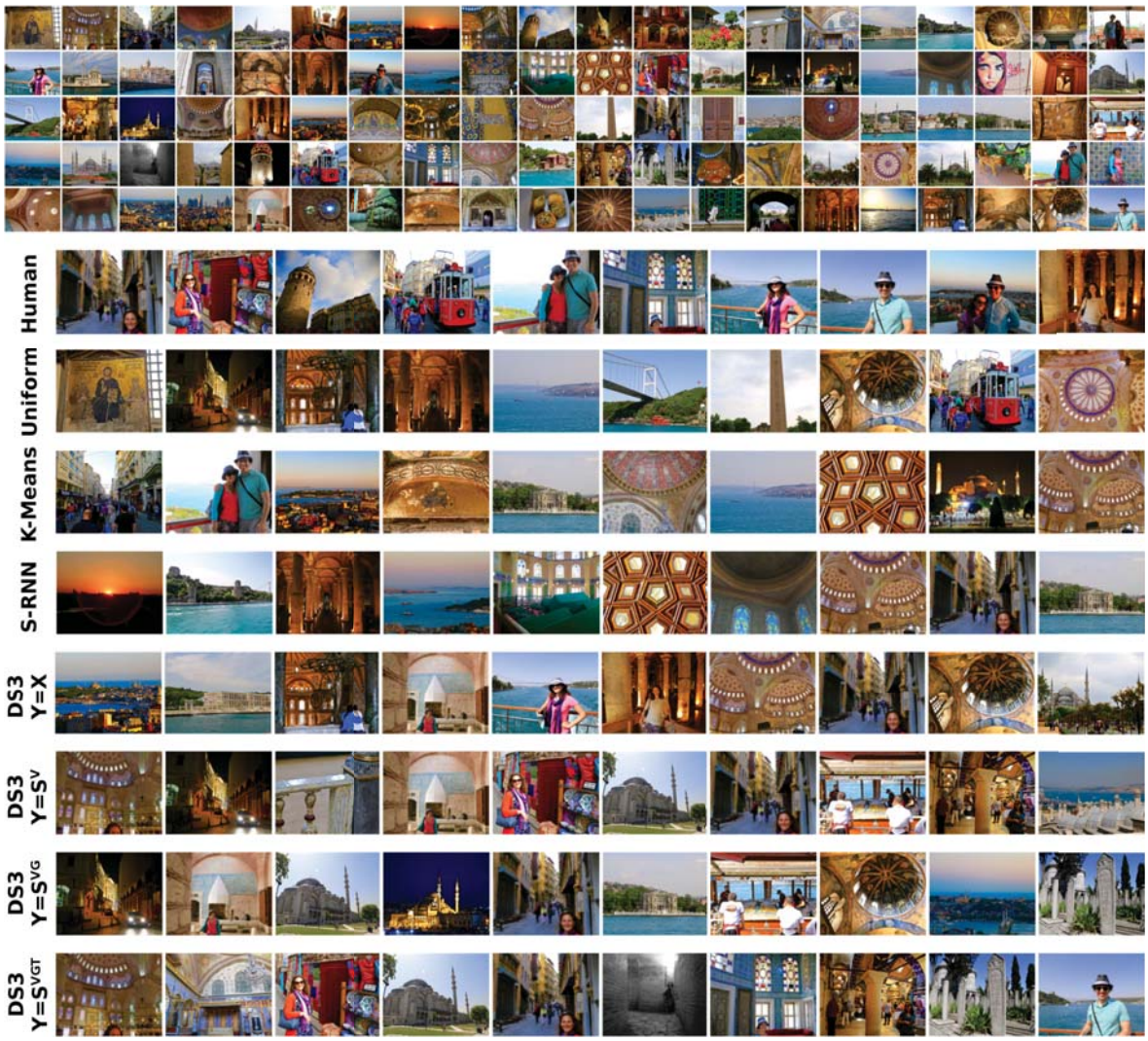


Figure 5.2. Summarization results of city Istanbul. Top: Input photo album. Bottom: Visual summaries done by a human, the baselines approaches Uniform Sampling, K-Means clustering, and S-RNN [3] along with the ones obtained via the DS3 method using self summarization ($\mathbb{Y} = \mathbb{X}$), the story graphs constructed with visual features ($\mathbb{Y} = \mathbb{S}^{V}$), both visual and GPS features ($\mathbb{Y} = \mathbb{S}^{VG}$) and all visual, GPS and textual features ($\mathbb{Y} = \mathbb{S}^{VGT}$).

framework produces better scores than the simple baselines even without additional metadata. Similarly, the quantitative results show that the summaries obtained by our approach are far better than the ones obtained by the state-of-the-art S-RNN model [3].

Figures 5.2.-5.7. show sample summarization results from the YFCC100M-CITIES dataset. Uniform baseline gives a low quality summary in that it includes similar and semantically uninteresting images. K-means baseline generates a summary that lacks a coherent story

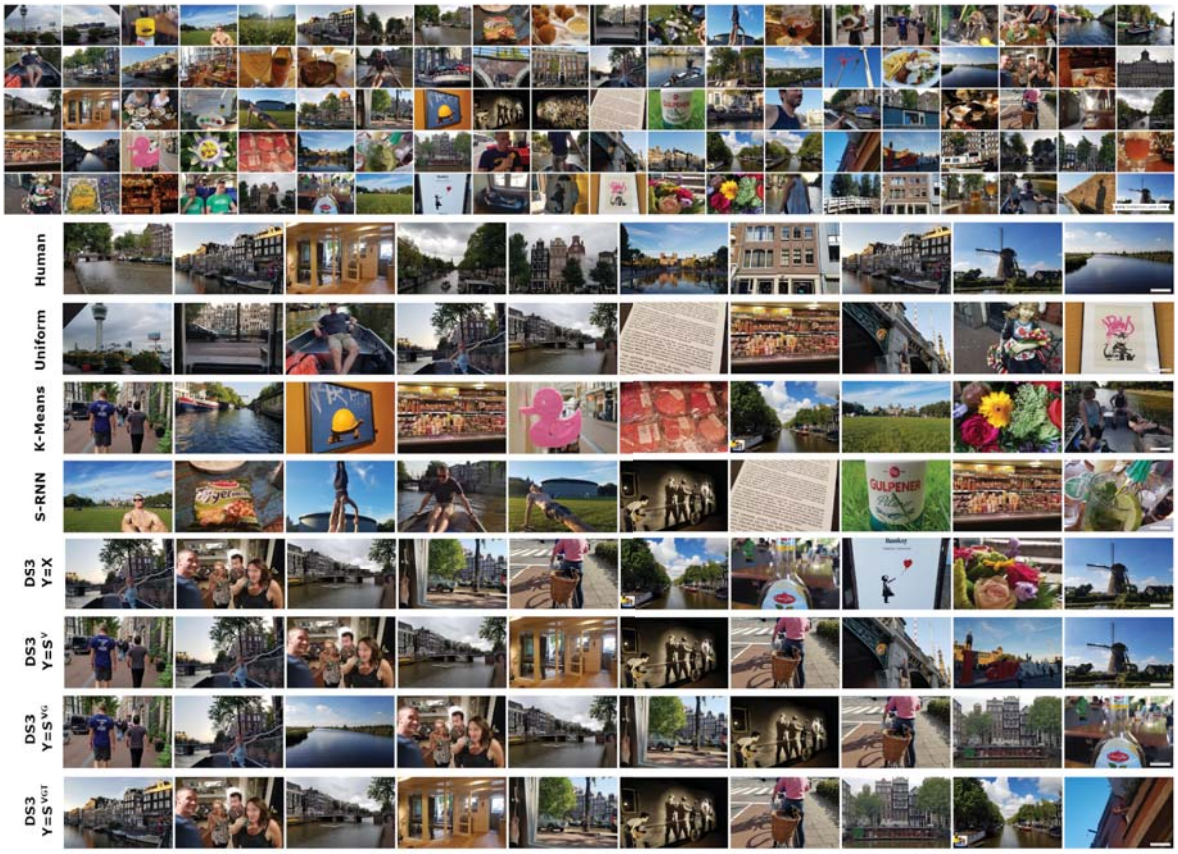


Figure 5.3. Summarization results of city Amsterdam. Top: Input photo album. Bottom: Visual summaries done by a human, the baselines approaches Uniform Sampling, K-Means clustering, and S-RNN [3] along with the ones obtained via the DS3 method using self summarization ($\mathbb{Y} = \mathbb{X}$), the story graphs constructed with visual features ($\mathbb{Y} = \mathbb{S}^{V}$), both visual and GPS features ($\mathbb{Y} = \mathbb{S}^{VG}$) and all visual, GPS and textual features ($\mathbb{Y} = \mathbb{S}^{VGT}$).

considering the content of the input photo album. Other approaches provide more diverse summaries, but $\mathbb{S}^{VGT}$ seems to provide the best result as the images selected for the summary cover the main events depicted in the input photo album, and they appear to be semantically more close the summary by a human. Overall, both of our qualitative and quantitative results show that photo album summarization can benefit from exploiting visual story graphs as a prior to encourage producing more coherent summaries.

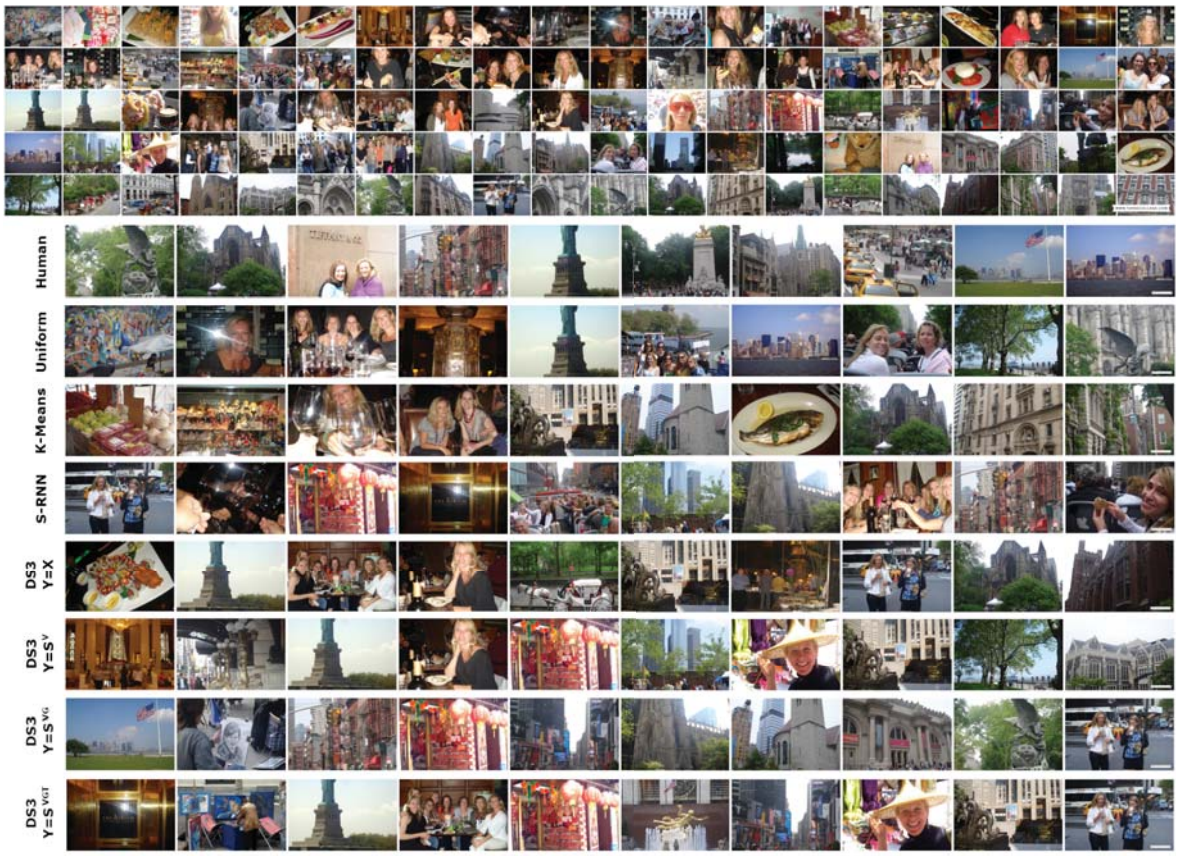


Figure 5.4. Summarization results of city New York. Top: Input photo album. Bottom: Visual summaries done by a human, the baselines approaches Uniform Sampling, K-Means clustering, and S-RNN [3] along with the ones obtained via the DS3 method using self summarization ($\mathbb{Y} = \mathbb{X}$), the story graphs constructed with visual features ($\mathbb{Y} = \mathbb{S}^V$), both visual and GPS features ($\mathbb{Y} = \mathbb{S}^{VG}$) and all visual, GPS and textual features ($\mathbb{Y} = \mathbb{S}^{VGT}$).

Table 5.3. V-ROUGE scores for the summarization experiments.

