# Assessment of photos in albums based on aesthetics and context

Dmitry Kuzovkin

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

# THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Informatique*

Par
## « Dmitry KUZOVKIN »

## « Assessment of photos in albums based on aesthetics and context »

**Thèse présentée et soutenue à Rennes, le 21.06.2019**
**Unité de recherche : IRISA Rennes et Technicolor R&I**

**Rapporteurs avant soutenance** :

| | |
|---|---|
| Henri NICOLAS | Professeur, Université Bordeaux, LaBRI Bordeaux |
| Nicolas BONNEEL | Chercheur CNRS, LIRIS Lyon |

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | Eric MARCHAND | Professeur, Univ. Rennes, IRISA Rennes |
| Examinateurs : | Frédéric DUFAUX | Directeur de recherche CNRS, L2S |
| | Karol MYSZKOWSKI | Directeur de recherche, Max-Planck-Institut für Informatik |
| | Tania POULI | Senior Scientist, Technicolor |
| Dir. de thèse : | Rémi COZOT | Maître de Conférences, Univ. Rennes, IRISA Rennes |
| | Olivier LE MEUR | Maître de Conférences, Univ. Rennes, IRISA Rennes |

**Invité(s)**

| | |
|---|---|
| Kadi BOUATOUCH | Professeur, Univ. Rennes, IRISA, Rennes |
| Jonathan KERVEC | Senior Scientist, Technicolor |

# ASSESSMENT OF PHOTOS IN ALBUMS BASED ON AESTHETICS AND CONTEXT

DMITRY KUZOVKIN

2019

THESIS ADVISORS
University of Rennes 1, IRISA: Rémi Cozot, Olivier Le Meur, Kadi Bouatouch
Technicolor: Tania Pouli, Jonathan Kervec

# ACKNOWLEDGMENTS

The journey of doing a PhD has been an incredible experience, an experience that has broadened my knowledge both in research and in life. The people that were around me during this journey made it very special.

I am deeply grateful to my thesis advisors, who were guiding me through the world of research, each in their unique way. Rémi Cozot, for inspiring many parts of this thesis. Our discussions about photography and thorough brainstorms were always producing new questions in my mind that kept me moving. Olivier Le Meur, for showing me how to combine curiosity and rationality in research. It is easy to get lost and diverged when there are so many paths to follow, but you helped me to be consistent. Tania Pouli, from whom I learned a whole lot, and who made this PhD possible, in many senses. I cannot imagine how much time you spent reviewing my works, but I know that it certainly made me a better researcher and writer. If you will find amusing the usage of articles (or brackets) on this page, let's consider that I did it for a comical effect. Jonathan Kervec, for always being a source of good mood. The joke of "you have one month" will certainly stay in my memories. Kadi Bouatouch, for the wisdom you generously shared, and for the support and encouragement during the manuscript writing. Your approach of explaining things is something that I will keep and convey to others.

During these years, I met fantastic friends and colleagues, who I would like to thank. First of all, our team of "the round table". Oriel, for our midnight strolls and your precious advice. Juan, for our cooking sessions and music nights for people without musical talents. Martin, for carbon nanotubes and all the fun stories we had. You guys are amazing. Fatma, for sharing the experience of surviving the start of a PhD. Jean, for the tea breaks and for going through those ski slopes together. Matthieu, for your love of people and videos about emus. Hadrien, for our language exchanges that made me believe in my French. Charlotte, for showing Bordeaux and for the infamous restaurant experiences that we can laugh about now. Liz, for making a great appearance at the end of this journey and for your voice messages about how to give a strong presentation.

I thank all people that I met at Technicolor, at the PERCEPT team of IRISA and at the SIROCCO team of Inria. Special thanks to our "La Carotte" reading group: you made me remember that the books are essential in my life. My heartfelt appreciation also goes to all my friends back in Russia: Natasha, Nastya, John, Lesin, Nika, and many others. I would like to thank our CIMET Master's and its people; that experience had a significant contribution in defining my research path. A special mention goes to the wonderful city of Rennes, which was a great place to do a PhD.

Finally, I wish to thank my family. My parents, Nikolay and Sofia: even when I had doubts about me, you never had any. My brother, Arseniy: we had fun no matter the circumstances. I am also very grateful to my grandparents and great-grandparents. Their collective efforts throughout many years brought me where I am now.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# RÉSUMÉ EN FRANÇAIS

La photographie a fait un bond en avant au cours des dernières années: d'une activité exercée principalement par des experts, elle est devenue un phénomène omniprésent, présent dans la vie de presque tout le monde. À l'ère moderne des smartphones avec un appareil photo et du stockage numérique, nous sommes en mesure de documenter chaque moment de notre vie, et les chiffres confirment cette tendance: environ 1 100 milliards de photos ont été capturées en 2016. [1]

La démocratisation de la photographie numérique a apporté des changements importants à notre approche de la photographie. N'étant plus limités par les contraintes imposées par le support de film, les utilisateurs ont tendance à accumuler de vastes collections de photos, dans lesquelles plusieurs prises de vues de la même scène sont effectuées, rendant difficile le choix des meilleures photos. Auparavant, une grande importance a été donnée à l'instant de la capture de la photo elle-même, lorsqu'un utilisateur effectue des évaluations différentes pour assurer la meilleure prise de photo possible. De nos jours, l'évaluation de la photo est souvent un processus "après réflexion", dans lequel la meilleure photo peut ensuite être sélectionnée parmi de nombreuses photos similaires, et peut en outre être améliorée par des techniques de post-traitement.

Les utilisateurs peuvent se soustraire à la responsabilité de prendre la meilleure photo lors de la prise de vue, mais bien que nous ayons la chance de pouvoir traiter des photos plus tard, de nouveaux défis apparaissent. L'augmentation du nombre de photos capturées entraîne une surcharge considérable en temps et en efforts pour organiser les collections de photos et sélectionner les images à conserver ou à rejeter. Par exemple, lors d'un voyage de vacances, un utilisateur typique crée des centaines, voire des milliers, de photos, dont beaucoup sont répétitives et redondantes. Avec de grandes quantités de stockage disponibles, de nombreux utilisateurs ne passent pas de temps à trier et organiser leurs photos, car cette tâche est très laborieuse. En accumulant plusieurs grands albums photo, les utilisateurs rencontrent des difficultés non seulement pour sélectionner les meilleures photos, mais également pour des actions plus élémentaires, telles que la navigation dans les collections et la monstration, ainsi que la récupération de photos spécifiques à partager avec d'autres utilisateurs. Il convient également de noter que des exigences similaires s'appliquent aux photographes professionnels: une partie essentielle de leur travail consiste en la sélection parmi un très grand nombre de clichés répétitifs.

Peut-on évaluer les grandes collections de photos capturées et sélectionner les meilleures photos efficacement avec l'aide des technologies modernes? Dans un scénario futur utopique, les albums de photos de chaque utilisateur sont automatiquement organisés et les points saillants correspondant aux meilleures photos sont sélection-

1 http://blog.infotrends.com/?p=21573

nés conformément aux préférences personnalisées de l'utilisateur. Cependant, l'état actuel des méthodes de calcul n'atteint pas cette performance personnalisée de manière robuste. De plus, la création de telles méthodes nécessiterait une grande quantité de données personnalisées correspondant à chaque utilisateur.

En l'absence d'informations personnalisées, peut-on aborder cette tâche en reformulant un problème? Lors du traitement automatique d'un album photo, nous ne disposons peut-être pas des données spécifiques à l'utilisateur, mais un album lui-même peut en effet contenir des informations utiles. Dans un album, la plupart des photos ne sont pas isolées, mais dans un contexte de toutes les autres photos dans le même album. Un tel contexte peut fournir des connaissances utiles à l'aide de diverses informations implicites. Des exemples de telles informations peuvent être la présence de photos similaires ou les caractéristiques des images à différents niveaux de regroupement: par exemple, les statistiques dérivées de toutes les photos au niveau de l'album pourraient fournir des indications sur les compétences et l'équipement du photographe.

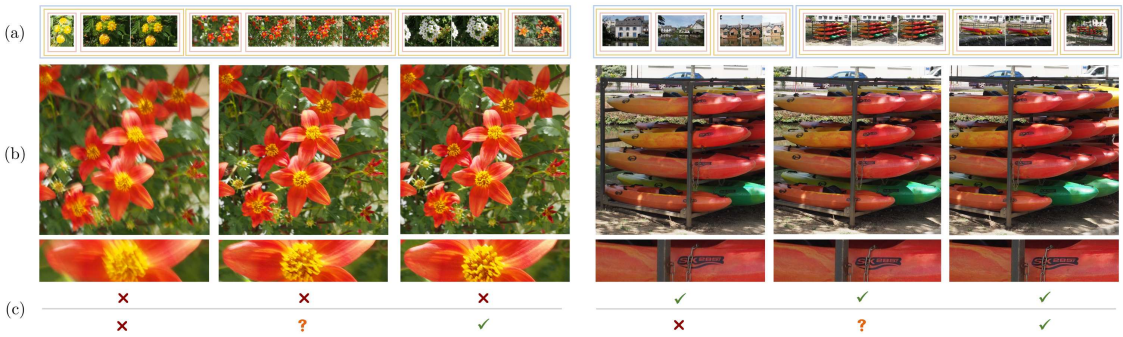
### OBJECTIFS ET MOTIVATIONS

Ces dernières années, un certain nombre de méthodes d'évaluation automatique des photos ont été proposées. Bien que certaines de ces propositions sont basées sur des mesures objectives de la qualité de l'image, d'autres utilisent des attributs subjectifs, ce qui représente la perception de la photo par un utilisateur. De plus, les solutions les plus récentes exploitent à la fois des caractéristiques objectives et subjectives, souvent intégrées à une approche d'apprentissage automatique.

Bien que certains progrès aient été réalisés dans ce domaine, le problème de l'évaluation de la photo et de l'esthétique des images reste ouvert. Une des raisons est que le raisonnement psychologique sous-jacent à la sélection de photos reste encore obscur et difficile à étudier. Des études psychophysiques supplémentaires et des analyses informatiques (telles que l'analyse de données massives) peuvent potentiellement aider cette tâche. De plus, parmi les domaines sous-explorés de l'esthétique et de l'évaluation de l'image, il y a un sujet d'évaluation personnalisée et contextuelle.

De nombreuses méthodes existantes abordent le problème de l'évaluation de la photo de manière indépendante: chaque image est considérée hors de son contexte, sans aucune information sur son origine. Néanmoins, chaque utilisateur a ses propres compétences en matière de création de photos, ainsi que ses préférences personnelles en matière de sélection de photos. De plus, nous évaluons et sélectionnons généralement les photos dans le contexte de la collection correspondante et en prenant en compte des photos similaires prises dans la même scène, ainsi que de la qualité moyenne des photos présentes dans tout l'album.

Bien que la modélisation des préférences personnelles de l'utilisateur reste une tâche très difficile qui nécessiterait une grande quantité de données personnalisées, la modélisation d'un contexte de photo à l'aide des informations disponibles sur les albums photo semble plus réalisable. Dans cette thèse, nous cherchons à approfondir le problème de l'évaluation photographique dépendant du contexte.




Lors de l'évaluation d'une photo, elle est généralement pas évalué de manière indépendante, mais dans le contexte d'autres images présentes dans une collection. Le contexte de la photo comprend les photos très similaires prises au même moment, des photos similaires de la même scène, ainsi que les photos de la collection complète. Dans cet exemple, nous montrons la sélection de photos en fonction de la netteté de l'image, réalisée indépendamment et adaptée au contexte de la photo. (a) Extrait d'une collection de photos et visualisation du contexte des photos. (b) Les photos prises dans la même scène et les gros plans des détails de l'image. (c) En haut: étiquetage d'image obtenu à partir d'une évaluation indépendante basée sur la netteté. En bas: étiquettes attribuées après l'adaptation contextuelle du score indépendant.

Plusieurs questions se posent sur une photo et son contexte. Quel est le contexte de la photo? Comment influence-t-il les décisions de sélection des utilisateurs? Peut-on modéliser un contexte informatiquement? Peut-on l'utiliser pour adapter une évaluation indépendante d'une image? Dans cette thèse, nous étudions comment ces questions ont été abordées plus tôt dans la littérature et nous aspirons à y apporter nos propres réponses, dans le but d'élargir les connaissances dans ce domaine.

Pour mieux comprendre les actions des utilisateurs lors de la sélection de photos au sein des albums et le niveau d'accord entre différents utilisateurs, nous menons trois études expérimentales avec les utilisateurs. Les deux premières études examinent la nature de la sélection d'images par les utilisateurs dans des albums photo. La troisième étude expérimentale explore une approche utilisateur du regroupement d'albums de photos. De plus, en combinant les résultats des expériences de regroupement et de sélection, nous pouvons tirer des conclusions utiles sur le contexte des collections de photos et son influence sur le processus de sélection.

Après avoir étudié la nature des décisions des utilisateurs, nous proposons une approche informatique pour modéliser le comportement des utilisateurs. Tout d'abord, nous présentons une solution de regroupement (clustering) hiérarchique pour les albums photo, qui fonctionne de manière adaptative en fonction du degré de similarité entre les photos au sein des albums. Cette approche de clustering peut être utilisée pour organiser une collection de photos et extraire le contexte de chaque photo défini par la présence d'autres photos similaires. Le contexte de la photo est utilisé dans la prochaine étape du modèle, où un score d'évaluation d'image, calculée de manière indépendante, est adaptée au contexte extrait. De cette façon, le score adapté fournit une meilleure

évaluation d'une photo dans son album photo correspondant. Dans le but d'adapter le contexte, nous proposons et comparons deux approches: une méthode simple basée sur des statistiques et une méthode basée sur une approche d'apprentissage automatique. Les performances des méthodes proposées sont estimées à l'aide des données utilisateur précédemment obtenues.

### CONTRIBUTIONS

Les contributions présentées dans cette thèse peuvent être résumées comme suit:

- Une étude expérimentale sur les décisions de sélection d'images dans des albums photo.
- Une étude expérimentale sur l'approche des utilisateurs pour le regroupement d'albums photo.
- Une nouvelle approche de regroupement (clustering) hiérarchique pour les collections de photos basée sur la similarité visuelle, avec une application dans l'extraction de contexte de photo et la navigation dans les collections.
- Une nouvelle approche pour l'adaptation de l'évaluation de la photo basée sur le contexte utilisant le regroupement calculé.

### STRUCTURE DE LA THÈSE

Le reste de la thèse est organisé de la manière suivante:

- **Chapitre 2** explore toute la durée de vie d'une photo numérique, du moment de sa création au moment de sa conservation, en passant par les aspects correspondants de l'évaluation de la photo. Nous donnons un aperçu des recherches menées dans ce domaine, en soulignant les principales contributions et en indiquant les principales limites des approches existantes. Dans ce chapitre, nous montrons également l'importance du contexte dans le processus d'évaluation.
- **Chapitre 3** présente les résultats de deux études expérimentales menées pour étudier la nature des décisions de sélection d'images prises par les utilisateurs au sein d'albums photo.
- **Chapitre 4** présente les résultats d'une autre étude expérimentale, qui explore une approche utilisateur du regroupement des albums photo. Cette étude nous permet de mieux comprendre le contexte dans des albums photo, et comment il affecte les sélections des utilisateurs.
- **Chapitre 5** propose de nouvelles méthodes pour le regroupement (clustering) d'albums photo. Nous introduisons une approche de clustering hiérarchique qui regroupe des photos similaires de manière complètement automatique. La performance du clustering est évaluée en utilisant les données acquises dans notre étude de l'utilisateur.

- **Chapitre 6** propose de nouvelles méthodes d'adaptation de l'évaluation de la photo utilisant le clustering calculée en tant que contexte de photo. Nous présentons une technique basée sur les statistiques et une technique basée sur l'apprentissage automatique, qui adaptent un score d'évaluation d'image pré-calculé au contexte de l'album photo. La performance d'adaptation est évaluée à l'aide des données de sélection acquises dans notre étude utilisateur.
- **Chapter 7** explore les caractéristiques esthétiques apprises par des méthodes basées sur les réseaux de neurones convolutionnels. Nous visualisons des caractéristiques profondes avec une technique de réduction de dimensionnalité, afin d'analyser l'influence de l'ensemble de données dans l'apprentissage en réseau. Nous étudions également le potentiel des réseaux génératifs et leur capacité à générer des échantillons d'images de différentes qualités esthétiques.
- **Chapitre 8** résume nos contributions, discute des difficultés rencontrées et de leurs limites et donne enfin un aperçu des perspectives de recherche future.

# OPENING

# 1

# INTRODUCTION

Photography has made a rapid jump over the last years: from an exclusive activity mostly done by experts it has become an ubiquitous phenomenon, which is present in almost everyone's life. In the modern era of smartphone cameras and digital storage, we are able to document each moment of our life, and the numbers indeed confirm this tendency: an estimated 1.1 trillion photos were captured in 2016 alone[1].

The democratization of digital photography has drastically changed our approach to the photography process. No longer limited by the constraints imposed by the film medium, users tend to accumulate large collections of photos, where often multiple repetitions of the same scene are captured, deferring the choice of the best photos to later. Earlier, a high importance was given to the instant of a photo capture itself, when a user performed different assessments to ensure the best possible shot was taken. Nowadays, the photo assessment is often an 'afterthought' process, where the best photo can be selected among numerous similar ones afterwards, and, furthermore, it can be additionally improved with post-processing techniques.

Users can now evade the responsibility of taking one best shot during the photo capture, but while we are fortunate to be able to deal with photos later, new challenges appear. The increase in the number of captured photos brings a significant overhead in the time and effort necessary for organizing one's photo collections and selecting which images to keep or reject. For instance, in a vacation trip, a typical user would create from hundreds to thousands of photos, many of which are repetitive and redundant. With large amounts of storage available, many users would not spend time sorting and organizing their photos, since it is a very laborious task. By accumulating multiple large photo albums, users face difficulties not only in the process of selecting the best photo, but in more basic actions, such as collection browsing and demonstration, as well as retrieval of specific photos to share with other users. Similar challenges apply to professional photographers, as essential part of their work consists in the selection among huge number of repetitive shots.

Can we assess large collections of captured photos and select the best ones efficiently with the assistance of modern technologies? In an utopian future scenario, the photo albums of each user are automatically organized, and the highlights corresponding to the best photos are selected, in accordance with personalized user preferences. However, the current state of computational methods does not achieve this personalized performance in a robust manner. Moreover, the creation of such methods would require large amount

1 http://blog.infotrends.com/?p=21573

of personalized subjective data corresponding to each user, which today are still hard to obtain.

In the absence of personalized information, can we approach this task by reformulating a problem? When dealing with a photo album, we might not have the user-specific data, but an album itself can contain useful information. In an album, photos do not exist in isolation, but rather in a context of all other photos in the same album. Such context can provide beneficial knowledge with the help of various implicit information. Examples of such information can be the presence of similar photos, or the characteristics of images on different levels of grouping: for instance, statistics derived from all photos on the album level could provide hints about a photographer's skills and equipment.

This idea of using the album context is the basis of the work conducted in this thesis. We study how the problem of the photo assessment is represented in the literature and how the context is addressed in recent methods. Then we conduct experimental studies on user behavior within photo albums, and, finally, we propose our own solutions that extract the context present within photo albums and adapt photo evaluations according to it.

## GOALS AND MOTIVATIONS

In recent years, a number of methods for automatic photo assessment have been proposed. While some of these proposals are based on objective image quality measures, others utilize subjective attributes, attempting to capture how users perceive a photo. In addition, the most recent solutions exploit both objective and subjective characteristics, often incorporated in a machine learning approach.

Although certain advances in this field have been made, the problem of photo assessment and image aesthetics remains open. One of the reasons is that the psychological reasoning behind photo selection still remains obscure and difficult to research. Both additional psychophysical studies and computational analysis (such as big data analysis) could potentially assist this task. In addition, among underexplored areas in image aesthetics and assessment, there is a topic of personalized and context-specific evaluation.

Numerous existing methods approach the problem of photo evaluation in an independent manner, where each image is considered out of its context, without any information regarding its origin. Nonetheless, each user has its own skills of photo creation, as well as personal preferences in photo selection. In addition, we typically assess and select photos in a context of its corresponding collection and in consideration of similar photos taken in the same scene, as well as average quality of photos present in the entire album.

While modeling personal user preferences is still a very challenging task that would require a large amount of personalized data, the modeling of a photo context using the available photo album information appears to be more feasible. In this thesis we research the problem of context-dependent photo assessment in more depth.

There are multiple questions that arise about a photo and its context. What is a context of the photo? How does it influence user selection decisions? Can we model a context computationally? Can we use it to adapt an independent score of an image? In this thesis we study how these questions were addressed earlier in the literature and we aspire to give our own answers to them, with an aim of extending knowledge in the field.

To better understand user actions within photo albums and the level of agreement between different users, we conduct three experimental user studies. The first and second studies investigate the nature of image selection by users in photo albums. The third experimental study explores how users approach the task of organizing or clustering the photos of a collection into coherent groups. Further, by combining the results of the clustering and selection experiments, we can draw useful conclusions about the context in photo collections and its influence on the selection process.

After studying the nature of user decisions, we propose a computational approach to model user behavior. First, we introduce a hierarchical clustering solution for photo albums, which performs in an adaptive manner depending on the degree of inter-photo similarity within albums. This clustering approach can be used to organize a photo collection and extract the context of each photo defined by the presence of other similar photos. The photo context is used in the next stage of the model, where an image score, which is computed in an independent manner, is adapted to the extracted context. This way, the adapted score provides a better assessment of a photo within its corresponding photo album. For the purpose of context adaptation, we propose and compare two approaches: a straightforward statistics-based method and a method based on machine learning. The performance of proposed methods is evaluated by comparing with the previously obtained user data.

Finally, we explore a different but complementary research direction. As convolutional neural networks (CNNs) attracted a lot of attention in recent years, they also found their application in the field of computational image assessment and aesthetics. Our score adaptation method also depends on the assessment results of such method, which can largely affect the scoring within photo albums. However, little consideration was given to the visualization of main characteristics learned by such CNNs. In the last part of our work, we study possible approaches to explore the 'aesthetic space' learned by CNN methods, and we also draw connections between the results of a recent state-of-the-art image assessment technique and the datasets used to train it. Furthermore, we explore the capabilities of generative adversarial networks (GANs) to generate image samples of different aesthetic quality.

## CONTRIBUTIONS

The contributions presented in this thesis can be summarized as follows:

- An experimental study on image selection decisions in photo albums, where we investigate how users select photos within albums based on different criteria and how much they agree in their decisions.


Figure 1.1: When assessing a photo, it is typically not assessed independently but in the context of other pictures present in a photo collection. The photo context includes the highly similar photos taken at the same moment (near-duplicates), similar photos from the same scene, and also the photos from the entire collection. In this example we show the photo selection based on image sharpness, performed independently and adapted to the photo context. (a) Extract from a photo collection and visualization of the photos' context. (b) Photos captured in the same scene and close-ups of image details. (c) Top: image labeling obtained from independent sharpness-based assessment. Bottom: labels assigned after the context-aware adaptation of the independent score.

- An experimental study on user approach to photo albums clustering, which investigates how users tend to group together similar photos in collections and what level of agreement is present between users. The obtained partitions are also used to study the influence of the context in photo selection decisions.
- A novel hierarchical similarity-based clustering approach for photo collections, which does not require input parameters and adapts to the varying granularity of inter-photo similarity. The proposed clustering solution can be applied in the photo context extraction and collection browsing.
- A novel approach for the context-based photo score adaptation using the computed clustering as the corresponding context. The proposed adaptation solution improves the scoring performance of independent photo assessment methods in photo albums.
- A research on the aesthetic space learned by CNN methods, its visualization, and the manipulation of aesthetic label using the conditional GAN.

## LIST OF PUBLICATIONS

Most of the described findings and proposals are published in the following proceedings of international conferences and journals:

- Kuzovkin, D., Pouli, T., Cozot, R., Le Meur, O., Kervec, J. and Bouatouch, K., 2017, July. Context-aware clustering and assessment of photo collections. In Proceedings of the symposium on Computational Aesthetics (p. 6). ACM. [80]

- Kuzovkin, D., Pouli, T., Cozot, R., Le Meur, O., Kervec, J. and Bouatouch, K., 2018, June. Image Selection in Photo Albums. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (pp. 397-404). ACM. [81]
- Kuzovkin, D., Pouli, T., Meur, O.L., Cozot, R., Kervec, J. and Bouatouch, K., 2019. Context in Photo Albums: Understanding and Modeling User Behavior in Clustering and Selection. ACM Transactions on Applied Perception (TAP), 16(2), p.11. [82]

### THESIS STRUCTURE

The rest of the thesis is organized in the following manner:

- **Chapter 2** explores the entire lifespan of a digital photo, from the moment of its creation to the moment of its preservation, along with the corresponding aspects involved in photo assessment. We give an overview of the research conducted in the field, where we highlight the major contributions and indicate main limitations of the existing approaches. In this chapter we also demonstrate the importance of the context in the evaluation process.
- **Chapter 3** presents results of two experimental studies that were conducted to investigate the nature of image selection decisions by users within photo albums.
- **Chapter 4** presents results of another experimental study, which explores a user approach to the photo albums clustering. This study provides us a better understanding of the context in photo albums, and how it affects users' selections.
- **Chapter 5** proposes new methods for the photo album clustering. We introduce a hierarchical clustering approach that groups similar photos together in a completely automatic manner. The clustering performance is evaluated using the clustering data acquired in our user study.
- **Chapter 6** proposes new methods for the photo score adaptation using the computed clustering as the photo context. We present a statistics-based technique and a machine learning based technique, which adapt a pre-computed independent image score to the photo album context. The adaptation performance is evaluated using the selection data acquired in our user study.
- **Chapter 7** explores the aesthetic characteristics learned by CNN-based methods. We visualize deep features with a help of dimensionality reduction technique, to analyze the influence of the training dataset. Also, we study the potential of generative adversarial networks and their capability to generate image samples of different aesthetic quality.
- **Chapter 8** summarizes our contributions, discusses encountered difficulties and limitations, and finally gives an overview of perspectives for future research.

## GLOSSARY

Throughout this thesis, we use different terms and notations in regards to photo assessment. There are no precise definitions of these terms in the literature, and some authors use certain terms interchangeably. Due to this, here we provide our definitions of the main concepts, which we adopt according to the definitions used in the majority of the literature. In the following chapters, these terms are used consistently, and their meaning can be consulted here.

IMAGE QUALITY defines a combination of technical factors that reflect how accurately the image captures the original scene, which often corresponds to the presence or absence of typical image artifacts. Some of the related image quality attributes: *sharpness*, *noise*, *dynamic range*. In some literature, the term of *Image Quality* also encompasses how pleasant and attractive an image for human observers. In this thesis, we limit the *Image Quality* to technical factors.

PHOTO AESTHETICS defines the overall attractiveness and beauty of a photo to viewers, which includes a number of subjective factors, such as photo composition, a subject of a photo, color and tonal characteristics of a photo, depth of field and so on. The *Photo Aesthetics* usually also includes objective characteristics of the *Image Quality*, and in this thesis the term of *Photo Aesthetics* comprises both subjective and objective factors of a photo.

PHOTO ASSESSMENT defines a process where a photo is evaluated by a human observer or a computational method and assigned a score (for example, from 1 to 10) or a categorical label (for example, *low* or *high* aesthetics). We also use the terms *Photo Evaluation*, *Image Assessment*, *Image Evaluation* interchangeably.

PHOTO ALBUM defines a set of photos related to a specific event (e.g. a family gathering, a travel, or a photo shoot). We use this term interchangeably with the *Photo Collection*.

INDEPENDENT PHOTO ASSESSMENT defines the assessment of a photo performed independently of the other photos from its original source, such as a photo album.

PHOTO CONTEXT defines a setting in which a photo is assessed. In our work, for a given photo the *Photo Context* is represented by a photo album of its origin and similar photos, which relate to the same captured moment.

# Part I

# NATURE OF PHOTO ASSESSMENT: LITERATURE REVIEW

Billions of photographs are taken every day, yet each of them goes a similar path, from the moment of photo creation to the moment of preservation in photo collections. In Part I we explore this subject through the prism of earlier conducted research, with a focus on photo assessment task: we highlight the major contributions in the field and indicate the points where the further research would be beneficial.

# 2

## THE LIFESPAN OF A DIGITAL PHOTO

In the modern world, the notion of photography largely corresponds to the digital photography domain. While traditional film photography has become a privileged activity of professionals and devoted enthusiasts, digital photography, on the contrary, is as widespread as ever. Nowadays, as digital photography is available to every person with a smartphone, the ubiquity of it has reflected on our every-day habits, changing some aspects of the photography process completely.

Every digital photo goes through similar stages, from the moment of its creation to the moment when it becomes a part of a larger photo collection: these principal stages are shown in Figure 2.1. Although actual presence and importance of each step can vary depending on a photograph's nature and a photographer's purpose, there is an essential component of the photo assessment, which is always explicitly or implicitly present in this process. In this chapter, we would like to explore the mentioned lifespan of a digital photo, in a direct connection with the subject of photo assessment and photo aesthetics.

The term of the photo assessment, as well as a corresponding term of photo quality, refer to different meanings in different photography applications. While in photo journalism the photo quality can refer to an informative quality of a photo, the most valuable photos in a family archive would be the ones that depict all family members, irrespective of the image quality. At the same time, the common way to assess a photograph is to assess the overall photo aesthetics. The photo aesthetics can be referred to a human ability to appreciate the beauty of a photograph. While this definition of aesthetics is broad and implies a great subjectivity in the task of photo aesthetics assessment, the notion of aesthetics is generally based on some implicit agreement between people that belong to a large homogeneous group. According to this notion, it is considered that certain visual properties can affect how aesthetically beautiful a given photo. The area of image and photo aesthetics is still actively explored by research in art, philosophy, psychology, as well as in computational research. As we will see, the aesthetics-based assessment can be based on different factors, ranging from objective image quality characteristics to subjective aspects based on human perception.

In what follows, we study the photo lifespan and a role of the aesthetics-based photo assessment in it, with the help of research previously conducted in the area, both to provide a deeper understanding of the subject and to highlight major contributions done in the field.

We divide this literature review into several main sections:

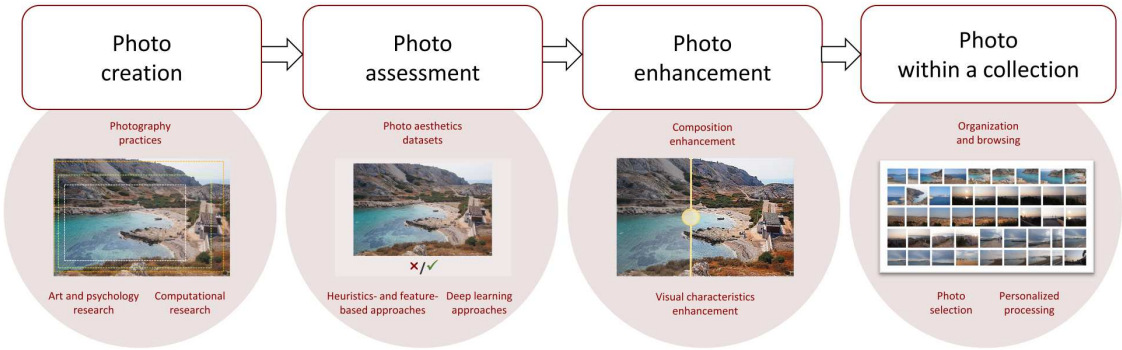



Figure 2.1: The lifespan of a digital photo. A typical photo passes through some (or all) of the shown stages. Along with these principal stages, we indicate relevant research directions.

- First, we study the moment of photo creation by reviewing common photography practices and the psychology behind the photo capture.
- Then, we discuss the works that approach the problem of computational aesthetics and address the task of automatic assessment for a taken photo.
- In conjunction with the works on photo assessment, we investigate the methods that aim to enhance some essential characteristics of a photograph.
- We study the next stage of photo processing, when a photo becomes a part of a photo collection and requires further high-level handling, which is then managed and organized.
- Finally, we indicate a gap between the stage of independent photo assessment and the stage of photo collection organization. We show that this gap is defined by the lack of context information within photo assessment, and we describe a few recently proposed methods that attempt to address this issue.

## 2.1 VISUAL AESTHETICS IN THE MOMENT OF PHOTO CREATION

Even in the era of digital photography, the instant of a photo capture remains a moment of high importance. This stage defines several important photo characteristics that cannot be modified in a post-processing. For many years, different recommended practices have been proposed in the books on photography, along with the research conducted on visual aesthetics in the field of psychology and arts.

The photography practices described in dedicated books reflect the common experience acquired since the invention of photography, as well as specific methods used by certain professional photographers. Some of these books provide a general overview of tips and advices that are applicable to any type of photography [30, 71]. Other photography books focus on specific aspects, such as achieving an ideal shot exposure [134] or setting up proper lighting conditions [60]. In addition, there are several books that focus on developing a creative vision in photography [38, 133], largely through applying different photo composition techniques.

To describe specific practices, we can address the book of Michael Freeman [38], which explains the author's scrupulous approach to photo framing. The book discusses the importance of the photographer's actions throughout the process of searching for a photo capture moment and gives examples of what comprises an ideal moment for the shot. A number of important photo characteristics and framing ideas are explained throughout the book. Here are some of these described characteristics (some of them are also illustrated in Figure 2.2):

- the framing has a direct influence on an image's impression: for instance, an interaction of objects created by framing can create either a dynamic tension or a harmonious balance;
- the placement of an object of interest helps to direct the viewer's attention: the frame filling can concentrate the focus entirely on the object, and the placement according to the golden section proportions (or the rule of thirds principle) can achieve a harmonious division of the image;
- guidance of viewer's attention can be also achieved through varying the perspective and depth-of-field, by means of using different viewpoints and lenses;
- the use of contrasting objects (such as contrasting in colors, textures, geometry, etc.) or rhythmic patterns can emphasize certain impressions;
- the employment of geometric shapes and lines created by objects in the scene can also stimulate a particular eye-path and provide an additional emphasis.

Overall, according to M. Freeman, the concept of an ideal photo is strongly connected to the frame composition and placement of objects within the shot. Photo composition is a crucial aspect, since the ability to correct the composition in the post-processing is often limited: most importantly, after the photo is taken, we cannot change the photographer's viewpoint or the placement of objects.

The importance of photo composition and image framing has been reflected throughout the history of the photography and it has also received a particular attention in the field of scientific psychology and in the research on art.

One of the most researched and debated topics in the field of visual aesthetics is the influence of the golden ratio (or the golden section). The concept of the golden ratio has been studied for many years in the field of art, however, up to this date, it is not confirmed whether the golden ratio has indeed a strong effect in aesthetics. A good overview of studies conducted in this field can be found in the work by Green [48].

In photography, the application of the golden ratio principle is often reflected in its simplified version, the rule of thirds. While research on the rule of thirds specifically is less extensive, there are certain studies that focus on its application in photography. For instance, a psychological study on aesthetic preference in spatial composition has been conducted by Gardner et al. [43]. In their experiments, users estimated a single object placement within a photographic frame, using a synthetically generated scene (see Figure 2.3). In tests with front-facing objects, a center bias was found, where the object placed at or near the center of the frame was preferred. In tests with left- or right-facing objects, the spatial asymmetry preference was found. Although partially


Figure 2.2: Photography composition templates suggested by M. Freeman [39]. The use of different composition approaches can guide a viewer and emphasize impressions. *Source:* M. Freeman.

confirming the rule of thirds principle, this preference was also characterized by a strong inward bias. That is to say, in a horizontal frame, objects facing towards the 'empty space' in a photograph's frame were preferred: as an example, an animal placed in the left part of the frame would look towards the right part. The authors also found similar patterns for vertical positioning of objects: for instance, a bowl facing upwards is more often placed in the bottom part of a picture. According to the results of Gardner et al., blindly following the rule of thirds might not lead to an optimal composition. Therefore, these factors should be taken into account during photo creation or when developing a computational modeling to imitate composition practices. Similar observations were done in the art research: specific frame points can define a perceived balance, and in this regard slightly off-center points are often preferred [129]. However, the importance of prior knowledge is also emphasized: knowing the facing direction of an object indeed influences the expected placement of it.

While photo composition is very important, as it is inherent to the moment of photo creation, a number of other factors also have a large influence on the aesthetic quality of a photo. Not surprisingly, multiple factors have been surfaced not only in research on photography, but in more general research on visual aesthetics and aesthetics in arts.

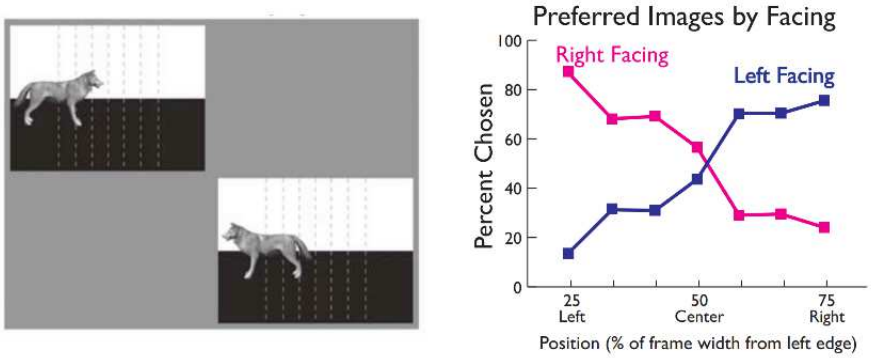



Figure 2.3: Objects placement and facing bias studied by Gardner et al. [43].
*Source:* Gardner et al.

In a recent study, Palmer et al. [129] provides a comprehensive overview of behavioral research on aesthetics in the visual domain. Beside the art-specific topics, they discuss the findings in the field of photography. The authors discuss certain color preferences and preferences for color combinations (often referred as color harmony). Although the concept of color harmony is sometimes debated, certain claims, such as preference for specific color pairs, are generally confirmed. The characteristics of spatial structures are also discussed: for example, the effect of spatial frequencies in paintings has been studied recently, to reveal that the power spectrum of paintings can be correlated with viewers' feelings of visual comfort or discomfort. At the same time, Palmer et al. indicate that the research on visual complexity and symmetry has not provided clear results: while generally people might prefer simple and symmetrical stimuli, there are large differences in these effects. Certain studies also have shown that people tend to prefer objects with curved contours. [129]

It can be noted that the aforementioned points are often related both to the moment of photo creation and to the successive moment of photo assessment and selection. That is, the composition rules that make an appealing photo are inevitably also considered when viewing it later on. In a next section, we turn our attention to the works that attempt to create a computational model to simulate human's behavioral response to the photos. Nevertheless, over the course of the entire chapter, we continue our discussion on the general concepts of photo aesthetics and photo assessment. In addition, we also revisit some of already mentioned visual aesthetics characteristics, by discussing the studies in computational aesthetics that brought an additional understanding of them.

## 2.2 COMPUTATIONAL AESTHETICS AND PHOTO ASSESSMENT

According to Palmer et al., the science of aesthetics is often doubted by some scientists, as "aesthetic response is subjective and whimsical" [129]. In response to this, Palmer and colleagues argue that while aesthetic responses are subjective, they can be studied objectively using behavioral methods. Certainly, we agree with this statement, and, additionally, we believe that the rise of interest to computational aesthetics has brought

a new dimension to such studies, along with appearance of novel experimental data and new approaches to the problem.

In 2005, the term of computational aesthetics was introduced by Hoenig [56]. According to him, "computational aesthetics is the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can". This aspiration — to estimate an aesthetic value of an artwork or a photograph, raises certain philosophical and ethical concerns — as one would expect [156]. However, the current state of computational techniques is still far from challenging art experts and critics, as well as from accurately modeling individual user preferences. Joshi et al. [67], in their survey on aesthetics, have noted that "the key challenges for researchers [in this field] are the loose and highly subjective nature of semantics associated with emotions and aesthetics, and the seemingly inherent semantic gap between low-level computable visual features and high-level human-oriented semantics". Although these challenges are still present, a certain progress has been already achieved since the introduction of this problem more than a decade ago, and the field continues to evolve.

### 2.2.1 *Photo Aesthetics Datasets*

The appearance of computational methods to model aesthetics of a given image first required an availability of experimental data to create and test such methods. The introduction of major datasets created for this purpose is inherently related to the proposals made in this field. Before proceeding to the methods description, we highlight the main public datasets that have been proposed up to this moment.

Two pioneering works in computational aesthetics have appeared in 2006: the work by Datta et al. [24] and the work by Ke et al. [70], both of which have proposed their own datasets. Datta et al. [24] have introduced a photo aesthetics dataset acquired from a peer-rated photo-sharing website *Photo.net*: the dataset is also often referenced under this name. This dataset initially consisted of 3,118 images with user-provided ratings from 1 to 7 (with at least ten ratings per image), and was later expanded to 20,728 images [22]. Ke et al. [70] have created a dataset, which is based on another peer-rated photography website *DPChallenge.com* — the dataset is often referenced as the *DPChallenge* or the *CUHK* dataset — where 12000 images were collected, with ratings from 1 to 10. Generally, the *DPChallenge* dataset is considered more challenging than the *Photo.net* dataset, as it contains a larger variety of photos.

A later dataset by Tang et al. [161], referenced as the *CUHK-PQ*, contains 17,690 images with a binary aesthetic score (of low or high aesthetics). The main novelty of this dataset was an introduction of seven semantic scene categories: landscape, plant, animal, night, human, static, architecture.

In 2012, Murray et al. have introduced the *AVA* dataset [122], which to this day remains one of the primary datasets in the field. The authors have claimed that the *Photo.net*, *DPChallenge* and *CUHK-PQ* datasets were not offering a difficult challenge anymore. More specifically, these datasets contain only images with a clear human consensus on their score, they do not provide a large variety of photos, and the per-image distributions

of scores are not available. Although the *AVA* dataset is also based on images retrieved from the photographers' social network *DPChallenge.com*, it is substantially different from previous proposals. The AVA dataset consists of 255,000 photos, with a rich annotation for each image that includes aesthetics scores distributions (where the scores range from 1 to 10, and the number of per-image votes ranges from 78 to 549), semantic labels and photographic styles. The presence of numerous ambiguous images with no consensus on their score have made this dataset noticeably more challenging than previously proposed ones. The total ratio of positive examples to negative examples is approximately 12:5.

The AVA dataset is still actively used in the research community, but some of its drawbacks should be taken into account. First, its origin in a photographers' peer-review website largely reflects on its content. A large number of photos are made by professionals or experienced amateurs and evaluated mostly by experts in the field, which leads to a certain bias for more professional-looking photos. Also, this is reflected in a fact that many photos are additionally post-processed. Second, a ratio of positive versus negative examples should be considered when, for example, creating a machine learning model.

Some of these drawbacks were addressed by Kong et al. in their recently introduced *AADB* dataset [75]. This dataset consists of 9,958 images, and it contains a more balanced distribution of professional and consumer photos. In addition to a score provided by annotators (5 per image), the annotators IDs are provided, along with eleven aesthetic attributes related to aesthetic judgments.

Despite these extensive efforts to create appropriate datasets for assessing image aesthetics, the concept of the photo context was often ignored until recently. The photos collected in the above-described datasets have no relation to each other and assessed in an independent manner. Thus, these datasets do not represent a typical consumer use case scenario, when photos inside photo collections are assessed. Some recent datasets have partially addressed this issue. Sadeghi et al. [143] have proposed a dataset based on Flickr vacation photo albums, acquired in five typical vacation and travelling topics. Their dataset consists of 63 albums (8,662 photos total), where for each album 4 annotators were asked to select five most representative photos. A similar problem of finding the most important photos for a specific event was explored by Wang et al. [175]. Their dataset is represented by 1883 Flickr albums from 23 event types. Each album was labeled by 3 annotators, where every photo received an importance score from four categories (very important, important, neutral, or irrelevant). The problem of photo aesthetics in context was more directly addressed by Chang et al. in their proposed *Photo Triage* dataset [15]. This dataset contains 5,953 series of photos taken in the same scene, with 2 to 8 images in each. The annotations consist of pairwise comparisons within series, provided by crowdsourced human decisions.