Photo Album	**Amsterdam Trip**	**Istanbul Trip**	**New York Trip**	**Paris Trip**	**Tokyo Trip**	**Venice Trip**
Uniform	0.31	0.38	0.48	0.33	0.45	0.45
K-means	0.45	0.26	0.39	0.37	0.39	0.29
S-RNN	0.30	0.39	0.41	0.35	0.42	0.33
DSS-S	0.38	0.41	0.39	0.38	0.39	0.24
DSS-D	0.40	0.44	0.49	0.39	0.52	0.27
DS3 ($\mathbb{Y} = \mathbb{X}$)	0.48	0.47	0.56	0.52	0.49	0.54
DS3 ($\mathbb{Y} = \mathbb{S}^V$)	0.48	**0.53**	0.61	0.44	0.52	0.57
DS3 ($\mathbb{Y} = \mathbb{S}^{VG}$)	0.46	0.42	0.50	0.47	0.53	0.58
DS3 ($\mathbb{Y} = \mathbb{S}^{VGT}$)	**0.56**	0.49	**0.67**	**0.56**	**0.63**	**0.66**

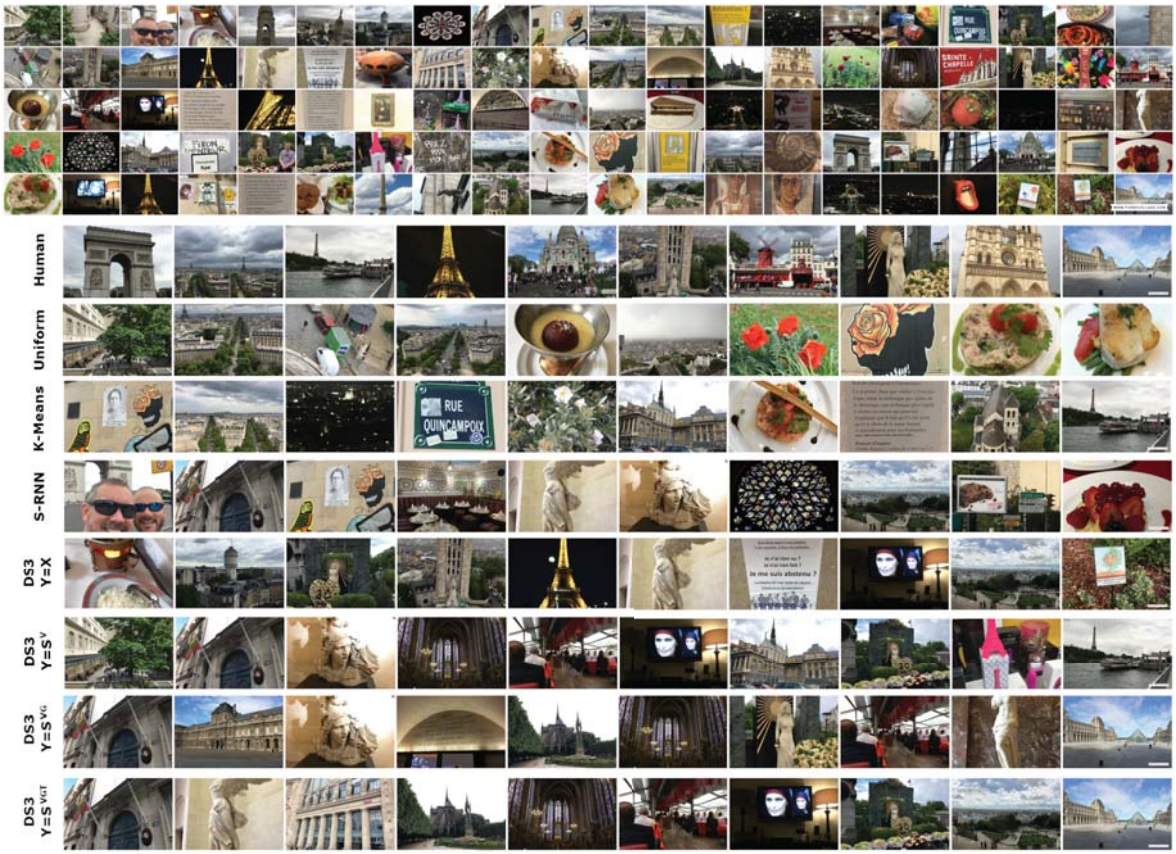


Figure 5.5. Summarization results of city Paris. Top: Input photo album. Bottom: Visual summaries done by a human, the baselines approaches Uniform Sampling, K-Means clustering, and S-RNN [3] along with the ones obtained via the DS3 method using self summarization ($\mathbb{Y} = \mathbb{X}$), the story graphs constructed with visual features ($\mathbb{Y} = \mathbb{S}^V$), both visual and GPS features ($\mathbb{Y} = \mathbb{S}^{VG}$) and all visual, GPS and textual features ($\mathbb{Y} = \mathbb{S}^{VGT}$).

Table 5.4. F-measure scores for the summarization experiments.

Photo Album	**Amsterdam Trip**	**Istanbul Trip**	**New York Trip**	**Paris Trip**	**Tokyo Trip**	**Venice Trip**
Uniform	0.02	0.10	0.17	0.05	0.11	0.13
K-means	0.12	0.05	0.06	0.09	0.12	0.05
S-RNN	0.05	0.10	0.11	0.07	0.08	0.08
DSS-S	0.07	0.07	0.09	0.16	0.08	0.16
DSS-D	0.12	0.11	0.17	0.15	0.13	0.18
DS3 ($\mathbb{Y} = \mathbb{X}$)	0.16	0.08	0.12	0.13	0.08	0.20
DS3 ($\mathbb{Y} = \mathbb{S}^V$)	**0.25**	0.10	0.14	0.17	0.04	0.17
DS3 ($\mathbb{Y} = \mathbb{S}^{VG}$)	0.15	**0.12**	0.16	0.19	**0.15**	0.21
DS3 ($\mathbb{Y} = \mathbb{S}^{VGT}$)	0.14	0.10	**0.19**	**0.21**	0.11	**0.23**

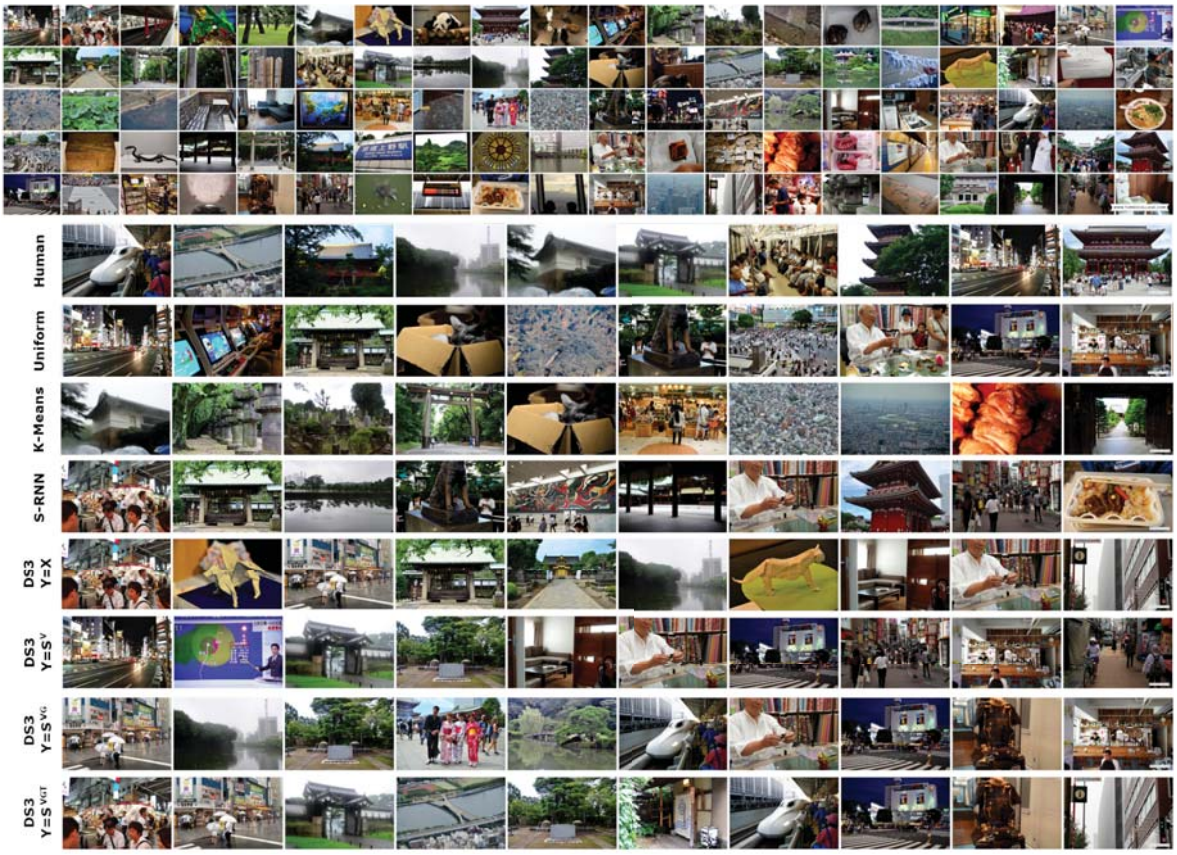


Figure 5.6. Summarization results of city Tokyo. Top: Input photo album. Bottom: Visual summaries done by a human, the baselines approaches Uniform Sampling, K-Means clustering, and S-RNN [3] along with the ones obtained via the DS3 method using self summarization ($\mathbb{Y} = \mathbb{X}$), the story graphs constructed with visual features ($\mathbb{Y} = \mathbb{S}^V$), both visual and GPS features ($\mathbb{Y} = \mathbb{S}^{VG}$) and all visual, GPS and textual features ($\mathbb{Y} = \mathbb{S}^{VGT}$).

5.3.3. Next Image Prediction

In our second experiment, we focus on the next image prediction task suggested in [1], which captures a story graph's ability in predicting what happens next given an input image. This task is related to evaluating coherence aspect of story graphs as the purpose is to identify how related the output image is to the query in terms of spatio-temporal continuity. We first select a small subset of canonical images for each city by simply clustering the entire set of photos into 50 clusters and retrieving the most photos that are close to the cluster centers. Given a query image, we localize the most similar photo in the reconstructed story graph and retrieve its next image in the corresponding chain. In the user study, subjects are presented

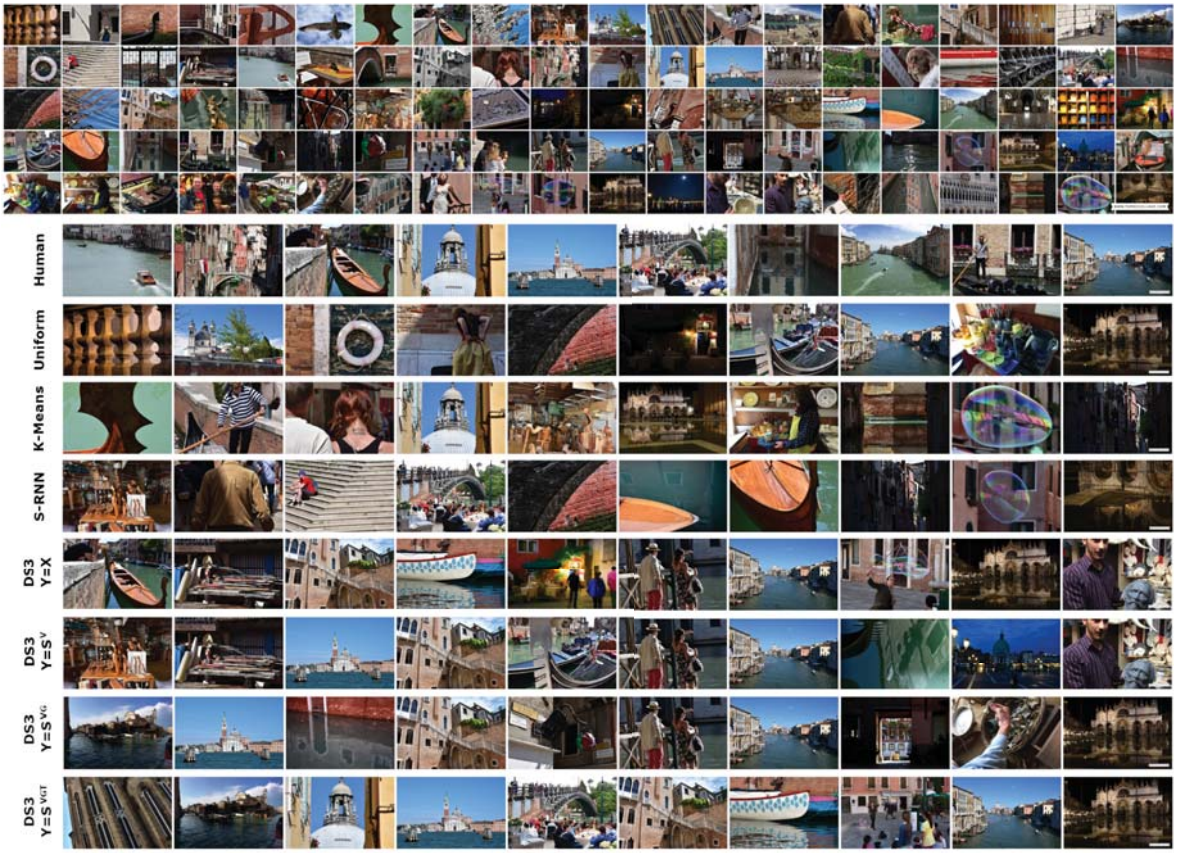
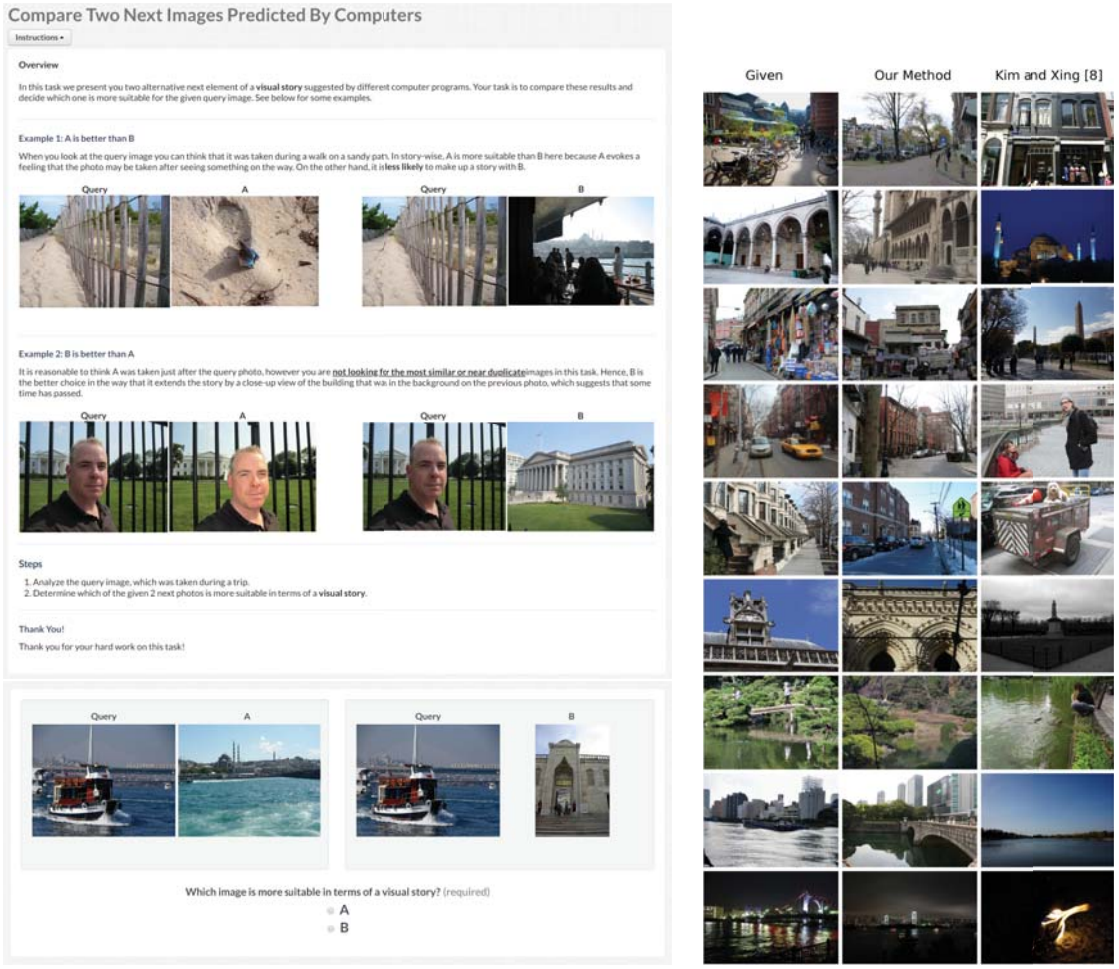


Figure 5.7. Summarization results of city Venice. Top: Input photo album. Bottom: Visual summaries done by a human, the baselines approaches Uniform Sampling, K-Means clustering, and S-RNN [3] along with the ones obtained via the DS3 method using self summarization ($\mathbb{Y} = \mathbb{X}$), the story graphs constructed with visual features ($\mathbb{Y} = \mathbb{S}^V$), both visual and GPS features ($\mathbb{Y} = \mathbb{S}^{VG}$) and all visual, GPS and textual features ($\mathbb{Y} = \mathbb{S}^{VGT}$).

with results obtained with our approach and with those by Kim and Xing's method [1] and are asked to choose the one which is the most likely sequence (Fig. 5.8.(a)). We perform the user study on Figure Eight platform[2] in which a total of 331 workers have participated. For each test question, we obtain responses from at least 10 users. Fig. 5.8.(b) shows examples of the next likely images predicted by our approach and the competing method. The results of the pairwise preference tests are given in Table 5.5.. On average, our predictions are favored 61% of the time.

[2]Figure Eight is a web-based data annotation company which can be accessed from https://www.figure-eight.com



(a) (b)

Figure 5.8. Next image prediction. (a) Screenshot of the user interface used in our experiments on the next image prediction task. (b) Example images predicted by our algorithm and the method of Kim and Xing [1].

Table 5.5. User study results for the next image prediction task. The preference rate denotes the percentage of comparisons in which the users favor one method over the other. On average, our predictions are preferred 61% of the time against the state-of-the-art method in [1].

	Amsterdam	Istanbul	New York	Paris	Tokyo	Venice	Average
Kim and Xing [1]	43.1	48.6	12.3	45.3	42.4	44.9	39.4
Ours ($\mathbb{S}^{VGT}$)	**56.9**	**51.4**	**87.7**	**54.7**	**57.6**	**55.1**	**60.6**

5.3.4. Coverage

In our last set of experiments, we compare the coverage of the story graphs generated by our approach and the method of Kim and Xing [1]. For each city, we first identified a diverse set of tags about the points of interest and attractions in that city via inspecting the user tags from YFCC100M dataset and additionally using the Google search engine. Table 5.6. shows these tags. For each tag we also provide an illustrative image just to give the workers an opinion about what that tag is about. In the user study, we then show the photos compiled from the reconstructed story graphs and ask users to select the tags that they think are relevant to one or more images displayed to them (Fig. 5.9.). For each tag we estimate the percentage of workers who selected the tag for that particular story graph. Then, we calculate the average selection rate through all the tags to get the final coverage rate of the story graph with respect to all the tags of that city. We perform the user study on Figure Eight platform in which a total of 238 workers have participated. For each test question, we obtain responses from at least 10 users. For each city, our story graph achieves a higher coverage rate than that of Kim and Xing [1]. On average, our proposed approach covers 46.3% of the tags whereas the method of Kim and Xing covers 34.8% (Table 5.7.). This demonstrates that the photos in the story graphs extracted by our method include points of interests and more interesting locations for a city, resulting in a more inclusive and covering visual narrative of a city.

5.4. Summary

In this chapter, We demonstrate that story graphs obtained with our approach we proposed in previous chapter 4. can be utilized for photo album summarization. In particular, our story graphs can be interpreted as a kind of prior that represent important concepts, landmarks and events depicted in the large photo collections, and hence, the images in the story graphs can serve as a measure of representativeness while extracting summary of a photo album of similar theme. Our experimental analysis reveals that the story graphs obtained by our approach allow to obtain better performances than the previous approaches for three different tasks including photo album summarization, next image prediction, tag coverage.

Table 5.6. Tags used in coverage experiments.

City	Tags
Amsterdam	*Anne Frank House, Canals, Church, Cycling, Dam Square, Fine arts, Food, NEMO Science Museum, Night life, Parks, Port of Amsterdam, Rijksmuseum, Royal Palace Amsterdam, Van Gogh Museum, Windmills*
Istanbul	*Basilica Cistern, Bath houses, Beyoglu Street, Bosphorus Bridge, City Walls, Galata Tower, Grand Bazaar, Maiden's Tower, Mosques, Museums, Obelisk of Theodosius, Palace, Sea tour, Turkish food*
New York	*Broadway, Brooklyn Bridge, Cathedral, Chinatown, Coney Island, Grand Terminal, Museums, NYC Subway, Parks, Public Library, Skyscrappers, Statue of Liberty, Times Square, Wall Street*
Paris	*Arc De Triomphe, Art, Cafes, Champs Élysées, Eiffel Tower, Fountains, Louvre Museum, Montmartre, Moulin Rouge, Musée d'Orsay, Notre-Dame de Paris, Pantheon, Parks and gardens, Versailles*
Tokyo	*Disneyland, Edo-Tokyo, Fish Market, Ginza Crossing, Japanese food, Kabuki Theatre, Mount Fuji, Museums, Parks, Rainbow Bridge, Roppongi, Sanrio Puroland, Skytree, Subway and trains, Temples, Tokyo Imperial Palace, Traditional clothes*
Venice	*Bridge of Sighs, Carnival Masks, Fine Arts, Glassworks, Gondola, Grand Canal, Venetian Lagoon, Lido, Museums, Palazzo Ducale, Rialto, San Marco, St Mark's Campanile, Venetian Churches*

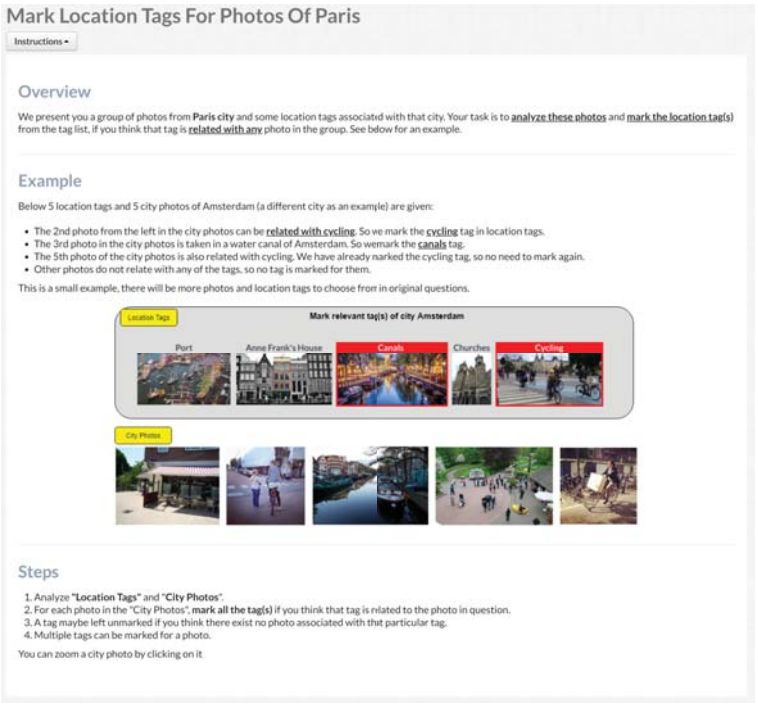


Figure 5.9. A screenshot of the user interface used in our experiments on the coverage task.

6. Intrinsic Properties

Until now, we made use of apparent or more concrete properties of photos when we are constructing our story graph generation framework. These properties generally come together

Table 5.7. User study results for the coverage task. The scores denote the average percentage of the tags selected by the workers for images included in the story graphs. On average, our story graphs cover 46% of the tags, providing a significantly higher rate than that of the state-of-the-art method in [1].

	Amsterdam	Istanbul	New York	Paris	Tokyo	Venice	Average
Kim and Xing [1]	34.7	24.3	30.0	41.9	26.7	50.9	34.8
Ours ($\mathbb{S}^{VGT}$)	**45.3**	**50.1**	**38.6**	**43.0**	**43.4**	**57.1**	**46.3**

with the photo in terms of metadata (timestamp, geological location, etc) or are straightforward to extract from the photo itself (visual patches). However, photos intrinsically shelter more abstract properties hidden beneath the apparent pixels which are more related with human sentiments, emotions or indirect perception. Some examples of this kind of intangible properties are *interestingness, aesthetics* - at which level people find an image interesting or aesthetically attractive, *popularity* - how popular an image is among human preference - or *memorability* - generally how easy people can retrieve the image from their memories when they see it again. Even though these properties seem to be more subjective due to their personal nature, scientific studies show that people statistically remember and/or forget particular kinds of photos and there is consensus on the photos they select if that photo seems attractive or dull.

We incorporated these intrinsic properties into our story graph generation framework to analyse their contribution to the quality and/or usability of the story graphs. However, first we will give the details of our work on image memorability, which we will utilize in our story graph generation in the next chapter.

6.1. Attention Related Memorability With Semantics

We humans have an astonishing ability to rapidly perceive and understand complex visual scenes. When exploring parts of a city that we have never visited before, glancing at the pages of a magazine or a newspaper, watching a film on television, or the like, we are constantly bombarded with a vast amount of visual information, yet we are able to process this information and identify certain aspects of the scenes almost effortlessly [65, 66]. We also

have an exceptional visual memory [67, 68] that we can remember particular characteristics of a scene with ease even if we look at it only a few seconds [69]. Here, what is being remembered is considered nothing like an identical representation of the scene itself but the gist of it [70, 71]. Although there is no general agreement in the literature about the contents of this "gist", the most common definitions include statistical properties of the scene such as the distributions of basic features like color and orientation, the structural information about the scene layout like the spatial envelope of Torralba and Oliva [29], and the image semantics such as existing objects and their spatial relationships.

Interestingly, we can recall some images surprisingly well while some are lost in our minds. Put simply, not all images are equally memorable. Isola et al. [4] were the first to carry out a computational study about this phenomenon, the so-called intrinsic memorability of images. They devised a Visual Memory Game experiment and utilized Amazon's Mechanical Turk service to quantify the memorability of 2222 natural images (see Figure 6.1.). In the course of these experiments, a total of 665 participants were shown a sequence of images, each of which was displayed for 1 second with a short gap in between image presentations. These subjects were then asked to provide a feedback any time whenever he/she thinks an identical image is displayed. By this setup, a memorability score for each image is calculated by the rate at which the subjects detect a repeated presentation of it. The authors showed that the memorability of an image is pretty consistent across subjects and under a wide range of contexts, which indicates that image memorability is in fact an intrinsic property of images. In addition, the authors explored the use of different visual features and interestingly showed that the intrinsic memorability of an image can indeed be estimated reasonably well by a machine. Since that seminal work, there has been only a few works that explore this difficult and interesting problem [6, 30–32, 72].