### 2.2.2 *Hand-crafted and Generic Features for Aesthetics Assessment*

Familiar with the background on photo aesthetics datasets, we can explore the history of the computational methods in this field itself. The first computational model for assessing photo aesthetics appeared in 2004, when Tong et al. [166] proposed a classification approach to distinguish between professional and amateur photographs. For this purpose, they adopted a set of low-level image features, which included contrast, colorfulness, saliency, blur measure, and other features related to texture and shape attributes.

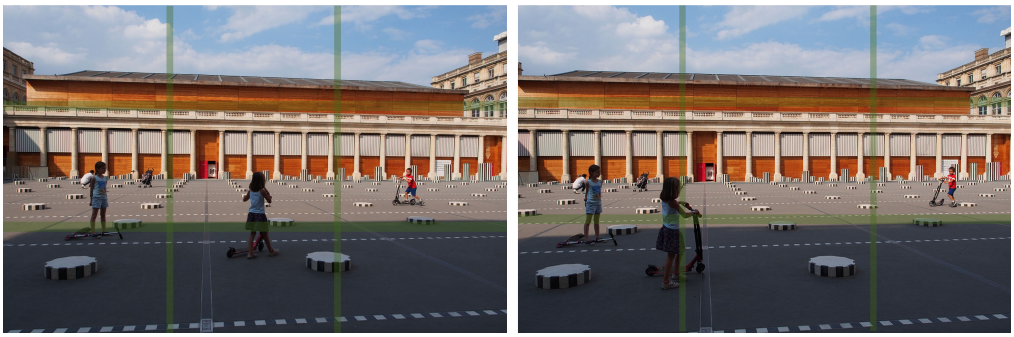
In 2006, two pioneering approaches that aimed to estimate image aesthetics were proposed by Datta et al. [24] and Ke et al. [70]. These two works introduced first significant contributions to the field. Along with introducing the first datasets on photo aesthetics, they proposed their own approach to compute image aesthetics, which have largely defined the common framework of numerous successive proposals. The general approach defined by them can be represented as follows: (1) a dataset of photographic images is collected, which consists of images with aesthetic labels or scores; (2) a set of image features is extracted from each photo; (3) the acquired features are used to train a classification or a regression system to predict an aesthetic assessment for unseen images.

Following this framework, Datta et al. [24] have collected their *Photo.net* dataset (as described above, in Section 2.2.1) and used it to train machine learning solutions. Given a photo as an input, they would compute a binary aesthetic classification, into *low* or *high* aesthetics, and a numerical aesthetic rating. Their approach was based on a number of hand-crafted visual features, which were employed to build a classifier using a support vector machine (SVM) and a linear regression model to predict a ranking score. Most of the features are based on the image data in the HSV color space, extracted from an entire image and different segmented regions. On their proposed Photo.net dataset, Datta et al. have achieved an average classification accuracy close to 70%. The research conducted in this work has later contributed to an online system that computes an automatic aesthetic rating for a given image, ACQUINE [23]. For completeness, we include the overview of features used in the Datta's work (some of the characteristics modeled by these features are illustrated in Figure 2.4):

- average pixel intensity across an image to characterize the use of light (exposure);
- colorfulness measure defined as a relation of color distribution to a hypothetical ideal distribution;
- average saturation;
- rule of thirds measure computed by area proportions for average hue, saturation and value (in HSV space);
- familiarity measure - how unusual an image is - based on similarity of color, texture and shape information to other images in dataset;
- texture feature - how smooth or grainy image is - computed from wavelet transform coefficients;


(a) Image exposure


(b) Image saturation



(c) Rule of thirds. In the left example the subjects of photos do not provide a clear composition, while in the right example one of the subjects is located according to the rule of thirds, on a so-called "power point"

(d) Depth of field. The photo on the right shows much shallower depth of field, which makes the subject stand out from the background.

(e) Simplicity/smoothness. The photo on the left has a more cluttered background with high frequency elements, while the photo on the right provides a simpler and cleaner look.

Figure 2.4: Examples of photo characteristics typically modeled by image features in photo assessment methods.

- aspect ratio of an image;
- region composition - a number of distinct regions of same properties - based on relative sizes of segmented regions;
- color harmony related features reflecting color spread around the color wheel and presence of complimentary colors;
- depth of field indicator, computed from a relation of wavelet coefficients corresponding to different image blocks;
- shape convexity feature - related to a notion that convex regular shapes produce a positive aesthetic response - computed from convex hull properties of image segments.

Another pioneering work by Ke et al. [70], which also appeared in 2006, has approached the task similarly. Using their acquired DPChallenge dataset, each photo was assigned a class of low or high aesthetics computed from an average photo rating. To discriminate between two classes of aesthetics, a naive Bayes classifier was trained on extracted photo features. Also largely inspired by the photographers' practices, the authors proposed to design high level semantic features, which were modeled using different image processing and computer vision techniques. The set of proposed features included:

- simplicity, captured by two features: the spatial distribution of the high frequency edges, and the hue count of a photo;
- color palette, modeled with an image color histogram and its relation to histograms of similar images;
- image blur level, computed from an estimation of smoothing parameter for a Gaussian filter modeling the blur;
- contrast and brightness features, assessed from image gray level histograms.

We have provided the detailed description of the works of Datta et al. [24] and Ke et al. [70] and the features employed in their proposals, as these works have largely influenced the subsequent research in image aesthetics. Later models followed a similar approach with a different selection of features. Nevertheless, in most cases the used features remained grounded in photography practices or research on visual arts. Some of these features were also underlined in a guideline proposal by Peters [132], where the importance of studying the aesthetics in the context of human-computer interaction was emphasized. Inspired by earlier works on art and psychology, they assumed that the aesthetic primitives defined by color, form, spatial organization, motion, depth and human body depiction define the human perception of computer images.

After the works of Datta et al. and Ke et al., several prominent methods based on handcrafted features have proceeded. Among these, the methods that base their assessment not on the whole image but on main subjects have received particular attention. These methods were generally employing similar selections of features, but varied in the manner of subject extraction. It was proposed to identify the subject region using a blur metric (based on the depth of field assumption) [107], using photo composition features defined by subjects' positions' and a visual weights [5], and using the graphlets

representing the most prominent objects [192]. In the latter approach of Zhang et al. [192], the photo ranking is trained on visual scan paths, which are assumed correlated with aesthetic scores.

Some approaches have added a notion of the content depicted in the image. For example, the global features, such as features based on color statistics and composition properties, can be complemented with the category-specific features relevant to the object or scene classes recognized in pictures [28, 106, 151]. Certain techniques focus only on the photos with people present and rely on features extracted from the facial regions, such as lighting, face expressions and poses [87, 140]. Instead of using typical image features, it was shown possible to assess a photo using only color harmony principles, where the color histograms from different local regions are assessed in a machine learning approach [127].

An appearance of different computer vision techniques has indeed influenced the field of computational aesthetics, with an introduction of generic image descriptors and bag-of-words methods. The attribute information from different detail levels can be accumulated using a multi-resolution technique and a bag-of-words approach, where the most discriminative features are then selected with the Adaboost technique [157]. Also, the assessment process can be simplified and achieve higher efficiency by using general statistics-based features [94].

A trend towards different feature extraction has resulted in the work of Marchesotti et al. [115], where they have proposed to use generic image descriptors, commonly used in image classification. They suggest that aggregation of local information from commonly used SIFT descriptors [99] into an image-level bag-of-visual-words or Fischer vector provides better implicit modelling and encoding of best photography practices. According to the shown results, proposed generic descriptors outperform previously proposed approaches by Datta [24] and Ke [70]. Authors observations indicate that SIFT descriptors can encode blur information and contrast between sharp foreground and out-of-focus background, while color descriptors are able to encode particular color combinations with higher aesthetics. The use of generic descriptors and their fusion with hand-crafted features was also explored in some later works [114, 117, 122].

The work of Marchesotti et al. [115] was one of the first to demonstrate that the heuristics-based features, defined by photography practices, might be not ideal, as they potentially can be replaced by generic image descriptors. At the same time, the entire field of computer vision, once defined by image descriptors, such as SIFT, SURF or HoG (still essentially hand-crafted), has been going through large changes recently. These changes have been brought with the arrival of convolutional neural networks (CNNs) and deep features. Nowadays, with larger generalization capabilities of CNNs, we are able to approach more complex vision problems. Following the general trend, the most recent methods for aesthetic assessment have also gradually abandoned the use of heuristics and hand-crafted features.

### 2.2.3 *Deep Learning For Aesthetics Assessment*

In 2012, the work of Krizhevsky et al. [78] has changed the field of computer vision, being the first work to confirm the potential of convolutional neural networks, with their proposal of the AlexNet network for image classification. The powerful representations learned from large-scale datasets have led to impressive generalization capabilities in various tasks, such as image classification, object segmentation and image retrieval.

CNNs are a category of neural networks: they are typically applied to image data, and consist of a large number of hidden layers of learnable weights and nonlinear transformations. The weighted network layers are analogous to a sequence of image filters with different convolution kernels, where these filters are learned with the use of training data. While in the task of image classification an input image is labeled with an object class, in the aesthetic classification an input image is labeled with an aesthetic label (e.g. binary label of *low* or *high* aesthetics).

The principal ways of adapting CNNs to different tasks can be generally divided into three categories: (1) training an entire network for a given task, using the available annotated task data; (2) re-training the last layers of an available CNN pre-trained on another task (also called transfer learning); (3) directly employing deep features computed with a CNN pre-trained on another task. The choice of an appropriate approach usually depends on the similarity of tasks and an availability of computational resources.

The first proposal to apply CNN architectures for the task of aesthetic classification can be considered the work by Lu et al. [101, 103]. One of the challenges in applying typical CNNs to the task of aesthetic assessment is a fixed-size constraint: all input images must be preliminary transformed to a pre-defined size with a square aspect ratio. These transformations inevitably lead to a loss of image details and photo composition, which might reflect on an image assessment. To cope with this challenge, Lu et al. have proposed a double-column architecture, where two identical CNN branches based on AlexNet [78] architecture are used (see Figure 2.5). The first column uses a global image view, and the second column uses one local image patch defining fine-grained details. In their experiments, the highest performance was achieved by using a warped image transformation as a global view, and a randomly extracted image patch as a local view. Additionally, the authors have incorporated the style attribute information using an extra CNN column, which has further improved the overall performance.

The problem of preserving both local and global information was further investigated in later research. First approaches to this problem were based on different patch combination techniques, where deep features were extracted from different regions separately and then used in a separate classifier or on a deeper CNN layer. In this case, the region extraction can be done in a spatial pyramid pooling manner [29] or by randomly sampling patches in an image [102]. Nevertheless, these approaches did not bring a complete solution for the multi-scale feature extraction without damaging image spatial composition, which appear to be crucial in photo assessment. Mai et al. [112] have decided to address this issue with a special network design: while using a

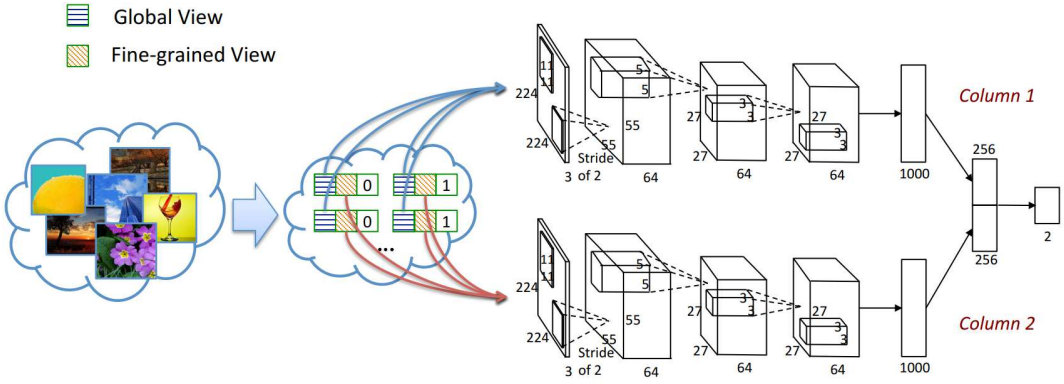



Figure 2.5: The convolutional neural network for aesthetic classification proposed by Lu et al. [101]. In their double column architecture, an input image is represented by its global and local views. *Source:* Lu et al.

VGG network architecture [152] as a base, they have added an adaptive spatial pooling layer introduced before the fully connected layers. This layer performs the pooling operation similarly to the conventional pooling layer, but it adjusts the size of receptive field according to the input dimension and always produces output of a defined size for subsequent layers, requiring fixed size data. Thus, even other CNN architectures can be modified in the same manner, to accept arbitrary-size input images.

Ma et al. [109] have further extended previous ideas [101, 103, 112] by proposing an adaptive layout-aware multi-patch network. To preserve the location information of patches, they utilize a graph approach that captures the layout of salient objects in an image. By training a CNN that combines this attribute graph modeling together with an adaptive patch selection technique, the authors achieve 82.5% accuracy and 0.92 F-Measure on the *AVA* dataset. Similarly to them, Sheng et al. [149] have proposed an attention-based approach, where the importance of extracted patches and corresponding features is weighted according with predicted attention. Their method achieves 83.03% accuracy on the *AVA* dataset. The methods by Ma et al. [109] and by Sheng et al. [149] currently can be considered as the state of the art approaches in photo assessment.

Some authors have focused on creating content-adaptive CNNs that are trained and respond differently depending on image semantics. Wang et al. [172] have modified the AlexNet, by adding a special group of scene layers into the architecture, to represent seven scene categories. These corresponding layers were pre-trained to initialize the model and improve aesthetic classification performance. Another proposal mentioned earlier, the composition-preserving deep network by Mai et al. [112], was also augmented with a scene-categorization sub-network, to improve the overall accuracy.

While these scene-aware CNNs were trained using only one common loss directly associated with an aesthetic label, some authors have explored a multi-task learning approach. For example, it was shown possible to learn the task of aesthetic quality assessment jointly with the task of semantic recognition, where two loss functions are employed at the same time [69]. Some later works have concluded that that two losses are difficult to calibrate simultaneously, and it is simpler to perform training in two

stages [121]. In this case, a fine-tuning is performed first to predict a scene semantic category, and then the aesthetic predictor is trained. The latter approach by Murray et al. [121] also attempts to predict an aesthetic score distribution from multiple raters (such training data is available in the *AVA* dataset). Another deep learning method to predict a distribution of scores instead of one score was proposed by Jin et al. [65]. In their technique, they have introduced a fine-tuning scheme with sample weights to better adapt to images from a wide range of possible origin and aesthetic quality.

Another notable work was done by Kong et al. [75]. Their work has introduced a novel dataset presenting a better balance between professional and consumer photographs (*AADB* dataset). Their joint learning architecture incorporates a Siamese network fine-tuned on pairs of images, combined with an attribute- and scene- adaptive networks. The estimation of specific photographic attributes (11 in the dataset) within their architecture serves as a regularizer for the final score prediction. At the same time, the prediction of photography content category and scene provides a further refinement of the score.

While we mentioned earlier that the terms of the technical image quality and aesthetic attractiveness are usually distinguished from each other, it can be argued that the majority of existing aesthetic assessment methods also incorporate image quality computation. In earlier approaches we could notice that some quality-specific hand-crafted features (such as, for example, blurriness or exposure) were explicitly taken into account. In more recent deep learning based methods, the notion of quality is often learned implicitly, from available data examples. A recent image assessment framework, NIMA (Neural Image Assessment) [160], proposed by Talebi et al., is designed to estimate both image quality and aesthetic attractiveness of an image. Thus, it is proposed to use both for assessment of images in aesthetic datasets, such as *AVA* dataset, or to rank identical images affected by different quality distortions. Similarly to earlier discussed proposals [65, 121], NIMA is trained to predict the distribution of human scores for an input image, rather than a fixed ranking score.

### 2.2.4 *Which Attributes Affect User Decisions?*

Along with the introduction of computational techniques, researchers began to wonder whether these methods can help to determine which image features in fact influence our decisions. Marchesotti et al. have approached this problem by employing the style attributes tags of the *AVA* dataset in a learning procedure [114, 115]. This way, the learned visual attributes are associated with specific tags, and the user preferences can be used to discover the most discriminative attributes. Their proposed method not only predicts aesthetics quality, but also suggests a set of image annotations as an explanation of what makes a given image attractive or not. Recently, this idea was pushed further by Chang et al. [16], where instead of an aesthetic score they generate an aesthetics captioning, which contains a comprehensive photo critique in a sentence-based form (examples are shown in Figure 2.6).

Several recent studies explored the influence of different geometrical properties within an image. Among other works, the influence of shapes on observers' perceptions [100],


Figure 2.6: Examples of photo critique captions generated by the approach of Chang et al. [16]. Instead of describing depicted objects, their approach provides comprehensive sentences explaining good and bad points of a photo. *Source:* Chang et al.

the global geometrical patterns in the landscape genre [124], and the effect of the image display scale [18] were explored.

A more global approach to the problem was taken in the work of Redi et al. [141], where the stylistic differences in photography between different countries were analyzed, based on a large collection of photos from the Instagram platform. Another analysis on typical web images has revealed the necessity of content-specific aesthetic models [139].

A recent introduction of CNN-based techniques opens a promising path for further investigation of features that affect our perception. Although it is often argued that deep learning models provide low interpretability, certain advancements in this direction have been made recently. Notably, CNN attention visualization was achieved using special back-propagation techniques. In the work of Murray et al. [121], a technique of adversarial examples is employed, where an image is gradually modified to achieve a higher or lower aesthetic score. By visualizing the modified regions with heatmaps, it is possible to observe which photo regions contribute to the aesthetic score the most. A global average pooling approach was proposed by Malu et al. [113], where the pooling layer provides attribute-specific activation maps, which highlight the most important regions for each aesthetic attribute.

### 2.2.5 *Other Relevant Works*

Although in our work we focus on the subject of photo aesthetics, it is also worth mentioning other relevant works, which study visual data perception.

It appears natural to extend image aesthetic assessment to video content. One of early works proposed an extension of their subject-focused photo assessment technique to video, by tracking the detected subjects throughout frames and extracting corresponding features [107]. These features combine aesthetics with video-specific characteristics based on motion vectors. A similar approach can be found in later works, where features

are combined on microshot level from multiple frames [120] or using a multi-level approach to represent different granularity of temporal structure in videos [6]. Some approaches utilize more advanced motion feature extraction, which combine optical flow and saliency computation [188]. Aesthetic assessment was also employed to select video thumbnails [155]: after a consecutive rejection of near-duplicate and low quality frames, a set of candidate frames is selected, where the frames with the highest representation and aesthetics scores are kept. A related concept of video interestingness was also investigated, using a mixture of visual and audio features [64].

A concept that is sometimes investigated in connection with aesthetics is an emotional response to an image. The work of Machajdik et al. [111] is based on psychology and art theory concepts and attempts to extract features that represent the emotional content of an image. By using a set of hand-crafted features similar to the ones commonly used in computational aesthetics — e.g. color statistics, color harmony principles, texture-based features and composition-related attributes — they attempt to classify a photograph or a painting into an emotional category. The task of emotional inference was also approached with CNN by Peng et al. [131]: they directly retrain the AlexNet on the AVA dataset for different abstract tasks, such as emotion classification, artist classification and artistic style classification, memorability prediction and interestingness prediction. They also showed that concatenating CNN features learned from different tasks can often enhance the performance in each task.

The concepts of memorability and interestingness have also received considerable attention with the rise of computational techniques. The term of memorability is linked with the likelihood that a user will recognize the same photograph after a certain time delay [63, 73]. The term of interestingness is linked with the ability of a certain image or video to draw attention of a user to its content and keep this attention for an extent of time [25, 51]. Although the methodology used to approach these problems is similar, such methods are primarily concerned with evaluating the effect a new, unseen image might have to a user.

FURTHER READING For a reader interested in the field of computational aesthetics, we recommend the following as a further reading: a profound research survey on aesthetics and emotions in images by Joshi et al. [67], a philosophic view on the problem of computational aesthetics by Spratt et al. [156], an experimental survey on existing datasets and assessment methods by Deng et al. [26], and another overview of existing computational techniques by Brachmann et al. [9].

## 2.3 ENHANCEMENT OF ESSENTIAL PHOTO CHARACTERISTICS

So far, we have discussed the stages of photo creation and photo aesthetics assessment. Certain photo characteristics are inherent to the moment of photo capture, and even minor mistakes in them might be impossible to fix, such as an incorrect scene exposure or a wrong person pose. At the same time, some characteristics, including the ones that affect subjective photo assessment process, can be corrected to a certain degree. In this

section we examine some of the techniques proposed for photo enhancement. It should be noted that photo enhancement is a vast subject by itself, and describing every aspect of it would be out of scope of this manuscript. Instead, we give a general overview of the field, with a focus on techniques that improve some essential characteristics related to the photo creation. Specifically, we pay particular attention to photo composition enhancement methods, since photo composition and framing has been previously shown to be important in the perception of a photo.

### 2.3.1 *Photo Composition Enhancement*

As we mentioned earlier, it is agreed both by professional photographers and researchers that a photograph's composition is an important factor in aesthetic quality. A well-composed photograph is usually a product of a number of rules and best practices combined together, often in an intuitive manner. While in previous sections we described the works that attempt to assess a photo as a whole, a number of works have focused only on photo composition assessment and possible approaches for its enhancement. They aim to create a computational model that could summarize these practices, rules and intuitions: this way, we would be able to perform automatic recomposition of an already captured photo, or offer live guidance while capturing the photo itself.

Composition enhancement is also referred to as photo recomposition, which might include image cropping and, sometimes, image retargeting. The cropping procedure extracts an image subregion, while the retargeting procedure adjusts the relative locations of objects in the image.

Until recently, the majority of works approached this task by first extracting main subjects in a photo (or any objects or regions of interest) and then searching for their best updated location, according to a set of hand-crafted rules and metrics. To extract the subjects, the image saliency map computation is used most commonly, sometimes aided by complementary computer vision segmentation techniques. The very first computational methods were focused on optimizing subject content area, while preferring the crop cuts passing through homogeneous regions, and allowing extra room around subjects [105, 146]. Later proposals have kept an optimization approach to maximize an overall aesthetic score, however, inspired by photography practices, they have computed additional features from foreground objects, such as rule of thirds, visual weight ratio and dominant lines [5, 93]. Liu et al. [93], apart from the traditional cropping, have also introduced a retargeting technique, which allows the change of relative locations between objects, to further improve the score.

Certain methods took a more advanced approach to analyze the overall structure and objects dependencies, by using graph structures. It was shown that graph cuts can be used to optimize the dependence between foreground objects and background regions and produce a recomposed photo using a crop-and-retarget technique [191]. A cropping problem can be also formulated as a graphlet-to-graphlet matching, where image segments are modeled as a set of graphlet regions, and an optimal graphlet for the cropping transformation is searched, using a Bayesian network [192].

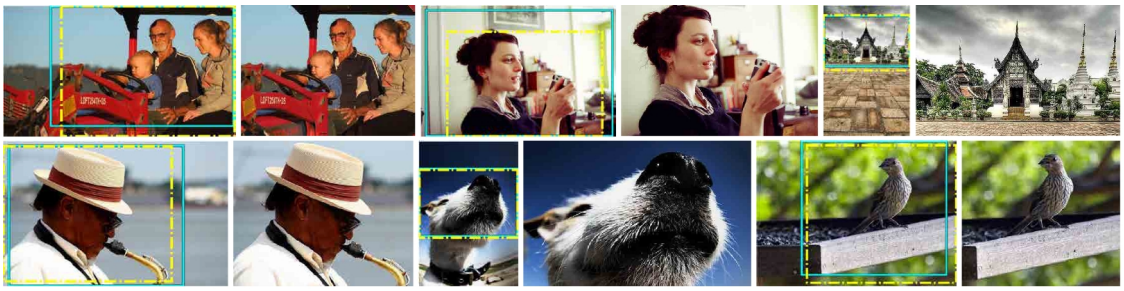


Figure 2.7: Examples of automatic cropping suggestions generated by the approach of Wei et al. [177] *Source:* Wei et al.

Some non-trivial solutions are also worth mentioning. Instead of focusing on the aesthetics of a final photo, Yan et al. proposed to consider the influence of removed and changed parts [183]. Samii et al. proposed to search for a well-composed example, semantically similar to a given image, and minimize a composition distance between two photos [144].

In the described class of approaches, the composition estimation is generally based on main subjects features computation and their interaction, rather than direct evaluation of an entire cropping proposal as an image candidate. However, such approach was not very common before the introduction of convolutional neural networks, apart from a few proposals (e.g. work by Nishiyama et al. [126], where an aesthetics classifier is applied directly on possible candidate crops). Nowadays, due to the nature of CNNs and their inherent object-attention properties, it is more straightforward (and often more beneficial) to directly calculate an aesthetic or a composition score of an entire crop candidate, without recourse to hand-crafted features and objects extraction.

Certain CNN-based methods for aesthetics assessment have proposed a straightforward adaptation of aesthetic scoring for automatic cropping, via a sliding window score computation [112]. Some authors have kept a regions-of-interest based approach: a predicted attention region can be used to generate a set of crop candidates around it, where the best candidate is selected using an aesthetics-based selection architecture [173]. Another approach to generate fewer candidate windows is to use reinforcement learning with an actor-critic architecture, which produces a reward leading to a higher aesthetics score [88]. Possible image transformations are defined within an action space, where an LSTM unit is employed to memorize previous states.

To help the research in photo recomposition, view pair datasets were proposed, consisting of pairs of manually scored crops of same images. Chen et al. [17] have created a dataset of 34,130 comparative view pairs (derived from 3,413 different images). Wei et al. [177] have introduced a dataset with over one million comparative view pairs (derived from 10,800 different images). In addition, Wei et al. have proposed an image cropping technique, where a view proposal network is combined with a view evaluation net, which acts in a pairwise comparison manner using a Siamese architecture as a base. Examples of automatic cropping suggestions by their method are shown in Figure 2.7.

The automatic recomposition methods that we have described so far allow us to enhance composition of an already captured photograph. However, similar research was conducted to propose a guidance system that could provide composition suggestions to user in real time, by analyzing an actual world scene. Although technically such approaches relate to the stage of photo capture, the underlying principles share numerous similarities with the discussed post-capture cropping techniques.

Certain proposals focus on providing guiding examples from a professional dataset, which can be relevant or similar to a current shot. Such composition and aesthetics feedback approach has been demonstrated for generic photos [186] or for specific genres, such as portrait photography [34].

Real-time composition guidance with viewfinder feedback can be found in a number of proposals, too. While some approaches tend to rely on a total aesthetic score from a standard set of handcrafted features [95], others apply more sophisticated composition-specific models [125, 158]. For instance, in the system by Su et al. [158], an optimal sub-view is estimated from a large scene view obtained with a use of wide-angle lens or panorama techniques.

Real-time techniques have also found applications in specific tasks. Aesthetic models for guiding a selfie portrait capture were created, using annotated synthetic datasets of a human 3D model positions [33, 90]. Recommendation systems for human positioning within a landmark-centric photos were proposed, based on datasets of well-composed photos [174, 182]. A viewpoint selection for architecture photography for famous landmarks was proposed in recent solutions, based on 2D images and 3D models [53] and on geographical location combined with crowdsourced data [190]. Finally, it was shown that an assisted photography framework can find its application to help visually impaired people in better aiming the camera at an object of interest [169].

To conclude with the topic of photo recomposition, we also address an interesting work on photo cropping from the psychology field. Different patterns of behavior in photo cropping were investigated by McManus et al. [118]. In their study they arrived to the following conclusions. Color appears to not significantly influence the cropping process, as users tend to crop color and black-and-white images in a similar manner. Experts and non-experts approach the cropping task differently, where experts take longer time and estimate a wide range of crops possibilities. Notably, the users who create better crops, are not necessarily experts, and experts' crops are not preferred more either by experts or non-experts.

### 2.3.2 *Visual Characteristics Enhancement*

Here we give an overview of techniques that aim to improve the overall visual perception of a photograph. In this section we do not aim to review different artifact-removal techniques, such as noise reduction or haze removal. Instead, we focus on methods that can increase an aesthetic value of a given image by a complementing transformation, such as by using color or tone enhancement methods.

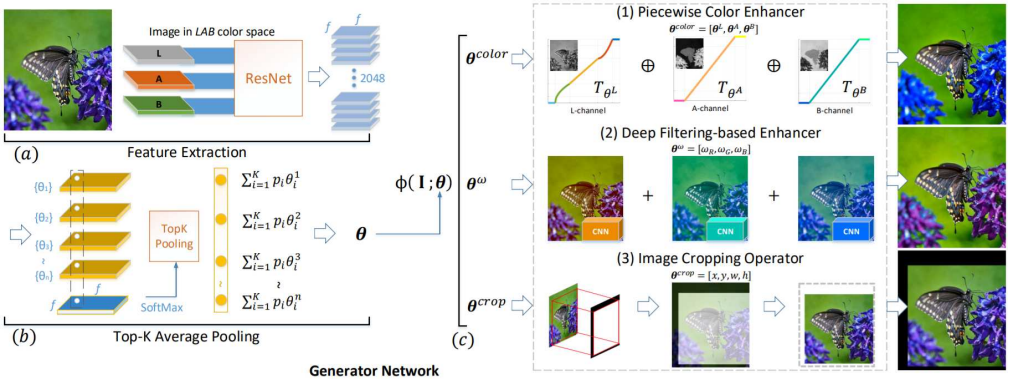



Figure 2.8: The architecture of the image enhancement method of Deng et al. [27]. Their method learns piecewise color enhancement, along with the image filtering and image cropping, using deep features and weak supervision (binary labels of image aesthetics). *Source:* Deng et al.

Frequently, a color enhancement task is formulated as the color transfer problem, where the desirable color or tone style is mapped from a supplied exemplar image [35]. For example, an input photo can be transformed to look as a given professional-looking photo [3]. As such color transfer methods are often too restrictive in the amount of user control and their results are largely defined by the choice of an exemplar image, some improvements were proposed. In some solutions, a set of different stylized outputs is computed and then suggested to a user, so he can perform further selection [84, 159].

Recent methods tend to avoid the exemplar-based approach altogether and try to automatically reproduce a photographer's actions. Yan et al. [184] attempts to imitate a typical photographer's workflow, where one proceeds in a sequence of intermediate steps during the photo processing (such as, for example, a saturation decrease followed by a contrast increase, followed by a further saturation decrease). Instead of simulating a step-by-step procedure, some approaches learn how to estimate commonly used transformations. In the work of Bychkovsky et al. [10], an extensive dataset of photo pairs before and after processing by different professional retouchers was created. It was further used to learn a tone style adjustment model, which could be adapted to a particular user with a subset of training photos. A later proposal by Deng et al. [27] employs a CNN GAN strategy of adversarial learning, where, instead of an image output directly, the trained generator model searches for best image transformation parameters (see Figure 2.8). In addition to piecewise color enhancements, their architecture produces general filtering suggestions and an optimal cropping.

Instead of learning parameters of some pre-defined transformation functions, it is also possible to learn an approximation of the unknown function itself, via deep networks. It was shown by Yan et al. [185] that the photo enhancement can be formulated with nonlinear transformation functions learned with CNNs in a pixelwise manner. To learn these transformations, a pixel feature descriptor is used as an input to the network, which is defined by pixelwise features, contextual features from a surrounding region and global image features. The most recent approaches demonstrate a direct image-

to-image transformation learning. A network architecture proposed by Ignatov et al. [61] learns a function to imitate DSLR quality from mobile photos, by improving color rendition and image sharpness. While this method required aligned example pairs of original and enhanced photos, the later proposal from the same authors has alleviated this issue by using an improved GAN-inspired architecture [62].

## 2.4 PHOTOS WITHIN PHOTO COLLECTIONS

Up till now, we have considered a photo as an independent entity, which is created, processed and assessed with no connection to other possibly related photos. At the same time, for a typical photographer (either amateur or professional) a photo rarely exists out of context. Commonly, a photo is a part of a larger collection unified by a topic or location, where other similar photos are present, and certain inter-relations between them can be found. Hereby, we discuss a role of a photo within photo collections or photo albums (we use these terms interchangeably).

A large part of photo treatment within a collection includes collection management and organization. Back in 2002, at the beginning of the digital photography era, people were not confident what the digital realm will bring to them, but they already expected that it would change their social activities and ease the remote photo sharing with relatives and friends [41]. A difference between traditional film photographers and new ones - active users of social networks - has become apparent especially in the aspect of photo sharing. For instance, in contrast with film photographers, users of social platforms are more eager to share even family photographs outside of their close circles [119]. Furthermore, the introduction of cameraphones has increased the frequency of photo acquisitions: people have incorporated a smartphone's camera into multiple every-day activities [168].

Since digital photography has become more widespread, people have sensed the shift towards a different way of dealing with photos: photo collections have started to increase in size, contain numerous similar pictures, and, overall, have become more complex and difficult to organize [74]. Studies on family habits in photo sharing and organizational strategies have indicated a problem of important photos retrieval, which is consequent to the difficulty of maintaining an organizational structure within personal photo albums [147, 179].

With the growth of their personal photo collections users have also started to make decisions about what photos they want to keep. Different factors affecting users' decisions were revealed: while some studies have shown that a personal importance is crucial in long-term photo preservation [12], other studies have also demonstrated an importance of photo aesthetics and quality [180]. It was also shown that users follow some organizational patterns in their decisions: they might follow a reduction strategy to remove unnecessary near-duplicates first, to reduce the decision space, and then they would follow a coverage-based strategy, to keep the most representative photos, justified by memories or some subjective reasons [123]. It was also shown that although a task of selecting photos within personal albums is more enjoyable and simple (in contrast with

unseen non-personal albums), users spend more time on selection in their own albums [170].

Mentioned studies have indicated the need for numerous assistance techniques in different subtasks of collection management and organization. Among others, these subtasks include: splitting a big photo collection into smaller events, arranging an album for easier browsing and retrieval, as well as album summarization and selection of the best photos.

The problem of photo album organization is often formulated as clustering a photo album into a set of sub-events. To find related photos, different information can be used. The first proposed approaches were generally based on temporal information (for example, photo timestamps extracted from EXIF metadata) [21, 136], often complemented with simple content-based information, such as color histograms [96, 135]. Modern cameras, especially cameraphones, can provide additional information, for instance geo-locations from GPS sensors, which can be combined with temporal information to create a certain hierarchy of events [11]. While mentioned techniques allow only for fine-granularity segmentation of a photo album into sub-events united by some temporal or visual information, more recent proposals approach this task as a more general problem of event segmentation, often with the aid of Hidden Markov models based on a combination of features [8, 46].

While it is important to find a way to organize photos, the process of collection browsing can also be aided with an appropriate visual interface. Although in our everyday activities we use a conventional image browser interface, alternative interfaces have been proposed. Directories or metadata-related images can be represented with data blobs filling the screen space, which might be useful to work with large personal collections [4]. A quadrant layout scheme can be applied for faster retrieval of related pictures [79]. Also, another rectangle-based technique, combined with a multi-depth approach was demonstrated [45]: an entire collection is clustered using image similarity or topic-based correspondence, and then represented as a set of representative photos. Each photo can be activated to reveal cluster contents, also packed into rectangular blocks. This approach is more applicable for content-based retrieval in large collections of mixed content.

In relation with organization and browsing strategies, Gozali et al. [47] have studied how people organize event-based photos. By testing different browser layouts and event organizations, they have concluded that users value the chronological order of sub-events, while order within these sub-events is not always important. Also, flexible non-conventional layouts have not introduced an additional value for users, in most cases. Overall, the sub-events based organization can help in typical tasks, such as storytelling or a search for a particular photo.

Although organizational techniques can be useful to deal with large photo archives, users also require an ability to quickly summarize a collection in a set of photos, which could be especially useful to share with other users. In this sense, the concepts of the best photos selection and the photo album summarization are inter-related, where the

latter is sometimes used to define a selection of the most representative or informative photos.

Some proposals have focused on retrieving a coverage-based summary, which would represent key photos (not specifically of high quality): in these approaches, the main focus was on temporal events representation, as well as faces representation and photos uniqueness [89, 154]. Other approaches have combined event-clustering coverage modeling with a rejection of low quality images [96, 153, 167]. Different approaches to deal with near-duplicate photos were also proposed [15, 19, 66]. It was also shown that the selection of photos can be efficiently predicted based on the eye gaze data [170].

Certain summarization methods go a step further and propose to generate new content out of representative and related photos, such as photo collages. Such methods range from a single-collage proposal systems that utilize auto-cropping and enhancement techniques together with a user input [181] to more complex automatic authoring solutions to create complete printable photobooks [55, 137].

It is also possible to introduce the notion of a storyline into the summarization task, creating a personal photo-based story, which might be not even constrained by time and representation criteria. Such approach has been shown on vacation photographs, where a story is formed from 5 to 10 photos [143]. In this case, the importance of a resulting sub-album is defined by photos' aesthetic appeal and representation of main scenes and people, which are computed with a use of different visual features. A more sophisticated approach, inspired from the storytelling structures in dramaturgy and cinematography (such as *acts, scenes, shots* and *characters*), was also recently proposed [143]. Elaborating through these event structures with a use of temporal and visual similarity clustering, the proposed solution employs face-centric aesthetics metrics to produce a consistent story. The heavy use of face-related computations is motivated by a primary application for social networks albums. Some methods focus on learning a particular order of images in photo albums, which, for example, can be used to model a typical "Paris vacation" experience, reflecting an order in which monuments are visited, or a typical ski trip, as a more general example [150]. Recently, it was also shown possible to create a story out of randomly permuted images and their captions [1].

## 2.5 A CONTEXT GAP IN PHOTO ASSESSMENT

The field of computational aesthetics and independent image assessment in particular have received considerable attention during recent years, as we have seen in Section 2.2. However, a typical photo aesthetics model usually reflects the learned preferences of an 'average user'; such models can be applied in a number of applications, but they might be not suited to assess photos in personal collections. While the existing aesthetic assessment methods try to predict how attractive is one image on an abstract independent scale, the actual evaluation procession largely depends on the context of a photo.

In fact, each photo in a collection exists in a particular context, which can be defined on multiple levels. First, the photo exists in the immediate context of highly similar

photos taken at the approximately same moment. Then, the photo exists in a larger context of generally similar photos taken in the same place or scene. Also, as we already mentioned, the photo belongs to a collection, which itself can convey certain characteristics, e.g. related to shooting conditions or equipment. Finally, the highest context level, on which the photo is perceived and assessed, is the individual user experience: each user has his own judgments and subjective preferences.

Can we model this context of individual preferences? And if this is impossible in the absence of personalized data, can we learn something useful from the lower level photo context of similar images and the collection itself? Since an assessed photo in a personal collection always exists in the presence of other related photos, can these photos be used to model an assessment context, to increase the evaluation performance?

In this section, we discuss these questions, which constitute the problem of the context gap in photo assessment – the problem that limits the effective application of the computational aesthetics techniques for photo collections. Although until now this field was not largely studied, a few important proposals indicating possible research directions have been made (some additional works on dealing with the semantic and the context gap in photo collections can be also found in a survey by Sandhaus and Boll [145]).

### 2.5.1 *Personalized Photo Assessment*

A fully personalized photo assessment can be considered as an ultimate goal for computational aesthetics. Indeed, a computational model that could perform photo album organization and the best photos selection according to the taste of a user or a group of users would be an ideal solution in the most cases. However, even recent research has not achieved this state yet. The need of a huge quantity of personalized data is one of the main obstacles. At the same time, we can expect that the data accumulated in different social networks can give a strong boost to this research in the future. Some recent studies implicitly confirm these assumptions.

After collecting information from Flickr on multiple users' "likes", Lovato et al. [97] were able to identify individual users based on these preferences, even from a limited set of image "likes". In a further study, they also claimed that a personal aesthetic taste can be used as a behavioral biometrical trait, specific to a person, and even adopted in forensic technologies [98]. In another study by Guntuku et al. [50] mappings between textual tags and deep visual features were learned, which were further used to train a general model able to predict images that a user might like, and, vice-versa, a user who would like particular images. Although these studies raise some concerns regarding user privacy, they also confirm the potential of creating user-specific image preference models.

One of the first attempts to introduce a personalized aspect into photo ranking was the work of Yeh et al. [187]: in their approach, a general aesthetics scoring model is learned, where the influence of each extracted feature (related to color, texture properties, composition rules, etc.) can be weighted by a user's settings, either adjusted manually or

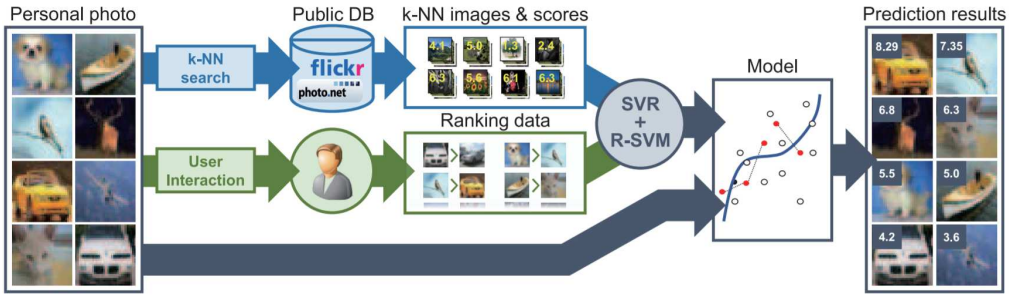



Figure 2.9: A framework of personalized aesthetic assessment by Park et al. [130]. Given a set of photos, the similar photos with the known scores are found in social photo platforms. Additionally, user performs ranking of a small image subset. By combining these sources of data, a joint regression model is learnt, which is able to predict scores for user photos. *Source:* Park et al.

defined with a selected photo example from a database. Recently, with an introduction of deep features, new proposals have appeared. Two recent methods, by Park et al. [130] and Ren et al. [142] approach this task similarly: first, a general model is learned on a large public dataset, and then a small set of images annotated by a user is used to compensate the generic aesthetics score with a personalized input. Both techniques use deep features and the SVM-based models to perform such adaptation. The framework proposed by Park et al. is illustrated in Figure 2.9.

### 2.5.2 *Context-aware Photo Assessment*

In the absence of individual data, useful photo context characteristics can be determined by considering each photo within the context of related similar photos. Although this idea also applies to the photo collections assessment, it is not exclusive to it.

Some works in the field of image aesthetics assessment have followed the assumption that a photo should be assessed in a group of related photos. Apart from the already discussed works that employed content-specific features [28, 106, 151], other works are specifically aimed at the relative photo assessment. For example, Tian et al. [164] have assumed that similar images should share similar assessment rules and proposed a query-dependent model: for a given image, a specific training subset of similar images is retrieved (with known labels) and an SVM model is trained on this refined subset, using deep features. While not purely context-aware, this approach stands out from a typical independent assessment framework.

Another approach would be to use relative features that reflect the characteristics specific to the analyzed group of images. When dealing with traditional hand-crafted features, it was proposed to use the difference-based relative features [189]. The most prominent solutions that apply such approach are found among the CNN-based techniques. For instance, instead of a typical architecture trained on direct pairs "image versus score/label", it is possible to learn a ranking model on pairwise comparisons. In one of the first proposals, Lv et al. [108] has demonstrated such approach by using

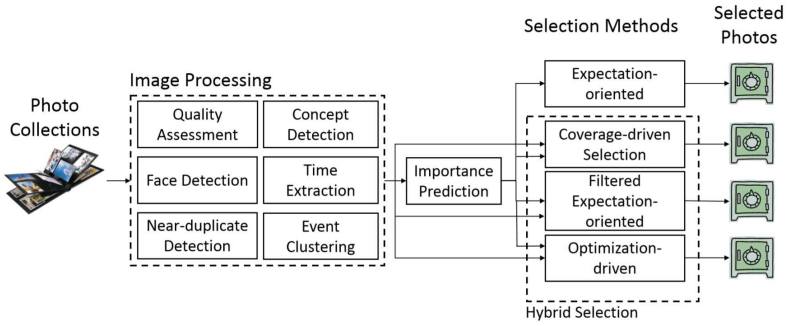



Figure 2.10: A photo selection approach for photo collections by Ceroni et al. [13]. A number of features is extracted from each image and the entire collection, and then used to train a selection solution. *Source:* Ceroni et al.

deep features in the SVM ranking model. Later, researchers adopted a Siamese network principle, to train pairwise models directly [15, 75]. Some recent approaches have also experimented with learning relative encodings within triplets of images [148]. While we can expect that such models might be more suitable for the assessment of similar content, their performance is still influenced by the training data used: commonly, the pairwise training is performed within large collections of unrelated photos.

The Siamese network in the method of Chang et al. [15] was trained specifically to perform assessment within photo series taken in the same scene. They detect the series from 2 to 8 similar images, using the SIFT descriptors, and then perform a relative ranking using the machine learning solution. In their work, they have experimented with different features (including hand-crafted features combined with deep features), but the highest performance was achieved by using the end-to-end training with the VGG Siamese network. Although their proposed method is well adapted for the task of best photos selection among similar photos, it doesn't attempt to solve more general problem of a photo selection within an entire photo collection.

Groups of images of very high similarity are addressed in a recent method for real time photo selection within a burst sequence [171]. To deal with this type of near-duplicate photos, Wang et al. propose a deep model to learn a latent relative attribute space, which is able to represent subtle image differences. After training on a large number of burst sequences, their feature extraction architecture learns a particular emphasis on sharpness, eye openness, body pose attractiveness and overall composition, as principal selection factors for a series of very similar photos.

How the context-aware assessment is represented in proposals dealing with photo albums? The photo summarization methods that we mentioned earlier [19, 89, 96, 154] usually act in the following manner: a photo album is clustered into groups of related photos (by similarity or other criterion) and then the most representative photos are selected. The selections can be performed using a combination of face and time criteria [89, 154], image quality [96] or by finding the photo with the most representative region of interest [19]. In these approaches more global collection level information is usually not taken into account, since the assessment is performed only within the clustered groups of similar photos.

Some methods dealing with assessment within photo collections were also proposed. An earlier mentioned story creation method for vacation photographs [143], which selects from 5 to 10 photos, can be considered one of them. Their proposed machine learning technique, apart from face and image quality features, takes into the account a number of scene features, and makes a selection on the global collection level. Sinha et al. [153] have also addressed the task as photo album summarization, but they formulated it as the multi-objective optimization problem, where a subset with optimal properties of quality, diversity and coverage is searched.

Similarly, instead of using the two-step of clustering and subsequent selections within the clusters, Ceroni et al. [13] directly trains a prediction model for personal importance prediction, based on a set of image and cluster features. The image features are represented by image quality and face properties, and the cluster features are represented by cluster sizes and average image quality within them. Continuing Ceroni's work, Fu et al. [42] have introduced a feedback mechanism for continuous training: after a user's collection is assessed with a general model, the user can provide new annotated data to refine the system's decision. Given the new data, their model is re-trained from scratch.

As we just mentioned, the photo group features can be considered as low-level information for the context-aware assessment, and the collection-level features can be considered as high-level information. It was also shown possible to rely on high-level information as the complementary data. To solve the problem of finding the event-specific most important photos, Wang et al. [175, 176] complement a Siamese pair comparison architecture with the event type label. In their case they predict one of four categorical importance score (very important, important, neutral, or irrelevant) for an image. Despite an interesting approach, the importance-based concept might be not always applicable to the selection within personal photo albums. In addition, the absence of low-level context information in this case, such as knowledge of other duplicate photos, can be crucial when selecting between similar photos.

Although a noticeable progress has been made to address a context gap in photo assessment, current context-aware methods for photo collections are still imperfect in their performance. From our perspective, the field could benefit from further research in a few specific directions. First, a deeper understanding of user behavior in dealing with photos within collections is necessary. In particular, we would like to investigate the influence of photo context on users' selection decisions within photo albums. In addition to the general album context, we aim to study the influence of the immediate photo context, defined by other similar photos taken in the same scene. Our research on user behavior through experimental studies is given in Part II. Second, we aim to model the analyzed user behavior with a computational approach: that is, model a user approach to the context extraction (with a clustering-based approach, as we will see in the following) along with a user approach to the process of image assessment and selection. The description of our approach and its performance is given in Part III.

## 2.6 SUMMARY

The research in the field of photography encompasses a vast range of subjects related to different moments of photo lifespan. At this point, we have a set of good practices that guide how to create a better photo, but we do not fully understand how we actually perceive and assess photos, and why we decide if a given photo is beautiful or important for us.

Nevertheless, the computational research might lead us to new insights in this field. Recent progress in the field of machine learning, the availability of training data and the introduction of convolutional neural networks have significantly contributed to the field of computational aesthetics. Currently, the photo assessment and photo enhancement methods for individual images continue to improve in their performance.

However, a typical consumer photo does not exist out of context, in an isolated manner. Commonly, it is a part of larger collection, and, with current storage capacities, such collections can grow uncontrollably. Although the problem of photo albums organization is connected with the photo assessment and selection, only few methods attempted to address it in a joint way.

While most photo aesthetics assessment methods act in an independent manner, assessing the quality of each photo separately, a completely personalized solution that considers the preferences of an individual user would be ideal, but it would also require large amount of user-specific data. However, we can use the context-specific information available in the collection itself, which could help automate common photography tasks such as organizing photo collections and selecting the best photos. In the rest of the manuscript we explore this subject in more details.

# Part II

# EXPERIMENTAL STUDIES

In Part II we aim to investigate user behavior when dealing with a collection of photos. To achieve this, we have conducted two experimental studies on image selection within photo albums and one experimental study on photo album clustering. Further, to study the influence of photo context on the selection process, we augment the obtained selection data with clustering data and derive useful statistics about users' photo selection decisions within groups of similar photos inside the albums.

# 3

# IMAGE SELECTION IN PHOTO ALBUMS

Although a number of photo assessment datasets are publicly available [15, 24, 70, 75, 122, 175], they have certain limitations. The principal datasets, such as the *Photo.Net* dataset [24] and the *AVA* dataset [122], were created to address the task of independent image assessment: each photo was assessed without any context information about its origin (such as a corresponding collection and similar photos taken in the same scene). The absence of this information makes them irrelevant to explore the influence of the context. An absence of context information was addressed in some recent datasets. The most important event-specific photos were identified by users in the study by Wang et al. [175]. In their dataset, the concept of photo selection is not focused on aesthetics or quality assessment, but rather on the importance of a particular photo for representing a given event. The most closely related study can be considered the work by Chang et al. [15], which focuses on selections within photo series of related images. However, since their dataset is based on pairwise comparisons and encompasses photo series not larger than 8 images, it does not represent the larger collection-level context. The collection context can be beneficial to reflect additional characteristics, such as intentions and skills of a photographer.

In addition, as many of existing datasets are based on different photo sharing web resources, it is worth noting the following. The photographers' peer-review social networks largely comprise of high-quality photos made by professionals or experienced amateurs and processed with a use of professional software. Such photos represent the final outcome of a photographer's work and neglect the preceding process of careful photo selection from the raw photo material by a photographer himself. The same can be said about generic photo sharing platforms, such as FLICKR. Although it is not uncommon to find entire photo collections there, such collections are often scrupulously prepared and possibly post-processed, too. These particularities largely affect the usability of such online data sources in the context-aware assessment; thus they should be used with care. Consequently, we have decided to conduct our own experimental studies on user selection decisions within various types of photo collections. During these studies, we have created two small-scale datasets that were used to analyze the nature of user selections, the degree of user decisions agreement, and to evaluate the performance of automatic assessment methods (which will be addressed in Part III).

In Section 3.1 we start with a user study on a pre-selection stage, where image sharpness is used as a selection criterion. Since the sharpness only selection data has

limited applicability, in the Section 3.2 we proceed with a more general selection study: the study on selection without predefined criteria, where observers are given freedom to choose any photos they like. Further, we return to the nature of photo selection again in Chapter 4, where we analyze user selection decisions in conjunction with collection clustering information from users.

## 3.1 EXPERIMENTAL STUDY ON SELECTION BASED ON IMAGE SHARPNESS

In this study we aim to collect people's decisions at a very early stage of pre-selection, when a large collection of photos is just captured and needs to be organized. Here we have focused on detecting low quality images, since it is common to reject low quality images first, before proceeding to further collection organization. In particular, we use image sharpness as the underlying photo selection criterion. Among the criteria leading to image rejection, blur (i.e. lack of sharpness) in its different forms is one of the most important factors [114, 180], as it can affect both professional and amateur photographs and it is hard to remove in post-processing. At the same time, the sharpness requirements largely vary depending on content type and user intentions, hence the need for context-aware assessment.

Figure 3.1: Interface of the user study on sharpness-based photo selection. Users could navigate through the collection freely, with a possibility of viewing all photos before making any selection. A currently examined image could be zoomed to view it at full resolution.

### 3.1.1 *Experimental Data and Procedure*

Our experimental study is based on 5 photo collections of different content type:

- *Travel collection* represents a common scenario of vacation photos, where photos of landscapes are mixed together with photos of people posing in front of a landscape. The photos are taken with a semi-professional camera, and multiple

highly similar photos of the same moment are taken (more than 5 photos on average).

- *Wedding ceremony* consists of wedding photos, captured indoors in difficult lighting conditions, and containing a noticeable number of blurred pictures. From 2 to 5 photos of the same moment are taken on average.
- *Sport event* album covers a volleyball match, where multiple pictures present motion blur due to the players' movements. From 2 to 4 photos of the same moment are taken on average.
- *Halloween party* album consists of photos during a Halloween party and presents cases of out-of-focus and motion blurred photos. Number of repetitions is high: more than 5 photos of each moment are typically taken.
- *Professional session* is a photo session conducted by a professional photographer, with a consistent level of photo quality and many repetitive photos with small differences (more than 5 photos on average).

All collections, except the professional collection, were acquired from the photo albums in YFCC100m dataset [68, 122], based on FLICKR data. We have manually selected the albums that did not present any signs of pre-selection, organization, or photos post-processing. The fifth collection, a professional photo session, is acquired directly from a photographer: the photos are extracted after a photo shooting, before any processing was applied to them.

As the initial photo albums are of different size, from each source album we extract 100 consecutive photos were selected, maintaining the structure of the original collection but limiting the duration of the study. Each image is resized, so that the longest side of the photograph is 1920 pixels. In this manner, five collections, each of 100 photos, are created. An overview of all collections can be found in Figure 3.2.

PARTICIPANTS The user study has been performed with 15 participants (9 male and 6 female), where each participant regularly takes personal photos in every-day life and is familiar with the task of photo selection and organization. Two of the participants occasionally use professional cameras, and can be considered as experts.

TASK Every participant was presented with each of the collections (collections chosen in a random order), and was given the task of labeling sharp and non-sharp photos. The users were asked to take the role of the photographer of each collection: while their current task would be to perform a photo pre-selection via choosing sharp enough photos, they would need to keep in mind the ultimate task of creating a photo album, containing the best and most representative photos. The user could assign one of three labels: *Accept* if he considered a photo sharp enough to keep, *Reject* if a photo was too blurry and not worth keeping, and *Maybe* if a photo was not absolutely sharp but still worth keeping or if they could not otherwise make a decision.

USER STUDY DESIGN In this study we have used an image browser layout, where all photos of each album were simultaneously visible as thumbnails, and users could navigate within the collection freely, with a possibility of viewing all photos before making any selection (see Figure 3.1). A larger version of the currently examined photo was also shown, where users could zoom to view the images at full resolution. Each album had to be completed before moving to the next one.

### 3.1.2 *Results Analysis*

As photo selection is a subjective process, the provided user selections can and do vary between users for each album. To assess the consistency of the selections of different users, the inter-agreement between observers for such data can be computed in different ways. A common choice to estimate the inter-rater agreement is the Kendall's W coefficient [72] or Cohen's kappa statistics [20]. The Kendall's W coefficient is generally used for ordinal ratings, while kappa statistics are applied for nominal ratings, which is our case (*Accept/Reject/Maybe* labels). As in our study fifteen users rate each album, we employ the modified Fleiss' kappa measure [37], which is a generalization of the original kappa for more than two raters.

*Fleiss' kappa measure*

The Fleiss' kappa measure is calculated as follows. Let $N$ be the number of rated items (images in a photo collection in our case), let $n$ be the number of ratings per item (number of users in our case), and let $k$ be the number of rating categories (number of possible image labels). For every item $i = 1, 2, ..., N$ and possible category $j = 1, 2, ..., k$ the $n_{ij}$ defines the number of raters that assigned the category $j$ to the item $i$.

First, the proportion of assignments to each category $p_j$ is calculated:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{3.1}$$

Then, the extent of agreement for each item $P_i$ (the proportion of how many rater pairs are in agreement) is calculated:

$$P_i = \frac{\sum_{j=1}^{k} n_{ij}^2 - n}{n(n-1)}. \tag{3.2}$$

The computed $P_i$ and $p_j$ values are used to obtain $\overline{P}$ and $\overline{P_e}$:

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i, \tag{3.3}$$

$$\overline{P_e} = \sum_{j=1}^{k} p_j^2. \tag{3.4}$$

Finally, the kappa measure $\kappa$ is defined as

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}. \tag{3.5}$$

According to Landis et al. [83], values of $\kappa$ can be interpreted as follows:

| $\kappa$ | Interpretation |
|---|---|
| <0 | Poor agreement |
| 0.01 – 0.20 | Slight agreement |
| 0.21 – 0.40 | Fair agreement |
| 0.41 – 0.60 | Moderate agreement |
| 0.61 – 0.80 | Substantial agreement |
| 0.81 – 1.00 | Almost perfect agreement |

Table 3.1: Kappa values interpretation [83].

A summary of the assigned labels as well as the corresponding Fleiss' kappa value for each collection are given in Table 3.2. According to the interpretation of Fleiss' kappa measure (Table 3.1), the level of user agreement ranges from fair to moderate on most of the collections, with the exception of the professional session, where a $\kappa = 0.172$ indicates only slight agreement between observers. This is not a surprising result, as this collection contains very little blur as well as many repetitive images. This observation was supported by participants' feedback as well, many of which mentioned that they had trouble assessing this particular collection. Similarly to the Professional session, the Halloween party collection also presents quite low agreement ($\kappa = 0.256$), as users' opinions can become divided between very similar photos with hardly perceptible differences in quality.

| | *Accept* | *Maybe* | *Reject* | Fleiss' kappa agreement |
|---|---|---|---|---|
| Travel collection | 46% | 26% | 28% | 0.470 |
| Wedding ceremony | 47% | 27% | 26% | 0.423 |
| Sport event | 24% | 30% | 46% | 0.336 |
| Halloween party | 35% | 34% | 31% | 0.256 |
| Professional session | 47% | 36% | 17% | 0.172 |

Table 3.2: Dataset characteristics computed from our user study on photo selection based on image sharpness criterion: label selection percentage and level of agreement between observers, according to Fleiss' kappa measure [37].

To visualize and explore user selections from multiple users, we transformed all selection decisions into a normalized *user preference score* for each image:

$$s_{u_i} = \frac{N_{A_i} + \frac{N_{M_i}}{2}}{N_{users}}, \tag{3.6}$$

where $N_{A_i}$ is the number of times image $i$ was assigned *Accept* label, $N_{M_i}$ is the number of times image $i$ was assigned *Maybe* label, and $N_{users}$ is the total number of users. Higher values of the user preference score indicate that an image was selected more often, with a value of one identifying images that were selected by all users. In this case, we divide the number of *Maybe* labels $N_{M_i}$ by 2, because these labels indicate uncertain decisions, which contribute equally to the acceptance and rejection of an image.

For an additional visualization of agreement between users, we compute a *user confidence score* which indicates how decisive the preference score for the particular image. The confidence score is calculated using the inverse of the triangular function over the preference score

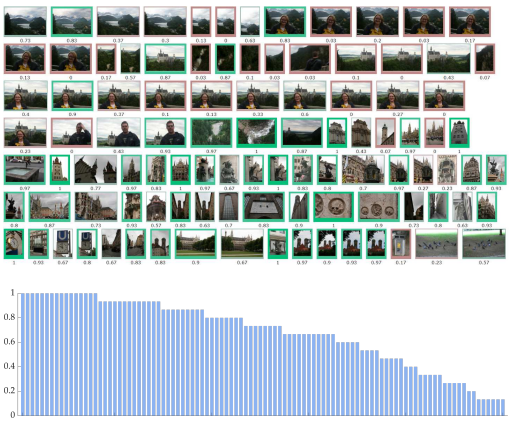
$$c_{u_i} = \frac{1}{\mathrm{tri}(2s_{u_i} - 1)}, \tag{3.7}$$

which produces higher confidence when the preference is closer to 1 (every user has selected the image) or to 0 (no user has selected the image). The per-image preference and confidence scores for albums used in our study can be seen in Figure 3.2.

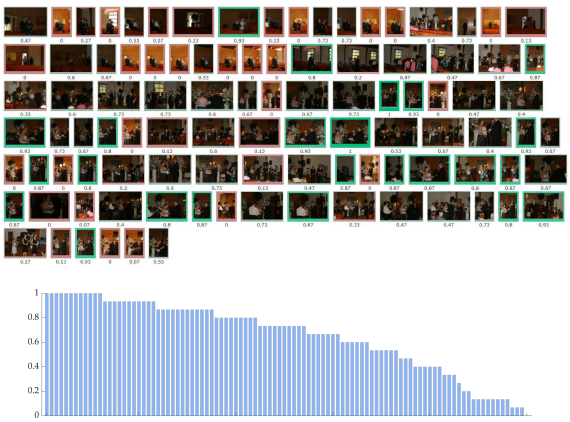
It can be noted that the confidence visualizations correspond with the computed per-album kappa values. The Travel collection contains the highest number of confidently assessed photos, which were selected or rejected by the majority of users, whereas Professional session presents much fewer confidently assessed photos. Our earlier observation on low agreement for the photos of nearly equivalent quality in Professional session and Halloween party album is also supported by the produced visualizations. For such photo series containing almost identical photos, we can notice a split of user agreements (when a user preference score is close to 0.5), hence low selection confidence.

At the same time, although the Travel album also contains a large number of repetitive photos, the sharpness differences are more perceptible within the series of photos, which leads to a higher user agreement. In addition, for the Sport event album a fair agreement is achieved since the majority of users have chosen to reject most of photos, while keeping only few. It is not surprising, as numerous photos in this album are depicting active movements and are noticeably blurred.

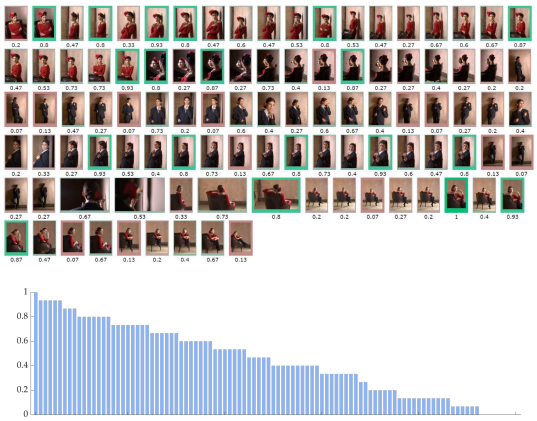
Regarding the context influence in selection decisions, it can be certainly found for certain scenes in all albums, however, its effect is not evident in some cases. As an example of context effect in selection process, in scenes with people from the Travel collection and Wedding ceremony, it is common that only one or two photos are generally preferred in the entire scene. Also, since the Wedding ceremony photos contain numerous low-quality blurry examples, it is commonly the case that only one photo is preferred among similar ones (which can be interpreted as "the best among the worst" choice for some severely blurred pictures). In general, it can be noted that the users follow the principle of keeping all representative photos of a collection that



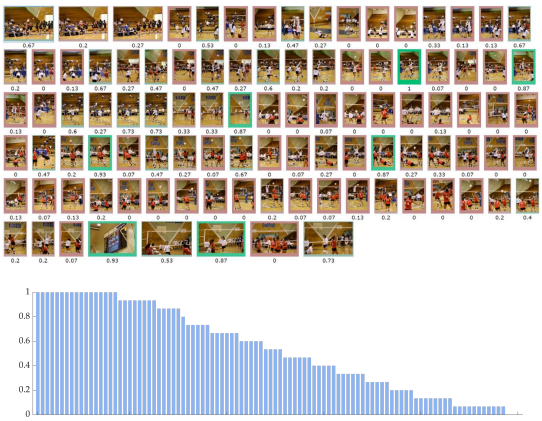
(a) Travel collection



(b) Wedding ceremony



(c) Professional session



(d) Sport event

(e) Halloween party

Figure 3.2: Demonstration of albums from the user study on sharpness-based photo selection. A user preference score is shown under each image. Confidence of user decisions is visualized with a bar chart, where per-image confidence values are sorted from highest to lowest value. *Note: thicker green or red borders can help to visually identify photos that were accepted or rejected by the majority of users.*

meet certain quality standards. At the same time, by following this approach, users sometimes opt to keep multiple almost-duplicate photos of the same scene. For instance, it can be observed in the Travel collection, where for certain architecture objects the majority of users have kept multiple photos of very high similarity.

When considering the effect of context in this user study, we should keep in mind the task performed by participants: selections were primarily motivated by sharpness, which is unlikely to be the only criterion used for selecting which photo to keep in a collection. Hence the observed behavior: users select only few scene-representative photos when dealing with low-quality photos and they select multiple repetitive photos when dealing with photos of high quality. According to our assumption, more clear context-aware patterns in user decisions should be observed in a different setting, where a user is free to select photos not limited by artificial constraints. Therefore, we have decided to conduct a differently formulated experimental study, where selection is performed without predefined criteria.

## 3.2 EXPERIMENTAL STUDY ON SELECTION WITHOUT PREDEFINED CRITERIA

As we have concluded in the previous section, a photo selection study based on a predefined assessment criterion could lead to certain limitations. First, we found that such constrained task can be challenging for non-experts, especially when faced with many similar images, as is the case in collections acquired from professional photographers. Second, the notion of selection context becomes ambiguous when analyzing user study results, since the original task was limited by a predefined criterion of image sharpness, thus the context would be also limited by the relative sharpness between similar photos. Third, the obtained data has limited applicability in a broader field of photo assessment: sharpness-based selections only partially represent overall image aesthetics.

To address these issues, we have conducted another experimental study. In this study we again collect people's selection decisions within photo albums, but with no predefined selection criteria. Inspired by an experimental approach similar to our sharpness-based user study, we now asked users to select photos freely according to their taste, along with a few additional user-friendly simplifications.

### 3.2.1 *Experimental Data and Procedure*

For the purpose of this user study, we selected six photo albums covering a variety of typical scenarios where amateur photographers may opt to take a large number of photos.

Photo albums were selected from several different sources, including PEC dataset [8], YFCC100M dataset [163], CUFED dataset [175] and personal albums of the authors. We have limited our search to collections that were not altered by image processing software, and where no evident pre-selection was applied before, thus possibly con-

Figure 3.3: Interface of the user study on photo selection without predefined criteria. User can freely browse through the entire collection and perform an image selection on key press.

taining multiple similar and near-duplicate photos, and reflecting a typical modern photo album taken with a digital camera or a smartphone. From each source album, 50 consecutive photos were selected, maintaining the structure of the original collection but limiting the duration of the study. The albums are shown in Figure 3.4.

The analyzed albums demonstrate different characteristics, which allow us to study various real-life scenarios and identify particular behavior of assessed methods:

- *Family Event 1* is a wedding photo album, with a moderate number of repetitive photos (from 1 to 3 photos for each captured moment on average), where the pictures are taken with a semi-professional camera.
- *Family Event 2* is another wedding photo album captured with a semi-professional camera, but with a higher number of repetitive photos (in some cases, more than 5 photos for each same moment).
- *Family Event 3* album represents a family birthday gathering taken mostly indoors, with a point-and-shoot camera. This collection presents a large number of fuzzy shots, where the points of interest are not well defined, with a moderate-to-high number of repetitive photos (from 3 to 6 photos on average).
- *Travel Album 1* represents a common scenario of vacation photos, where photos of landscapes are mixed together with photos of people posing in front of a landscape. The photos are taken with a semi-professional camera, and multiple highly similar photos of the same moment are taken (more than 5 photos on average).
- *Travel Album 2* consists only of landscape photos, taken with a semi-professional camera, and the number of repetitions is moderate: 2 to 3 photos are taken for each captured moment.

- *Travel Album 3* represents photographs taken during an amusement park visit. In this album, no people are present: it consists of multiple pictures of architecture, landscapes and objects (usually from 2 to 5 photos per same scene). The photos are taken with a point-and-shoot camera and various cases of blurred or under-exposed photos are present.

PARTICIPANTS In total, 30 participants took part in our study (23 male and 7 female), with ages ranging between 24 and 55 years. Each pair of albums was evaluated by 10 different users. All participants could be characterized as amateur or casual photographers, with varying levels of photographic experience and interest.

TASK Each user was presented with a pair of albums, where one given album represented a typical family event, such as wedding or birthday, while the second represented a travel photo collection. The users were tasked with putting themselves in the role of the photographer of that collection to select the *best, more representative, or most important photos* in their opinion. No limit was placed on the number of photos selected in each album. Before the start of the experiment, each user was presented with two practice collections of 12 photos each, in order to get familiar with a task and the interface. Then, once the user was ready, they could proceed to selecting photos in the two complete collections assigned to them.

USER STUDY DESIGN For each shown photo album, users were presented with a browser-based interface as shown in Figure 3.3. All photos of each album were simultaneously visible as thumbnails, while a larger version of the examined photo was also shown. Users could navigate within the collection freely, with a possibility of viewing all photos before making any selection. Each album had to be completed before moving to the next one.

It can be noted that, comparing to our previous study, we have changed several aspects. First, we have reduced the number of per-user albums and the number of photos in each album. This has been done to reduce the experiment's duration and lessen the users' fatigue. Second, we have concluded that the presence of *Maybe* selection label was often ambiguous and confusing for users, and could introduce unnecessary noise in the experimental data. For that reason, we removed this label and reformulated the task as simple selection of preferred photos (which is equivalent to binary *Accept*/*Reject* labels). In effect, after the changes introduced, the average user study duration also reduced to 20 minutes, as compared with 40 minutes average of our previous study.

### 3.2.2 *Results Analysis*

The results analysis is performed similarly to our study on sharpness-based selection, and the equations of corresponding computed measures can be consulted in Section 3.1.

To assess the per-album user agreement on selection decisions, we use the Fleiss' kappa measure, as given in Equation 3.5. The computed Fleiss' kappa values for each

album are given in Table 3.3. As explained earlier in Table 3.1, values for this measure can be interpreted as follows [83]: $\kappa < 0.2$ indicates slight agreement, $0.2 \leq \kappa < 0.4$ indicates fair agreement, $0.4 \leq \kappa < 0.6$ indicates moderate agreement.

| | *Accept* | *Reject* | Fleiss' kappa agreement |
|---|---|---|---|
| Family event 1 | 57% | 43% | 0.393 |
| Travel album 1 | 32% | 68% | 0.472 |
| Family event 2 | 33% | 67% | 0.334 |
| Travel album 2 | 39% | 61% | 0.179 |
| Family event 3 | 33% | 67% | 0.351 |
| Travel album 3 | 31% | 69% | 0.210 |

Table 3.3: Dataset characteristics computed from our user study on photo selection without predefined criteria: label selection percentage and level of agreement between observers, according to Fleiss' kappa measure [37].

For the purpose of visualization and performance analysis of proposed computational methods (we address automatic photo assessment later, in Section 6), we also compute a normalized *user preference score* for each image. The user preference score is computed similarly to the equation 3.6, by averaging user selections across the ten users assessing each album:

$$s_{u_i} = \frac{N_{sel_i}}{N_{users}}, \tag{3.8}$$

where $N_{sel_i}$ is the number of times image $i$ was selected, and $N_{users}$ is the total number of users. As before, higher values of $s_{u_i}$ indicate that an image was selected more often, with a value of one identifying images that were selected by all users. For visualization of the selection confidence we also compute a *user confidence score*. The computation is done equivalently to Equation 3.7 in previous section, using the inverse of the triangular function over the preference score. The per-image preference and confidence scores for the albums used in our study can be seen in Figure 3.4.

Most albums assessed lead to a fair to moderate agreement between users, with the exception of two albums where slight agreement was found (*Travel album 2* and *Travel album 3*). Looking at the content of the albums, several interesting conclusions may be drawn. We observe that albums with higher agreement contain a larger number of people portraits, with repetitive similar photos of the same person or group (including *Travel album 1*, which contains multiple people portraits taken in front of landscapes). At the same time, *Travel album 2* and *Travel album 3* do not contain people's photos and consist mostly of landscapes and architecture photos. These latter albums demonstrate a larger variance in user selections: the notion of an attractive landscape appears to vary much more than the understanding of a well-captured portrait or group photo.

This suggests that for users it may be easier to perform photo selection of people's photos within an album, even when the presented people are unknown. In fact, closer observation of users' selections during the study reveals that facial expressions were a critical factor guiding their decisions when multiple photos of the same people were

(a) Family event 1

(b) Travel album 1

(c) Family event 2

(d) Travel album 2

(e) Family event 3

(f) Travel album 3

Figure 3.4: Demonstration of albums from the user study on photo selection without predefined criteria. A user preference score is shown under each image. Confidence of user decisions is visualized with a bar chart, where per-image confidence values are sorted from highest to lowest value. *Note: thicker green or red borders can help to visually identify photos that were accepted or rejected by the majority of users.*

present. On the other hand, unique photos of people were almost always selected, irrespective of the quality or expression present.

Further, in *Travel album 2* we find a particular example of a photo sequence, where the concept of best photo selection may not be directly applicable: this photo sequence presents a panoramic-like capture of the surrounding landscape (seen in the second and third rows of this album in Figure 3.4). In this scenario, it is unlikely that a user would want to keep a single photo, as the series is intended for a particular type of post-processing. Indeed, in this case we note that users have shared their selection across the series with no particular photo showing higher selection preference. Another similar scenario could occur when capturing a bracketed series for later construction of a high dynamic range image. Such use cases can frequently occur given the general availability of advanced photo processing tools even on mobile devices.

Another challenging example where user selections become divided is when nearly identical photos are present, with no discernible differences in quality. Such behavior was already observed in our sharpness-based selection study, where the confidence of user decisions was largely decreased for almost identical photos. Again, the user votes are approximately equally shared between the photos in question, meaning that no single image led to a higher preference, despite users wanting to keep at least one representative image of such scenes.

In comparison with our earlier study on sharpness-based photo selection, the notion of context is more apparent in the data acquired in the current study. From the preferences overview in Figure 3.4 it can be noted that in numerous scenes of related photos, only one photo is preferred by the majority of users. It is especially noticeable in the series of people's photos, where one photo is often selected with high confidence. Certain choices can be also found in landscape scenes. Contrary to the earlier study, we do not observe photos of moderate similarity but equal technical quality selected together. However, the near-duplicate photos with barely noticeable differences can lead to the split of selections, as we have just discussed. Nevertheless, the context influence on user choices is more evident in the acquired dataset, and we will proceed with a more comprehensive analysis of it in the Section 4.2, where we augment the obtained selection data with the albums' clustering by users.

## 3.3 SUMMARY

Since the datasets focusing on unprocessed photo albums are lacking in the research, we have decided to conduct our own experimental studies to collect the relevant data. We have conducted two studies on photo selection within albums: a study on selection based on a pre-defined criterion (image sharpness) and a study on selection without predefined criteria.

Both studies have shown that the task of image selection in photo collections is not straightforward for users, and the level of user agreement can largely vary depending on the album content. Certain cases are especially challenging: for example, nearly identical photos with no discernible differences lead to a split of user selections. The study on

sharpness-based selection indicated that the selection with a pre-defined criterion can be difficult for non-experts, and that such constrained task might not reveal specific context selection patterns. The notion of the context is more apparent in the second study, where selection is performed without predefined criteria. Often, in certain series of photos (especially people's photos) only one photo is selected with high confidence. Overall, in no-criteria selections users tend to agree more when dealing with people's photos, comparing to landscape-focused albums, where agreement is lower.

As the context aspect of photo selection requires further analysis, in the following chapter we proceed with another experimental study on the same data, where we examine how users cluster photo collections. By combining the clustering and selection studies, we aim to explore how the clustering-defined context affects the subsequent photo selection process.

# 4

# CONTEXT IN PHOTO ALBUMS

When organizing a photo album, a user often explicitly or implicitly performs a task of grouping similar photos together, which can aid the browsing or selection process. In this chapter we aim to complement the previously acquired photo selection decisions with the clustering partitions produced by users in the same photo albums. Such partitions could be beneficial in different applications, and they could be used to formulate the context of a photo. In addition, while we can visually examine the previously obtained selection data to detect possible context patterns, the user-provided clustering can improve the precision of our analysis.

Although there were some earlier attempts to combine image selection decisions with clustering information in photo albums [13, 96], there is no publicly available dataset of such type, to this moment. Thus, we have decided to conduct a new experimental study, which would help to identify how users tend to cluster similar photos in albums, to what extent different users agree in their clustering decisions, and to investigate how the clustering-defined photo context affects the photo selection process.

In this chapter, we first describe the design of our experimental study on photo albums clustering and analyze the users' agreement in Section 4.1. Since our clustering study has been conducted on the same photo albums as the ones used in our selection study (see Section 3.2), we combine the acquired data from the two studies and perform an analysis of selection decisions within user-defined clusters, in Section 4.2. It is also worth mentioning that later in this work we benefit from the acquired clustering data again, when we explore automatic methods for photo albums clustering and their performance, in Part III.

## 4.1 EXPERIMENTAL STUDY ON CLUSTERING

In this study we investigate how users group similar photos together and determine the extent of users' agreement in their grouping decisions, for different types of content. We also search for common user traits that could be helpful in users' behavior modeling. We first present the details of the photo albums clustering experimental setup, and then we analyze the collected data and the level of agreement between the clustering partitions provided by users.

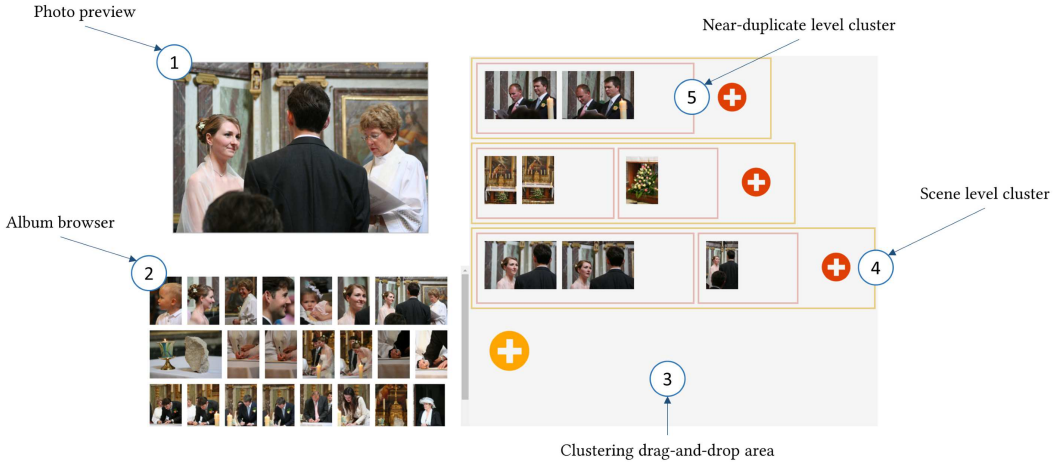



Figure 4.1: Interface of the user study on photo albums clustering. Users can browse the entire album on the left side and by using drag-and-drop gestures they can move images to the right side and assign to the clusters of their choice. New clusters of different depth can be created with special + buttons (these buttons can either be pressed directly or activated implicitly: if an image is dragged upon the button, a new cluster is created and the image is automatically assigned to it).

### 4.1.1 *Experimental Data and Procedure*

With the aim of combining the output of two experimental studies together, we have designed this study on the same photo albums that were used in our selection study without predefined criteria, as described in Section 3.2. The selected six albums represent different typical scenarios of photos taken by amateur photographers, where multiple similar and near-duplicate photos are present in each collection, and the number of repetitive photos varies for each collection. Likewise, each user was presented with a pair of albums (each of 50 photos) related to different photo scenarios: one album represented a typical family event, such as wedding or birthday, and the second album represented a travel photo collection.

PARTICIPANTS In total, 30 participants took part in our study (25 male and 5 female), with ages ranging between 23 and 57 years. Each pair of albums was processed by 10 different users. All participants could be characterized as amateur or casual photographers, with varying levels of photographic experience and interest.

TASK Each user was asked to put himself in the role of the photographer of each photo collection, whose ultimate task would be to create a curated photo album in which the best and the most representative photos are selected. With this in mind, the user was presented with the task of grouping similar photos from a photo album together, according to two levels of similarity: scene clusters and near-duplicate clusters. The following descriptions were given for each level:

- *Scene clusters.* "Photos depicting the same scene (for example, identified by the same background), but with significant changes present. Examples of possible changes: viewpoint changes, new persons or objects appearing throughout the photos. A scene cluster is formed by multiple smaller groups of high similarity (near-duplicate clusters)."
- *Near-duplicate clusters.* "Photos depicting the same objects or persons in the same scene. Although some variations between photos are possible, usually they do not largely change the scene. Examples of possible changes are: pose changes, small viewpoint changes, quality changes (e.g. blur or exposure changes)."

Although it is possible that users might exhibit a different behavior when assessing their own photos, we have opted for this experimental design in the interest of collecting statistical tendencies of users' behavior.

The motivation behind the two level clustering structure is the following. The photos from the same scene (usually depicting the same place or moment) are often perceived together in the process of album browsing. At the same time, groups of near-duplicate photos provide an immediate photo context when selecting which photos to keep, as users are likely to compare photos with one another to identify the best one Furthermore, in the selection task, the scene context can be used for a more refined decision after selecting among near-duplicates, to keep only few photos from the entire scene.

USER STUDY DESIGN The essential challenge in the design of such a user study is to provide the ability of multi-level clustering for users, where they could assign an image both to the scene or near-duplicate cluster. An example of the desired output clustering structure can be found in Figure 4.4. Since the near-duplicate clusters are logically enclosed into the higher level scene clusters, the user study interface is created in a similar two-level manner, based on a drag-and-drop principle. The interface is demonstrated in Figure 4.1: user can freely browse the entire photo collection and then perform clustering of images by a simple dragging gesture to the right part of the screen. A photo can be either assigned to the existing cluster of photos, or a new cluster can be created with the images automatically assigned to it.

Before the start of the experiment, each user was presented with two practice collections of 12 photos each, in order to get familiar with the task and the interface. Then, once the user was ready, they could proceed to cluster photos in the two complete collections assigned to them. No expected number of clusters was defined for users, neither the approximate number of images within a cluster was suggested. Each album had to be clustered before proceeding to the next one, but no time constraint was defined. On average, every user took around 30 minutes to complete the task for the assigned pair of albums.

Overall, users perceived the given task positively, while the following and similar comments were given by a few observers: *"I have to do such a grouping myself quite often when going through many photos after a travel"*, *"I might not explicitly put the photos into some folders, but in my mind I select the photos within the clusters of similar photos like this"*. These comments further reinforce our decision to consider photo in two levels of similarity.

### 4.1.2 *Results Analysis*

After all images in the album are processed by a user, each image is assigned clustering labels indicating its scene and near-duplicate clusters, allowing us to analyze the inter-user agreement on their clustering decisions This data is then used to perform an analysis of the inter-user agreement on their clustering decisions.

The analysis is performed using the Adjusted Rand Index (*ARI*) [58], which provides a measure of similarity between two data clusterings. The Rand Index (*RI*) considers all possible pairs of data elements and counts the number of pairs that are assigned to the same or different clusters between two given partitions. The Adjusted Rand Index is the corrected-for-chance version of the original Rand Index, which is adjusted by the expected value of *RI*.

*Adjusted Rand Index*

Let $S = \{i_1, i_2, ..., i_N\}$ be a set of initial elements (images in a photo collection before clustering in our case). Suppose two different partitions of this set are obtained after clustering by different users or methods: $X = \{X_1, X_2, ..., X_P\}$, a partition of $S$ into $P$ subsets, and $Y = \{Y_1, Y_2, ..., Y_R\}$, a partition of $S$ into $R$ subsets.

Now, we define the following quantities:

- $\boldsymbol{a}$, the number of pairs of elements in $S$ that are in the same subset in $X$ and in the same subset in $Y$
- $\boldsymbol{b}$, the number of pairs of elements in $S$ that are in different subsets in $X$ and in different subsets in $Y$
- $\boldsymbol{c}$, the number of pairs of elements in $S$ that are in the same subset in $X$ and in different subsets in $Y$
- $\boldsymbol{d}$, the number of pairs of elements in $S$ that are in different subsets in $X$ and in the same subset in $Y$

The Rand Index, $RI$ is computed as follows:

$$RI = \frac{a + b}{a + b + c + d} \tag{4.1}$$

Here, $a + b$ can be interpreted as the number of agreements between two different partitions, and $c + d$ can be interpreted as the number of disagreements. Thus, the Rand Index represents the frequency of agreements over all pairs in the set.

However, for two random partitions the expected value of the Rand Index is not a constant. To ensure that the Rand Index takes a value close to 0 for random labeling, the Adjusted Rand Index was proposed [58]. The Adjusted Rand Index

is the corrected-for-chance version, where correction is done using the Expected Index Value.

The calculation of the Adjusted Rand Index is based on a contingency table, which summarizes the overlaps between two groupings $X = \{X_1, X_2, ..., X_P\}$ and $Y = \{Y_1, Y_2, ..., Y_R\}$. In the contingency table each entry $n_{ij}$ denotes the number of objects in common between each subset $X_i$ and $Y_i$ ($n_{ij} = |X_i \cap Y_j|$):

| $X \backslash Y$ | $Y_1$ | $Y_2$ | ... | $Y_R$ | Sums |
|---|---|---|---|---|---|
| $X_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1R}$ | $a_1$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2R}$ | $a_2$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $X_P$ | $n_{P1}$ | $n_{P2}$ | ... | $n_{PR}$ | $a_R$ |
| Sums | $b_1$ | $b_2$ | ... | $b_P$ | |

Finally, the Adjusted Rand Index $ARI$ is calculated as

$$ARI = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{RI}} - \overbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index Value}}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}. \tag{4.2}$$

The per-album user agreement is demonstrated in Figure 4.2, and more detailed results can be found in Table 4.1. The $ARI$ is calculated for the clustering provided by each possible pair of users, and the values in the table represent the obtained mean and standard deviation values across all users. The Adjusted Rand Index is close to 0 for random labeling and equal to 1 when the clusterings are identical.

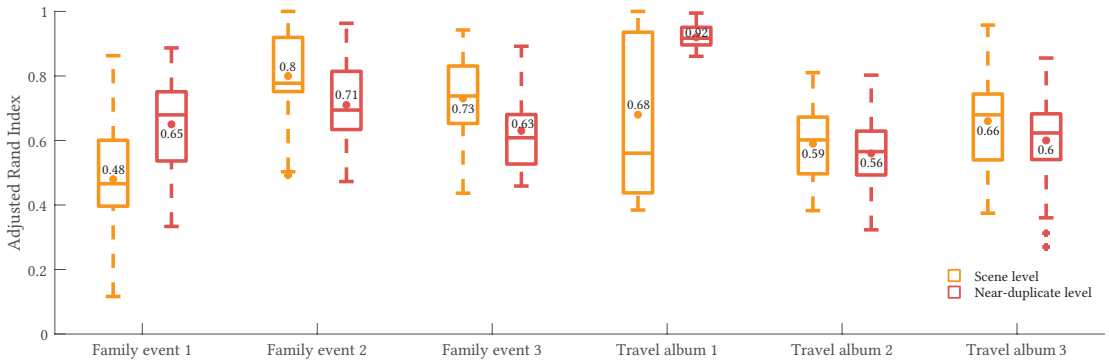



Figure 4.2: Per-album clustering agreement between users, according to the Adjusted Rand Index [58]. The box plots represent the per-album distributions of $ARI$, where a dot and the associated value indicate the average per-album $ARI$.