Our first goal in this part is to explore the role of visual attention in understanding image memorability. We humans use attentional mechanisms to efficiently perform higher level cognitive tasks by focusing on a small and relevant bits of the visual stimuli. Figure 6.2. illustrates this function of visual attention in selecting important features from images. Suppose that we are exposed to these three natural images, each having different visual contents,

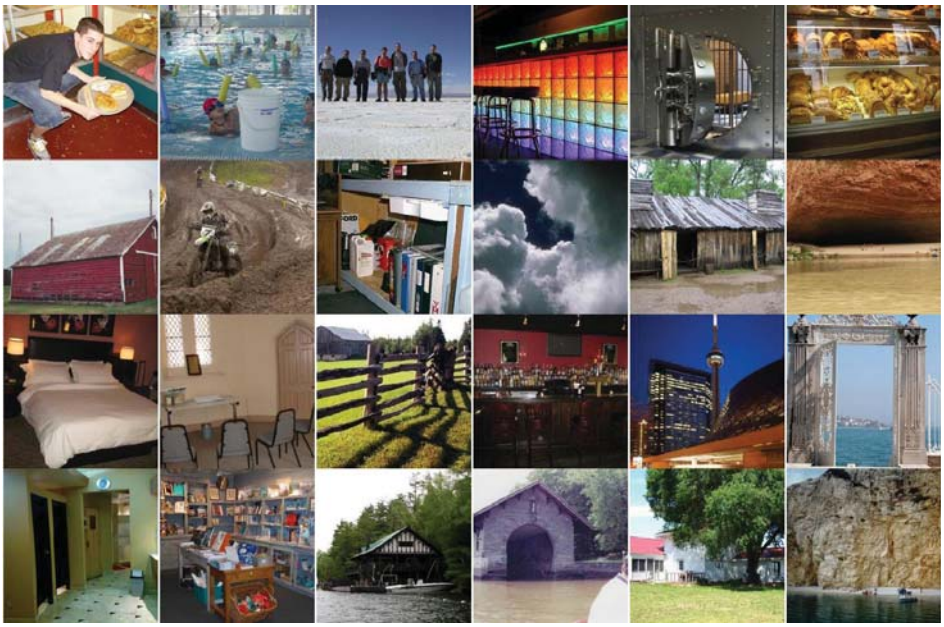


Figure 6.1. Sample images from the MIT memorability dataset [4]. The images are sorted from more memorable (top left) to less memorable (bottom right).

i.e. different objects, scene characteristics. Intuitively, our visual system focuses on certain regions that attract our attention as modeled here by a bottom-up saliency model. In this work we propose a visual attention-driven spatial pooling strategy to select important features from images. Our approach makes use of two complementary feature pooling schemes related to visual attention. First, we investigate selecting features from the most salient regions of the images determined according to a recently proposed bottom-up visual saliency model [5]. Our second scheme, on the other hand, considers a top-down definition of visual attention and employs an object-centric spatial pooling scheme. To our interest, a body of research in cognitive sciences promotes that attention plays an important role in understanding natural scenes and enhancing visual memory [71, 73–76]. However, none of the previously proposed memorability models make use of any attentional mechanisms for feature selection, and only [30, 32] use saliency maps but as additional image features.

Apart from the global dense image features, some previous studies on image memorability [4, 6, 30, 31] have also investigated the use of high-level semantic information about images. They consider objects-related features [4, 31], presence of certain object and scene

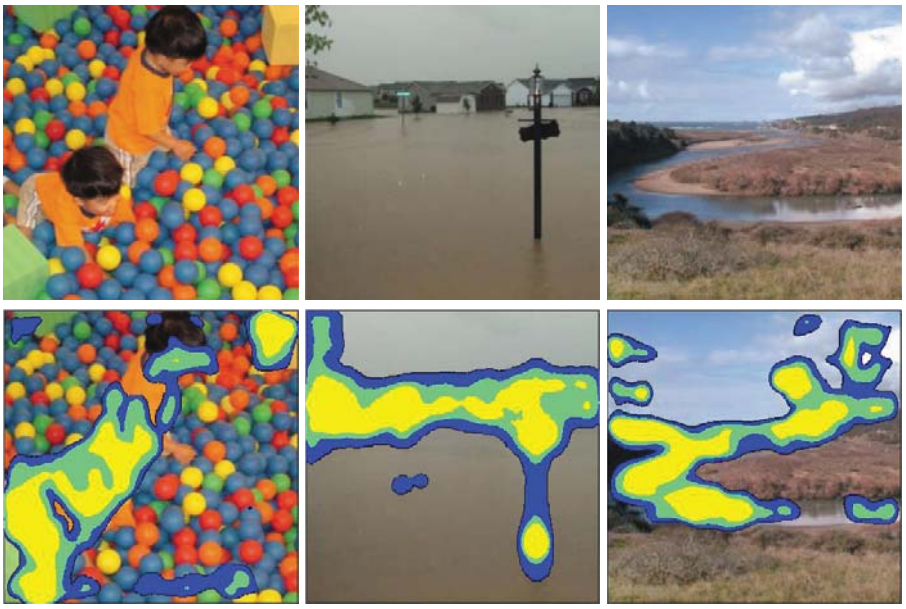


Figure 6.2. Top: Examples for the most memorable (left), typically memorable (middle), least memorable (right) images in the MIT memorability dataset. Bottom: Salient regions of the images extracted by the method in [5]. The color coding shows the strength of saliency with yellow, green and blue regions corresponding to top 10%, 20%, %30 most salient parts, respectively.

categories [4, 6, 30], and their attributes [6], which are all based on manual annotations produced by humans. Figure 6.3. illustrates some sample images from the MIT memorability dataset along with the semantic features that are manually collected from the human subjects [6]. As illustrated here, an image can be semantically represented in terms of objects, scene information and related attributes.

In addition to our attention-driven feature selection strategy, our second focus in this part is to investigate the use of a diverse set of recently proposed semantic features which encode meta-level object categories [48], scene attributes [7], and invoked feelings [8] for predicting image memorability. Compared to the features considered in the former studies [4, 6, 31], these semantic features can be directly extracted from the images, eliminating the need for manual annotations. Using these features thus decreases the complexity of the prediction process and makes the prediction model to work in a fully automatic manner. Moreover,

Object: *person, seat, bottle, chair, floor*

Scene: *indoor, casino, sports and leisure*

Attribute: *has_person, attractive, pleasant, individual, routine, sitting, clear_glasses, ...*

Object: *person, wall, chandelier, ceiling lamp*

Scene: *indoor, shopping and dining, bakery/shop*

Attribute: *has_person, standing, people go, is_interesting, group, routine, ...*

Object: *mountain, sky, tree, natural elevation*

Scene: *outdoor natural, mountains hills desert sky*

Attribute: *peaceful, is_interesting, hang_on_wall, exciting, famous, ...*

Figure 6.3. Top: Examples for the most memorable (left), typically memorable (middle), least memorable (right) images in the MIT memorability dataset. Bottom: Sample human annotated attributes as collected in [6].

compared to prior work, these features encode semantic properties of images from a perspective or scale that has not been investigated before. The Meta-class descriptor [48] encodes image semantics based on a hierarchical structure of object categories (concepts) by capturing the relationships among them. The SUN Scene Attributes [7] represents an image by means of responses of a comprehensive list of attribute classifiers that relates to different scene characteristics such as affordances, materials and surface properties. The SentiBank features [8] are the responses of a set of classifiers trained to detect adjective-noun pairs (attributes - objects), and used to associate certain sentiments with images.

In order to validate our approach, we performed a series of experiments on the MIT memorability dataset. To show the effectiveness of the attention-driven pooling strategy, we used the dense global features employed in [4], namely SIFT [45], HOG [46], SSIM [47] and we analyzed the gain when the features pooled over the salient regions are concatenated to the feature vectors obtained with spatial pyramid pooling [77]. Moreover, regarding our second goal, we performed experiments with the high-level semantic features [7, 8, 48] and tested their performances on predicting image memorability. Lastly, we compared our combined

model, which uses both semantic features and dense global features pooled over salient regions and spatial pyramids, to the state-of-the-art models in the literature.

Our main contributions are: (1) an attention-driven pooling approach to put special emphasis on the interesting parts of the images in the computations, (2) a systematic analysis of a diverse set of semantic features on predicting image memorability, and (3) experiments demonstrating that the combination of these ideas provides significant improvement over the existing fully-automatic models.

6.1.1. Attention-driven Spatial Pooling

The memorability model by Isola et al. [4] and the follow-up studies [30, 32] all employ spatial pyramid (SP) based pooling [77] for dense global features (Section 6.1.1.). In this study, we propose a complementary visual attention-driven spatial pooling scheme for image memorability, which allows us to select features from the salient image regions. In particular, these regions are estimated by considering two different saliency maps. While one of them is estimated via a bottom-up saliency model , the second one is derived from a complementary object-level saliency map which captures information about foreground objects in the images . We will give the details of our proposed attention-driven pooling strategy in the remaining part of this section.

The common pipeline of modern visual recognition tasks uses spatial pooling in order to construct compact representations and achieve robustness to noise and clutter. After extracting local or global low level features from images, feature vectors are encoded to codewords using a descriptive vocabulary. Then, histograms of these codewords are computed in order to get the fixed-length exemplar vectors of the predefined subregions of the image. Final representation is formed by simple concatenation of all histogram vectors obtained in this way. Boureau et al. [78] showed various factors that affect the performance of pooling strategies and demonstrated the importance of the step. For example, Isola et al. [4] used simple 2-level spatial pyramid pooling strategy in their work. However, in this study, we approach the

pooling step by further incorporating visual attention mechanisms with the inspiration that visual attention is considered highly related with memorability [71, 73–76].

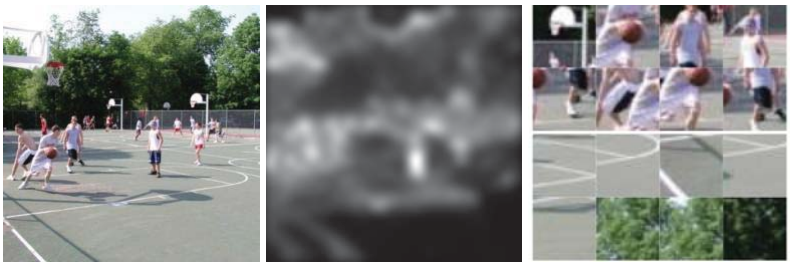
Visual Saliency: In recent years, there has been an increasing interest in computational models of visual saliency estimation and their use for several computer vision tasks. Starting from the seminal work by Itti, Koch, and Niebur [79], most of the existing models consider a bottom-up strategy. First, center-surround differences of various features at multiple scales are computed for each feature channel. Then the final saliency map is formed by linearly combining feature maps after a normalization step. For a recent survey, please refer to [80]. In our experiments, we employed the publicly available implementation of a recently proposed saliency model [5][3], which examines the first and second-order statistics of simple visual features such as color, edge and spatial information.

Consider Figure 6.4.(a) where we present the result of the bottom-up saliency estimation for a sample image. From the saliency map given in the second column, we randomly sample a number of image patches (rightmost four columns). Those sampled within the top 10% salient locations are given in the top two rows whereas the bottom two rows show sample patches from the bottom 20% salient locations. As can be seen, the saliency values are the strongly correlated with the interestingness of the regions [81, 82] in the sense that while the most salient patches captures the children, the least salient ones mostly correspond to background or those regions which have little importance in terms of image content.

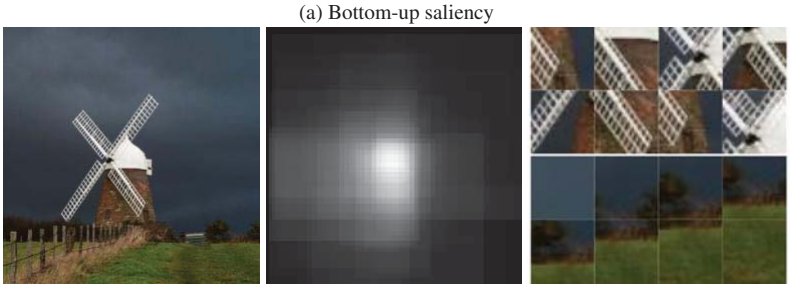
Objectness Measure: In [83], Alexe et al. introduced a generic (category-independent) objectness measure[4] to quantify how likely an image window contains an object. In more detail, the authors first analyzed several image cues, namely multi-scale saliency, color contrast, edge density (near window borders) and superpixel straddling, each of which were shown to be an indicator of objectness, but to a certain degree. Then they proposed a learning framework to combine these four cues to distinguish object windows from background. It was demonstrated that the approach is very general and can detect objects of novel classes

[3]The source code is available at `http://web.cs.hacettepe.edu.tr/~erkut/projects/CovSal/`

[4]The code is publicly available at `http://groups.inf.ed.ac.uk/calvin/objectness/`



(a) Bottom-up saliency



(b) Object-level saliency

Figure 6.4. Interesting and uninteresting patches extracted from two natural images based on visual attention. From the images, 8 image patches are sampled randomly from the top 10% salient locations (top 2 rows) and 8 others from the bottom 20% salient locations (bottom 2 rows) according to (a) a bottom-up visual saliency map and (b) an object-level saliency map, respectively.

not seen during training. As compared to the visual saliency model reviewed in the previous section which solely depends on bottom-up visual cues, the generic objectness measure can be used to estimate object-level saliency of images and provide top-down high-level information.

Figure 6.4.(b) shows some sample patches sampled from the object-level saliency map as we did for the bottom-up saliency. Similarly, the rightmost top two rows of patches taken from salient regions mostly correspond to the mill in the image, which is the most salient object. Other non-salient patches correspond to unimportant areas such as the sky or the field.

Instead of using a fixed pooling layout like the spatial pyramid structure used in [4], we propose an image-driven pooling strategy by considering salient regions of the images. For this purpose, we both utilize the bottom-up and object-level saliency maps described in the previous subsections. In this way, our pooling method adaptively focuses solely on the image regions that attract attention, ignoring not important, non-attractive parts of the images.

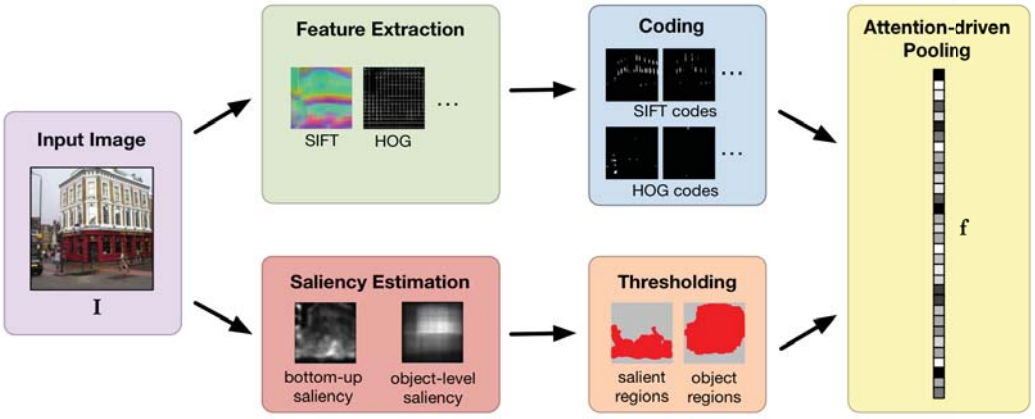


Figure 6.5. The proposed visual attention-driven spatial pooling pipeline for image memorability.