Our findings suggest that users do not always achieve a high agreement, however a certain level of agreement is present for all albums. The obtained results demonstrate that the difficulty of album clustering depends on its content and the presence of particular features. For example, the lowest scene clustering agreement is found for the wedding album *Family event 1*, where the average ARI is equal 0.481 (also, the standard deviation is rather high in this case and equal 0.19). This can possibly be explained by the sparsity of particular scene landmarks, as the entire album is related to the wedding ceremony in the church, and also due to the presence of multiple close-up shots, which do not provide many indications about the general setting. At the same time, the agreement for the near-duplicate level of clustering in this album is higher, as the ARI is equal 0.645, which can be explained by the presence of multiple repetitive shots that are easy to identify.

Almost perfect user agreement can be found for near-duplicate clustering (average $ARI = 0.92$) of the *Travel album 1*, which contains numerous almost identical shots, picturing people in front of landscapes. Also, the scene clustering agreement is relatively high for this album (average $ARI = 0.68$). According to the acquired results and users' remarks after the experiment, the albums with people present in photos are easier to cluster, as the boundaries between the captured moments are more clear. On the contrary, the albums represented by landscape or object photos make the clustering task more difficult. For instance, in the *Travel album 2* we can find a photo sequence that presents a panoramic capture of the surrounding landscape. While the scene level concept is applicable here, division into near-duplicate clusters does not achieve a considerable agreement by users, with individual user clusterings largely varying in this case.

| | Family event 1 | Family event 2 | Family event 3 | Travel album 1 | Travel album 2 | Travel album 3 | **Average** |
|---|---|---|---|---|---|---|---|
| Scene level (SC) | 0.481 (±0.19) | 0.802 (±0.14) | 0.732 (±0.13) | 0.682 (±0.24) | 0.592 (±0.11) | 0.657 (±0.16) | **0.658 (±0.16)** |
| Near-duplicate level (ND) | 0.645 (±0.15) | 0.715 (±0.12) | 0.625 (±0.11) | 0.922 (±0.03) | 0.561 (±0.11) | 0.598 (±0.14) | **0.678 (±0.11)** |

Table 4.1: Per-album clustering agreement between users, according to the Adjusted Rand Index [58]. *SC* denotes scene level clustering results, *ND* denotes near-duplicate level clustering results. The first value represents the mean *ARI* across all users. The second value (given in parentheses) represents the standard deviation of the *ARI*.

## 4.2 ANALYSIS OF JOINT RESULTS ON CLUSTERING AND SELECTION BY USERS

The acquired user clustering decisions can provide additional useful insight on the nature of user selections within photo albums. In our previous study described in Section 3.2, we have analyzed how users select the best or the most important photos in a photo collection. Since the present clustering study was performed on the same albums, we intend to combine the acquired information and analyze the possible influence of photo clusters in the selection process. Even though the selection study was performed on

plain albums without photo clusters indications, we believe that users might implicitly consider groups of similar photos, when they make their selection decisions.

### 4.2.1 *Ensemble Clustering from Multiple Partitions*

The scheme of data aggregation from the two studies is given in Figure 4.3. As the current clustering study and the earlier selection study were performed by different users, the aggregation of data cannot be performed in a direct manner. To overcome this, as the first step we create one generalized per-album clustering and then we combine the obtained clustering with multiple user selections inside each album.

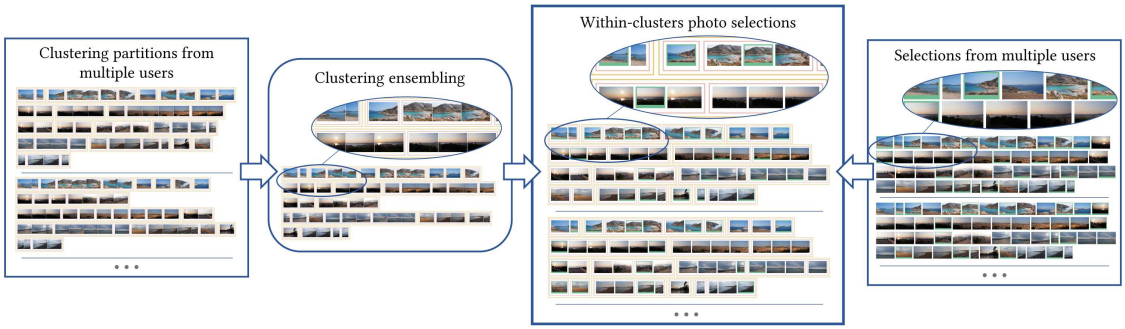


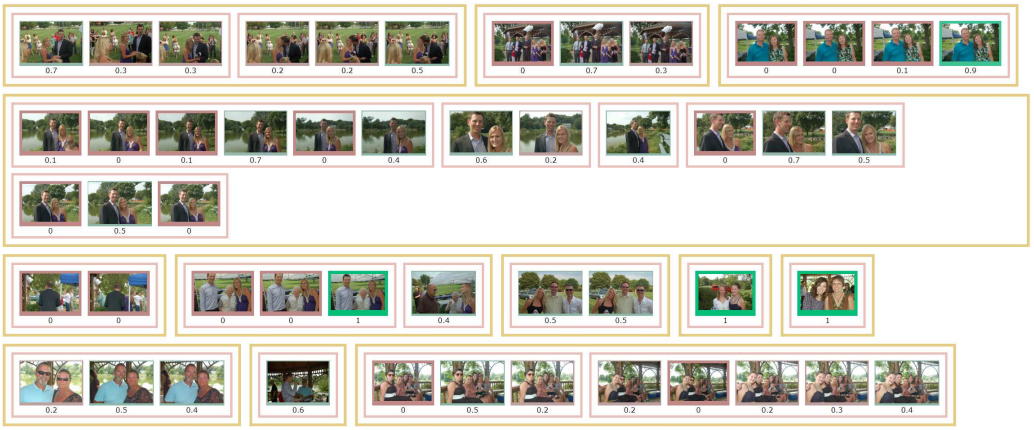

Figure 4.3: Data aggregation scheme for the combined study on clustering and selection. Multiple clustering partitions are transformed into one ensemble clustering using [104]. Photo selections acquired in our earlier selection study are incorporated into the computed ensemble clustering, to achieve within-clusters photo selections.

To create one aggregated clustering per album from the multiple user partitions obtained in the user study, we apply the ensemble clustering approach by Lu et al. [104]. Their solution creates one combined clustering by searching for a consensus partition via an optimization scheme, which maximizes the measure of agreement with all partitions. Examples of the acquired ensemble clustering are demonstrated in Figure 4.3 and Figure 4.4.

### 4.2.2 *User Selections within Clusters of Photos*

Once a per-album ensemble clustering is computed, we incorporate the user selection decisions into it. The photo selections acquired in our earlier study (Section 3.2) represent the decisions provided by multiple users for each album. These decisions are directly combined with the computed ensemble clustering, to obtain within-clusters photo selections for different users. Examples of the output clustering structure along with user selections from one user are demonstrated in Figure 4.3. By analyzing the photo selections within such context, we can determine the possible influence of groups of similar photos in decision making.

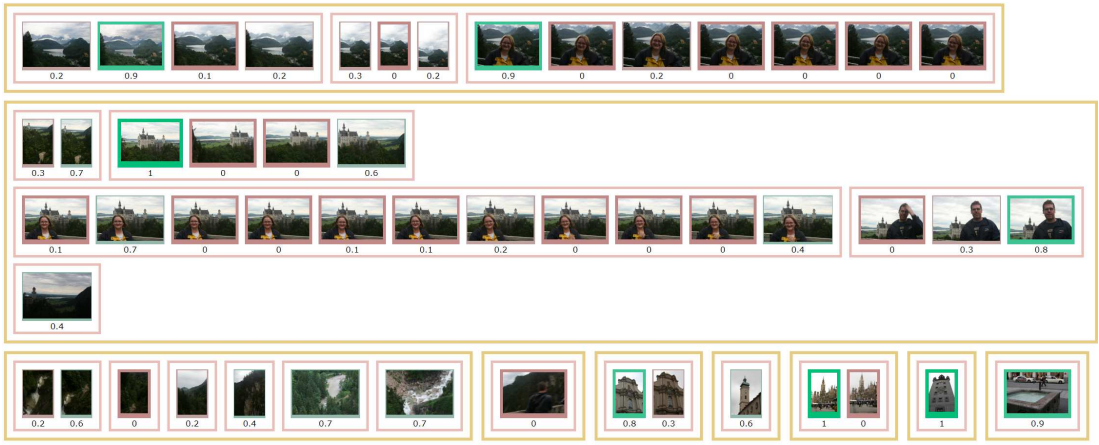
(a) Family event 1



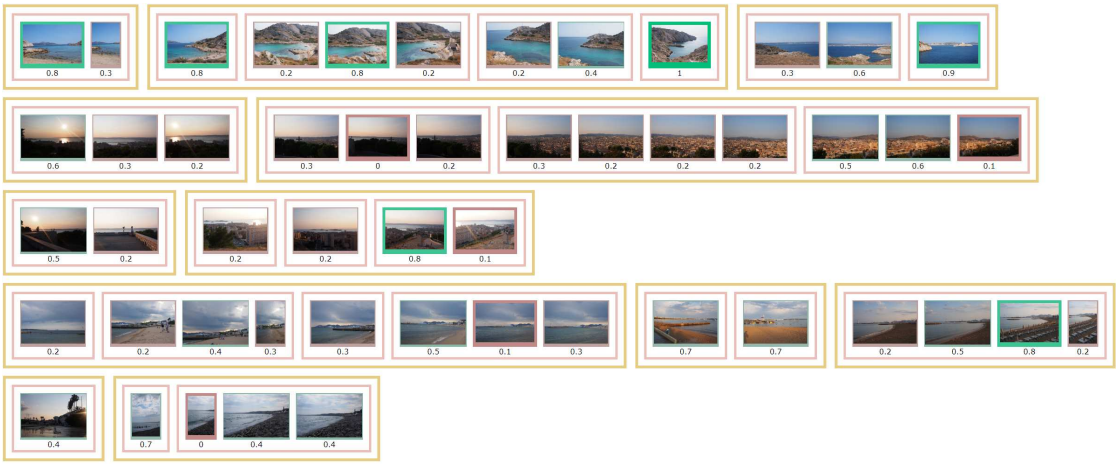
(b) Family event 2

(c) Family event 3

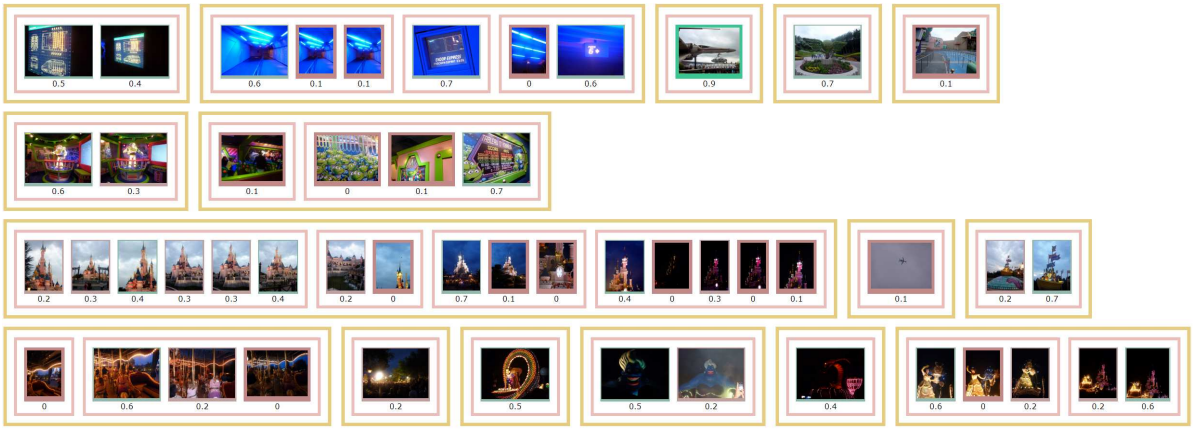
Figure 4.4: Demonstration of ensemble clustering together with user preferences from the user study on photo selection without predefined criteria. Yellow borders indicate scene level clusters, and red borders indicate near-duplicates clusters. A user preference score is shown under each image.



(d) Travel Album 1



(e) Travel Album 2



(f) Travel Album 3

Figure 4.4: Demonstration of ensemble clustering together with user preferences from the user study on photo selection without predefined criteria. Yellow borders indicate scene level clusters, and red borders indicate near-duplicates clusters. A user preference score is shown under each image.

| | Ratio of selected photos | Ratio of selected scene clusters | Ratio of selected near-duplicate clusters | Ratio of selected singleton (one-image) clusters |
|---|---|---|---|---|
| Family event 1 | 0.570 | 0.931 | 0.868 | 0.700 |
| Family event 2 | 0.326 | 0.883 | 0.779 | 0.333 |
| Family event 3 | 0.328 | 0.729 | 0.553 | 0.550 |
| Travel album 1 | 0.322 | 0.822 | 0.705 | 0.275 |
| Travel album 2 | 0.390 | 0.875 | 0.704 | 1.000 |
| Travel album 3 | 0.308 | 0.669 | 0.568 | 0.471 |
| Average | 0.374 | 0.818 | 0.696 | 0.555 |

Table 4.2: Selection ratio within obtained clusters in photo albums. Ratio of selected photos represents overall ratio of selected photos. Ratio of selected clusters represents the ratio of clusters where at least one photo is selected.

The results of the first analysis are given in Table 4.2. In this table we show ratios of selection for different entities within the albums:

(1) ratio of selected photos indicates a ratio between the number of selected photos within an album and the total number of photos in the album;

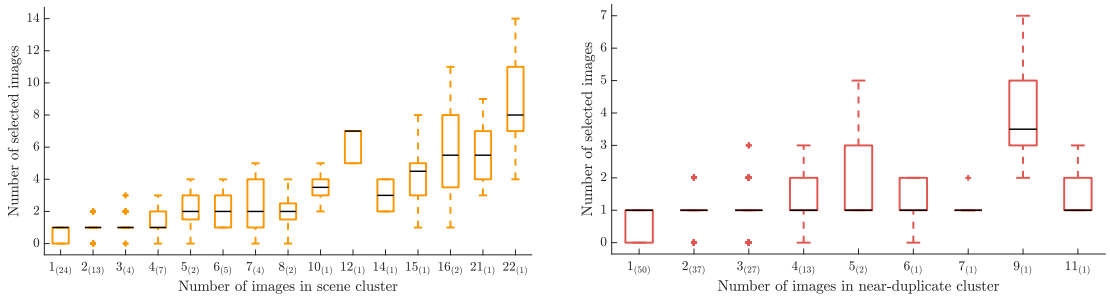
(2) ratio of selected scene clusters indicates a ratio between the number of the scene clusters where at least one photo was selected and the total number of scenes;

(3) ratio of selected near-duplicate clusters indicates a ratio between the number of the near-duplicate clusters where at least one photo was selected and the total number of near-duplicate clusters;

(4) ratios of selected singleton clusters indicates a ratio between the number of clusters consisting of only one image that were selected and the total number of such clusters.

Several conclusions can be drawn from these results. First, on average, users have selected around 37% of the photos. It is important to notice that in the album *Family event 1* the ratio of selected photos is higher than in others and equal to 0.57, which can be explained by the fact that this album has less repetitive content and contains a number of unique portraits of people not reappearing in other photos. Second, we can observe that the average ratio of selected scene clusters differs from the average ratio of selected near-duplicate clusters (0.818 versus 0.696). It appears natural, as the scenes generally contain wider range of photos, which leads to a higher chance that at least one image will be selected within them. Finally, the ratio of selected singleton clusters (consisting of only one image) largely varies for different albums, which indicates that there is no direct dependency that a unique photo of an object or a person will have a higher chance to be selected, perhaps contrary to intuition. Departing from this latter conclusion, the next step would be to investigate the dependency between a number of selected images in a cluster and the total number of images in the cluster.

The average number of selected images per cluster of different size is given in Figure 4.5. It can be observed that while for the scene clusters the median value of selected photos gradually increases with the size of the cluster, the same is not true for the near-duplicate level clusters. Their median value of selected images per cluster holds




(a) (b)

Figure 4.5: Dependency between a total number of images in a cluster and a number of selected images within it. Data is aggregated across all albums, and the distributions represent number of selections by different users. Number in parenthesis indicates how many clusters of corresponding size present in the ensemble clustering data. (a) Selections made within scene clusters. (b) Selections made within near-duplicate clusters.

around 1, except some larger clusters (which do not provide reliable statistics, since there are only few large-size clusters are found across the albums).

Although the subject of within-clusters photo selection merits further study, the above observations suggest that users tend to discard a considerable amount of content within clusters. As users tend to capture multiple repetitive photos to ensure a good result, they also inherently create a lot of redundancy in their collections, increasing the time and effort required to manually review them. Consequently, automatic approaches to both cluster and rate photos in an integrated manner might be beneficial.

## 4.3 SUMMARY

To analyze the nature of selections within photo context, we have conducted an experimental study on photo albums clustering and augmented our study on photo selection (no-criteria assessment) with the acquired data. The users were presented with the task of clustering an album into scene and near-duplicate clusters. The users achieve rather consistent level of agreement for both clustering levels. The obtained results suggest that the albums with people present in photos are easier to cluster, whereas the boundaries between captured moments are less clear for landscape photos. These observations are similar to our earlier findings in the photo selection study, where users also had higher agreement when dealing with people's photos.

After aggregating the data from the clustering and selection studies, we have observed that the number of selected photos increases together with the size of scene clusters but holds around 1 for near-duplicate clusters. At the same time, unique photos do not show a higher chance of being selected.

Our findings suggest that users do take context into account and discard repetitive and redundant photos during the selection process. In the next part of the manuscript

we attempt to model these aspects computationally, to create an automatic approach for clustering and selection in photo albums.

Part III

# COMPUTATIONAL MODELING OF CONTEXT IN PHOTO ASSESSMENT

The conducted experimental studies have revealed important patterns in clustering and selection user decisions within photo albums. Now, a natural question arises: can we model the discovered user behavior automatically? In this part we attempt to answer this question by proposing our own computational solutions. Since the clustering study results have demonstrated the importance of clustering partitions in context modeling and photo selection process, we start this part with the research on automatic photo albums clustering. As a result of the conducted research, we propose a photo collection clustering approach based on the hierarchical clustering technique. Subsequently, we propose to use the computed partitions to extract the corresponding photo context information and adapt automatic photo assessments according to it. With the aid of previously acquired user data, we assess the performance of the proposed solutions and discuss the potential of automatic modeling of user decisions within photo albums.

# 5

# MODELING USER BEHAVIOR IN PHOTO ALBUMS CLUSTERING

One of the subtasks in photo album organization is the grouping of similar photos: even if not always performed explicitly, it is implied in other tasks, including the process of photo selection. In Chapter 4 we have studied how users group similar photos, where we have found that users achieve certain agreement in their clustering decisions. Also, our findings suggest that the context information defined by clustering partitions could aid the assessment and selection process. In this chapter we aim to obtain such partitions automatically, and for this purpose we propose a solution based on a hierarchical clustering approach.

We propose a photo album clustering method that groups together similar photos related to the same captured moments, while preserving a linear time organization. This method serves two purposes: (a) provide a user interface which facilitates album browsing for photo selection, and (b) automatically extract the context of each photo, which can be further used for an album-based adaptation of independent image scores. Our proposal is based on a hierarchical clustering approach. To adapt this approach to the task of photo albums clustering, we introduce several necessary modifications, specifically pertaining to the image distance computation and the hierarchical tree construction.

In Section 5.1 we describe our hierarchical clustering method and its corresponding stages, along with the techniques that were explored in this work. In Section 5.2 we evaluate the clustering performance of the introduced techniques, with the aid of the data acquired in our user study described in Chapter 4.

## 5.1 HIERARCHICAL CLUSTERING OF PHOTO ALBUMS

In our proposed approach, the clustering of an album is performed in three steps, as illustrated in Figure 5.1:

(1) As a preliminary step, a temporal grouping is performed, which facilitates subsequent similarity-based clustering in photo albums: the result of it is a set of time windows. This step is described in Section 5.1.1;

(2) An image similarity distance is computed for every pair of images inside each temporal group, as described in Section 5.1.2;

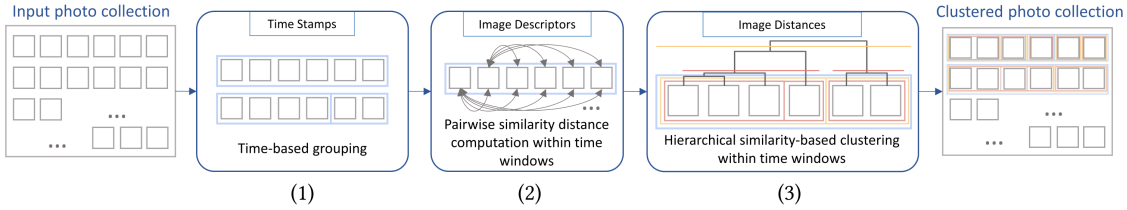



Figure 5.1: A framework of the proposed approach for photo albums clustering.

(3) The computed distances are used to create a hierarchical tree representing the structured relationship between images. The obtained hierarchical trees within each time window are cut at two levels, in order to obtain a two-level clustering: the resulting scene and near duplicate clustering levels represent different level of visual similarity between photos. The hierarchical clustering procedure and its application in photo albums clustering are described in Section 5.1.3.

The output of the above steps, organized into clusters, is similar to the ones produced by users, as demonstrated in Figure 5.9.

### 5.1.1 *Temporal Grouping*

In our approach we aim to preserve the time linearity of the album, i.e. only group the similar images that approximately relate to the same capturing moment and preserve the sequential timeline between different clusters. To better preserve such time linearity and avoid clustering similar scenes from different time occasions together, we begin with a prior time-based grouping of the collection.

The entire album is split into sequential temporal windows using the available photo time stamps, extracted from the EXIF data. If no reliable timestamps are available, the entire album is treated as a single temporal window in the succeeding steps. The temporal windows are computed using the method proposed by Platt et al. [136]. According to it, the temporal window boundary is created at the position $N$ between two images, when the time gap $g_N$ exceeds the average time gap in-between surrounding images, according to the following condition:

$$log(g_N) \geq K + \frac{1}{2d+1} \sum_{i=-d}^{d} log(g_{N+i}), \tag{5.1}$$

where $K = log(17)$ and $d = 10$, as proposed by the authors [136].

The described approach presents a useful property for us: as the window boundaries adapt according to time differences across the image neighborhood, the granularity of time clustering is automatically accommodated to the time span of the album. If, for example, the input album spans one hour, where photos from the same time window are separated by seconds, the different time windows are split apart by minutes. In contrast,

if the input album represents a day trip, where photos in the same time window are separated by minutes, the time window gap can approach one or several hours.

### 5.1.2 *Image Distance Computation*

The obtained temporal windows serve as the basis for further steps of the hierarchical clustering. The hierarchical clustering technique requires a distance metric computed for each image pair. In our case, the distance is based on the visual similarity between two images, which is computed for each image pair inside the temporal window.

We propose three different techniques to compute a visual similarity distance between images. We first explore two variations of a distance based on the descriptors computed with the scale-invariant feature transform (SIFT). As the SIFT-based distance presents certain limitations, we introduce an image distance that relies on a CNN-based image similarity metric.

#### *SIFT-descriptor Based Similarity Distance*

The first distance metric that we explore is based on the SIFT descriptors [99]. The SIFT descriptors are chosen due to their advantage in identifying the image matches even in presence of distortions, such as rotation and scaling, which often appear in series of similar photos. Here, we describe two variations of this distance metric.

Before computing the distance itself, for each pair of images $I$ and $J$, we extract the SIFT keypoints and perform the keypoint matching. We perform the matching using the second closest-match algorithm of Lowe et al. [99]. The potential match is estimated from the distance ratio of Euclidean distances between keypoint descriptors in two compared images:

$$R(a_I, b_{J(I)}) = \frac{||a_I - b_{J(I)}||_2}{||a_I - b_{J'(I)}||_2}, \tag{5.2}$$

where $a_I$ denotes a descriptor in the image $I$, while $b_{J(I)}$ and $b_{J'(I)}$ correspond to the closest and the second-closest descriptors in the image $J$, respectively. The matches exceeding a given threshold are rejected. The use of the second closest match in this case reduces a number of ambiguous matches. Typically, the threshold for $R(a_I, b_{J(I)})$ is set between 0.4 and 0.8 [99], where a higher value leads to more matches detected; we set its value to 0.5.

However, the described matching process is asymmetrical, that is, a match found for a direction $I \to J$ might be rejected for the reverse direction $J \to I$, and vice versa, due to the second closest-match condition (the descriptor match relies not only on the direct distance, but on the distance with other descriptors). Consequently, the number of found matches can vary for two directions of matching. To address this asymmetry, we introduce a measure of pair consistency between matches $P(I, J)$, which is based on the number of matches computed for both directions, $m_{I \to J}$ and $m_{J \to I}$:

$$P(I,J) = \frac{min(m_{I\to J}, m_{J\to I})}{max(m_{I\to J}, m_{J\to I})}. \tag{5.3}$$

In addition, we compute the average number of matches $M(I,J)$:

$$M(I,J) = \frac{m_{I\to J} + m_{J\to I}}{2}. \tag{5.4}$$

Then, our first proposed distance metric between images $I$ and $J$ is defined as an inverse proportion of the number of matches, weighted by the matches consistency:

$$d_{SIFT_1}(I,J) = \frac{10}{M(I,J)\cdot P(I,J)}. \tag{5.5}$$

In Equation 5.5 the nominator is set to 10, to scale the output distance to approximately a $0-1$ range for most images. However, as it is demonstrated later, this scaling proves to be inconvenient in the tree cutting stage of hierarchical clustering, since it requires defining fixed thresholds. To address this issue, we have also proposed another way to define the SIFT-based distance metric:

$$d_{SIFT_2}(I,J) = 1 - \frac{M(I,J)\cdot P(I,J)}{N(I,J)}, \tag{5.6}$$

Now, the right part of the equation defines the similarity between images (a higher value means higher similarity), which we subtract from 1, to compute the opposite measure of image distance (a higher value means lower similarity), since in our further computations we will use a distance measure.

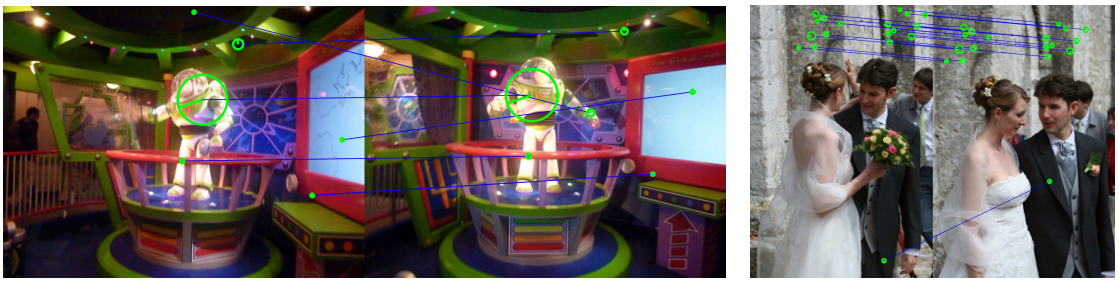
In Equation 5.6, $N(I,J)$ is the average number of detected SIFT descriptors in both images:

$$N(I,J) = \frac{n_I + n_J}{2}, \tag{5.7}$$

which serves as a normalization factor in the distance computation. This way, our distance metric is naturally bounded on the interval $[0,1]$.

*CNN-descriptor Based Similarity Distance*

The traditional SIFT descriptors are generally capable of identifying matches between images even in the presence of distortions and rotations, which is advantageous for the task of matching a series of similar photos. However, their capability is limited in the presence of strong blur, large in-scene rotations or significant pose changes of the objects (see example in Figure 5.2). Such changes can be potentially handled more effectively by the use of image descriptors based on activations of convolutional neural networks, as they are able to provide a more generalized representation of image features. Due to this, in our approach we have also explore a similarity metric based on the CNN-computed global image descriptors.



(a) (b)

Figure 5.2: Examples of SIFT-descriptor based matching limitations. Large transformations, such as camera motion around the object in the photo can cause a reduction of reliable SIFT matches. (a) The matching can be further complicated by the presence of significant motion blur, which affects the image gradients and, in its turn, the keypoints' characteristics. (b) In case of considerable change in human poses, the most prominent keypoints can become hidden.

Radenović et al. [138] proposed a CNN fine-tuning scheme to train global image descriptors for the task of image retrieval (CNNR), where the network can be based on any popular CNN, such as AlexNet [78], VGG [152] or ResNet [54]. In their approach they proposed to use the generalized-mean pooling with learnable parameters for the feature processing in the final layer, instead of a typical fully-connected layer. They trained the network with a siamese approach using pairs of similar and dissimilar images, where the loss was based on the mentioned vectors of pooled features from the final layer. This way, if the input images are similar, the network learned to produce the feature vectors that are close to each other in the feature space. The resulting CNN can be used to produce the feature vector for any given image, which can be further used as a global image descriptor. The approach of Radenović et al. currently achieves state-of-the-art results in the task of image retrieval.

As the applications of global image descriptors are not limited by the image retrieval, we decided to employ their descriptor computation in our clustering approach. In our case, we use the Radenović's version of the fine-tuned ResNet network [54], which provides a feature vector $f$ of dimensionality 2048 for an input image. The vector $f$ is $l_2$-normalized, therefore similarity between two images can be evaluated with the inner product of their corresponding feature vectors. To obtain a distance measure from the CNNR-based feature vectors, we proceed as follows:

$$d_{CNNR}(I, J) = 1 - f_I^T f_J, \tag{5.8}$$

where $f_I$ and $f_J$ represent the feature vectors of compared images $I$ and $J$, and the computed distance $d_{CNNR}$ is also bounded on the interval $[0, 1]$. Similarly to Equation 5.6, we subtract the similarity measure $f_I^T f_J$ from 1, as in our clustering computation we require a distance measure, in which a higher value would mean lower similarity.

### 5.1.3 *Hierarchical Clustering Algorithm*

After the pairwise image distances are computed, we can proceed with the clustering step. For this purpose, we have chosen the agglomerative (also called bottom-up) hierarchical clustering approach, as it presents several important properties that are useful in photo albums clustering. First, to help understand these properties, we explain a general procedure of a hierarchical tree construction and its use for clustering.

*Hierarchical Tree Construction and Clustering Computation*

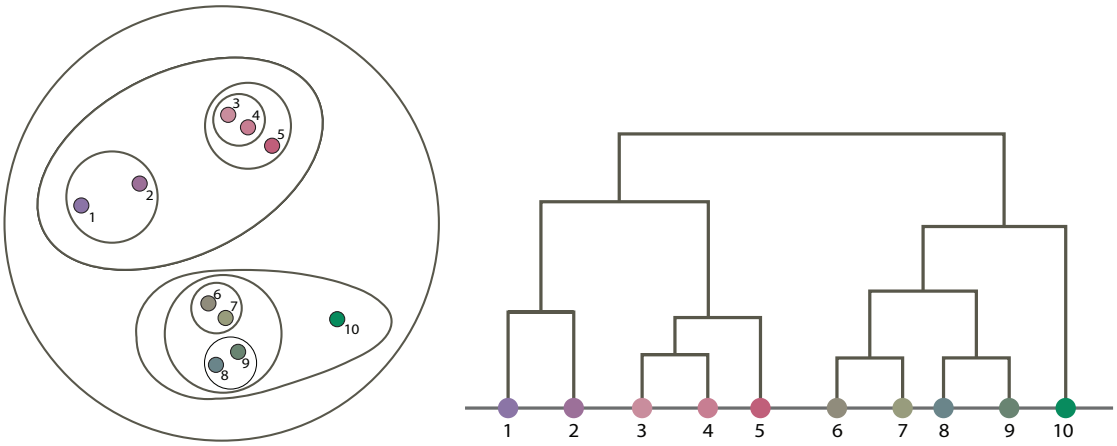



Figure 5.3: Hierarchical tree construction. For easier interpretation, similarity between objects is illustrated by relative distances in the graph and by color similarity. (a) A set of objects grouped into nested clusters, based on pairwise distances. (b) A dendrogram representing a hierarchical tree organization of the obtained nested structure.

In the bottom-up hierarchical clustering, we start with a set of objects, where each object initially belongs to a cluster containing only the object itself. With each step they are merged into new clusters of objects, until all objects are agglomerated into a single large cluster (hence the naming of the clustering method). This agglomeration process is illustrated in Figure 5.3(a). At each step, the most similar objects, namely those with a minimal distance between them, are found and merged into a new object. For example, in Figure 5.3, at the first step it is the objects $\{3\}$ and $\{4\}$ that are joined into a new cluster object. Then, at some later step, this two-elements cluster $\{3,4\}$ will be merged with the cluster $\{5\}$ into another cluster object $\{\{3,4\},\{5\}\}$, thus creating a nested cluster structure.

The merging decision is based on the distance between clusters. At each step, the two clusters with the shortest distance between them are merged. When two compared clusters include multiple objects, the distance is typically defined as the shortest distance over all possible pairs of objects between these two clusters. In our example, to compare distances between two-elements clusters $\{6,7\}$ and $\{8,9\}$, all pairwise distances are

computed, and then the shortest distance is used to decide if these clusters are getting merged at the current step. In this case, the decision would be based on the distance between elements $\{7\}$ and $\{9\}$. This way strategy for cluster merging, based on the shortest distance between objects pairs, is also known as the *single linkage criterion*.

After all clusters are merged together, a nested structure is obtained, where all clusters are hierarchically enclosed into each other, up to the largest cluster containing all elements, which is illustrated as the largest circle in Figure 5.3(a). This hierarchical organization is most easily understood with a dendrogram representation, demonstrated in Figure 5.3(b). Here, the relationships between objects are represented in a hierarchical tree structure, where the height of tree branches reflects the similarity between objects and clusters of objects.

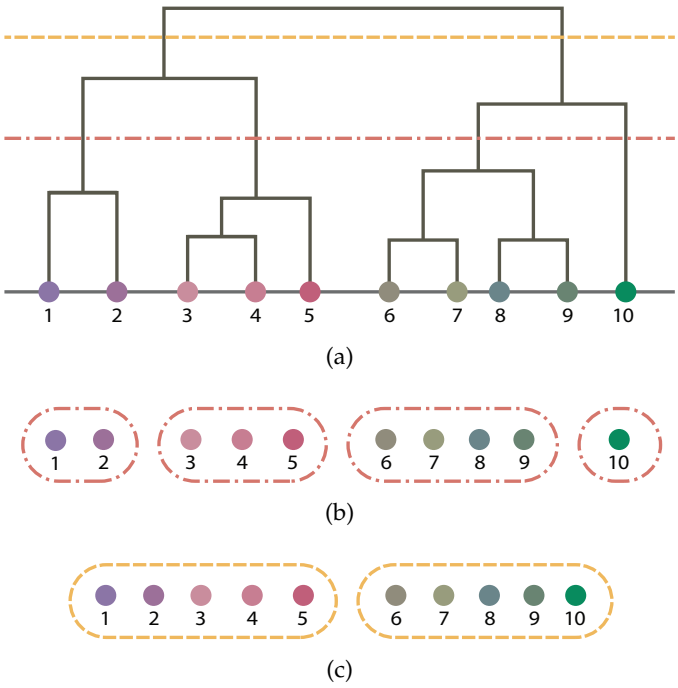



Figure 5.4: Clustering output produced by using different cuts. For easier interpretation, similarity between objects is illustrated by color similarity. (a) A hierarchical tree dendrogram with the lines representing cut levels. The height of tree branches corresponds to cluster similarity. (b) Lower level cut produces clusters of high granularity, where only objects of very similar colors are grouped together. (c) Higher level cut produces clusters of low granularity, where objects are grouped by more general hue similarity.

To produce the desired clustering output, the obtained hierarchical tree can be cut at a predefined level, as shown in Figure 5.4. Given a cut height, the tree is traversed top-down, and all connected nodes below the defined cut are assigned a single cluster. As the height of tree branches corresponds to the distance between objects, we can vary the cut level to obtain the output clusters of different inter-similarity.

Computationally, hierarchical clustering is usually based on a distance matrix, which is used for cluster linkage calculations. In such a matrix, the value at a location $ij$ defines

the distance between the elements $i$ and $j$. A toy example of such a distance matrix based on visual similarity between images is shown in Figure 5.6.

Above, we described a cluster merging procedure known as a *single linkage* clustering, which is based on the shortest distance between objects pairs. An alternative strategy for merging clusters is to consider the farthest distance between elements of the two candidate clusters. This approach is known as the *complete linkage criterion* and allows for a more conservative merging. The difference between the two approaches is illustrated in Figure 5.5. The *single linkage* approach often leads to the chaining phenomenon, when more distant objects are grouped together through a chain of intra-similar objects between them. On the contrary, the *complete linkage* approach tends to find more compact clusters. Both phenomena can be beneficial for photo albums clustering, and both approaches are employed in our method to different effects.

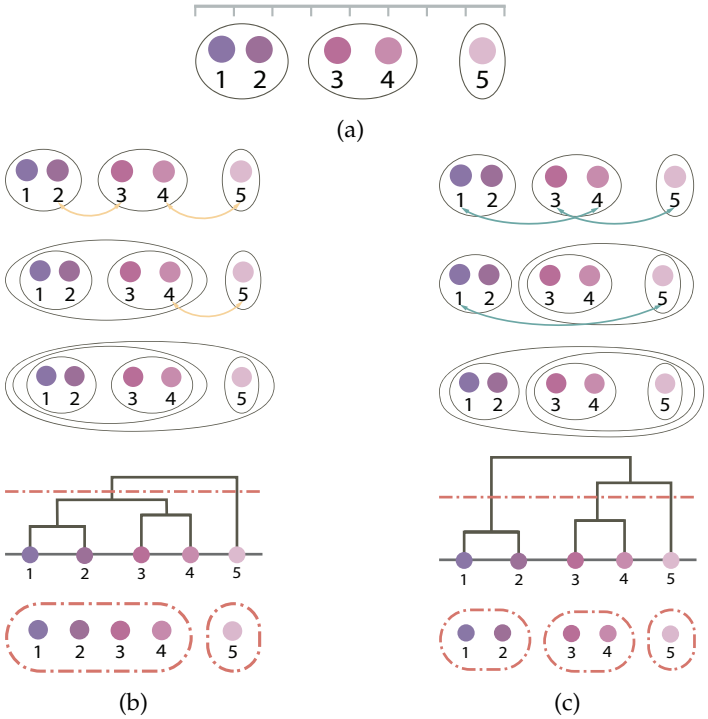



Figure 5.5: Comparison of linkage criteria in hierarchical clustering. By applying different linkage criteria, yet keeping the same tree cut level, we obtain different output clusters. (a) Original state: with a given set of objects and distances between them, both linkages would provide this intermediate partition. (b) *Single linkage*. In the *single linkage* clustering the distance between two clusters is defined as the shortest pairwise distance between elements in each cluster. The chaining phenomenon can be observed in this case, when the objects $\{1\}$ and $\{4\}$ end up in the same cluster, although originally distant from each other. (c) *Complete linkage*. In the *complete linkage* clustering the distance between two clusters is defined as the longest pairwise distance between elements in each cluster. The obtained hierarchical tree have taller branches (indicating larger distances between clusters), and the output clusters are more compact, preserving only close objects together.

### 5.1.4 *Applying Hierarchical Clustering to Photo Albums*

The hierarchical clustering approach, introduced in the previous section, can be applied to the task of clustering photo albums in a straightforward manner. The clustered objects in this case are images, and the distances between them are defined according to their visual similarity, explained in more detail in Section 5.1.2.

We have chosen the hierarchical clustering approach, as it has several useful properties when dealing with photo albums. First, the created hierarchical tree structure represents the similarity connections between objects in an organized manner: in our case, the height of tree branches would reflect the visual similarity between images. Second, by cutting the hierarchical tree at different heights, we can obtain output clusters corresponding to different levels of similarity. As we will see, this way we can create a nested clustering of photos, which represents image similarity on two levels simultaneously. Third, no predetermined number of clusters is required as an input, a common limitation of many clustering techniques, which is crucial when dealing with different photo collections, as no preliminary information on the potential number of clusters is available.

In case of photo albums clustering, we follow the same two-level strategy utilized in our clustering study (Section 4.1). As a result of the photo album clustering, we aim to obtain groups of similar photos, according to two levels of similarity: scene clusters and near-duplicate clusters. Scene clusters represent smaller sub-events within albums that depict the same scene but with considerable changes present. Near-duplicate clusters represent repetitive photos inside the scene clusters, with potentially very small changes.

An example of photo album clustering with the hierarchical approach is shown in Figure 5.6. To compute a similarity distance matrix in this example, we used the SIFT-based distance introduced in Equation 5.5. The distance matrix is used to perform a merging process and construct a hierarchical tree. It is worth noting that in case of photo collections, there will be pairs of images with no matches present between them, thus leading to infinite similarity distance (denoted as X in similarity matrix). In this case, the merging stops when one cluster is achieved or when the distance between all remaining clusters is infinite. Thus, it is possible to obtain multiple, unconnected hierarchical trees, as can be seen in Figure 5.6(c).

To produce the clustering output, the tree is cut as explained earlier. In the task of photo albums clustering, we cut the tree at two different levels, corresponding to two levels of similarity. The $c_{ND}$ cut is used to obtain near-duplicate level clusters, where photos of high similarity are grouped together. The $c_{SC}$ cut is used to obtain the scene level clusters, where photos depicting the same scene overall, but with larger changes in viewpoint, framing and so on, are grouped together. Rather than producing two separate clustering outputs, we create a nested clustering structure, where these outputs are combined. As $c_{ND} < c_{SC}$, the near-duplicate level clusters of higher similarity are enclosed into the scene level cluster. This way, the obtained clustering structure corresponds to the clustering data acquired in the experimental study and reflects the

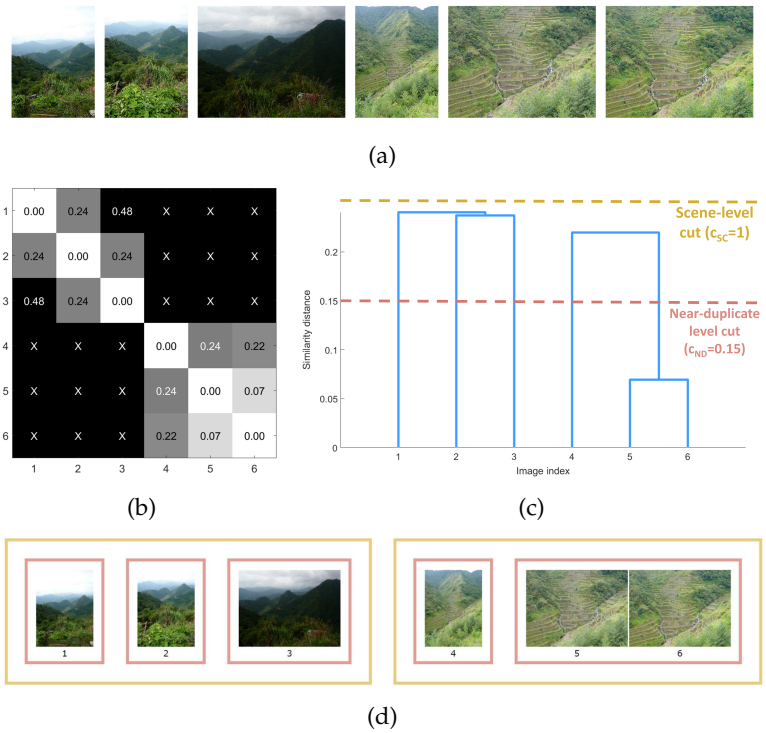



Figure 5.6: A simple example of photo album clustering. (a) Original set of images. (b) Similarity distance matrix between each pair of images. Crosses indicate that no matches between images are found. (c) A hierarchical tree constructed from the similarity distances. The height of tree branches corresponds to cluster similarity. In this case two unconnected trees are obtained, due to the absence of matches (hence infinite distance) between images $\{1,2,3\}$ and images $\{4,5,6\}$. The horizontal lines represent cut levels. (d) Output of clustering. Cuts at specific height have produced a nested clustering structure: images are clustered into two scenes (yellow borders), where the images $\{5,6\}$ from the second scene are grouped into a near-duplicates cluster (red borders).

natural organization made by photographers. An example of this nested hierarchical clustering is demonstrated in Figure 5.6(d).