The system diagram of the proposed pooling approach is given in Figure 6.5.. First, dense visual features are extracted from the input image. Low level dense features are then encoded into higher dimensions through vector quantization using a bag of features approach. In the meantime, bottom-up and object-level saliency maps are estimated from the image and then thresholded to obtain both the salient regions and those regions possibly containing important foreground objects. Next, to form histogram-based visual descriptors, the encoded vectors are pooled together over the extracted attention-driven spatial layouts.

For the prediction pipeline for spatial pooling, we used the following steps:

(1) Feature Extraction. For an image $\mathbf{I}$, we obtain a global description of $\mathbf{I}$ by extracting D-dimensional local features such as SIFT [45], HOG [46], SSIM [47] at N different locations, denoted with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$. The SIFT descriptor gives the local image structural information whereas the HOG descriptor provides rich local orientation information that can be related to the receptive fields found in early human vision areas. Lastly, the SSIM descriptor captures the local layout of geometric patterns.

(2) Coding. Assuming that we have a learned codebook of K visual words, denoted with $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K] \in \mathbb{R}^{D \times K}$, each local feature $\mathbf{x}_i \in \mathbf{X}$ is encoded into a code vector $\mathbf{c}_i = [c_1^i, c_2^i, \ldots, c_K^i]^T$ by applying vector quantization. After the coding step, $\mathbf{I}$ is represented by a set of codes $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N] \in \mathbb{R}^{N \times K}$.

(3) Bottom-up and object-level saliency maps. To obtain the attention-driven spatial layouts for the proposed feature pooling scheme, we make use of bottom-up and object-level saliency maps. The bottom-up visual saliency map of image $\mathbf{I}$ is computed by a recently proposed model [5], which was shown to provide state-of-the-art performance in predicting eye fixations. For the object-level saliency map, we randomly sample many windows from $\mathbf{I}$ and measure the objectness of these image windows by using the generic objectness measure proposed in [83]. Then we compute an objectness score for each pixel by averaging over all the scores of the windows which contain that pixel to obtain the generic objectness map of $\mathbf{I}$.

(4) Pooling. In the pooling step, instead of considering a fixed image-independent set of spatial regions, as employed in [4], here we propose to use image-specific spatial regions for feature pooling. Specifically, we locate the regions of interest by respectively segmenting the bottom-up and object-level saliency maps into salient/non-salient and object/non-object regions by thresholding. In our experiments, we varied the threshold value to find the optimum thresholds to determine salient and object regions in the images for spatial pooling of features. We found out that the mean works well for the bottom-up saliency maps whereas the best performance for the object-level saliency maps is achieved when the threshold is set to 0.25 times the maximum objectness value. Figure 6.6. shows some examples of these attention-driven regions. For each region of interest $\mathcal{R}$, we then perform average-pooling, i.e. compute a histogram (or take the average) of the codes over the region $\mathcal{R}$:

$$f(\mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum\nolimits_{i \in \mathcal{R}} \mathbf{c}_i \tag{17}$$

where $|\mathcal{R}|$ denote the number of dense features in $\mathcal{R}$. Moreover, the final feature vector $f(\mathcal{R})$ is renormalized to have L_1-norm of 1.

Figure 6.6. Visual attention-driven feature pooling scheme. For a given image a bottom-up saliency map and (b) an object-level saliency map are estimated and then the feature vectors are pooled over the salient regions of the images (depicted as bright areas in the images.

6.1.2. Semantic Features

[4] showed that memorability of an image is highly correlated with the semantic content of the image. For instance, only making use of manual annotation of object and scene labels is shown to give pretty good results. In a follow-up work [6], the authors collected attributes that humans used to describe images and explored their role in determining the intrinsic memorability of images. Motivated from these findings, here, our goal is to extend our framework to include automatically extracted semantic attributes. For that purpose, we propose to use three recently proposed semantic descriptors: The Meta-class descriptor [48] provides object-specific high-level information about image content (Section 6.1.2.). The SUN Scene Attributes [7], on the other hand, characterize the images by means of a set of hand-picked functional, material, surface and spatial properties (Section 6.1.2.). Finally,

the SentiBank features [8] are used to include feelings that are invoked in a viewer into the computations (Section 6.1.2.).

Meta-class Features: In computer vision, attributes typically denote properties that humans use to verbally describe the visual content such as individual objects, object classes, scenes. Besides, they can also indicate properties shared among different object classes. The Meta-class descriptor [48] falls under this category that it captures common visual properties of different sets of object classes and represents an image in terms of them. In essence, these abstract categories are linear combinations of multiple non-linear classifiers trained on different low-level features. The authors trained a tree of classifiers using a subset of ImageNet [84] dataset and with the help of predefined object classes from ILSVRC2010 and Caltech256 datasets. Each node in the tree correspond to a meta-class obtained by combining two previously determined meta-classes (i.e. a set of object classes) which makes them easy to distinguish from other sets of object classes. They demonstrate that this descriptor gives state-of-the-art results for object categorization against similar semantic representations such as Object-Bank [85] and PiCoDes [86].

In our work, we use Meta-class features, i.e. the responses of the learned tree of classifiers, to obtain a semantic representation of image content by means of the presence or absence of the meta-classes. Figure 6.7. demonstrates the importance of certain object classes in determining the memorability of an image on some sample images from the MIT memorability dataset. It can be easily observed that the most memorable images generally are those that contain close-up human faces. Interestingly, typical memorable images generally do have humans and/or human-made structures or objects at a distance. The least memorable images are mostly the images of natural scenes.

SUN Scene Attributes In [7], Patterson and Hays carried out a large scale experiment to form a scene attribute dataset by crowdsourcing. They collected 102 discriminative attributes related to different visual properties of a scene, namely affordance, material, surface and spatial envelope properties. Using these collected attributes as ground truth, they also trained a binary classifier for each attribute and proposed to use responses of these classifiers to obtain

Figure 6.7. Sample images from memorability database. Top row shows samples from most memorable images which mostly contain close-up human faces. Middle row shows samples from typically memorable images which generally have humans and/or human-made structures or objects at a distance. Bottom row shows least memorable samples which are mainly the images of natural scenes.

an attribute based scene representation. They showed that this intermediate level representation captures scene content information remarkably well and can be effectively used for different computer vision tasks including scene classification, automatic image captioning, semantic image retrieval.

In our framework, we use the confidence scores of the scene attribute classifiers as complementary semantic features for learning image memorability. Figure 6.8. illustrates some of the most confident scene attributes [7] that are extracted from some sample images having different memorability scores. We observe that the most memorable images are typically associated with the "*no-horizon*", "*enclosed-area*", "*cloth*" and "*man-made*" attributes whereas

cloth, enclosed area, no-horizon, electric/indoor lighting

man-made, no-horizon, enclosed area, natural light

far-away horizon, natural light, open area, grass, natural, vegetation

Figure 6.8. Sample images from memorability database for most memorable (left), typically memorable (middle) and least memorable (right) with their most confident scene attributes predicted by [7].

the least memorable ones mainly have "*open-area*", "*grass*", "*vegetation*" and "*natural*" attributes. These observations are in accordance with the findings reported in [4, 6] suggesting that the images of people and enclosed spaces are more memorable images than those of natural images.

SentiBank Borth et al. [8] recently proposed a large scale visual sentiment ontology based on the psychological theory of Plutchick's Wheel of Emotions [87]. To construct this ontology, the authors followed a data-driven approach and used a large set of tagged images and videos from the web to gather a list of adjectives and nouns based on their co-occurrences with each of the 24 emotions defined in [87]. They assigned certain sentiment values to these tags and employed them to form Adjective-Noun Pairs (ANPs) which reflect strong emotions and frequently appear together. Then, they trained a classifier for each ANP using some low and high-level visual features. They finally selected 1200 of those trained ANP classifiers that have a reasonable classification performance to build their visual sentiment analysis framework known as the SentiBank.

In our approach, we use the visual sentiment classifiers from the SentiBank to include emotion-based semantic features to our image representations. Figure 6.9. demonstrates some sample images with different memorability scores with the associated ANPs as predicted by the SentiBank classifiers. As can be observed, in each case, the classifiers accurately capture the feelings invoked in the viewers. Although there is no common pattern for

shy smile, traditional dress, innocent smile | *warm creek, smooth water, calm pond* | *beautiful garden, sunny trees, pretty scene*

Figure 6.9. Sample images from memorability database for most memorable (left), typically memorable (middle) and least memorable (right) with their most confident sentiment ANPs as predicted by [8].

ANPs associated with images from different memorability levels, we observe that in general, the most memorable images are linked with the emotions that can relate to humans (e.g., *shy smile*). Moreover, the typically memorable images invoke feelings related to man-made structures (e.g., *calm pond*) whereas the least memorable ones are associated with ANPs related to natural scenes (e.g., *beautiful garden*).

6.1.3. Experiments

In this section, we first give brief details about our experimental setup and then demonstrate the effectiveness of the proposed approach through a series of experiments.

Experimental Setup For the quantitative analysis we used Spearman's rank correlation measure (ρ). The performance was evaluated over 25 different splits of the dataset containing 1111 training and 1111 testing images (the same splits used in [4]). These train and test splits were scored by different halves of the participants, showing a human consistency of $\rho = 0.75$. Thus, the effectiveness of a computational image memorability model can be assessed by measuring how close the model's Spearman's rank correlation to this score. In addition, the performance of a model can be analyzed at different memorability levels by ranking the test images according to their predicted memorability scores and then computing the cumulative average of empirical memorability scores at different quantiles. For instance, a good image memorability model should have cumulative averages close to 1 for the top

most memorable images predicted by a model and close to 0 for the bottom least memorable images.

Results and Discussions In the first part of the experiments, we analyzed the performance of our proposed attention-driven pooling scheme in detail. We conducted our experiments on three global dense features, SIFT [45], HOG [46] and SSIM [47], which were used in [4]. Specifically, we analyzed the performance when features obtained with our attention-driven pooling strategy are concatenated to those derived by the standard spatial pyramid pooling. We examined the prediction accuracy of each dense feature separately. We also provided the results for the combination of these features. We separately trained different SVRs to map from the features pooled over these maps to memorability scores.

A summary of our results is given in Table 6.1.. As can be seen, the attention-driven pooling alone performs poorly as compared to the 1-level spatial pyramid (SP) based pooling. However, for each dense feature, there is a notable improvement with the inclusion of our attention-driven pooling scheme to the SP based baseline. More specifically, the SSIM feature has the most significant gain where the correlation moves from $\rho = 0.436$ to $\rho = 0.454$. Furthermore, we observed that the result of the combined features can be also improved when our pooling strategy is used. However, the amount of gain, from $\rho = 0.458$ to $\rho = 0.472$, is relatively smaller than those of single features. When the average memorability scores of the models are examined at top 20/100 and bottom 20/100 quantiles, we have similar observations. In conclusion, the combined pooling framework performs especially much better by assigning less memorable images lower scores. These results support our claim that the image regions which retain in human memory are correlated with the areas that attract our attention.

In our second experiment, we included the semantic features, namely the Meta-class features [48], the SUN scene attributes [7] and the SentiBank features [8] to the original feature pool (pixels, GIST, SIFT, SSIM, HOG-based image features), and performed a thorough analysis of the framework with all possible combinations of these features and pooling strategies.

Table 6.1. Comparison of pooling schemes (Spatial Pyramid pooling (SP Level-1) and Attention-based Pooling (AP Level-1)) using dense global features SIFT, HOG and SSIM. Results are given as the average empirical memorability scores reported for the top 20, top 100 highest and bottom 20, bottom 100 lowest predicted memorability scores and the Spearman's Rank Correlation (ρ) values.

		SIFT	HOG	SSIM	SIFT+HOG+SSIM
SP(L1)	Top 20	83.8%	83.3%	83.2%	85.0%
	Top 100	82.3%	81.9%	80.7%	80.5%
	Bottom 100	54.9%	56.0%	56.7%	54.6%
	Bottom 20	50.3%	47.9%	54.0%	50.1%
	ρ	0.430	0.431	0.436	0.458
AP	Top 20	87.6%	87.8%	84.9%	87.4%
	Top 100	81.8%	83.0%	83.4%	83.7%
	Bottom 100	56.6%	55.9%	56.7%	55.6%
	Bottom 20	58.2%	48.4%	56.4%	51.8%
	ρ	0.390	0.420	0.427	0.438
SP(L1) + AP	Top 20	86.0%	86.9%	86.8%	86.9%
	Top 100	83.3%	82.9%	81.0%	82.6%
	Bottom 100	55.7%	54.8%	53.6%	53.4%
	Bottom 20	49.9%	47.4%	48.5%	53.2%
	ρ	0.435	0.448	0.454	0.472

Table 6.2. demonstrates the results obtained by SSIM (best performing low-level image feature), our semantic features and their combination. One key observation is that the Meta-class features and the Scene Attributes provide fairly good predictions as compared to the SentiBank or any other low-level cues. In particular, the Meta-class descriptor alone achieves approximately $\rho = 0.49$ correlation value, which shows us that memorability of images are not only related to single object properties but also to inherent and shared characteristics of different object classes. Similarly, the Scene Attributes alone give nearly $\rho = 0.48$, illustrating the importance of scene properties over objects in the images for image memorability. We achieved the best performance when we combined all semantic features and SSIM with a combination of our proposed attention-driven pooling and 2-level spatial pyramid pooling for the dense features. With this model of ours, the Spearman's rank correlation between the ground-truth ranking and the predictions is estimated as $\rho = 0.515$. This correlation value

Table 6.2. Comparison of the best local dense feature (SSIM) and all semantic features. Results are given as the average empirical memorability scores reported for the top 20, top 100 highest and bottom 20, bottom 100 lowest predicted memorability scores and the Spearman's Rank Correlation (ρ) values.

	SSIM	Scene Attributes	Meta-class	SentiBank	All
Top 20	86.8%	86.4%	86.8%	85.7%	85.0%
Top 100	81.0%	83.7%	81.5%	82.5%	83.3%
Bottom 100	53.6%	54.2%	53.3%	54.8%	52.2%
Bottom 20	48.5%	51.3%	46.7%	47.1%	47.4%
ρ	0.454	0.477	0.487	0.449	0.515

is smaller than the correlation among humans ($\rho = 0.75$) but it is the best result reported in the literature so far by a fully automatic scheme that does not use any manual object, scene or attribute annotations. It also demonstrates the importance of high-level semantic features as incorporating them increases the rank correlation score from $\rho = 0.472$ (SP(L1)+AP) to $\rho = 0.515$ (All). Moreover, the increases in the top 20 and top 100 average memorability predictions support the hypothesis that the semantic content of images is highly correlated with their intrinsic memorability.

In Table 6.3., we compare the result of our proposed method with the methods of Isola et al. [4], Khosla et al. [30], Mancas and Le Meur [32] and more recent works which utilized deep learning models [34–36]. Our method has the best performance among pre-deep learning studies which makes use of hand-crafted features for memorability prediction. While Khosla et al. [30] achieved ρ = 0.50 with their global model which additionally considers memorability characteristics of the local image regions, our model achieves slightly better results with far less complexity. Moreover, another key observation from Table 6.3. is that most of the memorability prediction schemes predict top memorable images with high precision. For the top 20 and top 100 images, the models have obtained nearly the same average empirical memorability values, which are very close to the scores of human subjects. However, predicting whether an image is less memorable is a more difficult problem. In that respect, our model provides better predictions for the bottom 20 and bottom 100 images as

Table 6.3. The first four rows indicate average empirical memorability scores over different memorability levels. (ρ) is the Spearman's rank correlation between predictions of existing fully automatic models and the empirical results.

	Isola et al. [4]	Khosla global [30]	Khosla local+global [30]	Mancas & Le Meur [32]	Our Approach	Human subjects
Top 20	83%	84%	85%	–	85%	86%
Top 100	80%	80%	81%	–	83%	84%
Bottom 100	56%	56%	55%	–	52%	47%
Bottom 20	54%	53%	52%	–	47%	40%
ρ	0.46	0.48	0.50	0.48	**0.52**	0.75

compared to the previous models.

In order to demonstrate the effectiveness of our proposed combined model, we compare our result with those of the human annotations reported in [72]. For object semantics, the authors in [72] achieved $\rho = 0.47$ whereas we obtained a correlation value of $\rho = 0.49$ with the Meta-class descriptor that describes abstract object classes. This shows that fully automatic approaches can also capture object semantics to some extent to improve memorability predictions. On the other hand, the model based on the attribute annotations, gives a better correlation value of ρ = 0.52 as compared to those of SUN Scene Attributes and SentiBank features respectively corresponding to $\rho = 0.48$ and $\rho = 0.45$. Moreover, the model which considers the combined overall semantics (objects + scenes + attributes) has a correlation value of $\rho = 0.54$, which is higher than that of our proposed combined model having $\rho = 0.52$. However, we observe that our model provides better predictions especially for the least memorable images. For the bottom 100 and 20 images while the average ground-truth memorability scores are $\%55$, and $\%51$ for object, scene and attribute annotations, respectively, ours are $\%52$, and $\%47$ which are much closer to the human subjects. Overall, human annotations still have advantage over automatic attributes, however the gap is small. Considering the cost of gathering annotations from human subjects, our approach gives similar performance with much less computational effort.

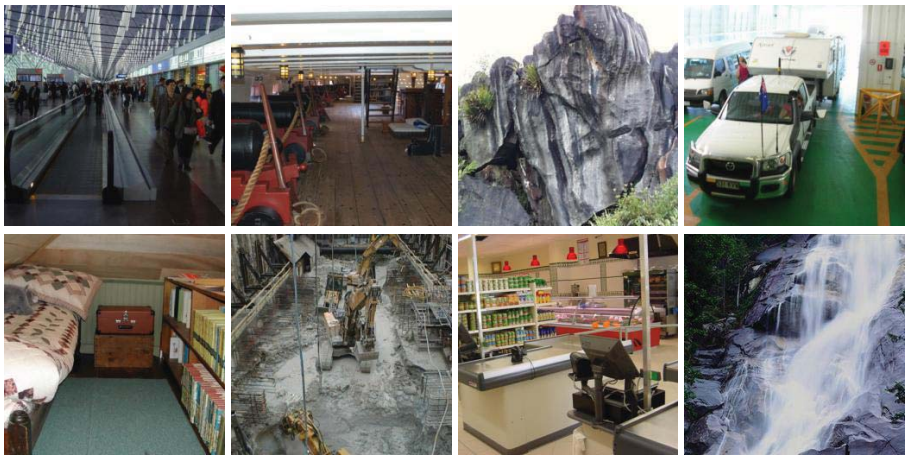
In Figure 6.10., we additionally show sample images for different memorability levels predicted by the proposed framework. Figure 6.11. shows some images on which the memorability predictions based on our approach are incorrect as compared to the empirical results.

Predicted most memorable (85%) Predicted typical memorable (68%) Predicted least memorable (40%)

Figure 6.10. Memorability predictions by the proposed strategy. Out of all test images, the 8 images in (a) are found to be the most memorable, the ones in (b) are predicted as typically memorable and the other 8 images in (c) are guessed as the least memorable. The numbers denote the average prediction scores of the given image sets. The images predicted as highly memorable contains highly distinctive visually salient elements as compared to other groups of images.

The reasons for this could lie in the inaccurate predictions of the semantic content or focused regions of images. In Figure 6.12.(a)-(b), for example, we provide the bottom-up and object-level saliency maps of two of the images from Figure 6.11. together with their memorability maps as computed by the protocol used in [4, 30]. In the memorability maps, the red regions illustrate the objects that contribute positively to the predicted memorability and the blue regions show the objects that contribute negatively to the predicted memorability. In an ideal case, the predicted salient regions need to correspond to the image regions that affect the memorability scores positively or negatively. For the image in Figure 6.12.(a) whose memorability rank is overshot by the proposed prediction scheme, our pooling method can not correctly identify the object regions that correlate with the image memorability. For the image in Figure 6.12.(b) whose memorability rank is undershot by the proposed scheme we observe a similar behavior for the detection of important object regions that affects the memorability predictions negatively. These imperfect predictions of the important image regions make the features collected via our attention driven pooling scheme cover the image content in an inaccurate way, affecting the estimated memorability scores.

Finally, it is important to note that at that time there was still a large gap between our result and that of human subjects in predicting the less memorable images. In Table 6.4. we show the more recent approaches utilizing deep neural networks on visual memorability prediction after our work. It is clear that they brought noticeable improvement over our approach. However, these studies trained their models using LaMem dataset [33] which is a larger set



Predicted too high (+793/1111)

Predicted too low (-873/1111)

Figure 6.11. Sample images on which our proposed scheme failed to capture the memorability. The memorability ranks are predicted too high for the images in (a) and too low for the ones in (b), as compared to their empirical memorability ranks. The numbers in the parentheses show the mean rank error between the predicted and the empirical ranks across each group.

specifically collected for memorability task and tested on the SUN dataset [54] which is the set that all previous non-deep learning memorability prediction works used. The most recent work of Fajtl *et al*. [36] achieved $\rho = 0.65$ the closest performance to human subjects. Although deep neural networks closed the gap between human and machine predictions, the results show that there is still room for improvement on the visual memorability prediction task.

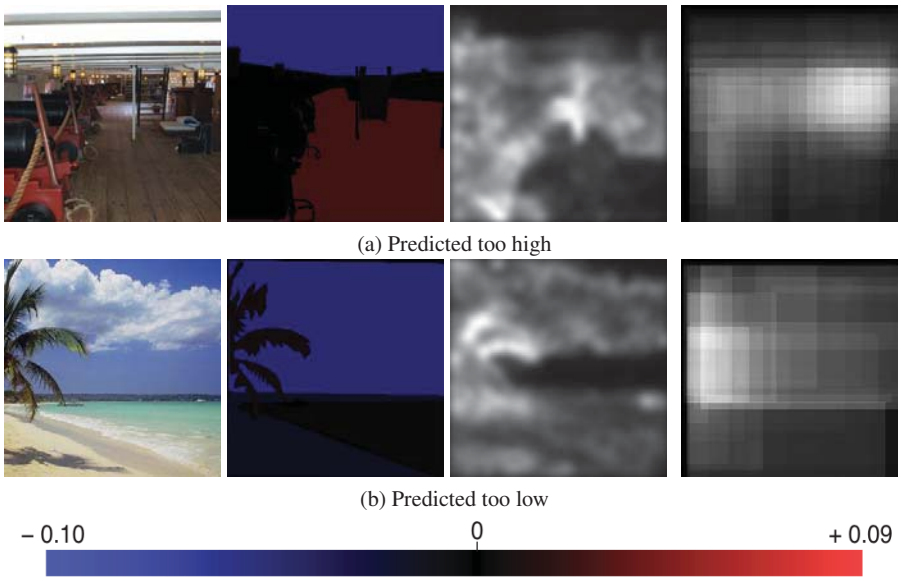


Figure 6.12. Memorability maps versus bottom-up saliency and object-level saliency maps of two of the images from Figure 6.11..

Table 6.4. Memorability scores of our framework and more recent methods using deep learning approaches. (ρ) is the Spearman's rank correlation between predictions of existing fully automatic models and the empirical results.

	Our Approach	Khosla LaMem [33]	Baveye et al. [34]	Zarezadeh et al. [35]	Fajtl et al. [36]	Human subjects
ρ	0.52	0.63	0.64	0.62	**0.65**	0.75

6.2. Summary

In this chapter, we describe our efforts to develop a new fully automatic model for estimating image memorability, which benefits from a novel feature pooling strategy based on visual attention and a set of semantic features that encode meta-level object categories, scene attributes, and invoked feelings.

Our proposed feature pooling strategy is derived from the observation that main memorable areas of an image are the ones that attract the most attention. Different from the fixed pyramidal structure as in [4, 30, 32], our regression model learns memorability scores of images by additionally taking into account the features pooled over the saliency maps. In our pooling scheme, we employed two saliency maps, one obtained by a bottom-up saliency model [5]

and the other by a generic objectness model [83], respectively modeling bottom-up and top-down attentional influences on image memorability. Our experiments demonstrated that for the global dense features the combination of classical SP based pooling with the proposed pooling scheme improves the prediction quality.

Moreover, we investigate the use of three recently proposed semantic features, namely the Meta-class [48], the SUN Scene Attributes [7] and SentiBank [8] features, all of which can be automatically extracted from the images. These high-level features are used to describe the presence of certain abstract object categories, attributes related to functional, material, surface properties of scenes, and the emotions induced by the images as captured by the specific adjective-noun pairs. The inclusion of these semantic features into the computations greatly improves the prediction performance that we obtained superior results on the MIT Memorability dataset than those of the fully automatic pre-deep learning studies.

For highly memorable images, the existing approaches to predict image memorability can yield estimates close to the ground-truth scores from human subjects. However, their performances on determining whether an image is unmemorable is currently far from empirical scores. Even though our model provide the best results reported in the literature for predicting the memorability of less memorable images, it is not as accurate as desired. This opens up possibilities to design or learn new types of features to especially understand less memorable images.

7. Story Graphs with Intrinsic Properties

7.1. Memorable Story Graphs

As we explained in Section 6.1., after we achieved state-of-art performance on memorability prediction by utilizing visual attention mechanism, deep learning studies improved the performance by large margin. Because of this, in order to have better memorability scores we used the current state-of-art memorability prediction framework which is the work of Fajtl *et al*. [36] and is called *Attention based Memorability estimation Network (AMNet)*. Similar to our work, authors predicted the memorability scores of photos together with utilizing an attention mechanism. However, due to the neural network's improved prediction accuracy, they are able to achieve improved results. Compared with the other neural network based memorability prediction studies, they obtained the state-of-art visual memorability predictions. This shows that their work is on the same manner with ours in terms of attention-based memorability prediction and we can use their proposed prediction scheme for our experiments.