Although our clustering can directly operate on complete photo collections, a more robust and performant clustering can be achieved when applied within temporal groups of photos. If reliable time stamps are available, we perform a pre-clustering based on temporal information, as explained in Section 5.1.1 (also see the general framework in Figure 5.1), which is then followed by the hierarchical clustering applied within each time window. The simple example demonstrated in Figure 5.6 can be considered as an example of an image set from one time window within a larger collection.

In Equation 5.5 we introduced a non-normalized SIFT distance, which is approximately scaled to a range $0-1$ for most images, but not bounded on a precise interval. For images with few or no matches, the distance could take very large or infinite values. As a result, it would be challenging to define the cut thresholds automatically in this case. We have experimentally found that setting the cut thresholds as $c_{SC} = 1$ and

$c_{ND} = 0.15$ led to a clear separation into scene level and near duplicate level clusters for the collections tested. The example demonstrated in Figure 5.6 is based on the SIFT distance using these predefined thresholds.

Although this strategy can lead to an effective clustering in many cases, the fixed cut values limit the flexibility of the approach. To address this issue and automatically adapt the output clusters, we propose an adaptive cut approach.

### 5.1.5 *Photo Albums Clustering with an Adaptive Cut*

One of the observations arising from the experimental studies presented in Chapter 4 is that when presented with a scene of only few repetitive similar photos, even if the photos are not exactly near-duplicates of each other, users tend to keep them in the same near-duplicate cluster. On the contrary, when presented with many repetitive photos from the same scene, users tend to find small differences between photos and split them into more clusters of higher granularity (an illustration is given in Figure 5.7).

**ALGORITHM 1:** Hierarchical clustering via adaptive cut computation

**Input :** Set of pairwise image distances $d(I, J)$ in album
$D_A = \{d(1,2), d(1,3), ..., d(I_A, J_A)\}$, set of temporal clusters $T = \{t_1, t_2, ..., t_i\}$
**Output:** Set of scene clusters $S = \{s_{i_1}, s_{i_2}, ..., s_{i_j}\}$, set of near-duplicate clusters
$N = \{n_{i_{j_1}}, n_{i_{j_2}}, ..., n_{i_{j_k}}\}$

*Compute scene and near-duplicate level clusters within each temporal cluster*:
$C_{SC} = \overline{D_A}$;
**foreach** *temporal cluster* $t_i$ *in* $T$ **do**
  $D_{t_i} = \{d(1,2), d(1,3), ..., d(I_{t_i}, J_{t_i})\}$;
  $tree(t_i) \leftarrow$ `ConstructHierarchicalTree(`$D_{t_i}$`);`
  $S_i = \{s_{i_1}, s_{i_2}, ..., s_{i_j}\} \leftarrow$ `PerformCutIntoSceneClusters(`$tree(t_i)$, $C_{SC}$`);`
  **foreach** *scene cluster* $s_{i_j}$ *in* $S_i$ **do**
    $D_{s_{i_j}} = \{d(1,2), d(1,3), ..., d(I_{s_{i_j}}, J_{s_{i_j}})\}$;
    $tree(s_{i_j}) \leftarrow$ `ConstructHierarchicalTree(`$D_{s_{i_j}}$`);`
    $C_{ND} = \overline{D_{s_{i_j}}}$;
    $N_{i_j} = \{n_{i_{j_1}}, n_{i_{j_2}}, ..., n_{i_{j_k}}\} \leftarrow$ `PerformCutIntoNearDuplicateClusters(`$tree(s_{i_j})$, $C_{ND}$`);`
  **end**
**end**

To simulate this user behavior, we define the cut thresholds in an adaptive manner. In our adaptive approach, the average distance between images is taken into account to define the cut thresholds, as it already implicitly encodes how similar are images within album. The outline of main steps is given in Algorithm 1. First, the scene clustering cut threshold $C_{SC}$ is defined as the mean value $\overline{D_A}$ of the image distances in the entire album. Then, a hierarchical tree is constructed inside each temporal window $t_i$, and the scene clustering cut $C_{SC}$ is applied on this tree to create separate scene clusters. At the second stage, each scene cluster $s_{i_j}$ is further split into near-duplicate clusters in a similar manner, where the cut $C_{ND}$ is defined as the mean value $\overline{D_{s_{i_j}}}$ of the image

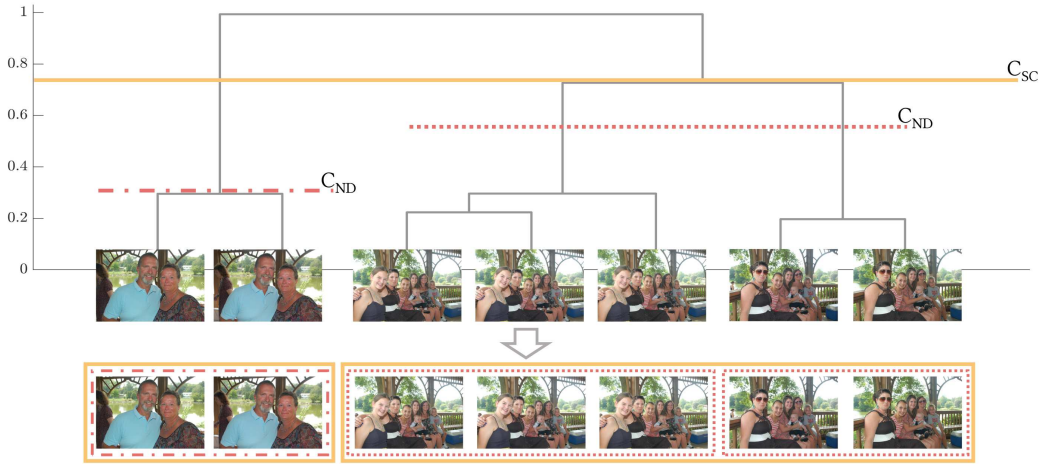



Figure 5.7: Example of adaptive granularity in users' decisions and a corresponding clustering cut adaptation. Photos from the second scene (five photos on the right) contain many repetitions, so users tend to find possible differences between them, creating two separate near-duplicate clusters. Our adaptive cut approach performs in a similar manner. After creating a distance-based tree, this tree is cut on two levels. First, using the collection-defined threshold $C_{SC}$ to obtain division into scene clusters. Second, the near-duplicate clustering threshold $C_{ND}$ is computed within each scene, to further divide it into the clusters of near-duplicate photos.

distances in the processed scene $s_{i_j}$. An illustration of the described process can be found in Figure 5.7. This way, we can obtain an effect similar to users' actions, where the granularity of their clustering depends on the perceived differences within the images.

This approach can be applied using either the SIFT-based distance or the CNN-descriptor distance. In either case the distances are bounded on the interval $[0, 1]$. Even in the absence of matches, the mean values of image distances $\overline{D_A}$ and $\overline{D_{s_{i_j}}}$ will also be bounded on the same interval.

In addition, considering possible instabilities of SIFT distance in lack of reliable image matches, for the SIFT-based approach with an adaptive cut we have defined the scene cut threshold as $C_{SC} = \overline{D_A} - \sigma_A/2$, when the time stamps are not available.

CLUSTERING LINKAGE IN PHOTO ALBUMS Above, we explained a difference between *single linkage* clustering and *complete linkage* clustering (see Figure 5.5). The initial experiments, based on non-normalized SIFT distance and fixed thresholds, were conducted using only the *single linkage* criterion in the entire tree construction. When experimenting with an adaptive cut approach, we have found that both linkage criteria might be beneficial in albums clustering. As shown in Figure 5.5, the shortest distance principle used in *single linkage* can lead to a chaining phenomenon, when more distant clusters are grouped together, through a chain of intra-similar elements between them. We have observed that this phenomenon is beneficial for the SIFT-based clustering on the scene level, as often the lower number of matches may be not sufficient to

link images with a smaller visual overlap. However, for the clustering based on CNN features, the higher abstraction level of these descriptors can lead to the over-merging of scene clusters. In this case, the *complete linkage* is more applicable, since it is based on the longest distance between two compared clusters, thus creating more compact clusters.

With this in mind, for the scene tree construction in our adaptive approach, we use the *single linkage* method in case of SIFT descriptors and the *complete linkage* method in case of CNN descriptors. For the near-duplicate tree construction inside each scene, we use the *complete linkage* for both types of descriptors. In this case, the compactness property aids in our adaptive process, where users have the tendency to find clusters of higher granularity on the near-duplicate level.

For the reference, we provide an overview of proposed clustering approaches in Table 5.1, where we outline main differences and variations between them.

| | | | | Clustering linkage | |
|---|---|---|---|---|---|
| | Similarity distance | Temporal pre-clustering | Tree cut | Scene level | Near-duplicate level |
| SIFT | $d_{SIFT_1}$ | ✗ | fixed | single | single |
| Time-SIFT | $d_{SIFT_1}$ | ✓ | fixed | single | single |
| Adaptive SIFT | $d_{SIFT_2}$ | ✗ | adaptive | single | complete |
| Adaptive Time-SIFT | $d_{SIFT_2}$ | ✓ | adaptive | single | complete |
| Adaptive CNNR | $d_{CNNR}$ | ✗ | adaptive | complete | complete |
| Adaptive Time-CNNR | $d_{CNNR}$ | ✓ | adaptive | complete | complete |

Table 5.1: Overview of the proposed clustering approaches that were analyzed in our experiments.

## 5.2 RESULTS AND DISCUSSION

To assess the performance of the proposed clustering solutions, we take advantage of the results of the clustering user study described in Section 4.2. We compute the performance of each clustering method similarly to the computation of users' clustering agreement. The Adjusted Rand Index, which provides a measure of similarity between two data clusterings, is computed between the computed clustering partitions and the partitions provided by each user, and then the average value is computed.

First, we compare the average performance of our clustering approaches with two state-of-the-art approaches, as shown in Table 5.2. For this purpose, we have selected two clustering methods that, similarly to our method, do not require a prior estimate of the potential number of clusters and thus can operate completely automatically. The first is the time-based clustering by Platt et al. [136], which we have described in Section 5.1.1. The second is the affinity propagation clustering proposed by Frey et al. [40]. Their approach iteratively searches for the most representative exemplars, while the associated data points are used to define cluster boundaries. This approach was applied to the task of clustering images of human faces and to perform image data summarization [31]. Since their approach works with any similarity measure, we employ it using the earlier

described CNNR-descriptors based similarity, as it has shown the best performance in our tests. Additionally, as each of these two methods does not provide a specific multi-level structure and produces one clustering output, we evaluate this output both against our scene and near-duplicate user partitions.

| | User Agreement | Time-based clustering [136] | Affinity propagation [40] | SIFT | Time-SIFT | Adaptive SIFT | Adaptive Time-SIFT | Adaptive CNNR | Adaptive Time-CNNR |
|---|---|---|---|---|---|---|---|---|---|
| Scene level | 0.658 | 0.372 | 0.459 | 0.471 | 0.533 | 0.481 | 0.551 | 0.573 | 0.580 |
| Near-duplicate level | 0.678 | 0.116 | 0.407 | 0.411 | 0.411 | 0.606 | 0.629 | 0.610 | 0.624 |

Table 5.2: Average *ARI* performance of the analyzed clustering methods. For our proposed methods, the prefix *Time-* indicates a preliminary temporal clustering applied before the similarity-based hierarchical clustering.

The time-based clustering provides the lowest performance compared to the other methods tested. It is not unexpected, since the temporal information alone is generally not sufficient for such task. The affinity propagation clustering based on visual similarity shows reasonable performance for both clustering levels, close to our proposed SIFT-based hierarchical clustering, when temporal information is not used. Nevertheless, as this method was not designed for multi-level clustering, its performance is lower than our proposed approaches.

The detailed per-album results are given in Table 5.3 for clustering without time information available and in Table 5.4 for temporal-aided clustering. Also, an overview of the results for the entire dataset is shown in Figure 5.8. We estimate the clustering performance for three methods and their temporal-aided variations: SIFT-descriptors based clustering with fixed cut thresholds (*SIFT* and *Time-SIFT*), SIFT-descriptors based clustering with adaptive cut thresholds (*Adaptive SIFT* and *Adaptive Time-SIFT*) and CNNR-descriptors based clustering with adaptive cut thresholds (*Adaptive CNNR* and *Adaptive Time-CNNR*). The prefix '*Time-*' signifies the case when the temporal information is available and the temporal pre-clustering is performed first. The best obtained results, by the *Adaptive Time-CNNR* approach, are demonstrated in Figure 5.9.

In Table 5.3 and Figure 5.8(a) we consider a test case when no time information is available, such that no temporal pre-clustering can be performed, and the entire collection is clustered directly. In this case, the performance of the SIFT-based methods is largely lower for the scene level, comparing to the time-aided clustering, confirming the limited capability of SIFT-based matching to find matches of higher abstraction, which are necessary for scene level matches. At the same time, no significant deterioration in performance is observed for the near-duplicate level in absence of time information. A performance of the CNNR-based method does not noticeably change between two cases. It suggests that the CNNR-based clustering is more robust overall and is able to provide relevant results even when no time information available (although the time linearity of the photo album might be not preserved in this case).

In general, it can be seen that for both cases the approaches based on adaptive cut largely outperform SIFT methods based on fixed cut. It is also interesting to note that the Adaptive Time-SIFT generally outperforms the Adaptive Time-CNNR method for the albums focused on people's photos, while the CNNR-based method shows better

| | User Agreement | | SIFT | | Adaptive SIFT | | Adaptive CNNR | |
|---|---|---|---|---|---|---|---|---|
| | Scene level (SC) | Near-duplicate level (ND) | SC | ND | SC | ND | SC | ND |
| Family event 1 | 0.481 (±0.19) | 0.645 (±0.15) | 0.338 (±0.18) | 0.269 (±0.09) | 0.301 (±0.11) | 0.590 (±0.10) | 0.421 (±0.16) | 0.465 (±0.07) |
| Family event 2 | 0.802 (±0.14) | 0.715 (±0.12) | 0.880 (±0.12) | 0.735 (±0.14) | 0.761 (±0.10) | 0.623 (±0.12) | 0.660 (±0.07) | 0.614 (±0.07) |
| Family event 3 | 0.732 (±0.13) | 0.625 (±0.11) | 0.366 (±0.09) | 0.143 (±0.06) | 0.558 (±0.12) | 0.494 (±0.07) | 0.585 (±0.13) | 0.520 (±0.09) |
| Travel album 1 | 0.682 (±0.24) | 0.922 (±0.03) | 0.422 (±0.10) | 0.916 (±0.04) | 0.655 (±0.16) | 0.886 (±0.02) | 0.742 (±0.18) | 0.930 (±0.02) |
| Travel album 2 | 0.592 (±0.11) | 0.561 (±0.11) | 0.495 (±0.05) | 0.241 (±0.11) | 0.389 (±0.11) | 0.547 (±0.05) | 0.537 (±0.08) | 0.517 (±0.12) |
| Travel album 3 | 0.657 (±0.16) | 0.598 (±0.14) | 0.325 (±0.06) | 0.162 (±0.07) | 0.225 (±0.04) | 0.493 (±0.11) | 0.492 (±0.07) | 0.616 (±0.11) |
| Average | 0.658 (±0.16) | 0.678 (±0.11) | 0.471 (±0.10) | 0.411 (±0.09) | 0.481 (±0.11) | 0.606 (±0.08) | 0.573 (±0.12) | 0.610 (±0.08) |

Table 5.3: Per-album user agreement and performance of the analyzed clustering methods without temporal pre-clustering. *SC* denotes scene level clustering results, *ND* denotes near-duplicate level clustering results. The first value represents the mean ARI across all users. The second value (given in parentheses) represents the standard deviation of the ARI.

| | User Agreement | | Time-SIFT | | Adaptive Time-SIFT | | Adaptive Time-CNNR | |
|---|---|---|---|---|---|---|---|---|
| | Scene level (SC) | Near-duplicate level (ND) | SC | ND | SC | ND | SC | ND |
| Family event 1 | 0.481 (±0.19) | 0.645 (±0.15) | 0.338 (±0.18) | 0.269 (±0.09) | 0.546 (±0.20) | 0.716 (±0.12) | 0.470 (±0.15) | 0.567 (±0.08) |
| Family event 2 | 0.802 (±0.14) | 0.715 (±0.12) | 0.880 (±0.12) | 0.735 (±0.14) | 0.880 (±0.12) | 0.605 (±0.10) | 0.710 (±0.07) | 0.624 (±0.11) |
| Family event 3 | 0.732 (±0.13) | 0.625 (±0.11) | 0.366 (±0.09) | 0.143 (±0.06) | 0.600 (±0.15) | 0.537 (±0.07) | 0.495 (±0.12) | 0.464 (±0.05) |
| Travel album 1 | 0.682 (±0.24) | 0.922 (±0.03) | 0.795 (±0.25) | 0.916 (±0.04) | 0.795 (±0.25) | 0.889 (±0.01) | 0.742 (±0.18) | 0.930 (±0.02) |
| Travel album 2 | 0.592 (±0.11) | 0.561 (±0.11) | 0.495 (±0.05) | 0.241 (±0.11) | 0.481 (±0.14) | 0.617 (±0.08) | 0.569 (±0.09) | 0.545 (±0.11) |
| Travel album 3 | 0.657 (±0.16) | 0.598 (±0.14) | 0.325 (±0.06) | 0.162 (±0.07) | 0.006 (±0.00) | 0.414 (±0.05) | 0.492 (±0.07) | 0.616 (±0.11) |
| Average | 0.658 (±0.16) | 0.678 (±0.11) | 0.533 (±0.13) | 0.411 (±0.09) | 0.551 (±0.14) | 0.629 (±0.07) | 0.580 (±0.11) | 0.624 (±0.08) |

Table 5.4: Per-album user agreement and performance of the analyzed clustering methods, when the temporal pre-clustering is performed. *SC* denotes scene level clustering results, *ND* denotes near-duplicate level clustering results. The first value represents the mean ARI across all users. The second value (given in parentheses) represents the standard deviation of the ARI.

performance in landscape-focused albums. A possible explanation is that the CNNR descriptors provide more abstract image representations, and, in some cases, images of general resemblance might match (e.g. see an example in the last row of *Family Event 1* in Figure 5.9).

We have also observed that *Travel album 3* presents large difficulties for SIFT-based methods. This album is particularly challenging, since it does not contain reliable time information to perform the first pre-clustering and facilitate further similarity-based clustering steps. Also, it contains numerous low quality blurred shots and it demonstrates a number of captured objects or landmarks with the large viewpoint changes present. In this case, the CNNR descriptors generally achieve better matching performance.

Overall, we can observe that the proposed automatic approaches are capable of modeling user clustering decisions and, effectively, can be used in clustering-dependent applications. Based on the hierarchical clustering method, we have achieved a flexible solution that can find visually similar photos within photo albums and group them with an adaptive level of granularity. At the same time, no preliminary information on expected number of clusters is required. The output clustering provides a convenient

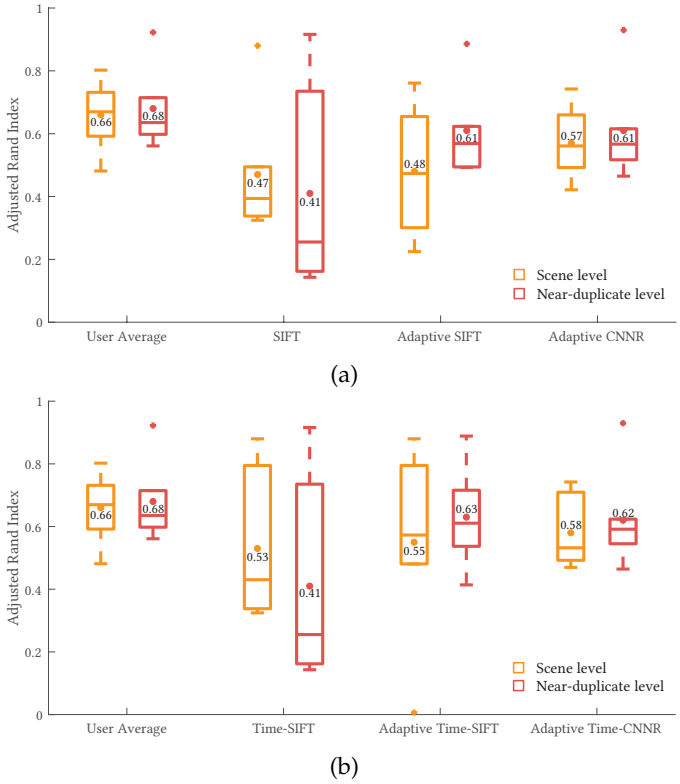



Figure 5.8: Performance of clustering methods. The box plots represent *ARI* distributions across the albums, where a dot and the associated value indicate the average *ARI* for each method. (a) Performance with no time information available. (b) Performance with an aid of preliminary temporal clustering.

visualization and navigation structure. In the next chapter we explore how the computed clustering structure can be used to improve photo assessment within photo collections.

### *Clustering Visualization*

As an additional point of interest, we can discuss the created clustering visualization. The nested clustering structure created with our approach can be used to facilitate browsing of a photo collection and navigation within it.

Previous approaches that rely on hierarchical clustering for organizing photographs present images in a tree browsing structure, where one representative image replaces the contents of the cluster [32, 77]. However, in the typical photo viewing and managing software (e.g. Picasa, Lightroom) flat representations are preferred, with flags and other identifiers used to label photographs.

Our clustering solution provides album visualizations similar to the outputs of our user study. With our approach, an input photo album is automatically clustered and then a flat representation is constructed, where the enclosing borders can be used to identify clusters at different context levels (temporal, scene-level, near-duplicate), as shown in Figure 5.9. This simple visualization can help the user to quickly understand

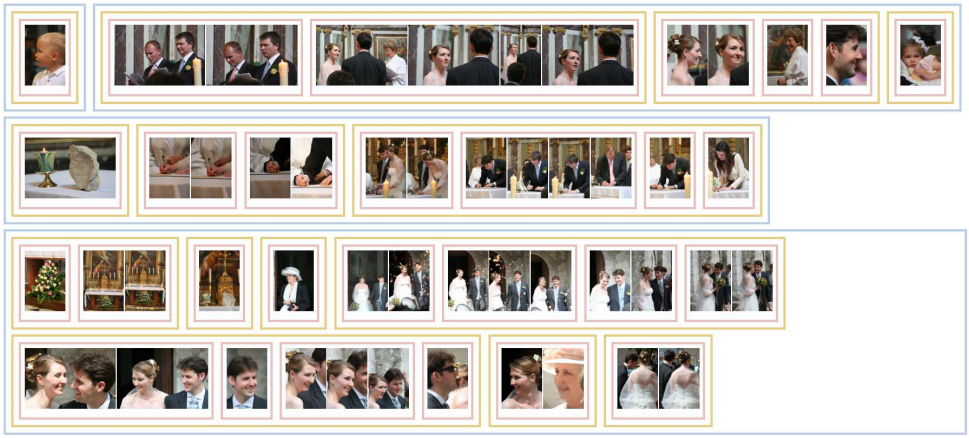
the structure of their collection and perform their selections, without having to navigate several levels within a tree structure.

## 5.3 SUMMARY

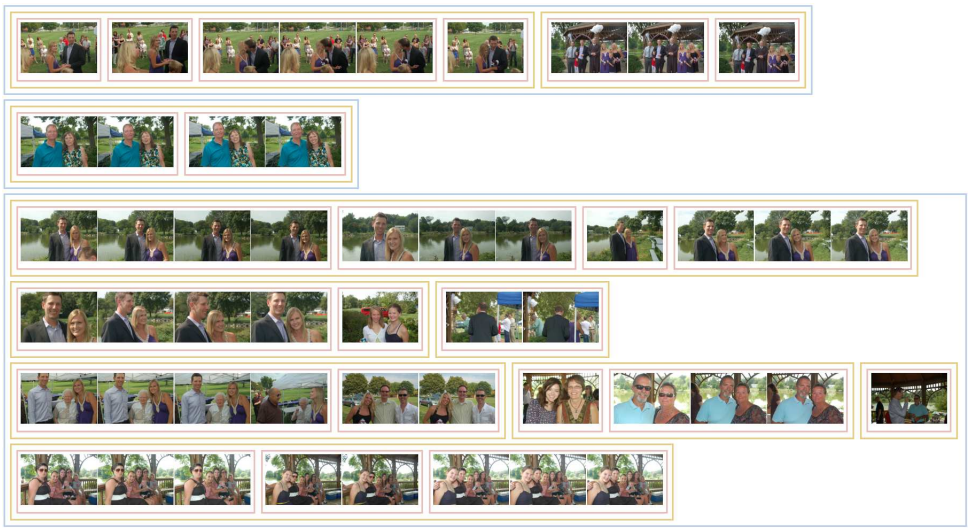
The ability to model photo album clustering with an automatic approach can be beneficial for different tasks, such as organization of user collections for simpler browsing experience, or extraction of the photo context consisting of similar related photos. For this purpose, we proposed a hierarchical clustering solution based on image similarity.

The hierarchical clustering approach does not require an input information on expected number of clusters, and our proposed adaptive cut technique removes the necessity for predefined cut thresholds. After testing different visual descriptors and manners of clustering, we have found that the most robust performance is achieved by using the CNNR descriptors together with an adaptive cut technique, where a temporal pre-clustering applied before. At the same time, the SIFT descriptors provide a comparable performance in absence of strong artifacts or viewpoint changes.

Our results have shown that the proposed solution can cluster photo albums in an automatic manner and achieve high agreement with user partitions. As we succeed to model user clustering decisions, the next step would be to use them in a score adaptation for image assessment within photo albums.



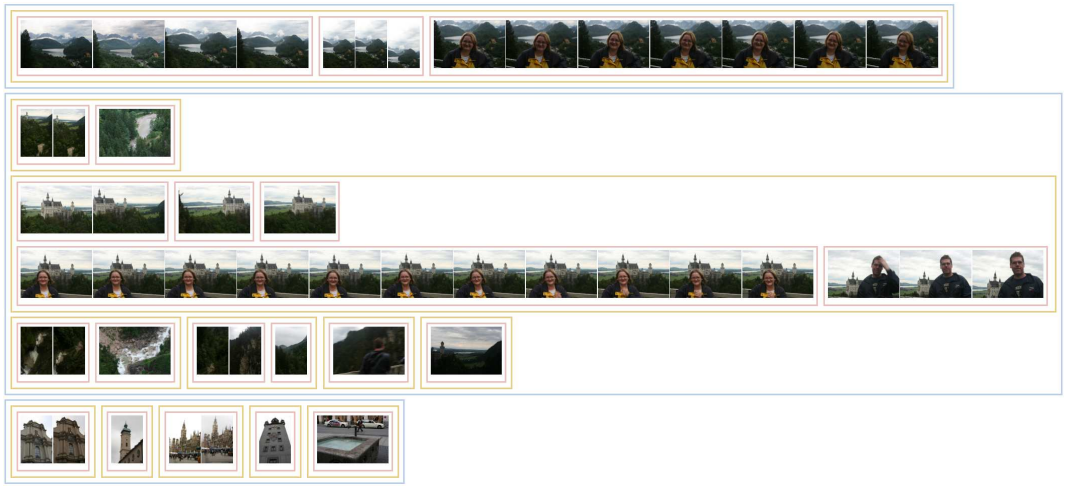
(a) Family event 1



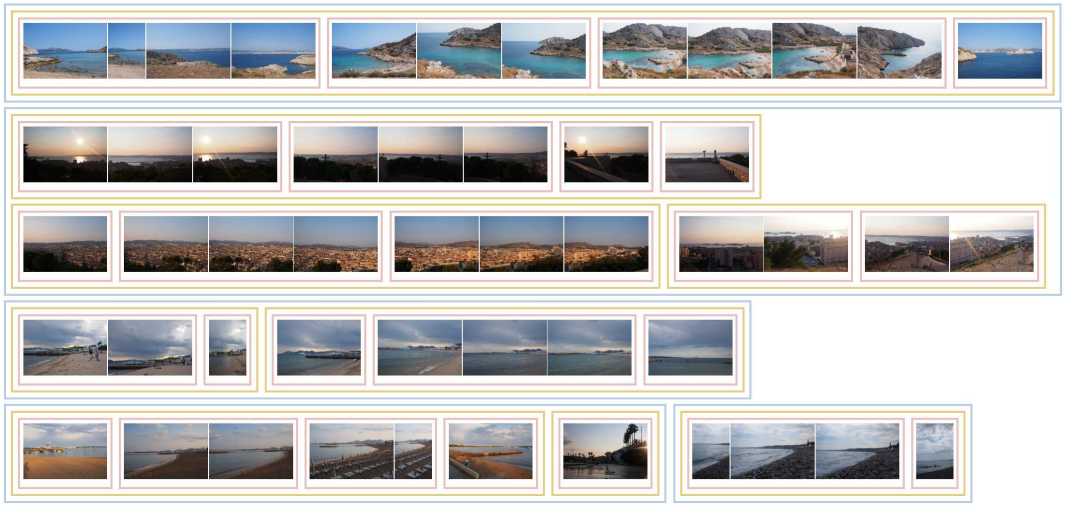
(b) Family event 2

(c) Family event 3

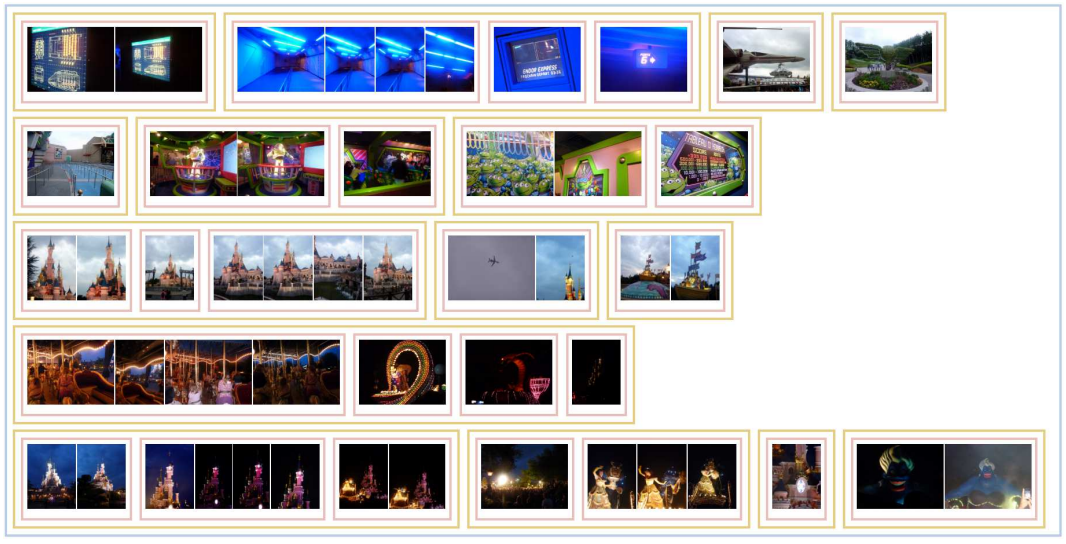
Figure 5.9: Demonstration of clustering results by our proposed method (the results of Adaptive Time-CNNR approach are demonstrated). Blue borders indicate temporal clusters, yellow borders indicate scene level clusters, and red borders indicate near-duplicates clusters.



(d) Travel Album 1



(e) Travel Album 2



(f) Travel Album 3

Figure 5.9: Demonstration of clustering results by our proposed method (the results of Adaptive Time-CNNR approach are demonstrated). Blue borders indicate temporal clusters, yellow borders indicate scene level clusters, and red borders indicate near-duplicates clusters.

# 6

# MODELING USER BEHAVIOR IN IMAGE ASSESSMENT IN PHOTO ALBUMS

When assessing or selecting photos within an album, users consider the implicit context of each photo, which can only be determined when viewing each photo within the surrounding content in the photo album, as we observed in the user study presented in Section 4.2. The clustering approaches described in the previous chapter offer an automatic way of extracting this context, in a manner consistent with how users organize photo collections (Section 5.2). In this chapter, we go a step further: we aim to improve the automatic photo assessment within photo albums, by representing a photo context using the computed clustering.

The goal of our approach is to simultaneously organize and assess the quality of photographs belonging to a photo collection, by considering each image not independently but in its surrounding context. The starting point of our approach is any photo quality metric or score, computed on each image of the collection independently. To adapt such an independent photo score to the context of each photo, we consider several approaches and rely on statistics extracted from the computed multi-level clustering. The obtained adapted photo score can be used to re-assess photos within albums. Further, the scores can also be transformed into selection decisions, to keep only a subset of the best photos.

The framework of our approach is demonstrated in Figure 6.1. We base our approach on adapting an image score computed with a conventional photo assessment method, which does not take context into account (a). The hierarchical clustering is used to define the similarity-based context of each image and create the intuitive collection visualization through the context-aware representation (b). The per-image independent score computed initially is then re-estimated using the information derived from the clustering (c). The result of context adaptation can be directly used as an output score or can be employed to obtain a selection label.

We proceed with an explanation of the adaptation approach as follows. In Section 6.1, we present the independent photo assessment methods used in our adaptation experiments and we further illustrate a demand for context-aware assessment. In Section 6.2 we describe our proposed score adaptation approaches based on clustering information. In Section 6.3 we demonstrate the performance of the proposed adaptation approaches and discuss the obtained results.

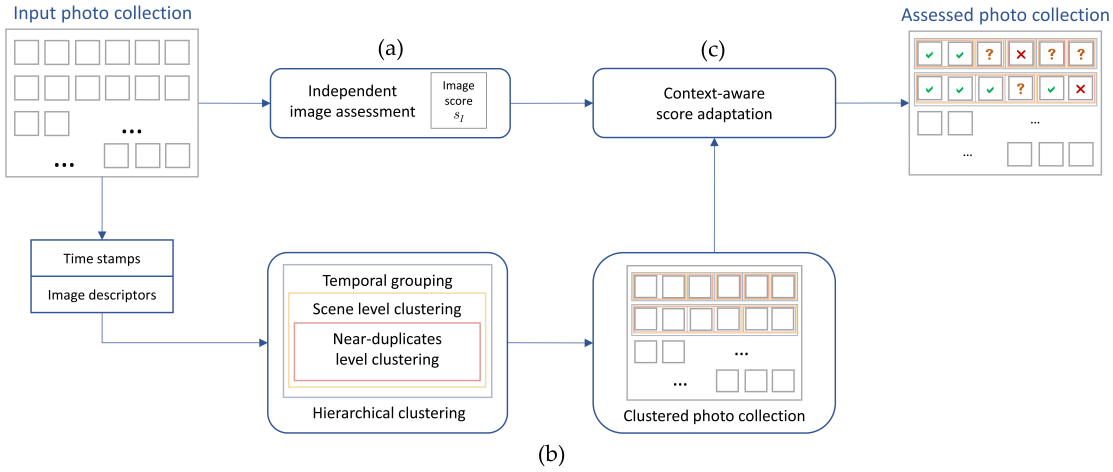



Figure 6.1: An overview of our context-aware framework. (a) An independent image scoring is performed based on the selected assessment criteria. (b) Using photo collection time stamps and computed image descriptors, photos are clustered in a hierarchical manner. Obtained clustering is used to define the context for each photo and create the context-aware visualization of the collection. (c) An independent image score is re-estimated for the three context levels, leading to a final scoring and labelling of the photo collection.

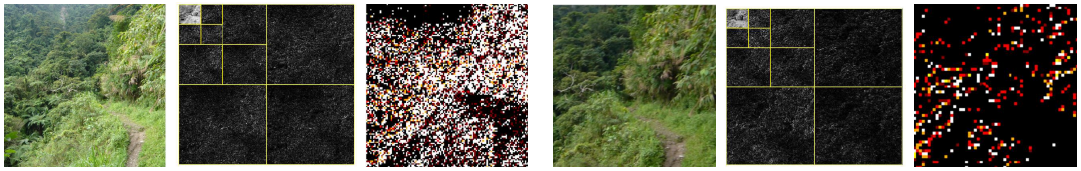
## 6.1 INDEPENDENT IMAGE ASSESSMENT IN PHOTO ALBUMS

In accordance with our experimental studies, we perform the automatic photo assessment in two manners: using a single simple criterion (image sharpness in this case), and based on general aesthetics assessment without predefined criteria. Rather than defining completely new metrics for this purpose, we have decided to adopt existing state-of-the-art techniques. In this section we introduce these techniques and also demonstrate some of their limitations. Apart from the described techniques, our method can be also adapted to other techniques providing independent photo evaluations.

### *Photo Assessment Based on Image Sharpness*

Existing techniques are capable of assessing sharpness or blur in individual images, but they cannot be easily adapted to the nature of particular collections. For example, a collection captured in difficult illumination conditions by an amateur photographer might exhibit a large number of blurred photos. However, typical blur assessment methods are pre-trained on a wide range of photos, thus they can underestimate and reject many photos in a low-quality collection.

In our sharpness-based method, we adopt the wavelet-based blur detection by Tong et al. [165]. In their approach, they benefit from the multi-resolution analysis ability of the wavelet transform and its property to discriminate different types of edges. The Haar wavelet transform is applied to an image until the decomposition level 3, and edge



(a) (b)

Figure 6.2: Demonstration of an edge map computation in the image sharpness evaluation of Tong et al. [165]. An input image is processed with the Haar wavelet transform, and then the edge maps from three scales are combined into the final edge map. Here, the edge map structures are color coded: white and orange colors indicate sharp *Dirac* and *Astep* structures, respectively, and red and dark red colors indicate blurred *Roof* and *Gstep* structures, respectively. (a) A typical sharp image. (b) A typical blurred image. It can be noted that sharp edge structures tend to disappear in presence of blur.

maps $E_i$ are generated on three scales, using the pyramid images $LH$ (vertical details), $HL$ (horizontal details), and $HH$ (diagonal details):

$$E_i = \sqrt{LH_i^2 + HL_i^2 + HH_i^2} \quad (i = 1, 2, 3). \tag{6.1}$$

Non-maximum suppression is applied to process all maps into the equivalent scale, using scale-dependent local windows. Depending on the edge properties across scales, each pixel is assigned an edge or non-edge value, and each edge pixel is further assigned one of four edge structure types: *Dirac*-structure, *Roof*-structure, *Astep*-structure and *Gstep*-structure, as illustrated in Figure 6.2. In particular, *Dirac*-structure and *Astep*-structure pixels are considered sharp, as they tend to disappear in presence of blur), and their total number $N_{da}$, along with the total number of edge pixels $N_{edge}$ is used to compute a final sharpness score:

$$s_I = \frac{N_{da}}{N_{edge}}. \tag{6.2}$$

The obtained sharpness score is used as an individual image score in our adaptation approach.

*Photo Assessment Without Predefined Criteria*

Although it is common to assess only the technical quality of images, recent photo assessment methods have expanded their focus, evaluating the overall aesthetics. This way, in addition to the image quality, different aspects defining the attractiveness of a photo are also taken into account. These more general photo assessment models represent average user preferences, by training on a large number of photos, where each image is assessed independently.

Given the promising advances in this field with recent deep learning based methods, we opt for evaluating the following CNN-based methods, which were mentioned in Chapter 2:

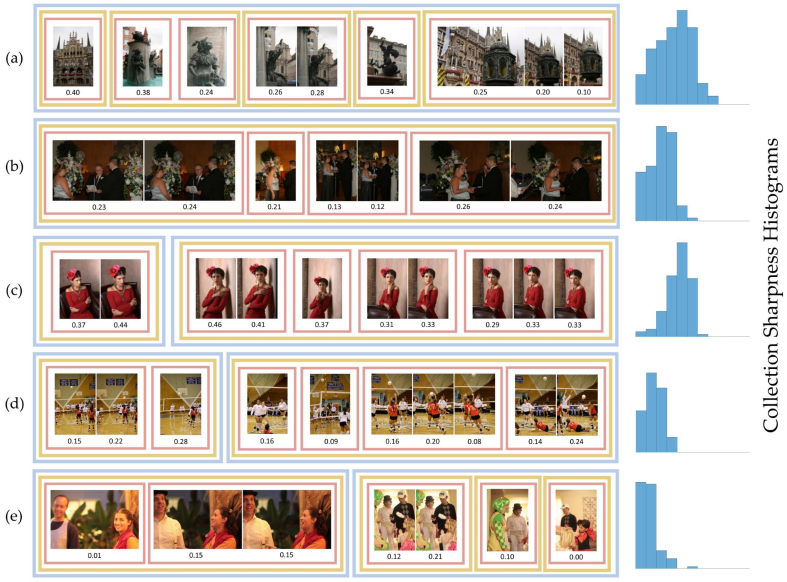



Figure 6.3: Demonstration of sharpness scores [165] in the collections from our sharpness-based selection study. We also demonstrate examples of the proposed automatic clustering. Blue borders indicate temporal clusters, yellow borders indicate scene level clusters, and red borders indicate near-duplicates clusters. (a) Travel collection. (b) Wedding ceremony. (c) Professional session. (d) Sport event. (e) Halloween party. On the right: histograms of sharpness values corresponding to all photos in the collections containing presented clusters.

- The approach by Kong et al. [75], which has recently shown the state-of-the-art performance in the independent image aesthetics assessment. Using the AlexNet [78] as a base, they fine-tuned it with a joint loss strategy: an Euclidean regression loss was complemented with a pairwise ranking loss in a Siamese network and an additional Euclidean loss based on per-attribute activations.

  The per-attribute activations of their network provide the encoding of 11 informative attributes, such as *Color harmony*, *Depth of field*, *Lighting*, *Object emphasis*. The use of these activations in the training is aimed to regularize the weights and give higher relevance to the attributes that might be more important in an input image. Also, their proposed AADB dataset, used in the training, is supposed to contain a more balanced distribution of professional and consumer photos, in comparison with the AVA dataset [122].
- The method proposed by Jin et al. [65], which aims to predict distributions of user scores, along with a total image score. Their approach is based on the VGG16 architecture [152] and the AVA dataset.

  To compensate for the imbalance of scores in the AVA dataset, they introduced a weighted loss function, which increases the importance of the under-represented data (such as images with very low or high scores). In addition, the authors demonstrated an application of their method for image composition enhancement,

where the best crop is searched for an input photo. As this application deals with composition assessment of highly similar photos, we could expect this approach to cope better with similar repetitive images often present in photo albums.

- The NIMA method proposed by Talebi et al. [160], which was designed to estimate both technical image quality and aesthetic attractiveness of an image. In our experiments we use their architecture based on the MobileNet [57] network.

  In their approach, Talebi et al. also predict a distribution of user scores, which then can be converted to a total image score. A distribution is predicted using a 10-neuron layer representing a histogram of scores. To compute the loss function in this case, they use the Earth Mover's Distance [86] between the cumulative distribution functions of a ground truth and a prediction. Their network has also shown an application in visual enhancement operations, such as the parameters tuning for tone and contrast enhancement techniques.

*Limitations of Independent Assessment in Photo Albums*

As we have found earlier in Chapter 3, photo selection is a highly subjective task, where, depending on the type of images and collections assessed, different users may have very different opinion and approaches. As such, automating this task is a daunting challenge. In our literature review (Section 2.5), we have introduced the problem of context gap in photo assessment: when assessing the quality or aesthetics of photos, we can be influenced in our decisions by the context provided by the collection where the photos belongs to. Although a wealth of approaches exists for rating or assessing images, only few of them take the origin and surrounding context of an assessed photo into account.

To assess the performance of general image assessment methods when faced with complete photo albums, we compare their assigned image scores with our experimental findings. We further illustrate the need of context for effective image assessment with some representative examples.

We compare the per-image ranking scores computed by the three analyzed methods (Kong et al. [75], Jin et al. [65] and NIMA method [160]) against the preference scores from our study on selection without predefined criteria (Section 3.2). To estimate the performance of each analyzed method, we compute the Pearson correlation coefficient $r(S, U)$ between the method's photo assessment scores $S$ and the user preference scores $U$:

$$r(S,U) = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{S_i - \mu_S}{\sigma_S} \right) \left( \frac{U_i - \mu_U}{\sigma_U} \right), \tag{6.3}$$

where $\mu_S$ and $\sigma_S$ are the mean and standard deviation of computed scores, respectively, and $\mu_U$ and $\sigma_U$ are the mean and standard deviation of user preferences. A correlation has a value between $-1$ and $+1$, where an absolute value of 1 indicates a perfect linear relationship, and a correlation close to 0 indicates no linear relationship.

This way, even if the scores from each method are not computed on the same scale, we can estimate the extent of correlation between the computed independent image scores and the user evaluations. The computed correlation values are shown in Table 6.1.

| | Travel album 2 | Travel album 3 | Family event 2 | Family event 3 | Family event 1 | Travel album 1 |
|---|---|---|---|---|---|---|
| Kappa agreement | 0.179 | 0.210 | 0.334 | 0.351 | 0.393 | 0.472 |
| Kong et al. [75] | 0.302 | 0.120 | 0.175 | 0.263 | 0.384 | 0.397 |
| Jin et al. [65] | 0.258 | 0.009 | 0.080 | 0.075 | 0.236 | 0.721 |
| NIMA [160] | 0.269 | 0.196 | 0.115 | 0.044 | 0.128 | 0.260 |

Table 6.1: Performance comparison of independent image assessment methods. The kappa user agreement values are given in the first row (the albums are sorted from the lowest to the highest user agreement). The performance values for each method are computed as the Pearson correlation coefficient between user preferences and scores provided by analyzed methods.

Several observations arise from this correlation analysis. The method by Jin et al. [65] performs poorly for most photo albums tested, despite its potential for dealing with image crops. Nevertheless, for *Travel album 1* it shows the best correlation with the user preferences among all the methods. This album consists of a number of very similar photos of landscapes and landscapes with people, which is possibly the scenario where the approach by Jin et al. [65] demonstrates its best performance. In addition, their approach appears to respond to the presence of blur in images, preferring sharp photos, which corresponds well to user selections in this album.

The NIMA method [160] performs worse in albums where people's photos are present, but it shows its best performance in the landscape-focused *Travel album 2* and *Travel album 3*. However, the degree of inter-observer agreement is relatively low for these collections, therefore no certain conclusions can be made.

On average, the best performance is demonstrated by the approach of Kong et al. [75]. Despite their primary aim of addressing general aesthetics scoring of photos, their produced score demonstrates a noticeable correlation with the user preferences. We hypothesize that this is due to their proposed content-dependent weighting scheme, which provides better estimation of different types of the image content. Thus, their approach could be potentially suitable for aiding in the task of photo selection within photo albums.

As we have observed, the applicability of independent image assessment methods is limited when addressing complete photo collections. Although in some cases this could be explained by the method's occasional failure on certain images with a complex content, another important reason is that the photo context is not considered by such methods. An illustration of this point is given in Table 6.2, where photos from different scenes in the same album often receive non-comparable scores, making difficult to rank photos in the album relative to each other.

A direct approach of using independent image scores in a photo album would be to rescale and normalize the scores linearly, in accordance with other scores computed in

| |  |  |  |  |
|---|---|---|---|---|
| Kong et al. [75] | 0.55 | 0.58 | 0.54 | 0.53 |
| Jin et al. [65] | 3.13 | 4.59 | 4.86 | 5.06 |
| NIMA [160] | 4.87 | 5.01 | 4.76 | 4.54 |

Table 6.2: Scoring of different photos in the same album by the analyzed image assessment methods. In this case, one pair of similar images is always scored higher than another pair, by each tested method. Thus, a global ranking of photos in an album would not provide an expected selection, if the presence of different scenes is not taken into account.

the album. However, we have just observed that the original scores given by methods are often poorly correlated with user preferences. One reason for this could be the elimination of the context present in photo albums, as each photo is assessed independently, without consideration of its surrounding photos. For example, even if multiple pictures of the same scene were taken, suggesting that the user found that scene important, each picture from the scene could receive a low score and be potentially rejected.

When analyzing joint results on clustering and selection by users in Section 4.2, we have also observed that the users' selection decisions were influenced by the context, represented by the clustering data. Therefore, it appears natural to employ the automatically computed clustering data (as described in previous chapter) to adapt the discussed independent scores and improve their overall assessment performance.

## 6.2 CLUSTERING-BASED IMAGE SCORE ADAPTATION

The independent assessment scores described in the previous section are used in our adaptation approach, which is based on the computed clustering described in Section 5.1. As shown in Figure 6.1, the adaptation is performed in the following manner. Given a collection clustering, we define the multi-level context of each photo. Then, after the per-image independent scores are obtained, the defined context information is used to adapt the scores. For this purpose, we have adopted two approaches. The first approach relies on statistics computed at the different context levels and employs the z-score measure. The second approach utilizes a machine learning technique, where the computed z-scores are combined with the additional data features computed from the obtained clusters.

### 6.2.1 *Photo Context Definition*

Our proposed adaptation approaches vary in the context data they use, but both of them are essentially based on the same context representation. The core of this representation is the clustering structure obtained with our clustering method, where all photos are clustered into a nested structure of the scene and near-duplicate clusters. The information from these clusters is supplemented with the collection level information, and finally the context is modeled on three levels, with the following intuition behind each level:

1. **Collection level** is used to reflect the general features of the capturing devices and skills of a photographer.
2. **Scene level** focuses on images of the same general scene but not necessarily completely overlapping. Thus, it aims to reflect environment properties, such as influence of illumination, or changes in composition.
3. **Near-duplicate level** focuses on images that depict the same scene or object with minor variations due to occasional user mistakes and other arbitrary changes between similar images. This information should serve to aid the scoring of very similar photos, which is often the case in photo albums.

Here, the scene and near-duplicate levels directly reflect the structure we have employed in our clustering study and our automatic clustering method.

In addition, although our clustering computation employs a temporal clustering into time windows, we do not use their corresponding data in the adaptation step. The time windows primarily serve to facilitate the visual similarity based clustering, hence the context-relevant information from them is inherently represented by the scene and near-duplicate clusters.

### 6.2.2 *Z-score Based Adaptation*

The first proposed score adaptation approach uses the statistical technique of z-scores [76]. As demonstrated in Figure 6.4, the z-score represents the distance of a data value in a distribution from the mean value, which is measured in the number of standard deviations. Thus, the z-score normalizes values around the mean, which effectively makes the data measured on different scales comparable.

Following our context structure, the context of each photo is defined on three levels: collection level, scene cluster level and near-duplicate cluster level. In the following we refer to them using the $C, SC, ND$ identifiers, respectively. Based on the assumption that the independent scores follow a Gaussian distribution, we adapt them using z-scores, as follows. Let $s_I$ be the independent image score of a given photo. To provide an estimation of how good the photo is on different levels of context, we compute the z-score at each level:

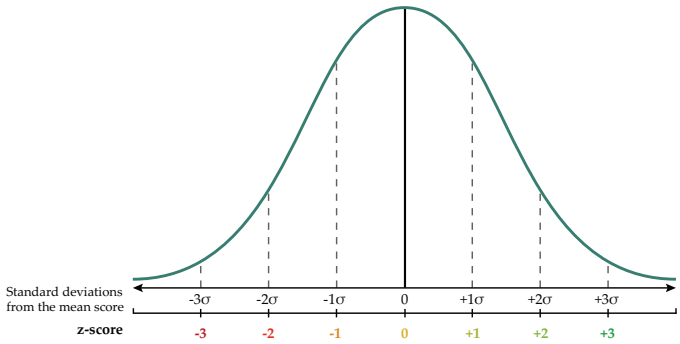



Figure 6.4: Z-scores on a normal distribution. The z-score evaluates the number of standard deviations from the mean data value.

$$z_{I_L} = \frac{s_I - \mu_L}{\sigma_L}, \tag{6.4}$$

where $\mu_L$ is the mean of image scores on the level $L \in C, SC, ND$, and $\sigma_L$ denotes a standard deviation on the level $L$. It is worth pointing out that each image belongs to only one cluster from each level (an example can be seen in Figure 6.3).

If the image is unique at its level (for example, no other similar photos have been taken in the same scene), then no other images are available for the computation of the statistics information, and we consider its z-score as undefined.

The global score $Z_I$ of an image is computed as an average of the adapted scores from three levels:

$$Z_I = \frac{z_{I_C} + z_{I_{SC}} + z_{I_{ND}}}{n_z}, \tag{6.5}$$

where $z_{I_C}$ is an image score on the collection level, $z_{I_{SC}}$ is a scene level score, $z_{I_{ND}}$ is a near-duplicates level score, and $n_z$ is the number of defined z-scores for this image (number of levels where image is not unique). Undefined z-scores for a given image are set to 0 for this calculation. We can note that we have also experimented with other ways of z-scores weighting from different levels, including learning the optimal weights with the linear least squares regression approach, but in our preliminary experiments we did not arrive to clear conclusions.

A toy example of the score adaptation is demonstrated in Figure 6.5. Three photos from the same near-duplicate cluster (Images 1-3) initially receive low independent assessment score, if compared with other photos in the collection. However, after the z-score based adaptation, the output scores $Z_I$ allows selecting one of the photos from this cluster.

### 6.2.3 *Neural Network Based Adaptation*

Although the z-score provides useful rescaling properties that allow to re-evaluate the independent score on different context levels, the z-score based solution is often not sufficient for our problem. By its nature, the z-score is essentially a linear transformation

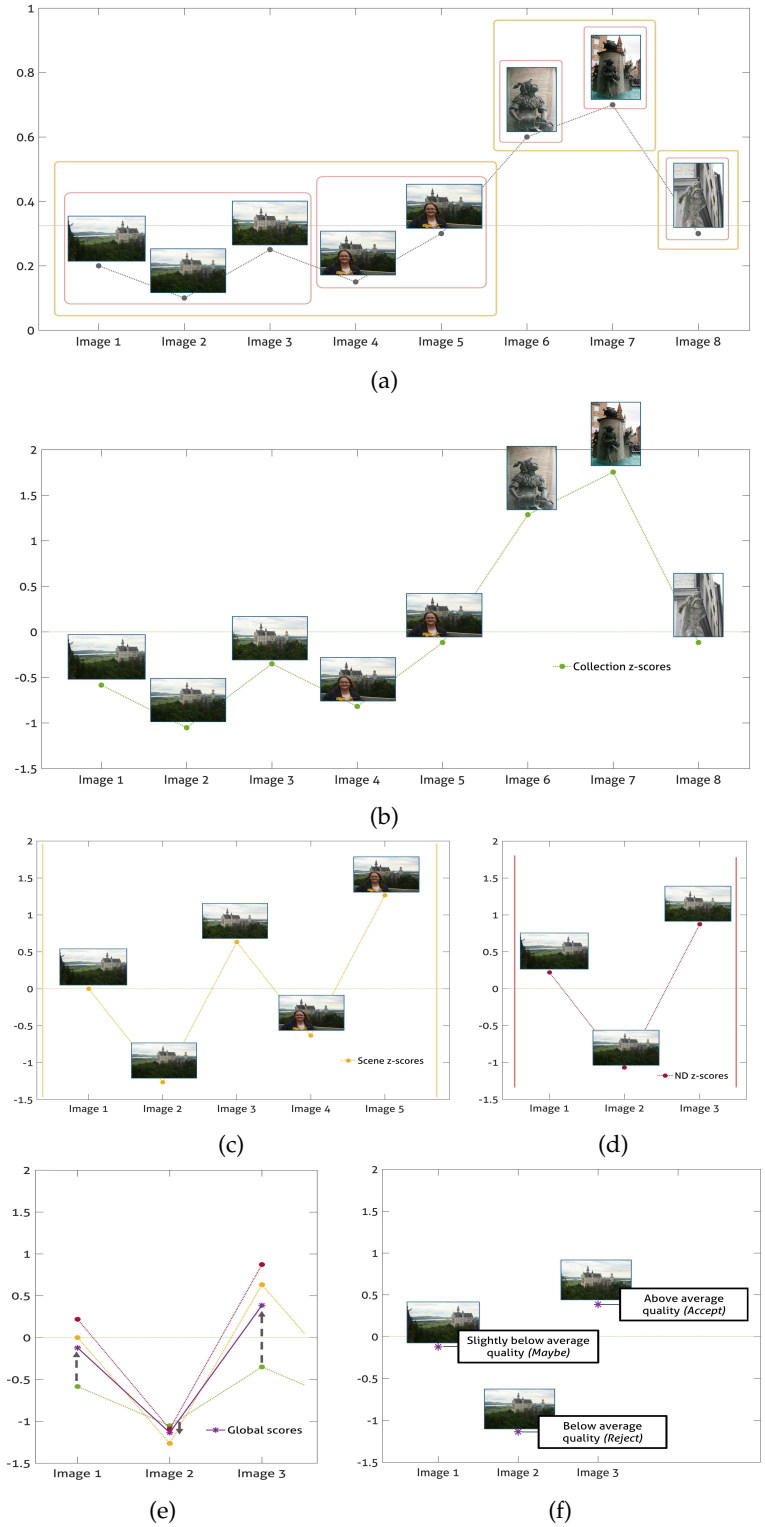



Figure 6.5: Demonstration of the z-score based adaptation. The score adaptation is shown for Images 1-3. (a) Independent collection scores before adaptation (the sharpness scores are used), along with the computed clustering. (b) Collection level z-scores $z_{I_C}$. (c) Scene level z-scores $z_{I_{SC}}$. (d) Near-duplicate level z-scores $z_{I_{ND}}$. (e) Global scores $Z_I$, with a transformation from the initial collection level z-scores indicated. (f) Possible photo labeling derived from the output scores. While all three images would be rejected if guided only by independent scores, after the adaptation one of the images is recommended to be selected.

for the scores at each level. The global score, by combining the z-scores from multiple levels, introduces some degree of nonlinearity, which depends on the characteristics of the corresponding clusters. Nevertheless, the z-score based approach can be too simplistic in its modeling performance.

Our initial experiments, as we will see in the next section, have confirmed that the z-score based approach is often not sufficient for the context adaptation. To address this issue, we have proposed a machine learning based solution, which could provide the necessary nonlinearity in the scores modeling. Notwithstanding, in the proposed approach we do not abandon the use of z-scores, but rather adopt them in a new manner.

Our approach is based on a shallow neural network trained for a regression task, to predict user preferences scores from the input features. The utilized network is a multilayer perceptron with one hidden layer consisting of ten neurons and the tan-sigmoid function used before the output layer with the linear function. The network's architecture is demonstrated in Figure 6.6. As an optimization approach, we use the Levenberg-Marquardt algorithm [85], which effectively is a combination of two minimization methods: the gradient descent method and the Gauss-Newton method [7].

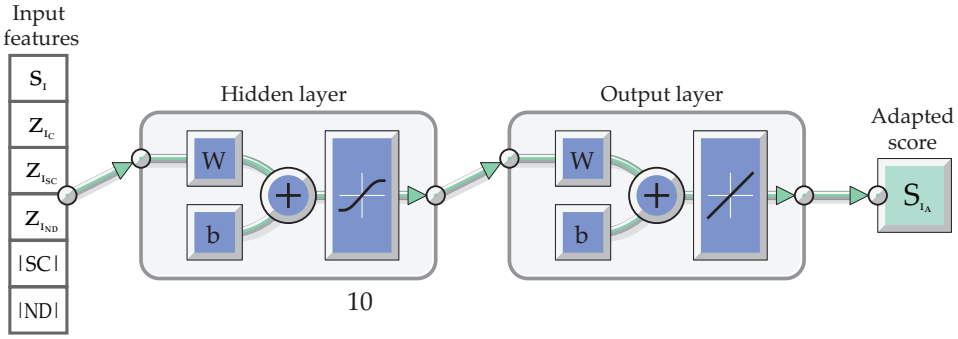



Figure 6.6: Architecture of the neural network used for score adaptation.

As the input features, we enrich the independent photo score by the data computed from the corresponding cluster partitions and extract the 6 following features: (1) original image score $s_I$, (2) collection level z-score $z_{I_C}$, (3) scene level z-score $z_{I_{SC}}$, (4) near-duplicate level z-score $z_{I_{ND}}$, (5) number of images in the scene cluster $|SC|$, (6) number of images in the near-duplicate cluster $|ND|$.

For the train-test process we use the nested cross-validation procedure. In the nested cross-validation, an outer k-fold loop is used to divide the data into training and test folds. Then, each training fold is further split into training and validations folds on an inner loop, to select the best model parameters.

In our case, on each outer loop, five albums are used for the training, and one album is set apart as a test fold, to test model's performance on it. Within the inner loop, the training is performed as follows: three albums are used for the training itself, and the remaining two albums are used in the validation fold, to select the best layer weights. After the best model is found inside the inner loop, its performance is evaluated on the one test album from the outer loop, thus avoiding possible model overfitting. This way,

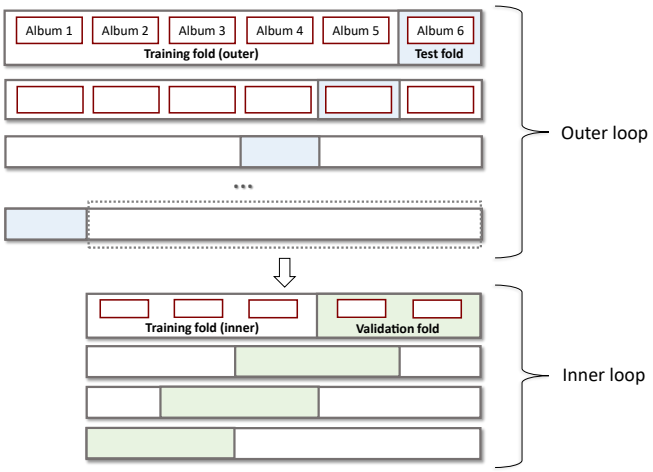



Figure 6.7: Nested cross-validation procedure.

on the outer loop, we receive a per-album model, which is trained on other albums in the dataset. These per-album results are reported and discussed in the next section.

## 6.3 RESULTS AND DISCUSSION

In this section we assess the final output of our proposed framework, to answer the question how well the context-adapted scores can improve over independent assessment. We combine different proposed clustering approaches with different adaptation techniques and compare them with the user preferences acquired in the experimental study, presented in Chapter 3.

### 6.3.1 *Performance of the Sharpness Score Adaptation*

As a starting point, we have conducted experiments on the sharpness score adaptation. Although sharpness by itself does not provide a complete representation of image quality, it is an objective characteristic, allowing us to assess the behavior of different adaptation solutions. Following typical photography practices, we began our evaluation by assigning discrete labels to photos representing selection decisions (i.e. *Accept*, *Reject*, *Maybe*). To obtain more in-depth conclusions, we expand our experiments by directly considering the correlation between adapted image scores and our previous user study results.

#### *Preliminary Performance Evaluation: Labeling Performance*

The output of our framework (given in Figure 6.1) is an updated context-adapted score for each photo, irrespective of the clustering and adaptation approaches used. The obtained score can either be used directly for sorting and displaying the photo collections, or it can be transformed into a user-friendly selection label.

In our early experiments on the sharpness score adaptation, we have investigated only the *Time-SIFT* clustering combined with the z-score based adaptation. In these experiments, we have decided to proceed with the labeling performance evaluation: the output labels of our method have been compared with the labels given by each participant in the sharpness-based selection study.

Similarly to the user study, we assign three types of labels : *Accept*, *Reject* and *Maybe*. To obtain labels for each photo according to the original sharpness estimation method by Tong et al. [165], the output score of their method was thresholded. The thresholds were estimated as follows: sharpness scores were computed and averaged over 1600 different images, leading to an average score of 0.34. Half a standard deviation ($\sigma/2 = 0.1$) was considered above and below this average, obtaining the following thresholds and corresponding labels: an image is labeled as *Reject* if sharpness $s_I < 0.24$, *Maybe* if $0.24 \leq s_I \leq 0.44$, and *Accept* if $s_I > 0.44$.

For the z-score adapted scores, we defined thresholds $Z_I \geq 0$ for the label *Accept*, $-0.5 \geq Z_I < 0$ for the label *Maybe*, and $Z_I < -0.5$ for the *Reject* label. This way the labels reflect the nature of z-scores, where $Z_I = 0$ corresponds to the mean of the scores across three levels of the context, and images with scores below zero are worse quality-wise than other neighboring images from their corresponding context. The examples of sharpness-based labeling within the context are shown in Figure 6.8.


Figure 6.8: Examples of the sharpness-based selection labeling within a photo context. Our context-aware framework performs the assessment and labeling of images in photo collections by considering image quality in the context of the collection as well as photos captured in the same scene. (a) Extract from a photo collection and visualization of clustering. (b) Photos captured in the same scene and close-ups of image details. (c) Top: image labeling obtained from independent sharpness-based assessment. Bottom: labels assigned after our context-aware adaptation of the independent score.

We evaluate the labeling performance using the metrics of accuracy, precision, recall and F-measure [36]. For each of three selection labels, we calculate the number of *TP: true positives*, *FP: false positives*, *TN: true negatives* and *FN: false negatives*. If we consider an analyzed label as a positive class and two other possible labels as a negative class, the number of *true positives* defines the number of images correctly labeled with the positive class, the number of *false positives* defines the number of images incorrectly labeled with the positive class, the number of *true negatives* defines the number of images correctly

labeled with the negative class, and the number of *false negatives* defines the number of images incorrectly labeled with the negative class. Then, the performance metrics are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.6}$$

$$Precision = \frac{TP}{TP + FP} \tag{6.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.8}$$

.

Finally, the F-measure is computed as the harmonic mean between precision and recall:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6.9}$$

.

The described metrics were calculated for each label and user separately, and average values of these metrics were computed across all participants. The performance comparison is given in Table 6.3. While the accuracy is not always a reliable metric [36], both precision and recall values are important to estimate the correctness of label prediction in our task. In this case the F-measure is a good indicator of the overall performance, due to its properties of the harmonic mean between two metrics.

| | Travel collection | | Wedding ceremony | | Sport event | | Halloween party | |
|---|---|---|---|---|---|---|---|---|
| | Independent | Context-dependent | Independent | Context-dependent | Independent | Context-dependent | Independent | Context-dependent |
| Accuracy | 0.552 | 0.692 | 0.479 | 0.605 | 0.595 | 0.568 | 0.466 | 0.536 |
| Precision | 0.495 | 0.569 | 0.228 | 0.448 | 0.279 | 0.452 | 0.185 | 0.384 |
| Recall | 0.482 | 0.559 | 0.431 | 0.471 | 0.368 | 0.463 | 0.313 | 0.408 |
| **F-measure** | **0.435** | **0.560** | **0.288** | **0.452** | **0.297** | **0.432** | **0.178** | **0.387** |

Table 6.3: Comparison of labeling performance based on independent blur assessment scores using the method by Tong et al. [165] and context-dependent scores obtained using our approach, for image collections used in the sharpness-based experimental studies. Higher value of each measure indicates better performance, where F-measure can be used as an indicator of the overall performance.

As shown in Table 6.3, when we assess the labeling performance, the adaptation of image quality scores using our framework offers a significant improvement on all collections tested, relative to the results obtained with the independent method of Tong et al. [165]. In collections containing more challenging photographs and conditions (Wedding ceremony, Sport event and Halloween party), our method showed the most improvement. As these collections no longer fit the implicit criteria of the sharpness estimation method, falling well below the expected quality, most of the photos are rejected according to the original non-adapted score. On the contrary, the user study participants adapt to the album quality level and the context, and tend to accept even

not perfectly sharp photos that might be important for the coverage of such an event, especially if no better alternatives are available in the album. Our method manages to adapt to the quality level in a similar way and provides labeling results, which are more consistent with the users' decisions. At the same time, noticeable improvements were measured even for the travel collection, where the overall image quality was relatively high and few blurry photos were present, showing the ability of our approach to adapt to the context of each collection and image.

An additional performance analysis was conducted using receiver operating characteristic (ROC) curves, which were computed for each of three labels over all evaluated collections. The shape of the curves, shown in Figure 6.9, as well as the Area Under the Curve measures, demonstrate better performance of our method for each label type. It can be noted that the independent method tends to correctly identify photos pre-labeled by users for rejection, while it does not show high performance on the two other labels. This agrees with our intuition that the non-adapted method is likely to reject the lower quality photos, while users try to find acceptable photos even among imperfect ones.

Another interesting observation that can be made from the ROC curves concerns the *Maybe* labels. Both the independent method and our context-dependent method suggest performance that is close to labeling by chance for that label. Indeed, the introduction of the *Maybe* labels makes the task of automatic pre-labeling more challenging. Initially, we have considered that the *Maybe* label can serve as a middle ground of selection suggestions for users, so users could decide by themselves if a photo marked with this label is worth keeping. However, we can observe that the introduction of this label is arguable, as the corresponding performance is low. In addition, during our user study, we have also found that the *Maybe* label was often ambiguous for the users and might have introduced undesired noise in the experimental data.

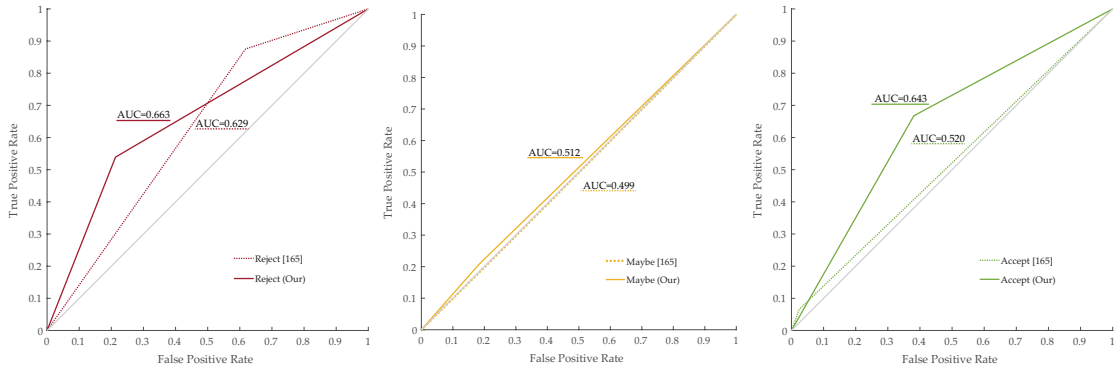



Figure 6.9: Receiver operating characteristic (ROC) curves for label prediction by the independent assessment method of Tong et al. [165] and our context-dependent method (computed over all collections). The closer the curve extends to the top left corner, the better performance for particular label is achieved. The diagonal line represents the performance of random labeling. The *AUC* value indicates the *Area Under the Curve* measure. Higher value indicates a better performance.

We performed a similar analysis using a second blur assessment method, namely the approach of Mavridaki et al. [116]. As their method relies on an SVM-based classification, we take the probability estimates as an output [14]. For the computed probability estimates, $P_i = 1$ indicates that the image is sharp, while $P_i = 0$ indicates that the image is blurred. The probability is used as an independent image quality score, and it is adapted the same way, using the z-score based adaptation. The output photo labeling for this method was performed in the following way: an image is labeled as *Reject* if $P_i < 0.4$, *Maybe* if $0.4 \leq P_i \leq 0.6$, and *Accept* if $P_i > 0.6$. Table 6.4 provides a comparison of the independent and context-dependent results for the different image collections. Similar to our analysis using the method of Tong et al., we find that in this case our approach also improves the initial results and leads to labels that better correspond to user assessments.

| | F-measure | |
|---|---|---|
| | Independent method by Mavridaki et al. [116] | Context adaptation by our method |
| Travel collection | 0.5444 | 0.5462 |
| Wedding ceremony | 0.5459 | 0.5771 |
| Sport event | 0.3735 | 0.4459 |
| Halloween party | 0.2562 | 0.4472 |

Table 6.4: Performance of the labeling obtained with independent blur estimation method by Mavridaki et al. [116], compared with the results obtained using context-adaptation by our approach.

*Scores Correlation Performance*

Although our approach has shown promising results in the labeling performance, there are certain implicit assumptions present, to obtain discrete labels from the adapted scores, which may bias our results. Each label corresponds to a fixed range of score values with predefined thresholds that were determined empirically. Although the proposed z-score based scoring can mitigate this issue, as all scores are brought to the same scale, it is still difficult to perform the performance comparison, since the evaluation of the original scoring also relies on the pre-defined thresholds. In addition, we have seen that some categorical decisions, such as *Maybe* label are challenging to deal with.

To exclude the influence of the labeling thresholds, we also evaluate the methods' performance by computing the Pearson correlation coefficient between the obtained photo assessment scores and the user preference scores (their acquisition is explained in Section 3.1.2). The per-album correlation is shown in Table 6.5: we demonstrate the performance of the same method used to receive the labeling results shown before in Table 6.3 and Figure 6.9.

While our z-score based adaptation has shown promising labeling results earlier, we now observe that its correlation performance is lower than the performance of the origi-

| | Professional session | Halloween party | Sport event | Wedding ceremony | Travel collection |
|---|---|---|---|---|---|
| Kappa agreement | 0.172 | 0.256 | 0.336 | 0.422 | 0.469 |
| Original scores [165] | 0.330 | 0.256 | 0.564 | 0.550 | 0.716 |
| Z-score based adaptation | 0.298 | 0.231 | 0.498 | 0.493 | 0.644 |

Table 6.5: Per-album correlation with user preferences of sharpness scores by Tong et al. [165] after z-scores based context adaptation. The adaptation is based on *Time-SIFT* clustering and follows the same procedure as for results in Table 6.3 and Figure 6.9. The kappa user agreement values are given in the first row (the albums are sorted from the lowest to the highest user agreement). It can be seen that the correlation performance indicates no real improvement over the original scores in this case.

nal scores. The performed z-score adaptation might have introduced scores rescaling, beneficial for the labels thresholding, but apparently the applied transformation is not enough to improve the scoring.

To address this issue and model the nonlinear nature of photo assessment, we apply our proposed neural network based adaptation. In this second approach, we learn the score adaptation from the photo z-scores and the clusters features. The average results of this adaptation based on different clustering approaches are demonstrated in Table 6.6.

Indeed, the neural network based adaptation achieves better performance. The average correlation substantially improves for the method of Tong et al. [165], for each clustering approach. At the same time, for the method of Mavridaki et al. [116] the original scores performance remains the best. To explore possible reasons of such performance, we can also analyze per-album correlation.

| | Original scores | Time-SIFT | Adaptive Time-SIFT | Adaptive Time-CNNR |
|---|---|---|---|---|
| Tong et al. [165] | 0.483 | 0.518 | 0.563 | **0.582** |
| Mavridaki et al. [116] | **0.553** | 0.548 | 0.468 | 0.549 |

Table 6.6: Correlation of the adapted sharpness scores with user preferences after neural network based context adaptation (computed over all collections). *Original scores* column represents original aesthetics assessment methods' results without adaptation applied. Other columns represent results of adaptation with automatically computed clustering.

In Table 6.7 we present the per-album correlation for the original scores and the best achieved adaptation results for each method (using the neural network based adaptation with *Adaptive Time-CNNR* clustering). We can see that a performance improvement is achieved for all albums, except the Travel collection, for both the original scores used in adaptation. Somewhat surprisingly, the Travel collection album presents the highest user agreement, yet the lowest adaptation performance. Nevertheless, such results may be explained by the following. Our training and testing are performed in a nested cross-validation manner, as explained in Section 6.2.3. It means that in the case of the Travel

collection the training is performed on other collections, which essentially present lower user agreement. In addition, as we acquired the photo albums representing different topics, it might be the case that they have different clusters structures, making the adaptation more difficult. Consequently, these issues can indicate the lack of training data, sufficient to model the desired properties.

| | Professional session | Halloween party | Sport event | Wedding ceremony | Travel collection |
|---|---|---|---|---|---|
| Original scores [165] | 0.330 | 0.256 | 0.564 | 0.550 | 0.716 |
| NN-based adaptation | 0.504 | 0.558 | 0.609 | 0.706 | 0.535 |
| Original scores [116] | 0.196 | 0.469 | 0.532 | 0.811 | 0.759 |
| NN-based adaptation | 0.250 | 0.570 | 0.563 | 0.819 | 0.544 |

Table 6.7: Per-album correlation with user preferences after neural network based context adaptation. The results represent the best adaptation results, achieved with *Adaptive Time-CNNR* clustering.

Regarding the performance of the sharpness score adaptation, several conclusions can be drawn.

The z-score based adaptation can be effective in rescaling scores for label thresholding, but the labeling performance might be not sufficiently reliable to evaluate the adaptation results. At the same time, the correlation performance evaluation demonstrates that the z-scores based adaptation does not improve the scoring performance. However, the proposed neural network based adaptation provides better results for most albums tested.

The remaining issues can be related to the amount of training data and also to the utilized photo assessment approach. Even aided by the context, our method largely depends on the independent score used as an initial quality estimator. It is not uncommon that the sharpness estimation method itself fails and provides a wrong score (for instance, a blurred image can get a high sharpness score due to the specific nature of blur). In case of such failure, the context-based adaptation cannot correct the original score. Furthermore, an incorrect score assigned to one image from a multiple-images cluster can negatively affect the score adaptation for the surrounding images.

In addition, in the original experimental study we have faced certain issues regarding user decisions, as we discussed in Section 3.1.2. As the task of sharpness-based selection by itself was often ambiguous for users, the acquired photo selection data might be noisy. Therefore, modeling such assessments can be challenging.

With these conclusions in mind, we proceed to the adaptation results for a general photo aesthetics score.

### 6.3.2 *Performance of the Aesthetic Assessment Adaptation*

To better predict general image quality and aesthetics, we employ several more general no-criteria photo assessment methods and adapt their scores as described in Section 6.2.

To evaluate the ability of our approach to model user preferences, in this section we compare the results of each method against the user assessment decisions obtained in our study presented in Section 3.2. We can expect that the modeling of these preferences will be also easier to analyze, since both clustering and selection studies have been done on the corresponding photo albums.

The independent photo scores are obtained with deep learning based methods for aesthetics assessment introduced in Section 6.1 and adapted using our two proposed adaptation approaches. Regarding the performance evaluation, given the limitations of the labeling based evaluation discussed previously, we follow only the correlation evaluation approach for the results of aesthetic score adaptation.

As a first context-adaptation approach, we use our method based on the multi-level z-score weighting. The performance of the z-score based adaptation is given in Table 6.8. Here, the Pearson correlation coefficient was computed between the computed scores and the user preference scores and averaged across all albums.

| | Original scores | User clustering | Time-SIFT | Adaptive Time-SIFT | Adaptive Time-CNNR |
|---|---|---|---|---|---|
| Kong et al. [75] | 0.274 | 0.232 | **0.261** | 0.248 | 0.217 |
| Jin et al. [65] | 0.230 | 0.254 | **0.260** | 0.252 | 0.243 |
| NIMA [160] | 0.169 | 0.248 | 0.217 | 0.185 | **0.219** |

Table 6.8: Average correlation with the user preferences after z-scores based context adaptation. *Original scores* column represents the original aesthetics assessment methods' results without adaptation applied. *User clustering* column represents the results of adaptation using the ensemble clustering acquired in the clustering study. Other columns represent results of adaptation with automatically computed clustering.

It can be seen that the z-score adaptation results only in a limited improvement over the original scores. In addition, the advantage of better clustering is not evident in this case, as often *Adaptive Time-SIFT* clustering and *Adaptive Time-CNNR* clustering provide lower correlation than the adaptation based on *Time-SIFT* clustering. Finally, instead of automatically computed clustering, we apply the ensemble *User clustering* (its computation is explained in Section 4.2.1) as an adaptation base, to estimate the overall feasibility of the z-score adaptation approach. The adaptation based on the *User clustering* also does not lead to a clear improvement of the results, which suggests that the employed adaptation approach may be too simplistic to model the user behavior, similarly to the case observed with the sharpness score adaptation.

Due to insufficient improvement provided by the straightforward z-score adaptation, we explore our second adaptation approach, based on a trained neural network. For each clustering approach, we retrain and test the network on its corresponding clustering results. The obtained results are shown in Table 6.9.

Compared with the original scores and the straightforward z-score based adaptation, this method achieves a significant improvement. The adaptation based on the *User clustering* provides the highest increase in correlation with the user preferences, which confirms the feasibility of the chosen adaptation. Moreover, all three automatic clus-

| | Original scores | User clustering | Time-SIFT | Adaptive Time-SIFT | Adaptive Time-CNNR |
|---|---|---|---|---|---|
| Kong et al. [75] | 0.274 | 0.546 | 0.485 | 0.482 | **0.489** |
| Jin et al. [65] | 0.230 | 0.517 | 0.440 | **0.463** | 0.439 |
| NIMA [160] | 0.169 | 0.543 | 0.415 | **0.459** | 0.457 |

Table 6.9: Average correlation with the user preferences after neural network based context adaptation. *Original scores* column represents the original aesthetics assessment methods' results without adaptation applied. *User clustering* column represents the results of adaptation using the ensemble clustering acquired in the clustering study. Other columns represent results of adaptation with automatically computed clustering.

tering approaches lead to improvement over the original unadapted scores. The score adaptation using the *Adaptive Time-SIFT* or the *Adaptive Time-CNNR* clustering methods provides similar performance. Overall, the best performance is achieved for the method of Kong et al. [75], using the *Adaptive Time-CNNR* clustering as an adaptation base. At the same time, for two other scoring approaches, their best performance is achieved using the *Adaptive Time-SIFT* clustering. Nevertheless, the adaptation results between these two clustering approaches are comparable, conforming with a similarity in their clustering performance.

Table 6.10 provides a more detailed per-album analysis of the best adaptation results obtained with the aesthetics scoring by Kong et al. [75] and the neural network adaptation using different clustering as a base. Each adaptation approach provides a significant correlation improvement over the original scores.

A probable reason of this failure is the specific structure of the album. The *Family event 1* album represents a wedding ceremony, and while it has noticeable groups of similar photos, they are different from other albums. This album has less repetitive content and contains a number of unique portraits of people not reappearing in other photos. In our study this album has obtained one of the lowest clustering agreements between users. At the same time, on average around 57% of photos were selected in this album, while in other albums the selection ratio was close to 33%. Therefore, the adaptation training performed on other albums might not succeed in this case, as it cannot effectively predict its different characteristics.

| | Travel album 2 | Travel album 3 | Family event 2 | Family event 3 | Family event 1 | Travel album 1 |
|---|---|---|---|---|---|---|
| Kappa agreement | 0.179 | 0.210 | 0.334 | 0.351 | 0.393 | 0.472 |
| Original scores [75] | 0.302 | 0.120 | 0.175 | 0.263 | 0.384 | 0.397 |
| User clustering | 0.587 | 0.506 | 0.610 | 0.353 | 0.539 | 0.684 |
| Time-SIFT | 0.554 | 0.272 | 0.585 | 0.407 | 0.432 | 0.661 |
| Adaptive Time-SIFT | 0.507 | 0.337 | 0.532 | 0.348 | 0.485 | 0.682 |
| Adaptive Time-CNNR | 0.421 | 0.495 | 0.523 | 0.396 | 0.402 | 0.697 |

Table 6.10: Per-album correlation with user preferences after neural network based context adaptation for the method of Kong et al. [75].

The visualization of the per-album correlation before and after the score adaptation for the three tested methods can be found in Figure 6.10. The adaptation for all methods

is performed using the neural network approach, based on the *Adaptive Time-CNNR* clustering. The observed absence of scoring improvement for the *Family event 1* album is also visible here, for all analyzed methods. Nevertheless, for the remaining albums we can observe a performance improvement for all the assessment methods after the adaptation.

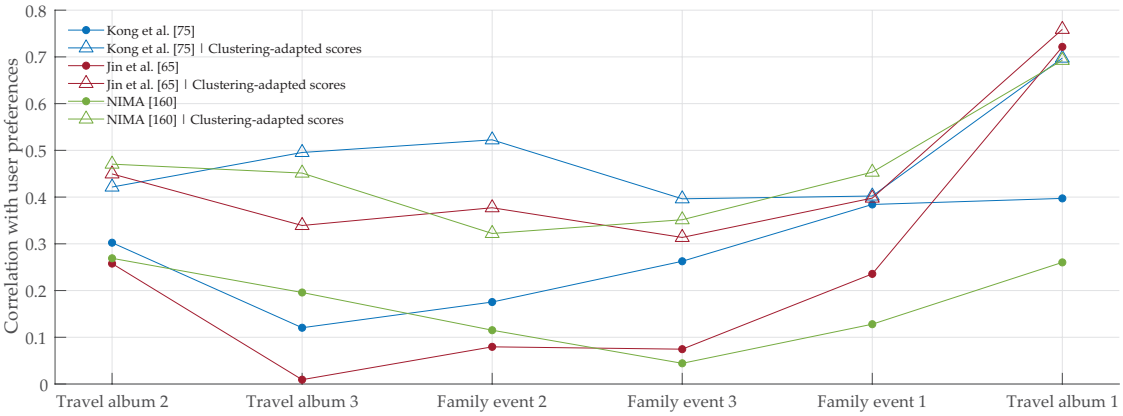



Figure 6.10: Per-album correlation between computed scores, their adaptation and user preferences. The adaptation is performed with the neural network based approach, using the Adaptive Time-CNNR clustering. The albums are sorted from the lowest to the highest kappa user agreement.

The obtained results confirm that the clustering information can be successfully used to improve the performance of the independent photo assessment solutions. Above all, the *User clustering* provides the largest improvement, since the clustering partitions are derived from user decisions in this case. Automatic clustering solutions lead to a similar performance, where a certain level of correlation with user preferences is achieved. Overall, the obtained results make us believe that, when applied to photo albums with a specific structure, the photo context defined via collection clustering can effectively assist the task of scoring and selection within albums, in absence of an additional user input.

ABLATION STUDY As an additional point of interest, we have explored another aspect of the proposed neural network adaptation approach. We have performed an ablation study, where our solution was trained and tested using only a subset of data features. The results of the ablation study are given in Table 6.11. It is indeed interesting to observe the influence of each feature introduced. We can observe that the introduction of scene and near-duplicate cluster sizes (number of images in a corresponding cluster) leads to an important contribution to the overall performance. This finding also corresponds with our user study conclusion on dependency between the total number of images in a cluster and a number of selected images within it, as described in Section 4.2. At the same time, while the z-scores based features provide lower improvement by themselves, they provide a significant increase in performance in case of a complete feature combination (especially when derived from *User clustering*). On the whole, the

ablation study demonstrates the importance of each data feature and confirms our earlier intuition on the photo selection nature.

| | Original scores [75] | $z_C$ | $z_{SC}$ | $z_{ND}$ | $\|SC\|$ | $\|ND\|$ | $(z_C, z_{SC}, z_{ND})$ | $(\|SC\|, \|ND\|)$ | All features |
|---|---|---|---|---|---|---|---|---|---|
| User clustering | 0.274 | 0.371 | 0.393 | 0.413 | 0.477 | 0.455 | 0.424 | 0.520 | 0.546 |
| Adaptive Time-CNNR | | 0.365 | 0.393 | 0.423 | 0.431 | 0.450 | 0.409 | 0.470 | 0.489 |

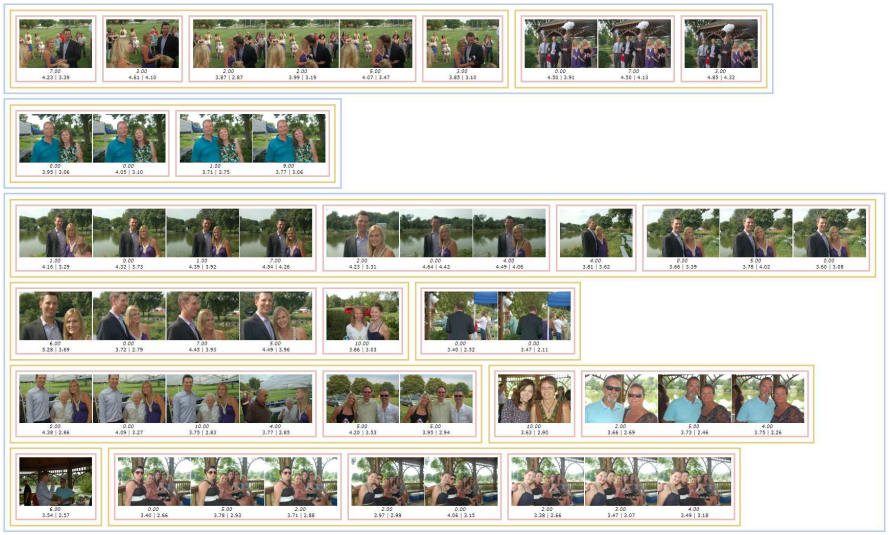
Table 6.11: Ablation study on influence of data features in the neural network adaptation approach. The average correlation is shown for the independent scores by Kong et al. [75] and the adaptation output, when the input is the original score supplied with indicated data feature(s). We can see that the best performance is achieved when all features are used together.