We calculated the memorability scores of photos of our YFCC100M-CITIES dataset (Section 5.2.) with AMNet model. In our story graph framework we described in Section 4., we extended the Coverage Equation (Equation 11) by adding a memorability factor as it is shown in Equation 18.

$$Coverage(\mathcal{S}) = \alpha \sum_{v \in \mathcal{V}} Coverage_{\mathcal{S}}(v) + (1-\alpha) \sum_{t \in \mathcal{T}} Coverage_{\mathcal{S}}(t) + Memorability(S) \tag{18}$$

Here, we calculated the $Memorability$ of the storyline S as the average memorability scores of each image in the story line. This way, when we are selecting story lines for the story graph by their incremental coverages as we described in Section 4.2.4., we encouraged the selection of more memorable images among high coverage story lines. We give sample

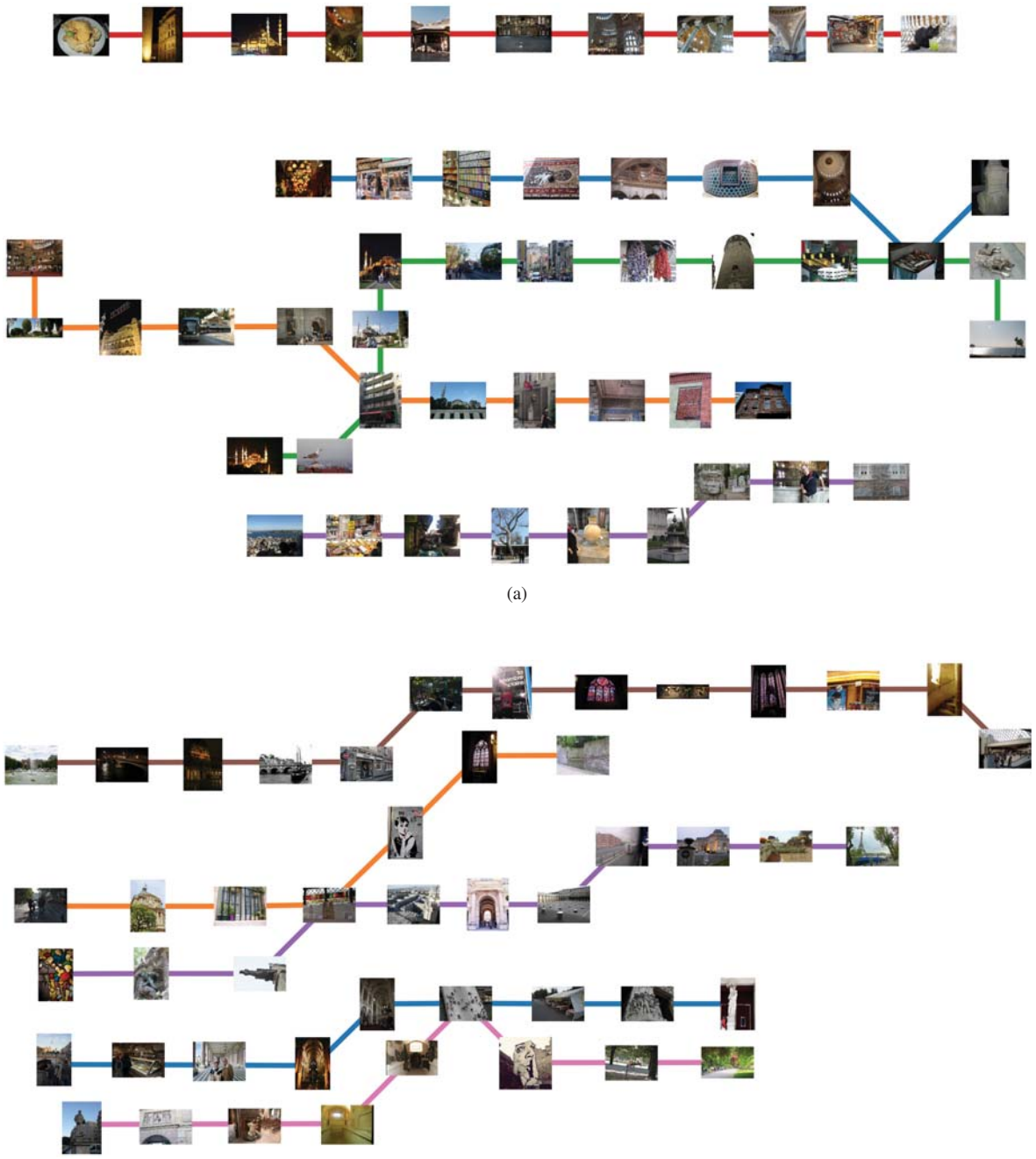
memorable story graphs of cities Istanbul and Paris on Figure 7.1.. When we compare with the story graphs without intrinsic properties (Figure 4.3.) because only the coverage part of the framework is slightly modified, we can see generally same photos are chosen. The most obvious difference is that there are less number of story lines on memorable story graphs. So together with the addition of memorable photos, less number of low memorable ones are elected for the story graph.

7.2. Aesthetic Story Graphs

Similar to the memorable story graphs, we used aesthetic scores of photos in our story graph generation framework and constructed aesthetic story graphs. In order to calculate aesthetic scores for all images of our YFCC100M-CITIES dataset we used the state-of-art visual aesthetic predictor model of Sheng *et al.* [44]. In their work they trained a deep model using aesthetic labels of Aesthetic Visual Analysis (AVA) dataset and utilized an attention mechanism to achieved state-of-art aesthetic prediction. We extracted the aesthetic scores of the photos of our dataset using their model. Then, similar to the memorable story graphs, we extended the Coverage Equation (Equation 11) by adding the aesthetics factor as shown in Equation 19. We give the aesthetic story graphs of cities Istanbul and Paris in Figure 7.2..

After the construction of the story graphs with intrinsic properties which are memorability and aesthetics, to have a solid quality measurements, we carried out the visual summarization experiments (Sec 5.3. on both memorable and aesthetic story graphs, which we will give the details on next section.

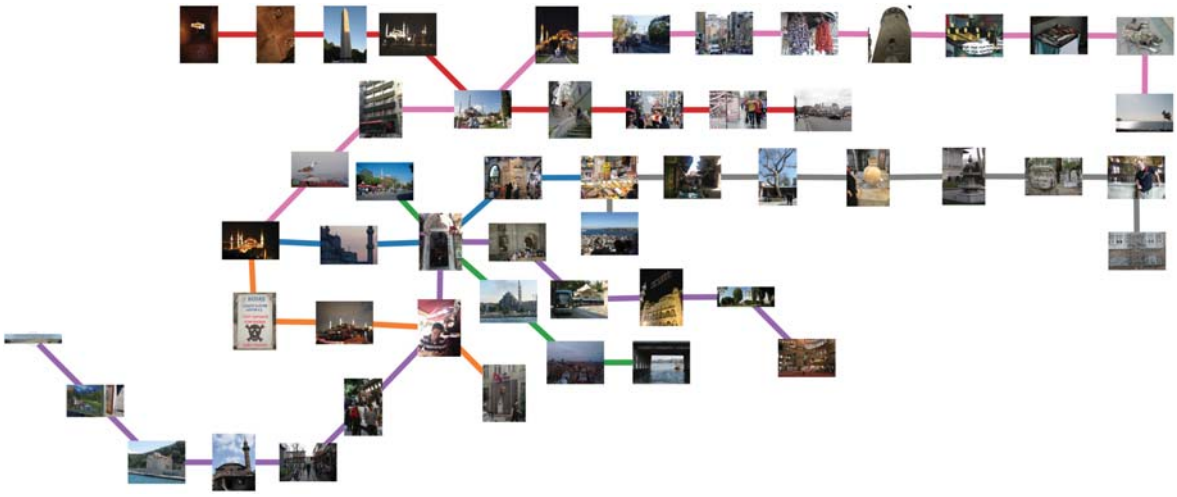
$$Coverage(\mathcal{S}) = \alpha \sum_{v \in \mathcal{V}} Coverage_{\mathcal{S}}(v) + (1-\alpha) \sum_{t \in \mathcal{T}} Coverage_{\mathcal{S}}(t) + Aesthetic(S) \quad (19)$$



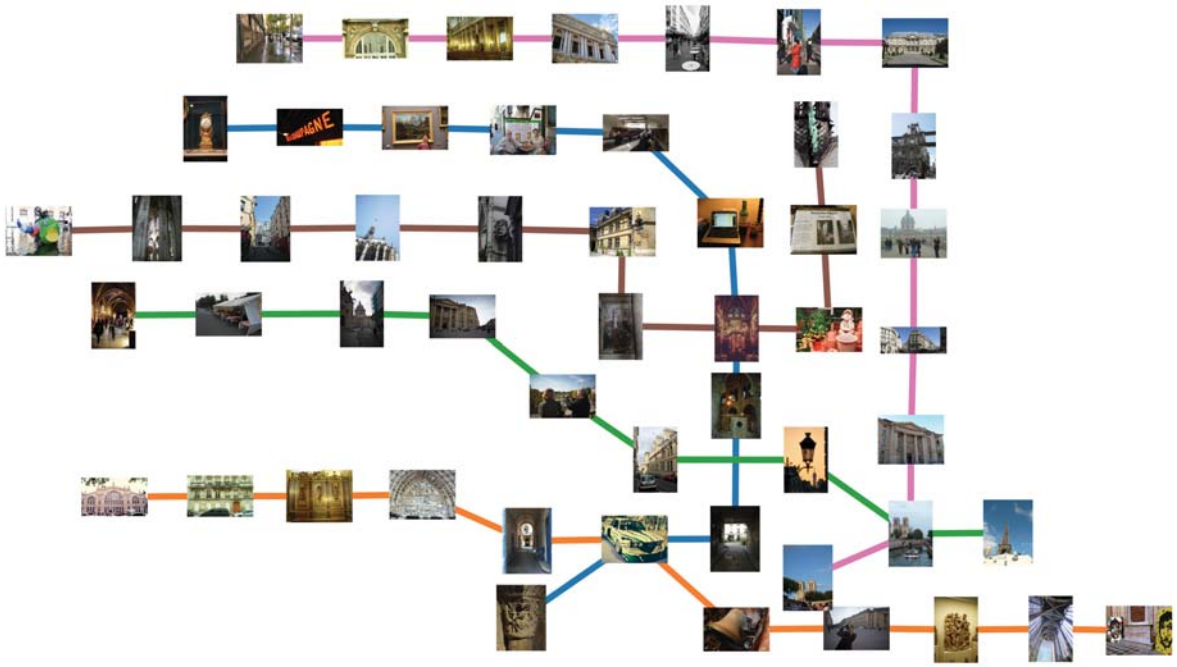
(a)

(b)

Figure 7.1. The memorable story graphs of (a) Istanbul and (b) Paris, which are based on travel photo albums collected from the web. The nodes (images) of the graphs are arranged based on the available timestamp information.



(a)



(b)

Figure 7.2. The aesthetic story graphs of (a) Istanbul and (b) Paris, which are based on travel photo albums collected from the web. The nodes (images) of the graphs are arranged based on the available timestamp information.

7.3. Summarization Experiments

After constructing memorable and aesthetic story graphs, we want to compare them with our previous story graphs without intrinsic properties. For this purpose and in order to make a

Table 7.1. V-ROUGE scores for the summarization experiments for aesthetic and memorable story graphs. $\mathbb{Y} = \mathbb{S}^{VGTM}$ denotes story graphs with the addition of memorability scores. $\mathbb{Y} = \mathbb{S}^{VGTA}$ denotes story graphs with the addition of aesthetic scores.

Photo Album	**DS3** ($\mathbb{Y} = \mathbb{S}^{VGT}$)	**DS3** ($\mathbb{Y} = \mathbb{S}^{VGTM}$)	**DS3** ($\mathbb{Y} = \mathbb{S}^{VGTA}$)
Amsterdam Trip	0.56	**0.57**	0.54
Istanbul Trip	**0.49**	0.47	0.48
New York Trip	0.67	0.63	**0.68**
Paris Trip	**0.56**	0.55	0.53
Tokyo Trip	**0.63**	0.62	0.60
Venice Trip	**0.66**	0.61	0.59

fair comparison and evaluation, we used the same summarization experiments we carried out in Section 5.3..

We give the V-Rouge and F-Measure scores of the summarization experiments in Table 7.1. and Table 7.2.. When we look at the V-Rouge scores, memorability and aesthetic properties seem to have a negative effect on story graphs on behalf of the summarization task. However, V-Rouge metric only measures the recall between human summaries and the ones we produced with our method. The F-Measure takes both recall and precision into account and provides a more stable metric against the outliers that the machine generated summaries may carry. For the 5 cities out of 6, aesthetic story graphs produced better summaries than the graphs without intrinsic properties. Similarly 4 cities out of 6 have better summarization scores on behalf of memorable story graphs. These results clearly indicates that intrinsic properties have positive impact on the construction of story graphs and increase the usefulness of the graphs for visual summarization tasks. We give the composed summary results of memorable and aesthetic story graphs together with the ones without intrinsic properties for comparison in Figures 7.3.-7.8..

7.4. Summary

In this chapter, we describe our efforts to extend our story graph generation framework by incorporating memorability and aesthetic properties of photos to generate memorable and

Table 7.2. F-Measure scores for the summarization experiments for aesthetic and memorable story graphs. $\mathbb{Y} = \mathbb{S}^{VGTM}$ denotes story graphs with the addition of memorability scores. $\mathbb{Y} = \mathbb{S}^{VGTA}$ denotes story graphs with the addition of aesthetic scores.

Photo Album	**DS3** ($\mathbb{Y} = \mathbb{S}^{VGT}$)	**DS3** ($\mathbb{Y} = \mathbb{S}^{VGTM}$)	**DS3** ($\mathbb{Y} = \mathbb{S}^{VGTA}$)
Amsterdam Trip	0.14	0.17	**0.20**
Istanbul Trip	0.10	0.16	**0.18**
New York Trip	**0.19**	0.18	0.17
Paris Trip	**0.21**	0.20	**0.21**
Tokyo Trip	0.11	**0.18**	**0.18**
Venice Trip	0.23	0.25	**0.29**

Figure 7.3. Summarization with aesthetic and memorable story graph results of city Istanbul. Top: Visual summaries using story graph constructed with visual, GPS and textual features. Middle: Visual summaries using memorable story graph. Bottom: Visual summaries using aesthetic story graph.

aesthetic story graphs. We created the story graphs by the extended method and carried out the same visual summarization experiments as we have described in Section 5.3.2.. We show that including the two intrinsic properties of photos which are memorability and aesthetics into story graphs positively effects the usability of the story graphs for visual summarization tasks. This further indicates that due to it's subjective nature, summarization encourages the utilization of personalization. Usage of more intrinsic properties of images together with interactivity with the user during the construction phase of story graphs opens opportunities for further research.

Figure 7.4. Summarization with aesthetic and memorable story graph results of city Amsterdam. Top: Visual summaries using story graph constructed with visual, GPS and textual features. Middle: Visual summaries using memorable story graph. Bottom: Visual summaries using aesthetic story graph.

Figure 7.5. Summarization with aesthetic and memorable story graph results of city Newyork. Top: Visual summaries using story graph constructed with visual, GPS and textual features. Middle: Visual summaries using memorable story graph. Bottom: Visual summaries using aesthetic story graph.