The visualizations of the computed scores for each album are given in Figure 6.11. We show the independent scores computed with the method by Jin et al. [65], along with the adaptation results, using our neural network method based on the *Adaptive Time-CNNR* clustering. For comparison, we also show user preferences for the same photos. We can observe that the structure of clusters certainly affects the adaptation output for a photo, which now depends on the presence of other similar photos and their corresponding scores. The effect is especially noticeable in the albums where we achieve higher performance, such as *Travel album 1*.

Earlier we also mentioned the possibility of user-friendly photo labeling, which could help to select the photos in albums. To obtain such labeling, a fixed threshold would have to be applied to the adapted images scores. Instead of manually defining thresholds, we can automatically determine them based on some input factors, such as the percentage of photos that the user wants to keep. In absence of user input, we could also employ the selection statistics acquired in our experimental study. For instance, we could aim to keep the best 30% of photos in an album, or photos could be selected from scene and near-duplicate clusters according to the corresponding statistics obtained from our study (Section 4.2).

Finally, the following conclusions can be made on the aesthetics scores adaptation. The obtained results demonstrate that the clustering information can help define the context of a photo, which can be used in the photo score adaptation. The highest performance was achieved using a machine learning approach, which employs the original score together with cluster-defined features. According to our ablation study, certain cluster features are more important in the adaptation process, especially the ones defining the cluster size. At the same time, the proposed approach does not perform equally well on all albums, which probably reflects its dependency on general structure of clusters ins the album. Also, the initial user selections lack agreement for some albums, which makes the task of scoring very challenging by itself. Lastly, the score adaptation heavily depends on the original score itself, which is often prone to errors in relative assessments between multiple similar photos. Thus, the adaptation of an initially incorrect score can introduce additional issues, since all scores are inter-connected through the data features, during the adaptation.

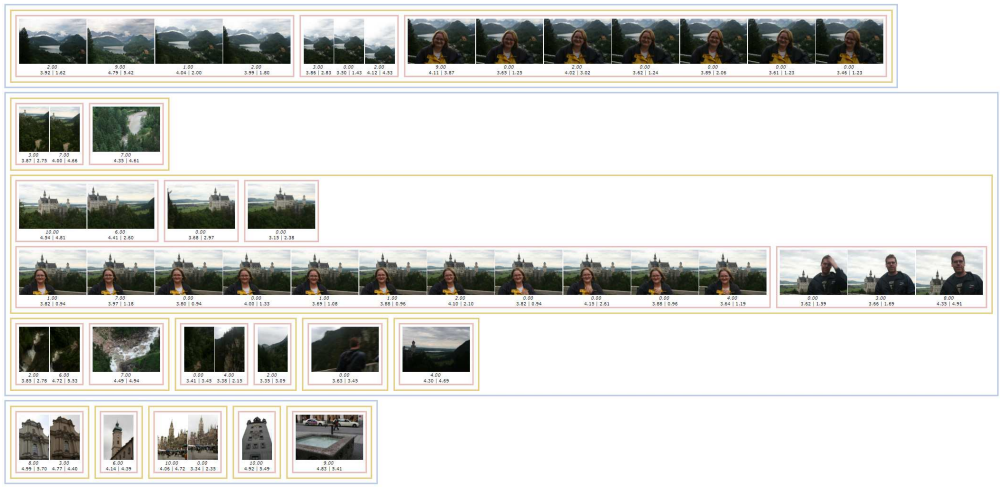
(a) Family event 1



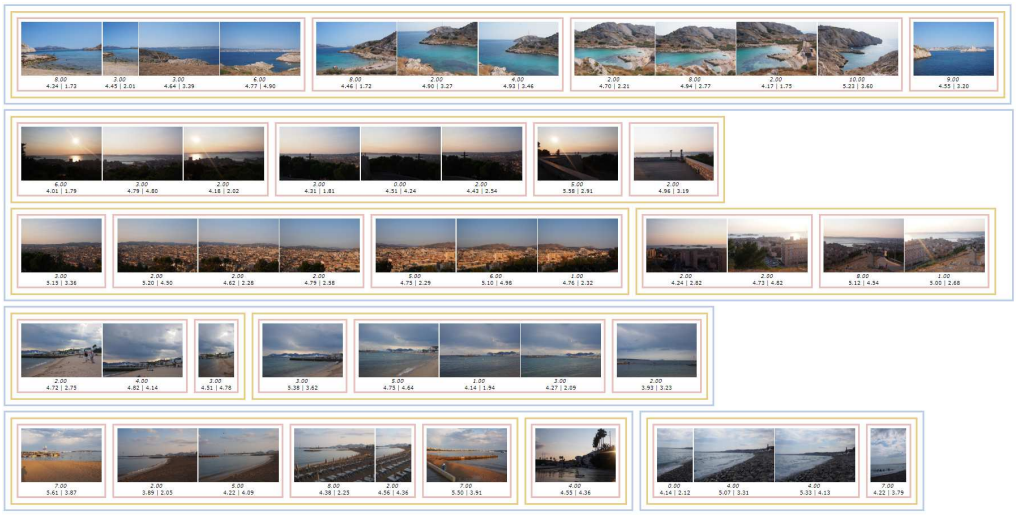
(b) Family event 2

(c) Family event 3

Figure 6.11: Demonstration of the aesthetic scores adaptation with a neural network in photo albums. The value on top represents user preferences acquired in our user study. The values on bottom represents the computed scores as follows: ORIGINAL SCORE | ADAPTED SCORE. All scores are scaled to the same range from 0 to 10, for comparison. The method by Jin et al. [65] is used to compute original scores. The adaptation is performed with the neural network based approach, using the Adaptive Time-CNNR clustering.



(d) Travel Album 1



(e) Travel Album 2

(f) Travel Album 3

Figure 6.11: Demonstration of the aesthetic score adaptation results by our proposed method. The value on top represents user preferences acquired in our user study. The values on bottom represents the computed scores as follows: ORIGINAL SCORE | ADAPTED SCORE. All scores are scaled to the same range from 0 to 10, for comparison. The method by Jin et al. [65] is used to compute original scores. The adaptation is performed with the neural network based approach, using the Adaptive Time-CNNR clustering.

## 6.4 SUMMARY

When applied directly to photo albums, the independent image assessment methods achieve low correlation with user preferences. To improve their performance in photo collections, we proposed two adaptation approaches that adapt an independent score to the surrounding context of the photo, where the context is defined from the earlier computed clustering.

The first adaptation approach, which is based on the z-score statistics measure, did not provide clear improvement over original scores. At the same time, the second adaptation approach, based on a shallow neural network and a set of context features, leads to a considerable increase in correlation with user preferences. In addition, we have found that the features of cluster size are important in the learned adaptation.

Nevertheless, certain limitations are inevitably present in the taken approach. As albums differ in their cluster structure, a trained solution cannot perform equally well on each album. Moreover, the level of user agreement is low for some albums, therefore a single model suiting all the users would be impossible to create. This fact also indicates the demand for the personalized solutions, which could adapt not only to the photo collection, but also to the individual preferences of a user.

Last but not least, as our approach depends on the original independent score, the results of the adaptation can be unpredictable for the clusters where the relative assessments were initially wrong. Studying all possible origins of such assessment mistakes would deserve another complete work. However, in the final part of this manuscript, we will explore some apparent tendencies present in typical image assessment methods.

Part IV

# CAN WE GO BEYOND? EXPLORING THE AESTHETICS LEARNED BY CNNS

In the preceding chapters we have introduced an approach to improve photo evaluations using the context present in photo albums. Nevertheless, an independent assessment method remains the heart of the proposed solution. Can we rely on such methods? What influences their decisions? Recent state-of-the-art approaches in computational aesthetics are typically based on deep learning techniques: these approaches are often trained on the same datasets, but using different CNN architectures. This fact allows us to get a more general picture of the photo assessment approaches and analyze what are possible limitations of them.

# 7

# VISUALIZING AND MANIPULATING COMPUTATIONAL AESTHETICS

When performing an adaptation of an independent photo score to the context of the surrounding photos, we should be confident in the computed score itself and its ability to provide reliable assessments for a wide range of photos. In the previous chapter we have observed that although the independent score can be improved using the photo context, the results are not always predictable and reliable. Certainly, some problems in this case are associated with the clustering and adaptation stages. At the same time, the heart of our solution remains the independent assessment. In our approach, we have relied on the CNN-based methods that estimate the aesthetics of a given image. In this chapter we would like to take a step back and analyze what might drive the decisions of computational aesthetics methods.

Since the recent methods for computational aesthetics are often based on convolutional neural networks (essentially a machine learning approach), we might expect that their behavior and performance largely depends on the utilized dataset. In Section 7.1 we try to explore tendencies in a photo assessment and an influence of the training dataset from the output of a CNN-based assessment technique and using a visualization technique for high-dimensional deep features. In Section 7.2 we approach the problem from another angle and use the generative adversarial networks for conditional image synthesis: by generating image samples with different aesthetic labels, we investigate what contributes to the photo aesthetics in learned GAN models.

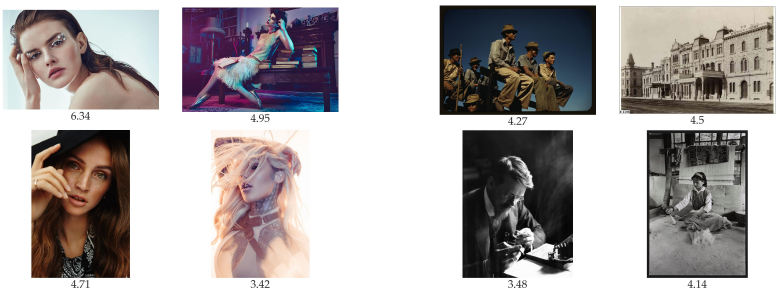
## 7.1 VISUALIZING TENDENCIES OF AESTHETIC ASSESSMENT

We base our analysis on one of the state-of-the-art CNN methods for photo assessment: the method of Kong et al. [75]. To shortly recall, this method is based on a modified and fine-tuned AlexNet architecture, where a joint loss strategy is used, with a combination of a regression loss, a ranking loss and a per-attribute activation loss. The training is performed on their proposed AADB dataset.

We have explored the performance of the analyzed method on the following datasets:

- *AADB* dataset [75]. It contains 9,958 images and it is supposed to have a balanced distribution between professional and consumer photos. This dataset was used to train the analyzed method.


(a) AADB (b) AVA (c) National Geographic




(d) Fashion (e) Classic Photographers

Figure 7.1: Examples of photos in different datasets and the scores computed with the method of Kong et al. [75].

- *AVA* dataset [122]. It contains 255,000 photos from the photographers' social network *DPChallenge.com*. A large number of professional-looking post-processed photos is found in this dataset.
- *Classic Photographers* dataset [162]. It contains 181,948 photos from well-known photographers, spanning the period from the early days of photography to the current time. A large number of the photos is black-and-white. Also, the photos' resolution largely varies.
- *National Geographic* dataset. It contains 3,457 photos from the winners of the National Geographic magazine and website contest, made by professional photographers.
- *Fashion* dataset. It contains 2,587 photos from editorial photo shoots for fashion magazines and websites, made by professional photographers.

### 7.1.1 *Score Distributions in Photo Datasets*

Using the method of Kong et al., we have computed the aesthetic scores of all images in each dataset. In Figure 7.1 we show a few examples of photos from each dataset and their corresponding scores Although it is not an exhaustive selection of all possible photos in these datasets, we can observe that in some cases it is difficult to conclude why a particular photo received a specific score. For instance, for the *National Geographic* and *Fashion* datasets, the scores are mediocre or low even for some aesthetically attractive photos.

To see a bigger picture of assessments in each dataset, we have plotted their scores histograms, which are shown in Figure 7.2 and Figure 7.3. For the datasets where the ground truth scores from users available, we have also plotted them for comparison.

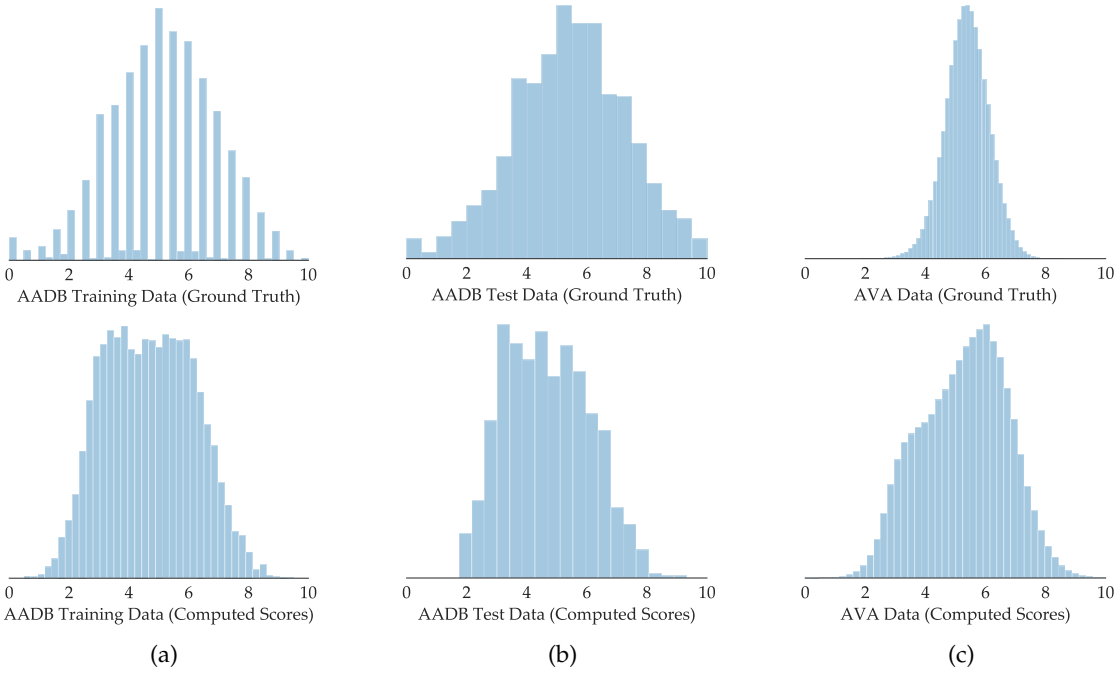



Figure 7.2: Histograms of dataset scores: datasets with the ground truth user scores available (top row).

First, let's consider the datasets with ground truth available (Figure 7.2). For the original *AADB* dataset, which was used for the method's training, we observe Gaussian-like score distributions, both in the ground truth and computed scores. The similar distributions can be observed for the *AVA* dataset, although the histogram of the computed scores is much more widely spread, comparing to the ground truth scores. The *AVA* dataset is known to have an unbalanced distribution of the scores, which we can observe in the ground truth histogram: the original scores are largely concentrated in the small area from 5 to 6.

For the other three datasets we do not have a user evaluation data, but we still can examine how the analyzed method assesses photos in them (Figure 7.3). The *National Geographic* dataset presents a similar Gaussian-like distribution, with the scores centered around the score of 5. At the same time, the observed estimations appear to be rather low for this dataset, even though the photos in it are high quality photos of nature and travel experiences, taken by professional photographers, and acclaimed by a prestigious contest. The histogram of the *Fashion* dataset is skewed towards right, with its mode located around the score 6.5. Nevertheless, in this case, also only few photos are scored higher than 8. Finally, the *Classic Photographers* dataset presents the lowest scores among all datasets: its distribution is strongly skewed towards left, with its mode around the score 3. Such low evaluation of photos from well-known photographers appear to be

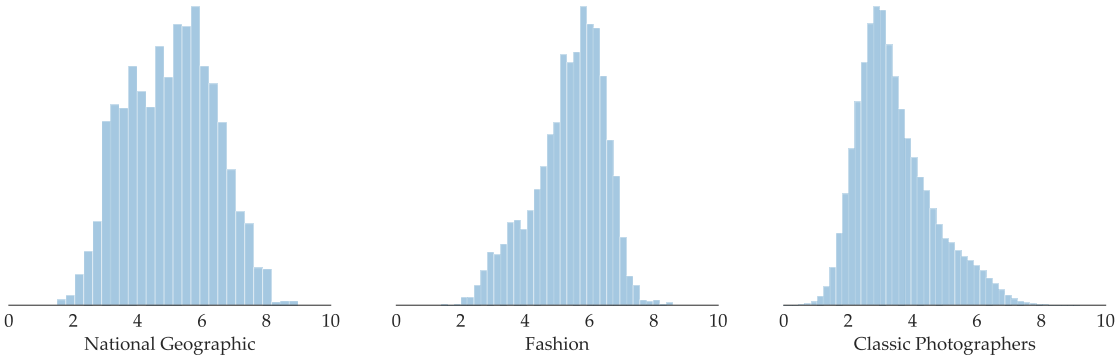



Figure 7.3: Histograms of dataset scores: datasets without the ground truth user scores available.

surprising at first, but there are several possible reasons of this. One of them could be the presence of numerous low-quality photos taken in the early years of photography. In addition, a number of distinguished photographers are represented with photos of low resolution in this dataset. Other possible reasons of such evaluations we will see in the next section.

The histograms of image scores provide some interesting hints as to how the analyzed methods assess photos. One observation is a possible lack of generalization for the trained method, since sometimes it does not perform in a predictable manner on the unseen data. The tendencies observed in this analysis suggest that certain biases may be present, which do not necessarily correspond to what would generally be considered as aesthetically pleasing. To go further in our analysis and explore these biases in more detail, we take advantage of the deep feature space inherent of the CNN-based method that we consider.

### 7.1.2 *Deep Features for the Aesthetics Space Visualization*

To explore what image characteristics can affect a method's scoring, we analyze the deep features computed by the Kong's network. While an original image exists in a space of extremely high dimensionality, deep features can provide a lower dimensional representation of it. The deep features are a set of responses extracted from a layer in a CNN, typically from one of the final layers, before the computation of the network's output. These layers combine the weighted contributions from all earlier levels, thus the features extracted from them can provide an abstract representation of the input image in the learned space. In case of Kong's network, we have extracted the deep features from the last fully-connected layer, which are represented by a 256-dimensional vector. Although such deep feature vectors are still of high dimensionality, it is feasible to process them further using dimensionality reduction techniques.

A performance of classic dimensionality reduction techniques, such as Principal Component Analysis (PCA), is limited for nonlinear manifolds within high-dimensional spaces, which are typically learned with deep networks. Currently, for a nonlinear dimensionality reduction, a widely used technique is the t-Distributed Stochastic Neighbor Embedding (t-SNE) [110], which allows embedding high-dimensional data in a space of two or three dimensions, which are suitable for visualizations. Due to its properties, it is also a common choice for deep features visualization.

Although a full explanation of the t-SNE would be out of scope of this work, we provide a short intuition of it.

*t-Distributed Stochastic Neighbor Embedding (t-SNE)*

The t-SNE is essentially a machine learning algorithm, which models high-dimensional points in a low-dimensional space, with a focus on preserving local relationships between points, while retaining a global structure of the data.

Given points $x_i$ and $x_j$ in the original space, their similarity is expressed with a probability $p_{ij}$ of picking $x_j$ as a neighbor of $x_i$, where the probability is defined with a Gaussian. In the low-dimensional space the mappings of these points are defined as $y_i$ and $y_j$, where the similarity of the points is expressed in a similar manner with a probability $q_{ij}$, which is defined with a Student t-distribution in this case. The locations of the mapped points $y_i$ (embeddings) are computed by minimizing the Kullback–Leibler divergence between $p_{ij}$ and $q_{ij}$:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{7.1}$$

Such optimization over the distributions forces the points that are close to each other in the original space to settle closer in the low-dimensional space, while the opposite happens for distant points.

The choice of the t-distribution for the probability $q_{ij}$ in a low-dimensional space comes as a solution to the crowding problem, which appears due to the curse of dimensionality: the points that are distant in the original high-dimensional space can crowd too close to each other after a projection into the low-dimensional space. The heavy-tail property of the t-distribution helps to alleviate this problem.

An overview of the t-SNE learning process from the features in a photo assessment network is shown in Figure 7.4. We apply the described t-SNE approach to the deep features extracted from all images in each analyzed dataset. Due to a large size of the *AVA* and *Classic Photographers* datasets, we randomly sample about 1/10 of images: for these datasets the t-SNE computation was performed using 27,500 and 19,000 deep feature vectors, respectively. We used the t-SNE algorithm with the learning rate set to 200 and the maximum number of iterations set to 2000. The perplexity parameter was

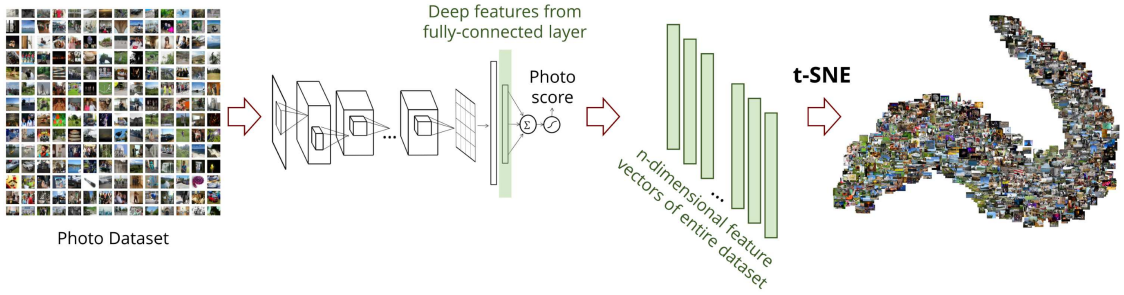


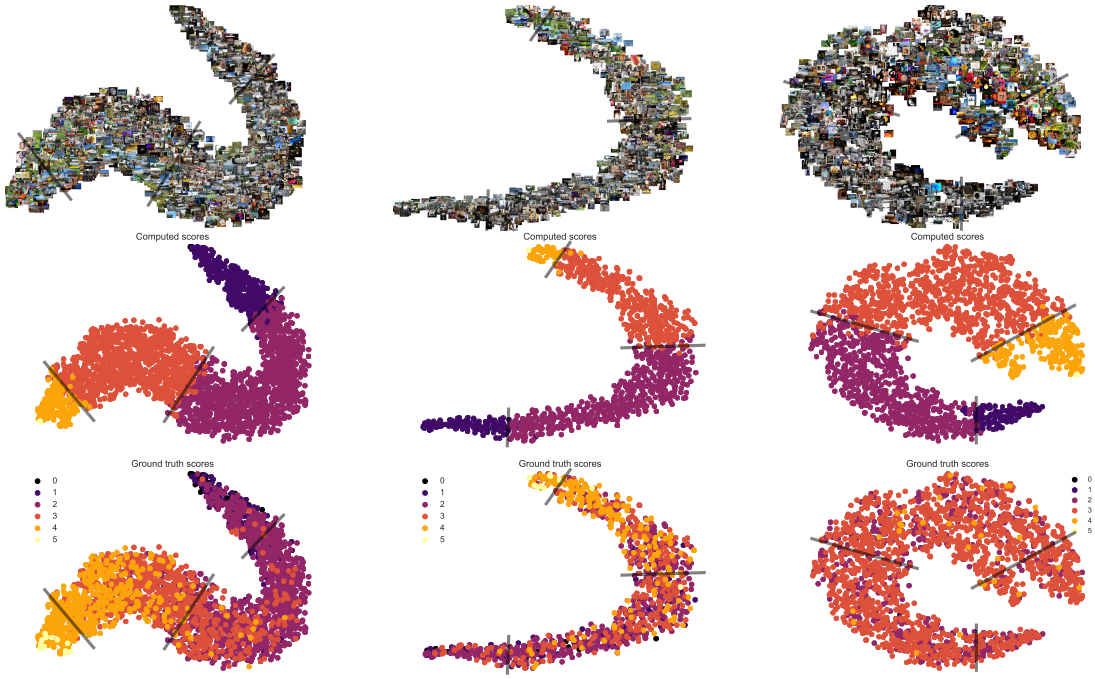

Figure 7.4: Learning t-SNE embeddings on a photo scoring deep network. Using deep features from the fully connected layer (computed for all photos in a dataset) we can learn t-SNE embeddings, which represent global relationships between images in the network's assessment space.

set to 30. The perplexity parameter affects a balance between learning more local or more global structures in the data: we have experimented with lower and higher values, but the chosen value of perplexity provided the most informative visualizations in our experiments.

The learned embeddings are visualized in Figure 7.5 and Figure 7.6. A close-up of the photo embeddings with low and high scores is shown in Figure 7.7. For each dataset we repeat the same 2D representations of deep features, but we visualize the obtained points with different data overlays — by using corresponding image scores or by using corresponding images directly. To improve legibility of visualizations, we limited a maximum number of plotted points and images by 2000 (randomly sampled for each dataset).

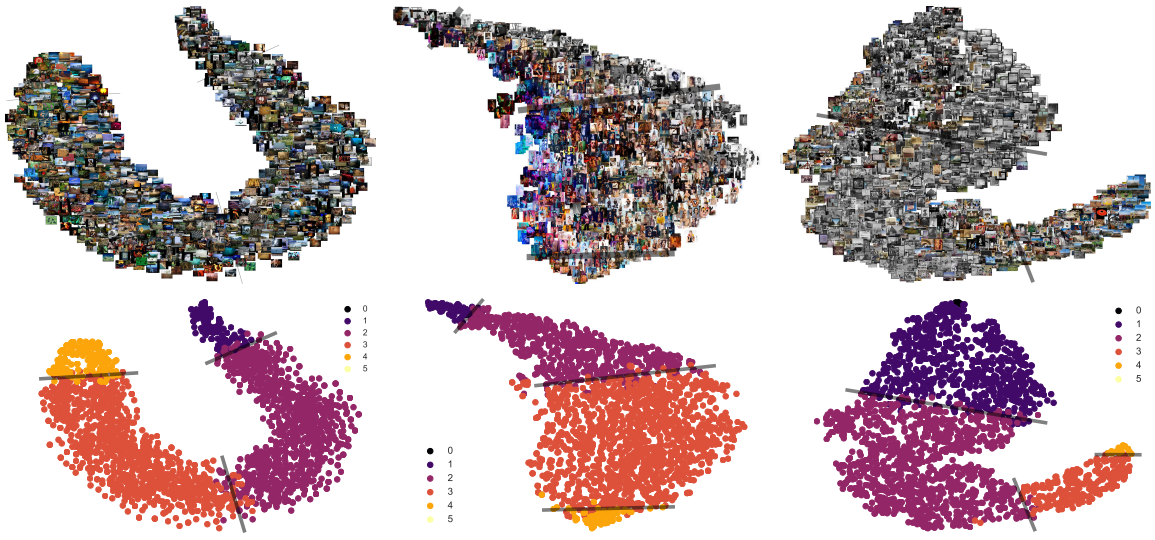
When examining the obtained embeddings, we should keep in mind that the computed deep features represent the original images not in their original space, but in the feature space learned by a specific method in a specific dataset. In our case, we observe the visualizations of features that were learned by the Kong's network from the AADB dataset. In effect, it allows the visualized embeddings to suggest an expected performance on each dataset and provide us insights what makes an image aesthetically appealing according to the analyzed method.

First, we explore the embeddings learned for the datasets where the ground truth scores are available (Figure 7.5). For the computed scores, we can observe a very clear score separation in each dataset, which is expected, since these embeddings essentially represent the assessment space learned by the CNN. If we compare visualizations of the computed scores and the ground truth scores, we can observe the following. For the *AADB* dataset, a quite distinct data separation is present for ground truth scores: albeit not perfect and more noisy, the approximate boundaries in the visualizations of the ground truth and computed scores are similar. In addition, we can observe that images with high or low scores are rare in the ground truth data (which corresponds with our earlier observation on the scores histograms), and the cases of images with a computed score of 5 or 0 are very rare.




(a) AADB train data (b) AADB test data (c) AVA

Figure 7.5: t-SNE embeddings of deep features in photo datasets: datasets with ground truth scores available. The plotted lines reflect approximate boundaries between scores computed by the analyzed method. The figure is best viewed electronically. *Note:* the original scores are continuous values — for the visualization purpose we rescaled the scores to a range from 0 to 5 and rounded them to integer labels.




(a) National Geographic (b) Fashion (c) Classic Photographers

Figure 7.6: t-SNE embeddings of deep features in photo datasets (datasets without ground truth scores available). The plotted lines reflect approximate boundaries between scores computed by the analyzed method. The figure is best viewed electronically. *Note:* the original scores are continuous values — for the visualization purpose we rescaled the scores to a range from 0 to 5 and rounded them to integer labels.

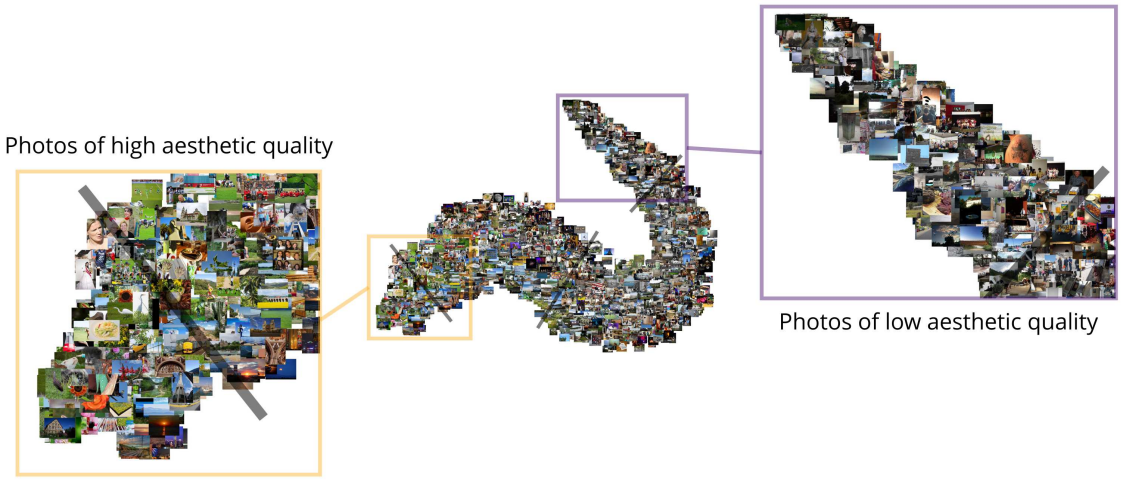
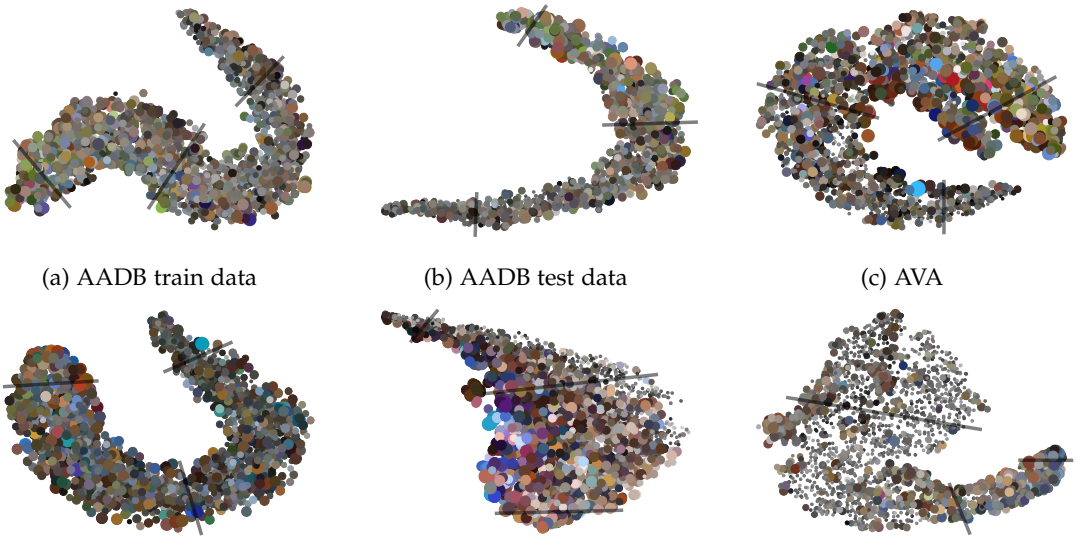



Figure 7.7: Examples of the t-SNE embeddings for photos with low and high scores in the training data of the AADB dataset. This figure represents a close-up of the photo embeddings in Figure 7.5(a).

At the same time, the embeddings learned for the *AVA* dataset do not have the same agreement between computed and ground truth scores. When visualizing the ground truth scores with the learned embeddings, the scores are in fact very mixed and appear noisy. Such discordance between scores visualizations and the lack of score boundaries in the visualization of ground truth might suggest a poor performance of the Kong's method on the AVA dataset and an overall lack of generalization. It is not surprising, since in the original paper [75] the authors have concluded that their model learned on the AADB dataset does not perform equally well on the AVA dataset.

As for other datasets, shown in Figure 7.6, it is more difficult to draw conclusions in the absence of the ground truth. Although we can also observe clear boundaries for the computed scores in this case, they do not indicate the expected performance on these datasets. Nevertheless, by juxtaposing the computed scores and image visualizations we can notice certain patterns in the method's assessments.

If we compare the areas of low and high scores, we can see that the low scores area are largely represented by non-saturated and dark images, while the areas of high scores are mostly represented by colorful and bright images (see Figure 7.7). This appears to be the case for all analyzed datasets, including the AADB and AVA datasets. In the original AADB dataset, which was used for training, we can note that colorful pictures of nature in general are rated higher than non-vivid images of various topics. To further analyze this finding, we plotted the t-SNE embeddings with another visualization overlay. In Figure 7.8 each image was plotted with a point of varying color and size, where the point color represents an average image color, and the point size represents an image colorfulness (computed with the method of Hasler et al. [52]). Here, we can observe that the images of low colorfulness almost never obtain high scores. At the same time, the largest points representing the most colorful images are mostly concentrated in the areas of high score. This effect is especially evident in the *Classic Photographers* dataset,



(a) AADB train data (b) AADB test data (c) AVA

(d) National Geographic (e) Fashion (f) Classic Photographers

Figure 7.8: t-SNE based visualization of photo colorfulness in datasets. The color of a point represents average image color, and the size of a point represents its computed colorfulness.

| | AADB | | | | AVA | | National Geographic | Fashion | Classic Photographers |
|---|---|---|---|---|---|---|---|---|---|
| | GT | | Computed | | GT | Computed | | | |
| | train | test | train | test | | | | | |
| Pearson correlation coefficient | 0.235 | 0.252 | 0.360 | 0.469 | -0.011 | 0.49 | 0.4 | 0.37 | 0.497 |

Table 7.1: Correlation between aesthetics scores and a colorfulness measure. The *GT* denotes the ground truth user scores from the original datasets. The *Computed* denotes the scores computed with the method of Kong et al. [75].

where a number of color photos is low: the observed score patterns suggest that the analyzed method is probably not applicable to assess black-and-white photos in general.

To produce a numerical analysis of the observed tendencies, we have computed a correlation between aesthetic scores and the obtained values of image colorfulness, using the Pearson correlation coefficient (see Equation 6.3). The results are shown in Table 7.1. We can observe that the correlation varies between datasets, but for certain data its value is close to 0.5. Interestingly, for the original *AADB* dataset that was used for the method's training, the colorfulness appears not to be substantial in original user decisions, but its importance increases for the trained method. We can hypothesize that although users originally consider a number of other factors in their decisions apart from colorfulness, the network does not succeed to capture these characteristics and the training process leads it to rely on color characteristics much more. The most significant change can be also noticed for the AVA dataset: while for the user ground truth scores even almost no correlation is observed, the assessment of the same photos with the method by Kong et al. appears to rely considerably on color characteristics.

Although the obtained results are not sufficient to conclude that the colorfulness is a crucial aesthetics factor for the analyzed method, we can certainly notice its influence on the computed scores. At the same time, we cannot be certain if it is the only bias present, as other influential characteristics are hard to isolate in the same manner. We have observed that the training process can lead a network to rely on particular characteristics more than users do, which might suggest that other factors are not properly learned by a network. In addition, while our findings are specific for the analyzed method of Kong et al., trained on the AADB dataset, we can expect that similar tendencies or biases (not necessarily color-specific) could be present for other methods. In conclusion, we would like to emphasize that we do not aim to undervalue the contribution of the specific analyzed method, but rather demonstrate that we should be aware that a use of any photo assessment method requires deeper understanding of the method's applicability.

## 7.2 AESTHETIC GANS: GENERATIVE ADVERSARIAL NETWORKS WITH AESTHETIC QUALITY CONDITIONS

In the previous section, we saw that a CNN-based image assessment method, despite being trained on an extensive set of images with ground truth user scores, showed a measurable bias toward colorful images. To further explore potential biases in the aesthetic assessment of images, we employ generative models which allow us to go a step further. By generating novel images with different levels of aesthetic quality, we can visualize the implicit visual characteristics encoded within the network.

### 7.2.1 *Auxiliary Classifier GAN*

Generative adversarial networks get their inspiration from game theory: inside a GAN, two networks compete with each other, where one tries to generate image samples that look as realistic as possible (Generator $G$), and another tries to determine if the samples are coming from the Generator or if they are real data (Discriminator $D$). These two networks are trained in an alternating manner, and both of them are supposed to get gradually stronger in their task.

The input of the generator network is a noise input $z$ and the output is a fake sample $G(z)$. The input of the discriminator network is either a real sample $x$ or a fake sample $G(z)$, and the output is $D(x)$ and $D(G(z))$, which are probabilities of the given sample being real or fake, correspondingly. Essentially, the original GANs are optimizing the following loss function in a minimax approach:

$$\begin{aligned} \min_G \max_D L_S(D,G) &= \mathbb{E}_{x\sim p_r(x)} \overbrace{[\log D(x)]}^{\substack{\text{Likelihood of real} \\ \text{data } x \text{ being real}}} + \mathbb{E}_{z\sim p_z(z)} \overbrace{[\log(1-D(G(z)))]}^{\substack{\text{Likelihood of fake} \\ \text{data } G(z) \text{ being real}}} \\ &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1-D(x)], \end{aligned} \tag{7.2}$$

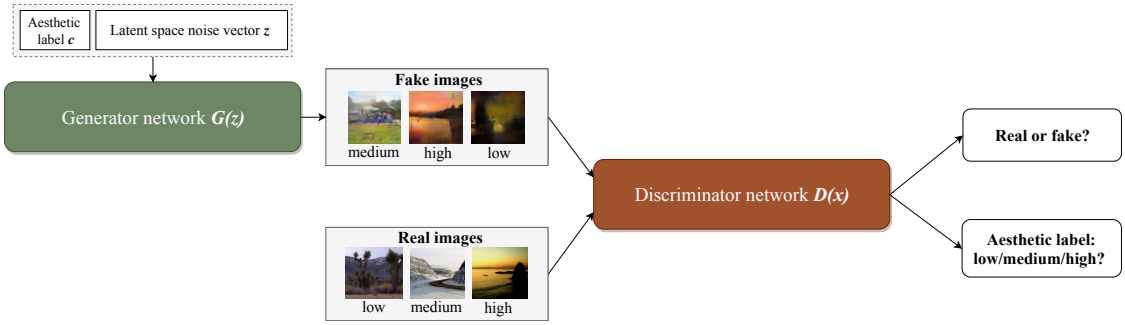



Figure 7.9: Architecture of the AC-GAN network with aesthetic label conditioning. Following the training process, the generator network $G(z)$ learns to produce realistically looking images with a given label of low, medium or high aesthetic quality.

where $p_r$ is an unknown complex distribution over real image samples $x$, $p_g$ is a generator's distribution over fake image samples, and $p_z$ is a distribution over noise input $z$. Ultimately, we would like to obtain a generator that can draw fake samples from the distribution $p_g$ as similar as possible to real samples from the unknown distribution $p_r$.

The original GAN architecture can learn to generate realistically looking image samples from the provided data, but it does not have control over the produced output. One approach to control the output is to use class-conditional GANs, which add a supervised part to a GAN. In our experiments we adopt the Auxiliary Classifier GAN (AC-GAN) [128].

The AC-GAN is a variant of traditional GANs, with an additional classifier loss introduced. Each input noise vector $z$ is complemented with a class label $c \sim p_c$, and the generator uses both inputs to generate a fake image $G(c, z)$. The real image samples $x$ received by the discriminator are also complemented by the class label (known from the dataset in this case). Now, the discriminator should determine not only if given samples are fake or real, but it should also learn to classify them into correct classes. To do this, the authors introduced an additional classification loss:

$$L_C(D, G, C) = \underset{\substack{x \sim p_r(x) \\ c \sim p_c(x)}}{\mathbb{E}} \overbrace{[\log D(c|x)]}^{\substack{\text{Likelihood of predicting} \\ \text{class } c \text{ for real sample}}} + \underset{\substack{x \sim p_g(x) \\ c \sim p_c(x)}}{\mathbb{E}} \overbrace{[\log(D(c|x)]}^{\substack{\text{Likelihood of predicting} \\ \text{class } c \text{ for fake sample}}} . \tag{7.3}$$

Essentially, this loss forces the discriminator to distinguish well between classes and the generator to generate samples with easily distinguishable classes.

In the AC-GAN paper [128] the traditional loss from Equation 7.2 is used as a likelihood of the correct source $L_s$, and the classification loss from Equation 7.3 is used as a likelihood of the correct class $L_c$. Then, the optimization is formulated in the following way. The discriminator $D$ is trained to maximize $L_s + L_c$, while the generator

$G$ is trained to maximize $L_c - L_s$. This way, the generator ultimately learns how to produce realistic samples with a given input label.

The original proposal of the AC-GAN aims to generate images of different visual classes, such as *butterfly*, *flower*, *dog*, *person*. In our experiments, we use the aesthetic label instead. Thus, the real images are supplied with a label of *low*, *medium* or *high* aesthetics, derived from the corresponding dataset. Following the general logic of the AC-GAN network, the generator will gradually learn how to produce realistic images with different aesthetics. A scheme of AC-GAN training with aesthetic labels is shown in Figure 7.9

As the last point, we should mention an additional consideration that can help to train a GAN network. The GAN networks are known for a number of problems: their *training instability*, the possibility of *vanishing gradients* when the discriminator becomes strong much faster than the generator, which causes the latter to stop learning, and also a *mode collapse*, when the generator get stuck in a small space of same generated image samples. These problems were considerably alleviated with an introduction of another loss function based on the Wasserstein distance [2, 49]. In essence, the proposed approach replaces the minimax objective from Equation 7.2 with the new objective:

$$\min_G \max_D L_S(D, G) = \mathbb{E}_{x \sim p_r(x)}[D(x)] + \mathbb{E}_{x \sim p_g(x)}[D(x)], \tag{7.4}$$

where the discriminator does not act as a direct evaluator of the data authenticity but rather as an estimator of the Wasserstein distance between real and generated data distribution. This allows for a smoother training procedure and also provides a meaningful loss function to track the training process. For a more detailed description of the Wasserstein distance approach in GANs, we recommend to refer to the original papers [2, 49].

In our experiments we have used the modified AC-GAN version where the Wasserstein distance was used as the source loss $L_s$. The full network architecture can be referred in the original paper [128]. We base our network training on the implementation by Nguyen [59]. The following parameters were used in most experiments: image dimension = 64x64, batch size = 128, optimizer = *Adam* ($\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$).