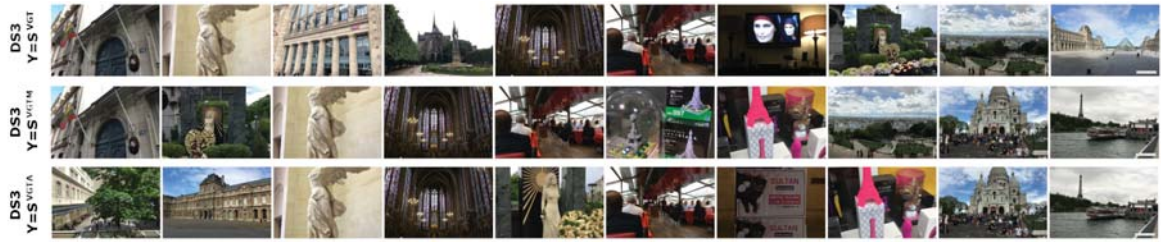


Figure 7.6. Summarization with aesthetic and memorable story graph results of city Paris. Top: Visual summaries using story graph constructed with visual, GPS and textual features. Middle: Visual summaries using memorable story graph. Bottom: Visual summaries using aesthetic story graph.

Figure 7.7. Summarization with aesthetic and memorable story graph results of city Tokyo. Top: Visual summaries using story graph constructed with visual, GPS and textual features. Middle: Visual summaries using memorable story graph. Bottom: Visual summaries using aesthetic story graph.

Figure 7.8. Summarization with aesthetic and memorable story graph results of city Venice. Top: Visual summaries using story graph constructed with visual, GPS and textual features. Middle: Visual summaries using memorable story graph. Bottom: Visual summaries using aesthetic story graph.

8. Conclusion and Further Directions

Today, handling and making use of the huge and continuously accumulating visual data is a challenging problem. Several approaches are developed and constructing informative visual story graphs is one of the effective but least explored one.

We believe that story graph is a useful and effective tool to capture and summarize the main concepts of a data collection. However constructing an ideal story graph is not an easy task and require detailed and carefully designed process.

In this work, we created an automated graph-based framework utilizing deep features of images to create visual story graphs. The automated approach makes it further convenient for updating the story graph as new data becomes available. The work grounds on the following statement:

> "We create visual information maps to enhance user experience over handling collaborative and massive photo collections."

Our main contributions in this work as follows:

- ⋆ **Formalizing the properties of a good story graph.** We identified and formalized the characteristics of a good story graph and combined them on an optimization function which determines the quality of the graph.

  We want the story lines in our story graph to have a theme and tells a story. So we formulated the *coherence* property around this intuition and used visual and textual elements as the connection points between photos. Second, we want our story graph to cover diverse and important concepts. *Coverage* property satisfies this feature by forcing to add visually and textually different photos from the data set to the story graph, encouraging diversity. Instead of having distinct story lines that are unrelated with each other, we want to identify common locations among different stories to capture more outstanding locations and to acquire the notion of a map.

- ⋆ **Collecting a new vacation data set.** Story graphs are best for summarizing a photo collection specific for a topic. Vacation to touristic cities around the world is a proper theme for this kind of problem. We collected a new photo set called "YFCC100M-CITIES" which consists of 132K photos from 6 touristic cities which are *Amsterdam, Istanbul, New York, Paris, Tokyo and Venice.*

- ⋆ **Utilizing the story graphs for visual summarization task.** The constructed story graphs captures important concepts from the data set. Because of this, they are convenient priors to be used on summarization tasks. We devised a visual summarization task and showed that they are capable to be used as a prior knowledge base for this kind of tasks.

- ⋆ **Implementation of user studies.** Due to the personal nature of vacation photos which are collected from multiple users, it is not easy to measure the quality of the story

graphs that are made up of those photos. We applied the general approach in the literature for this kind of problems which are user studies. We conducted two user studies ”next image prediction” and ”coverage measure” and compared with the similar approaches in the literature.

⋆ **Analyzing the effects of intrinsic properties of photos on story graphs.** As we previously mentioned, the personal nature of vacation photos points out that utilizing intrinsic and personal properties of photos intuitively affects the quality of story graphs. We incorporated aesthetics and memorability properties of photos in our story graph construction framework and evaluated the results. Additionally we proposed a novel unsupervised attention-driven memorability prediction scheme.

We first identified the formal properties of story graphs which are *coherence, coverage* and *connectivity*. Then after extracting the visual and textual elements from photos, we created a coherence graph based on the transitions over these elements. This graph is the structure where we analyze and extract the visually and semantically coherent short chain of photos. Grounding on a divide-and-conquer approach, by overlapping the coherent short chains, we obtained longer story lines which will form the routes of the story graph. Finally among those story lines we selected the ones that provide highest coverage of visual and textual elements together with connection points which emphasize prevailing locations. Some devised user studies showed that these story graphs are better than similar works in terms of coherency and coverage.

We devised a visual summarization experiment where the story graphs serve as a useful prior. For this purpose we built a novel travel dataset named YFC100M-CITIES consisting of six touristic locations among the world cities by querying the YFCC100M dataset. We created visual summaries with the story graphs we constructed and evaluated by comparing with baseline methods and a recent work utilizing neural networks. We showed that the story graphs are useful basis for this task by achieving higher scores than these methods in terms of V-Rouge and F-Measure metrics.

We worked on visual memorability which is an intrinsic property of photos. We proposed memorability prediction framework utilizing an attention-driven feature selection method. We achieved state-of-art results on memorability prediction showing that attention is closely related with visual memorability. Furthermore, we merged visual properties with semantic ones which are *meta-class, Scene Attributes* and *Sentibank*. These features further improved the memorability prediction accuracy showing that semantic attributes are also related with visual memorability.

Due to the subjective nature of story graphs, incorporating intrinsic features of photos to our story graph construction framework is promising. Thus, we integrated intrinsic image properties which are memorability and aesthetics into our story graph construction framework to improve the quality of story graphs. We constructed memorable and aesthetic story graphs. The results of the same visual summarization experiments on these story graphs showed that incorporating intrinsic properties further improves the quality of the story graphs.

Experimenting on a single dataset can be seen as an open section for this work. More experiments on similar diverse data sets different from YFCC100M-CITIES together with referencing works on similar subject would further evaluates the effectiveness of the story graphs.

For future directions we outline some promising directions:

- ⋆ As deep learning approaches achieving improved performance on many application areas, incorporating deep learning mechanisms into the framework would be a convenient approach.
- ⋆ It would also be interesting to include some kind of personalization to allow the users to enforce some preferences while constructing the story graphs. This opens a direction to generate personal story graphs where individual preferences are taken into account and can be utilized on diverse application areas.
- ⋆ We used hand-crafted formalization on *coherence, coverage* and *connectivity*. Further our optimization function is also a hand-crafted formula utilizing those story graph

features. Instead, *learning* an objective function further automatizes the construction of story graphs.

REFERENCES

[1] Gunhee Kim and Eric P Xing. Reconstructing storyline graphs for image recommendation from web community photos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3882–3889. **2014**.

[2] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *International Conference on Learning Representations (ICLR)*. **2016**.

[3] Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. Learning visual storylines with skipping recurrent neural networks. In *European Conference on Computer Vision (ECCV)*, pages 71–88. **2016**.

[4] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, **2011**.

[5] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):1–20, **2013**.

[6] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, pages 2429–2437. **2011**.

[7] G. Patterson, C. Xu, H. Su, and J. Hays. The SUN attribute database: Beyond categories for deeper scene understanding. *Int. J. Computer Vision*, 108(1–2):59–81, **2014**.

[8] 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**.

[9] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 623–632. **2010**.

[10] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1122–1130. **2012**.

[11] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721. **2013**.

[12] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Storygraphs: visualizing character interactions as a timeline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 827–834. **2014**.

[13] John C Platt, Michal Czerwinski, Brent Field, et al. PhotoTOC: Automatic clustering for browsing personal photographs. In *IEEE Joint International Conference on Information, Communications and Signal Processing (ICICS) and Pacific Rim Conference on Multimedia (PCM)*, pages 6–10. **2003**.

[14] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(3):269–288, **2005**.

[15] Tamara L Berg and Alexander C Berg. Finding iconic images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–8. **2009**.

[16] Gunhee Kim, Eric P Xing, and Antonio Torralba. Modeling and analysis of dynamic behaviors of web image collections. In *European Conference on Computer Vision (ECCV)*, pages 85–98. **2010**.

[17] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes Paris look like Paris? *ACM Transactions on Graphics*, 31(4), **2012**.

[18] Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. AverageExplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics*, 33(4):160, **2014**.

[19] Yanir Kleiman, Joel Lanir, Dov Danon, Yasmin Felberbaum, and Daniel Cohen-Or. Dynamicmaps: Similarity-based browsing through a massive set of images. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 995–1004. **2015**.

[20] Ian Simon, Noah Snavely, and Steven M Seitz. Scene summarization for online image collections. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. **2007**.

[21] Pere Obrador, Rodrigo De Oliveira, and Nuria Oliver. Supporting personal photo storytelling for social albums. In *ACM International Conference on Multimedia (ACM MM)*, pages 561–570. ACM, **2010**.

[22] Fereshteh Sadeghi, J Rafael Tena, Ali Farhadi, and Leonid Sigal. Learning to select and order vacation photographs. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 510–517. **2015**.

[23] 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**.

[24] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1336–1344. **2016**.

[25] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. Sequential prediction of social media popularity with deep temporal context networks. *arXiv preprint arXiv:1712.04443*, **2017**.

[26] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. " nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of International Conference on Multimedia Retrieval*, pages 385–391. **2014**.

[27] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 907–910. **2015**.

[28] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 1277–1286. **2018**.

[29] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, **2001**.

[30] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems*, pages 305–313. **2012**.

[31] J. Kim, S. Yoon, and V. Pavlovic. Relative spatial features for image memorability. In *ACM MM*, pages 761–764. **2013**.

[32] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 196–200. IEEE, **2013**.

[33] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398. **2015**.

[34] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction: the emotional bias.

In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 491–495. ACM, **2016**.

[35] Soodabeh Zarezadeh, Mehdi Rezaeian, and Mohammad Taghi Sadeghi. Image memorability prediction using deep features. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, pages 2176–2181. IEEE, **2017**.

[36] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6363–6372. **2018**.

[37] Shay Perera, Ayellet Tal, and Lihi Zelnik-Manor. Is image memorability prediction solved? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0. **2019**.

[38] Oleksii Sidorov. Changing the image memorability: From basic photo editing to gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0. **2019**.

[39] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer, **2006**.

[40] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE, **2006**.

[41] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998. **2015**.

[42] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 497–506. **2016**.

[43] Shuang Ma, Jing Liu, and Chang Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4535–4544. **2017**.

[44] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 879–886. **2018**.

[45] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, **2004**.

[46] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, **2005**.

[47] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, **2007**.

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

[49] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386. **2010**.

[50] 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**.

[51] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, **2007**.

[52] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, **2008**.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. **2016**.

[54] J. Xiao, J. Hayes, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. **2010**.

[55] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Trains of thought: Generating information maps. In *ACM International Conference on World Wide Web (WWW)*, pages 899–908. **2012**.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. **2014**.

[57] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367. **2010**.

[58] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, **1981**.

[59] Chandra Chekuri and Martin Pal. A recursive greedy algorithm for walks in directed graphs. In *IEEE Symposium on Foundations of Computer Science (FOCS)*. **2005**.

[60] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 52(2):64–73, **2016**.

[61] Goksu Erdogan, Bora Celikkale, Aykut Erdem, and Erkut Erdem. Summarizing personal image collections with intrinsic properties. pages 1225–1228, **2016**.

[62] Rishabh Iyer, Pratik Dubal, Kunal Dargan, Suraj Kothawade, Rohan Mahadev, and Vishal Kaushal. Vis-dss: An open-source toolkit for visual data selection and summarization. *arXiv preprint arXiv:1809.08846*, **2018**.

[63] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1413–1421. **2014**.

[64] Nancy Chinchor. Muc-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding*, pages 22–29. Association for Computational Linguistics, **1992**.

[65] M. C. Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5):509–522, **1976**.

[66] P. G. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5(2):195–200, **1994**.

[67] R. N. Shepard. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6:156–163, **1967**.

[68] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci. U.S.A.*, 105(38):14325–14329, **2008**.

[69] L. Standing. Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, 25:207–222, **1973**.

[70] Jeremy M. Wolfe. Visual memory: What do you know about what you saw? *Current Biology*, 8:R303–R304, **1998**.

[71] J. M. Wolfe, T. S. Horowitz, and K. O. Michod. Is visual attention required for robust picture memory? *Vision Research*, 47:955–964, **2007**.

[72] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1469–1482, **2014**.

[73] A. Hollingworth, C. C. Williams, and J. M. Henderson. To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review*, 8(4):761–768, **2001**.

[74] A. Hollingworth and J. M. Henderson. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):113–136, **2002**.

[75] M. A. Cohen, G. A. Alvarez, and K. Nakayama. Natural-scene perception requires attention. *Psychological Science*, 22:1165–1172, **2011**.

[76] K. Inoue and Y. Takeda. The role of attention in the contextual enhancement of visual memory for natural scenes. *Visual Cognition*, 20(1):94–107, **2012**.

[77] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. **2006**.

[78] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, pages 111–118. **2010**.

[79] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, **1998**.

[80] A. Borji. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, **2013**.

[81] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26, **2008**.

[82] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, **2008**.

[83] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, **2010**.

[84] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. **2009**.

[85] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, pages 1378–1386. **2010**.

[86] A. Bergamo, L. Torresani, and A. Fitzgibbon. PiCoDes: Learning a compact code for novel-category recognition. In *NIPS*, pages 2088–2096. **2010**.

[87] Robert Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, Publishers, **1980**.