### 7.2.2 *Generating Images of Different Aesthetic Quality using AC-GAN*

In our early experiments we were observing the overall potential of the AC-GAN to generate meaningful examples. In our analysis we have used the AADB dataset [75] and the AVA dataset [122]. Since the original datasets contain continuous scores on the range from 0 to 10 (not the categorical labels) and their distribution is close to normal, we have split each dataset into three approximately equal parts and assigned the labels of *low*, *medium* and *high* aesthetics correspondingly.

First, we based our network training on the complete AADB dataset. Examples of results, produced by the trained generator after 25,000 iterations, are shown in Figure

7.10. Here, only random samples were generated, with no specific label condition. Using the trained generator, we could not produce many realistically-looking images, which is not surprising, as the image structures drastically vary in the AADB dataset. At the same time, we have observed that the most meaningful images resembled the landscape-like structure.

Figure 7.10: Results of early trials of the AC-GAN training: random samples generated with GAN trained on AADB dataset. Aesthetic label generation was not considered at that point. Although the generated images were not meaningful for most samples, we have observed that the model was able to produce somewhat plausible looking images for landscape photos.

Then, we have decided to select specific image classes for our training. In the AVA dataset, apart from aesthetic scores information, a set of visual classes is available, which includes such classes as *animal*, *architecture*, *landscape*, *portrait*, *still life*. After our experiments with different classes, we have found that the most stable results are produced by training on images from the *landscape* class (all 5,000 photos of this class that are available in *AVA*).

Examples of random landscape samples, produced by the trained generator after 48,000 iterations, are shown in Figure 7.11. As we could observe a number of meaningful images (despite being very low resolution) among the generated samples, we decided to focus our analysis on this class, using the trained model.

Other examples of images generated with the same landscape photos generator can be found in Figures 7.12 and 7.13. One approach to investigate if the model overfitted on the training data is to explore the latent space between samples by interpolation. In Figure 7.12, we performed interpolation between two different noise vectors and generated intermediate image samples. As the intermediate regions in the latent space do not show very abrupt transitions in image space and the overall realism is mostly preserved, we can suggest that the model learned various types of landscape photos.

As a side note, in Figure 7.12 we show examples of interpolation using two interpolation approaches, linear and spherical. Although it is rarely highlighted in the literature, the spherical interpolation between input noise samples provides sharper intermediate samples with more clear structural changes, which is attributed to the shape of the latent space and a uniform prior on input noise data [178].

Armed with a reasonable confidence on the ability of our trained GAN to generate plausible landscape images, we can now analyze how aesthetics are encoded in the network and as a consequence in the generated images. To that end, we generate a

Figure 7.11: Random samples generated with AC-GAN trained on *AVA* landscape class. Aesthetic label generation was not considered at that point. As the image samples are randomly generated, many of them still do not look realistic. However, certain samples resemble real landscape photos.

series of images using the same noise sample but modifying the aesthetic label. Some example results are given in Figure 7.13

(a) Linear interpolation between data samples

(b) Spherical interpolation between data samples

Figure 7.12: Interpolation between data samples in the learned latent space. Given two images and their corresponding latent vectors, we can produce intermediate samples between two images, which show a gradual transition from one image to another. For the latent spaces of GAN models, the spherical interpolation is preferred, since it produces more distinct and sharp intermediate samples.

For numerous examples we have observed that the label change indeed affects the appearance of an image. Most importantly, we have observed that the label change mostly affects the color characteristics of an image: the images of *high* aesthetics typically receive more vivid colorful look. For some cases, the images get more dramatic, high contrast appearance with a change towards higher aesthetics. It is not surprising, since the training *AVA* dataset also shows similar tendencies, that the post-processed photos, with vivid colors and high contrast between dark and bright areas, have higher scores, especially for landscape photos. Nevertheless, it is interesting to observe the same effect learned by the generative network. In addition, this observation also corresponds with our earlier finding for the assessment method.

The observed model behavior suggests that it is potentially possible to generate a picture with the desired aesthetic quality. We could also imagine another scenario. What if we have captured a photo and would like to know what would make it more attractive, using our model? Then, we need to inverse the GAN process and find a representation of a given photo in the latent space of our generator. In this case, we know an image


Figure 7.13: Label change for the latent vector and the generated image data. For landscape-like photos generated with a trained model, a change of color characteristics is often observed along with the label change.

space sample $G(z)$ but we do not know a latent space vector $z$ that generated this image. If we are able to find such a vector $z$, we could simply use it as the generator's input, along with an aesthetic label $c$, to generate a new image of desired aesthetics.

A straightforward solution to recover latent vectors from image samples is to optimize directly over a noise vector reconstruction $z'$, so the generated sample $G(z')$ will be as close as possible to our input image $G(z)$ [92]. This can be formulated as an optimization problem in the following manner:

$$\min_{z'} ||G(z) - G(z')||_2^2, \tag{7.5}$$

where we minimize the $l_2$ norm between a given image sample $G(z)$ and a reconstruction $G(z')$.

First, we have performed such reverse search for an image produced by the generator, i.e. the image belonging to the generator's distribution $p_g$. As we can see in Figure 7.14, the optimization for such image sample is able to find a vector $z'$ that generates a reconstruction almost identical to the original image. Also, the label change both for the original and the reconstructed image produces a similar effect.

Then, we used an input image that is completely unknown to our generator model, as shown in Figure 7.15. While the image reconstructed from the found latent vector preserves certain geometrical patterns and colors of the original image, the results are less convincing. At the same time, with such low-resolution GANs, we cannot expect a very detailed result. Although the image details are not well defined, we can observe similar tendencies in label change: the results of *high* aesthetics present higher contrast and more saturated colors.


(a) (b)

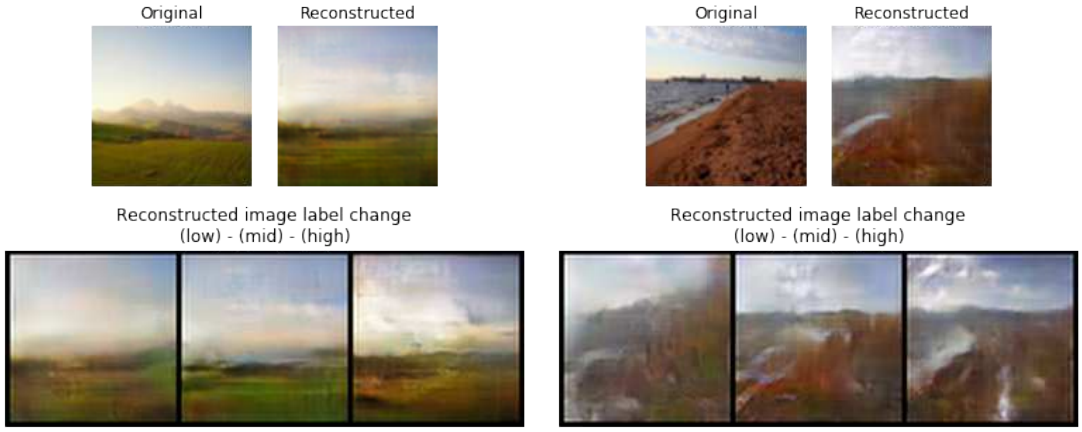
Figure 7.14: Latent vector search and label change for an image present in the generator space. The image reconstructed from the latent vector highly resembles original image, and the effect of the label change is also similar.

In our results we have found that although traditional GANs cannot produce images of high aesthetic quality to use their output directly, the generated images provide interesting observations for further analysis. First, the obtained results confirm the findings of the previous section, i.e. that colorfulness seems to be a major component in what automatic methods consider as aesthetically good. Second, we can imagine that a generator output could potentially serve as an indication or guidance of which properties to improve in an image (although the influence of the dataset on the generator training is still important in this case).

### 7.2.3 *CycleGAN for Changing Aesthetic Quality of a Photo*

We have also conducted additional experiments with another type of GAN networks. The Cycle-Consistent Adversarial Network (CycleGAN) [193] allows to learn a mapping that translates an image from a source domain $X$ to a target domain $Y$, in absence of paired examples between two domains. To do so, they couple a mapping $G : X \rightarrow Y$ with an inverse mapping $F : Y \rightarrow X$, and enforce a cycle consistency between two domains, so that $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$. In this case, two discriminators are present: $D_X$ forces $G(X)$ to produce outputs similar to $Y$, and $D_Y$ forces $G(Y)$ to produce outputs similar to $X$.

The unpaired image-to-image translations provide powerful transformations between different domains, and they have shown applications in different tasks, such as translations between paintings and photos and a change of object classes appearance [193]. For our purpose, we use it to learn a mapping between images of different aesthetics: we set the images of low aesthetics as a source domain $X$ and the images of high aesthetics as a target domain $Y$. In our experiments we have used the same *landscape* subset of the AVA dataset, where we have selected only images of low and high aesthetic labels.




(a) (b)

Figure 7.15: Latent vector search and label change for an image not present in the generator space. In this case, the reconstructed image has a resembling appearance, but image details are not well defined. The effect of label change is less evident than for images originally represented in the latent space.

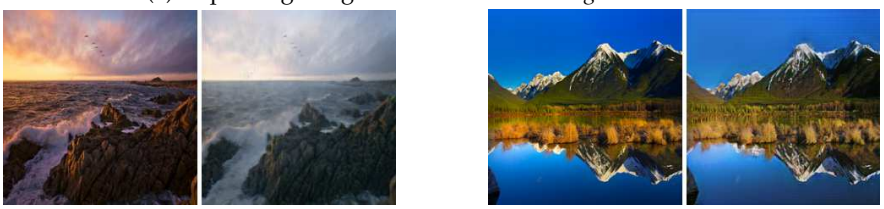
The example results of a trained network are shown in Figure 7.16. The architecture of CycleGAN allows the use of images of much higher resolution, compared to traditional GANs, and it also preserves the details of the original input. In the obtained results we show examples of the learned translations in both directions: from low to high aesthetics, and vice versa.

We can observe that for the $low \rightarrow high$ translation, the network has learned to apply certain image enhancements, mostly related to color and contrast adjustments. At the same time, the $high \rightarrow low$ translation provides an interesting effect. For one of the observed examples (sea landscape) the colors become completely desaturated, along with a large loss of sharpness. For another example (mountain landscape) the overall appearance does not change significantly, but a number of image artifacts appear after the translation.

Interestingly, in our experiments on AC-GANs and CycleGANs, we have not observed any noticeable changes in photo composition, although it is a common aesthetics factor according to photography practices. Currently, it is known that the GANs, especially CycleGANs, do not succeed to learn object shape changes [44, 91]. As an optimal composition search usually requires large changes in image structure, we possibly cannot expect a typical GAN to learn such transformations. However, as the research on GANs is still in its early stage, we can expect that the future studies might address this issue. At the same time, we have observed that the use of a dataset largely defines what we are able to learn with any network model, either discriminative or generative. Possibly, the introduction of more task-specific datasets for aesthetic assessment could also help in this regard.

Overall, our studies have shown the potential of generative adversarial networks for the exploration of image aesthetics, but this subject requires further investigation. In

(a) Improving image aesthetics: *low* → *high* translation.



(b) Deteriorating image aesthetics: *high* → *low* translation.

Figure 7.16: Examples of the aesthetic label translations learned with CycleGAN.

our experiments we have arrived to certain preliminary conclusions: we have observed that it is possible to generate meaningful images for certain photo classes and that the color characteristics are important when generating images of given aesthetics. The latter finding also aligns with earlier findings on photo assessment in Section 7.1. At the same time, the generative methods are currently limited in their capabilities, which does not allow us to explore all possible types of photos and which possibly limits the generated image transformations. Nevertheless, we believe that the GANs give a promising direction in further exploration of photo aesthetics.

## 7.3 SUMMARY

Our experiments with an independent photo assessment method showed that the choice of the training data largely defines the learned 'aesthetics space'. We observed that the scoring of photos from other datasets, which have not constituted the training or test data, often leads to rather inconsistent results. When visualizing the computed deep features using the t-SNE embeddings, a similar effect was found: the analyzed method achieves a good score separation for its native dataset, but the results for other datasets are less distinct. At the same time, we observed a tendency that more colorful and vivid images often get higher scores, while unsaturated and black-and-white photos usually get low scores.

A similar finding arose in our experiments with AC-GANs, where an aesthetic label is provided as an input, along with a noise vector. While current generation capabilities of GANs are limited, and they do not train well on mixed image types, they are able to produce meaningful images from certain classes, such as landscape photos. After training a generator on landscape photos, we observed that the label change leads to the generation of more colorful images with higher contrast. At the same time, we were not able to make conclusions regarding a possible link between photo composition and aesthetics, since GANs are limited in the structural changes they introduce.

Although this chapter reflects mostly preliminary studies done on the exploration of the aesthetics learned by CNNs, we believe that they provide interesting directions for the future research. Further studies of these subjects could improve the computational models of aesthetics and possibly bring a better understanding of human perception of photography.

Finally, returning back to the subject of photos in photo albums, we can highlight the relevance of discussed topics. First, by improving the photo assessment approaches in general, we can expect an improvement in the context-based photo assessment as well. Our proposed context-adaptation approach already shows a considerable improvement of the original scores, but a more reliable score would also provide a more reliable adaptation. Regarding the GAN models, besides providing us additional insights in computational aesthetics, they could essentially be used to improve a given photo. We can imagine a scenario, where the best photos are first selected in a photo album using a context-adaptive method, and then these photos are further improved using a generative approach. This way, even a full cycle of a digital photo lifespan could be covered. The only thing that would be required from the user would be to decide which moment to capture.

# CLOSING

# 8

# GENERAL CONCLUSION

## THESIS SUMMARY

"The best camera is the one that's with you" — the phrase attributed to the photographer Chase Jarvis — is a motto of many amateurs of the photography, and nowadays following it became as easy as never. Thanks to smartphones, we can always have a camera with us, and we can take as many pictures as we want. One question remains: who will help us to deal with all these photos later?

The field of automatic solutions that assist the process of photo management and selection continues to grow constantly. At the same time, the images in photo albums are often treated independently, when other photos from the same album are not taken into account. For instance, when we want to assess a photo in our personal collection, we consider it in the context of similar photos taken in the same scene. However, among the proposed automatic photo assessment approaches, the context is rarely addressed. In our work, we attempted to analyze this problem of the context gap and propose our own solutions to it.

More precisely, in this thesis we addressed two main topics. First, we conducted experimental studies on the nature of clustering and selection in photo collections, and we analyzed how the clustering-defined context affects users' decisions. Then, we have proposed new methods that can perform an automatic clustering of photo albums, where the computed clustering is further used as the context basis for an adaptation of an independent photo score.

During our experimental studies, the following findings have been made:

- In the study on image selection in photo albums, users achieve a certain selection agreement, but the agreement varies depending on the album's content. Users achieve higher agreement when dealing with people's photos, whereas they tend to disagree in landscape-focused albums. Also, very similar near-duplicate photos often lead to uncertain selections.
- In the study on clustering in photo albums, a certain level of agreement is present for all albums. The level of agreement often depends on the availability of particular landmarks, which can simplify the clustering task. Similarly to the selection task, the albums with people present in photos seem to be easier to cluster, as the moment boundaries are usually more clear.

- If we aggregate the clustering and selection data together, we can observe the influence of the context more clearly. We have found that, on average, users select around 37% of the photos in albums. Also, unique photos are not selected more often than others. Finally, the median value of selected image per cluster appears to increase along with the size of scene clusters, but holds around 1 for near-duplicate clusters.

Regarding the automatic clustering of photo albums, our proposals and conclusions can be summarized as follows:

- We proposed a hierarchical approach to cluster photo albums in a completely automatic manner: the proposed method does not require a pre-defined number of clusters, and the output structure is convenient both for the context modeling and for the collection browsing.
- Based on the observation that users tend to adapt the granularity of their grouping decisions, we proposed an adaptive cut approach, which depends on the average distance at the higher hierarchy level. The adaptive cut approach also removes the need for predefined cut thresholds in the clustering procedure.
- For our clustering approach, we tested different descriptors as the base for the visual similarity distance computation. We found that both SIFT and CNNR descriptors can be successfully applied to this task, although CNNR global descriptors are more robust overall.
- Using the data acquired in our clustering user study, we performed the performance evaluation of the proposed clustering methods. Overall, we achieved a high level of agreement with the users' partitions; for some albums, the performance is very close to inter-user agreement.

Regarding the context-based adaptation of a photo score, our proposals and conclusions can be summarized as follows:

- To improve the scoring performance of traditional photo assessment methods, we proposed to use the clustering information to define the photo context and adapt the pre-computed independent score according to it. For this purpose, we have introduced two adaptation approaches. The first adaptation approach is based on the multi-level z-scores computation. The second adaptation approach employs a shallow neural network, which is trained using a set of data features, extracted from the clusters' corresponding to the assessed photo.
- The performance of the proposed approaches has been assessed using the selection data acquired in the user study. The proposed machine learning based technique provides the highest performance: the adapted scores show a significant improvement in correlation with user preferences, in comparison with original unadapted scores.
- By performing an ablation study, we also found that the number of images within a scene and a near-duplicate cluster helps the training procedure; it conforms with our user study findings.

Additionally, we performed a set of experiments to explore and visualize the aesthetic characteristics learned by CNN-based methods, and we came to following conclusions

- After testing an independent assessment method on different data, along with examining the t-SNE based embedding visualizations, we have found that the learned 'aesthetic space' is largely defined by the original training dataset. Thus, the applicability of a trained method to other data is not particularly evident. Our findings also indicate that certain characteristics (such as color characteristics in the case of analyzed method) might be determining in the scoring process.
- By training a conditional GAN network on landscape photos, we were able to generate image with different aesthetic priors. Similarly to the analyzed photo assessment method, the photos of *high* aesthetics label, generated with our network, show more saturated and vivid colors.

## FUTURE WORK AND PERSPECTIVES

In addition to the presented contributions, we believe that this work opens some promising directions of future research.

CONTEXT ADAPTATION The context-based photo selection could benefit from further experimental studies, for example by doing a large scale version of the user studies conducted in this work. A larger available data could lead to additional insights on the context influence in photo albums and also provide a more reliable base for the training of the score adaptation solution. On the other hand, the concept of the context itself can be expanded. While in this work we have focused on the context mostly represented by similar photos in a collection, there are other possible definitions of it. The context within a photo album can be also defined on a higher semantic level, for example by a general topic of the depicted event, or by using extra knowledge, such as family or social connections between people in photo.

USER PREFERENCES PERSONALIZATION Even with a flawless album-adaptive score adaptation, it is impossible to create a singular solution that would suit each and every user. We have observed that users themselves can largely vary in their decisions, which suggests the need for the personalized automatic solutions. Such solutions could adapt not only to the nature of the photo collection, but also to the individual way of photo organization by a user. If it would be possible to obtain a large amount of personalized data, our adaptation approach could be also applied in a per-user manner, now treating only albums from a specific user.

INDEPENDENT PHOTO ASSESSMENT One of the main possible weaknesses of our approach is the original aesthetic score itself. As we heavily rely on the independent score computation, the failure of the independent score can lead to an incorrect adaptation result, and hence a possible propagation of the error across the scene of related photos. Not limited by our approach, achieving more robust independent photo assessment

would bring improvement to multiple applications related to computational aesthetics. As we have observed in our exploration of CNN-based methods, an additional attention should be paid to the training of photo assessment methods and utilized datasets. At the same time, further studies of generative networks could possibly provide useful insights on the computational aesthetics and possible directions of improvement.

PHOTO ENHANCEMENT We have seen that GAN models can present a novel direction for image manipulations. Apart from their exploratory applications, we can expect that in future the GAN-based methods will be more and more common in image processing and enhancement. As we have mentioned earlier, the following scenario is also possible: when a set of good photos in an album is obtained using a context-aware approach, we could go a step further and enhance these best photos with an aesthetics-based GAN approach.

BEST PHOTO CREATION FROM THE CONTEXT We can also imagine another application of our photo albums clustering. When a number of photos from the same moment are taken, they can be used to synthesize one 'ideal' photo representing this moment. For example, from a set of group photos, one best group photo can be created, where all people are optimally portrayed. Also, a combination of photos could be used to improve certain photo characteristics, e.g. by using a super-resolution approach. On the other hand, a set of photos of the same scene can be used for scene summarization: for instance, when a photographed scenery does not fit into a camera's field of view, we typically resort to take a panoramic image. As a panorama is not the most convenient way to visualize a large scene, a scene summarization photo could be created, that would represent the main captured objects in a compact way, possibly with an aid of recomposition and retargeting techniques. Such compact scene representation could be used for informative or creative purposes.

# BIBLIOGRAPHY


[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. "Sort story: Sorting jumbled images and captions into stories." In: *arXiv preprint arXiv:1606.07493* (2016).

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN." In: *arXiv preprint arXiv:1701.07875* (2017).

[3] Soonmin Bae, Sylvain Paris, and Frédo Durand. "Two-scale tone management for photographic look." In: *ACM Transactions on Graphics (TOG)* 25.3 (2006), pp. 637–645.

[4] Benjamin B Bederson. "PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps." In: *The Craft of Information Visualization* (2003), pp. 66–75.

[5] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. "A framework for photo-quality assessment and enhancement based on visual aesthetics." In: *Proceedings of the international conference on Multimedia - MM '10* (2010), p. 271.

[6] Subhabrata Bhattacharya, Behnaz Nojavanasghari, Tao Chen, Dong Liu, Shih-Fu Chang, and Mubarak Shah. "Towards a comprehensive computational model foraesthetic assessment of videos." In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM. 2013, pp. 361–364.

[7] Ake Bjorck. *Numerical methods for least squares problems*. Vol. 51. Siam, 1996.

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van. "Event Recognition in Photo Collections with a Stopwatch HMM." In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 1193–1200.

[9] Anselm Brachmann and Christoph Redies. "Computational and experimental approaches to visual aesthetics." In: *Frontiers in computational neuroscience* 11 (2017), p. 102.

[10] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. "Learning photographic global tonal adjustment with a database of input/output image pairs." In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 97–104.

[11] Liangliang Cao, Jiebo Luo, Henry Kautz, and Thomas S. Huang. "Annotating collections of photos using hierarchical event and scene models." In: *26$^{th}$ IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2008).

[12] Andrea Ceroni, Vassilis Solachidis, Mingxin Fu, Nattiya Kanhabua, Olga Papadopoulou, Claudia Niederee, and Vasileios Mezaris. "Investigating human behaviors in selecting personal photos to preserve memories." In: *2015 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2015* July (2015).

[13] Andrea Ceroni, Vassilios Solachidis, Claudia Niederée, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. "To keep or not to keep: An expectation-oriented photo selection method for personal photo collections." In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM. 2015, pp. 187–194.

[14] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines." In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011), 27:1–27:27.

[15] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. "Automatic triage for a photo series." In: *ACM Transactions on Graphics (TOG)* 35.4 (2016), p. 148.

[16] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. "Aesthetic Critiques Generation for Photos." In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE. 2017, pp. 3534–3543.

[17] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study." In: *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE. 2017, pp. 226–234.

[18] Wei-Ta Chu, Yu-Kuang Chen, and Kuan-Ta Chen. "Size does matter: How image size affects aesthetic perception?" In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM. 2013, pp. 53–62.

[19] Wei-Ta Chu and Chia-Hung Lin. "Automatic Selection of Representative Photo and Smart Thumbnailing Using Near-duplicate Detection." In: *Proceedings of the 16$^{th}$ ACM International Conference on Multimedia*. MM '08. ACM. Vancouver, British Columbia, Canada, 2008, pp. 829–832.

[20] Jacob Cohen. "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[21] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. "Temporal Event Clustering for Digital Photo Collections." In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 1.3 (Aug. 2005), pp. 269–288.

[22] Ritendra Datta, Jia Li, and James Z Wang. "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition." In: *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE. 2008, pp. 105–108.

[23] Ritendra Datta and James Z. Wang. "ACQUINE: Aesthetic Quality Inference Engine - Real-time Automatic Rating of Photo Aesthetics." In: *Proceedings of the International Conference on Multimedia Information Retrieval*. MIR '10. ACM, 2010, pp. 421–424.

[24] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. "Studying aesthetics in photographic images using a computational approach." In: *Computer Vision–ECCV 2006* (2006), pp. 288–301.

[25] Claire-Hélène Demarty, Mats Sjöberg, Mihai Gabriel Constantin, Ngoc Q. K. Duong, Bogdan Ionescu, Thanh-Toan Do, and Hanli Wang. "Predicting Interestingness of Visual Content." In: *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 2017, pp. 233–265.

[26] Yubin Deng, Chen Change Loy, and Xiaoou Tang. "Image aesthetic assessment: An experimental survey." In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 80–106.

[27] Yubin Deng, Chen Change Loy, and Xiaoou Tang. "Aesthetic-driven image enhancement by adversarial learning." In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 870–878.

[28] S. Dhar, V. Ordonez, and T. L. Berg. "High Level Describable Attributes for Predicting Aesthetics and Interestingness." In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1657–1664.

[29] Zhe Dong, Xu Shen, Houqiang Li, and Xinmei Tian. "Photo quality assessment with DCNN that understands image well." In: *International Conference on Multimedia Modeling*. Springer. 2015, pp. 524–535.

[30] David DuChemin. *Within the frame: the journey of photographic vision*. New Riders, 2009.

[31] Delbert Dueck and Brendan J Frey. "Non-metric affinity propagation for unsupervised image categorization." In: *2007 IEEE 11$^{th}$ International Conference on Computer Vision*. IEEE. 2007, pp. 1–8.

[32] Boris Epshtein, Eyal Ofek, Yonatan Wexler, and Pusheng Zhang. "Hierarchical Photo Organization Using Geo-relevance." In: *Proceedings of the 15$^{th}$ annual ACM international symposium on Advances in geographic information systems*. ACM. 2007, p. 18.

[33] Naihui Fang, Haoran Xie, and Takeo Igarashi. "Selfie Guidance System in Good Head Postures." In: *IUI Workshops*. 2018.

[34] Farshid Farhat, Mohammad Mahdi Kamani, Sahil Mishra, and James Z Wang. "Intelligent Portrait Composition Assistance: Integrating Deep-learned Models and Photography Idea Retrieval." In: *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM. 2017, pp. 17–25.

[35] H Sheikh Faridul, Tania Pouli, Christel Chamaret, Jürgen Stauder, Erik Reinhard, Dmitry Kuzovkin, and Alain Trémeau. "Colour mapping: A review of recent methods, extensions and applications." In: *Computer Graphics Forum*. Vol. 35. 1. Wiley Online Library. 2016, pp. 59–88.

[36] Tom Fawcett. "An introduction to ROC analysis." In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[37] Joseph L Fleiss and Jacob Cohen. "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability." In: *Educational and psychological measurement* 33.3 (1973), pp. 613–619.

[38] Michael Freeman. *The photographer's eye : composition and design for better digital photos*. Boston: Focal Press, 2007.

[39] Michael Freeman. *The Photographer's Mind*. Ilex, 2010.

[40] Brendan J. Frey and Delbert Dueck. "Clustering by Passing Messages Between Data Points." In: *Science* 315.5814 (2007), pp. 972–976.

[41] David Frohlich, Allan Kuchinsky, Celine Pering, Abbe Don, and Steven Ariss. "Requirements for photoware." In: *Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02* (2002), pp. 166–175.

[42] Mingxin Fu, Andrea Ceroni, Vassilis Solachidis, Claudia Niederee, Olga Papadopoulou, Nattiya Kanhabua, and Vasileios Mezaris. "Learning personalized expectation-oriented photo selection models for personal photo collections." In: *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. July. IEEE, 2015, pp. 1–6.

[43] Jonathan Sammartino Gardner. *Aesthetics of spatial composition: Facing, position, and context, and the theory of representational fit*. University of California, Berkeley, 2011.

[44] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. "Improving shape deformation in unsupervised image-to-image translation." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 649–665.

[45] Ai Gomi, Reiko Miyazaki, Takayuki Itoh, and Jia Li. "CAT: A Hierarchical Image Browser Using a Rectangle Packing Technique." In: *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE. July 2008, pp. 82–87.

[46] J. P. Gozali, M. Y. Kan, and H. Sundaram. "Hidden Markov Model for Event Photo Stream Segmentation." In: (July 2012), pp. 25–30.

[47] Jesse Prabawa Gozali, Min-Yen Kan, and Hari Sundaram. "How Do People Organize Their Photos in Each Event and How Does It Affect Storytelling, Searching and Interpretation Tasks?" In: *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '12. Washington, DC, USA, 2012, pp. 315–324.

[48] Christopher D Green. "All that glitters: A review of psychological research on the aesthetics of the golden section." In: *Perception* 24.8 (1995), pp. 937–968.

[49] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. "Improved Training of Wasserstein GANs." In: *Advances in Neural Information Processing Systems*. 2017, pp. 5767–5777.

[50] Sharath Chandra Guntuku, Joey Tianyi Zhou, Sujoy Roy, Weisi Lin, and Ivor W Tsang. "Understanding deep representations learned in modeling users likes." In: *IEEE Transactions on Image Processing* 25.8 (2016), pp. 3762–3774.

[51] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. "The interestingness of images." In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 1633–1640.

[52] David Hasler and Sabine E Suesstrunk. "Measuring colorfulness in natural images." In: *Human vision and electronic imaging VIII*. Vol. 5007. International Society for Optics and Photonics. 2003, pp. 87–96.

[53] Jingwu He, Linbo Wang, Wenzhe Zhou, Hongjie Zhang, Xiufen Cui, and Yanwen Guo. "Viewpoint Selection for Photographing Architectures." In: *arXiv preprint arXiv:1703.01702* (2017).

[54] 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*. 2016, pp. 770–778.

[55] Niels Henze and Susanne Boll. "Snap and share your photobooks." In: *Proceedings of the 16th ACM international conference on Multimedia*. ACM. 2008, pp. 409–418.

[56] Florian Hoenig. "Defining Computational Aesthetics." In: *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*. Computational Aesthetics'05. Eurographics Association, 2005, pp. 13–18.

[57] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." In: *arXiv preprint arXiv:1704.04861* (2017).

[58] Lawrence Hubert and Phipps Arabie. "Comparing partitions." In: *Journal of classification* 2.1 (1985), pp. 193–218.

[59] Hung Nguyen. *Improved Training of Wasserstein GANs in Pytorch*. URL: `https://github.com/jalola/improved-wgan-pytorch` (visited on June 21, 2019).

[60] Fil Hunter. *Light– science and magic*. Waltham, MA: Focal Press, 2011.

[61] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. "DSLR-quality photos on mobile devices with deep convolutional networks." In: *the IEEE Int. Conf. on Computer Vision (ICCV)*. 2017.

[62] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. "WESPE: weakly supervised photo enhancer for digital cameras." In: *arXiv preprint arXiv:1709.01118* (2017).

[63] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. "What makes an image memorable?" In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2011), pp. 145–152.

[64] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. "Understanding and predicting interestingness of videos." In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*. 2013.

[65] Bin Jin, Maria V Ortiz Segovia, and Sabine Süsstrunk. "Image aesthetic predictors based on weighted CNNs." In: *2016 IEEE International Conference on Image Processing (ICIP)*. Sept. 2016, pp. 2291–2295.

[66] Amornched Jinda-Apiraksa, Vassilios Vonikakis, and Stefan Winkler. "California-ND: An annotated dataset for near-duplicate detection in personal photo collections." In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2013, pp. 142–147.

[67] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. "Aesthetics and emotions in images." In: *IEEE Signal Processing Magazine* 28.5 (2011), pp. 94–115.

[68] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. "Real-time Analysis and Visualization of the YFCC100M Dataset." In: *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. ACM, 2015, pp. 25–30.

[69] Yueying Kao, Ran He, and Kaiqi Huang. "Deep aesthetic quality assessment with semantic information." In: *IEEE Transactions on Image Processing* 26.3 (2017), pp. 1482–1495.

[70] Yan Ke, Xiaoou Tang, and Feng Jing. "The design of high-level features for photo quality assessment." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1 (2006), pp. 419–426.

[71] Scott Kelby. *The digital photography book. the step-by-step secrets for how to make your photos look like the pros*. San Francisco, Calif: Peachpit Press, 2013.

[72] Maurice G Kendall and B Babington Smith. "The problem of m rankings." In: *The annals of mathematical statistics* 10.3 (1939), pp. 275–287.

[73] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. "Understanding and Predicting Image Memorability at a Large Scale." In: *2015 IEEE International Conference on Computer Vision (ICCV)* (Dec. 2015), pp. 2390–2398.

[74] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. "Understanding photowork." In: *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06* (2006), pp. 761–770.

[75] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. "Photo Aesthetics Ranking Network with Attributes and Content Adaptation." In: *Computer Vision – ECCV 2016*. Springer, 2016, pp. 662–679.

[76] Erwin Kreyszig. *Advanced engineering mathematics*. Wiley publishing, 2007.

[77] Santhana Krishnamachari and Mohamed Abdel-Mottaleb. "Image browsing using hierarchical clustering." In: *Proceedings IEEE International Symposium on Computers and Communications*. IEEE. 1999, pp. 301–307.

[78] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[79] Jack Kustanowitz and Ben Shneiderman. "Meaningful presentations of photo libraries: rationale and applications of bi-level radial quantum layouts." In: *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*. ACM. 2005, pp. 188–196.

[80] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch. "Context-aware clustering and assessment of photo collections." In: *Proceedings of the symposium on Computational Aesthetics*. ACM. 2017, p. 6.

[81] Dmitry Kuzovkin, Tania Pouli, Rémi Cozot, Olivier Le Meur, Jonathan Kervec, and Kadi Bouatouch. "Image Selection in Photo Albums." In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM. 2018, pp. 397–404.

[82] Dmitry Kuzovkin, Tania Pouli, Olivier Le Meur, Rémi Cozot, Jonathan Kervec, and Kadi Bouatouch. "Context in Photo Albums: Understanding and Modeling User Behavior in Clustering and Selection." In: *ACM Transactions on Applied Perception (TAP)* 16.2 (2019), p. 11.

[83] J Richard Landis and Gary G Koch. "The Measurement of Observer Agreement for Categorical Data." In: *Biometrics* 33.1 (1977), pp. 159–174.

[84] Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, and In So Kweon. "Automatic content-aware color and tone stylization." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2470–2478.

[85] Kenneth Levenberg. "A method for the solution of certain non-linear problems in least squares." In: *Quarterly of applied mathematics* 2.2 (1944), pp. 164–168.

[86] Elizaveta Levina and Peter Bickel. "The earth mover's distance is the mallows distance: Some insights from statistics." In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vol. 2. IEEE. 2001, pp. 251–256.

[87] Congcong Li, Alexander C Loui, and Tsuhan Chen. "Towards Aesthetics: a Photo Quality Assessment and Photo Selection System." In: *Proceedings of the international conference on Multimedia - MM '10*. New York, New York, USA: ACM Press, 2010, p. 827.

[88] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. "A2-RL: Aesthetics Aware Reinforcement Learning for Automatic Image Cropping." In: *arXiv preprint arXiv:1709.04595* (2017).

[89] Jun Li, Joo Hwee Lim, and Qi Tian. "Automatic summarization for personal digital photos." In: *ICICS-PCM 2003 - Proceedings of the 2003 Joint Conference of the 4th International Conference on Information, Communications and Signal Processing and 4th Pacific-Rim Conference on Multimedia* 3.December (2003), pp. 1536–1540.

[90] Qifan Li and Daniel Vogel. "Guided Selfies using Models of Portrait Aesthetics." In: *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM. 2017, pp. 179–190.

[91] Xiaodan Liang, Hao Zhang, Liang Lin, and Eric Xing. "Generative semantic manipulation with mask-contrasting GAN." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 558–573.

[92] Zachary C Lipton and Subarna Tripathi. "Precise recovery of latent vectors from generative adversarial networks." In: *arXiv preprint arXiv:1702.04782* (2017).

[93] Ligang Liu, Yong Jin, and Qingbiao Wu. "Realtime Aesthetic Image Retargeting." In: *Computational Aesthetics* 10 (2010), pp. 1–8.

[94] Kuo-Yen Lo, Keng-Hao Liu, and Chu-Song Chen. "Assessment of photo aesthetics with efficiency." In: *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE. 2012, pp. 2186–2189.

[95] Kuo-Yen Lo, Keng-Hao Liu, and Chu-Song Chen. "Intelligent photographing interface with on-device aesthetic quality assessment." In: *Asian Conference on Computer Vision*. Springer. 2012, pp. 533–544.

[96] Alexander C Loui and Andreas Savakis. "Automated event clustering and quality screening of consumer pictures for digital albuming." In: *IEEE Transactions on Multimedia* 5.3 (Sept. 2003), pp. 390–402.

[97] Pietro Lovato, Alessandro Perina, Nicu Sebe, Omar Zandonà, Alessio Montagnini, Manuele Bicego, and Marco Cristani. "Tell me what you like and I'll tell you what you are: discriminating visual preferences on Flickr data." In: *Asian Conference on Computer Vision*. Springer. 2012, pp. 45–56.

[98] Pietro Lovato, Manuele Bicego, Cristina Segalin, Alessandro Perina, Nicu Sebe, and Marco Cristani. "Faved! biometrics: Tell me which image you like and i'll tell you who you are." In: *IEEE Transactions on Information Forensics and Security* 9.3 (2014), pp. 364–374.

[99] D. G. Lowe. "Object recognition from local scale-invariant features." In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2. 1999, 1150–1157 vol.2.

[100] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. "On shape and the computability of emotions." In: *Proceedings of the 20th ACM international conference on Multimedia*. ACM. 2012, pp. 229–238.

[101] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang. "RAPID: Rating Pictorial Aesthetics using Deep Learning." In: *Proceedings of the ACM International Conference on Multimedia - MM '14* (2014), pp. 457–466.

[102] 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*. Dec. 2015, pp. 990–998.

[103] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. "Rating image aesthetics using deep learning." In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 2021–2034.

[104] Zhiwu Lu, Yuxin Peng, and Jianguo Xiao. "From Comparing Clusterings to Combining Clusterings." In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*. AAAI'08. Chicago, Illinois: AAAI Press, 2008, pp. 665–670.

[105] Jiebo Luo. "Subject content-based intelligent cropping of digital photos." In: *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE. 2007, pp. 2218–2221.

[106] Wei Luo, Xiaogang Wang, and Xiaoou Tang. "Content-based photo quality assessment." In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 2206–2213.

[107] Yiwen Luo and Xiaoou Tang. "Photo and Video Quality Evaluation: Focusing on the Subject." In: *Computer Vision – ECCV 2008*. ECCV '08. Springer, 2008, pp. 386–399.

[108] Hao Lv and Xinmei Tian. "Learning relative aesthetic quality with a pairwise approach." In: *International Conference on Multimedia Modeling*. Springer. 2016, pp. 493–504.

[109] Shuang Ma, Jing Liu, and Chang Wen Chen. "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment." In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*. 2017, pp. 722–731.

[110] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[111] Jana Machajdik and Allan Hanbury. "Affective Image Classification Using Features Inspired by Psychology and Art Theory." In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. ACM, 2010, pp. 83–92.

[112] 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*. June 2016, pp. 497–506.

[113] Gautam Malu, Raju S Bapi, and Bipin Indurkhya. "Learning Photography Aesthetics with Deep CNNs." In: *arXiv preprint arXiv:1707.03981* (2017).

[114] Luca Marchesotti and Florent Perronnin. "Learning beautiful ( and ugly ) attributes." In: *British Machine Vision Conference* (2013), pp. 1–11.

[115] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. "Assessing the aesthetic quality of photographs using generic image descriptors." In: *Proceedings of the IEEE International Conference on Computer Vision* (2011), pp. 1784–1791.

[116] Eftichia Mavridaki and Vasileios Mezaris. "No-Reference Blur Assessment In Natural Images Using Fourier Transform And Spatial Pyramids." In: *Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014* 1.October (2014), pp. 566–570.

[117] Eftichia Mavridaki and Vasileios Mezaris. "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography." In: *2015 IEEE International Conference on Image Processing (ICIP)*. Icip. IEEE, Sept. 2015, pp. 887–891.

[118] I. Christopher McManus, Fanzhi Anita Zhou, Sophie L'Anson, Lucy Waterfield, Katharina Stöver, and Richard Cook. "The psychometrics of photographic cropping: The influence of colour, meaning, and expertise." In: *Perception* 40.3 (2011), pp. 0–0.

[119] Andrew D Miller and W Keith Edwards. "Give and Take: A Study of Consumer Photo-sharing Culture and Practice." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, 2007, pp. 347–356.

[120] Anush K Moorthy, Pere Obrador, and Nuria Oliver. "Towards computational models of the visual aesthetic appeal of consumer videos." In: *European Conference on Computer Vision*. Springer. 2010, pp. 1–14.

[121] Naila Murray and Albert Gordo. "A deep architecture for unified aesthetic prediction." In: *arXiv preprint arXiv:1708.04890* (2017).

[122] Naila Murray, Luca Marchesotti, and Florent Perronnin. "AVA: A large-scale database for aesthetic visual analysis." In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[123] Wolfgang Nejdl and Claudia Niederee. "Photos to Remember, Photos to Forget." In: *IEEE MultiMedia* 22.1 (2015), pp. 6–11.

[124] Diep Thi Ngoc Nguyen, Hideki Nakayama, Naoaki Okazaki, and Tatsuya Sakaeda. "PoB: Toward Reasoning Patterns of Beauty in Image Data." In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 1786–1793.

[125] Bingbing Ni, Mengdi Xu, Bin Cheng, Meng Wang, Shuicheng Yan, and Qi Tian. "Learning to photograph: A compositional perspective." In: *IEEE Transactions on Multimedia* 15.5 (2013), pp. 1138–1151.

[126] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato. "Sensation-based photo cropping." In: *Proceedings of the 17th ACM international conference on Multimedia*. ACM. 2009, pp. 669–672.

[127] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. "Aesthetic quality classification of photographs based on color harmony." In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 33–40.

[128] Augustus Odena, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier GANs." In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 2642–2651.

[129] Stephen E. Palmer, Karen B. Schloss, and Jonathan Sammartino. "Visual aesthetics and human preference." In: *Annual Review of Psychology* 64.September 2012 (2013), pp. 77–107.

[130] Kayoung Park, Seunghoon Hong, Mooyeol Baek, and Bohyung Han. "Personalized Image Aesthetic Quality Assessment by Joint Regression and Ranking." In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017), pp. 1206–1214.

[131] Kuan-Chuan Peng and Tsuhan Chen. "Toward correlating and solving abstract tasks using convolutional neural networks." In: *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE. 2016, pp. 1–9.

[132] Gabriele Peters. "Aesthetic primitives of images for visualization." In: *Proceedings of the International Conference on Information Visualisation* (2007), pp. 316–325.

[133] Bryan Peterson. *Learning to see creatively : design, color & composition in photography*. New York: Amphoto Books, 2003.

[134] Bryan Peterson. *Understanding exposure : how to shoot great photographs with any camera*. New York: Amphoto Books, 2010.

[135] J. C. Platt. "AutoAlbum: Clustering digital photographs using probabilistic model merging." In: *Proceedings - IEEE Workshop on Content-Based Access of Image and Video Libraries, CBAIVL 2000* (2000), pp. 96–100.

[136] John C. Platt, Mary Czerwinski, and Brent A. Field. "PhotoTOC: automatic clustering for browsing personal photographs." In: *Proceedings of the 2003 Joint Conference of the 4$^{th}$ International Conference on Information, Communications and Signal Processing and 4$^{th}$ Pacific-Rim Conference on Multimedia* 1 (Dec. 2003), 6–10 Vol.1.

[137] Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. "Automatic creation of photo books from stories in social media." In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 7S.1 (2011), pp. 1–18.

[138] F. Radenović, G. Tolias, and O. Chum. "Fine-tuning CNN Image Retrieval with No Human Annotation." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), pp. 1–1.

[139] Miriam Redi, Frank Z Liu, and Neil O'Hare. "Bridging the aesthetic gap: The wild beauty of web imagery." In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM. 2017, pp. 242–250.

[140] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. "The beauty of capturing faces: Rating the quality of digital portraits." In: *11$^{th}$ IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE, May 2015, pp. 1–8.

[141] Miriam Redi, Damon Crockett, Lev Manovich, and Simon Osindero. "What Makes Photo Cultures Different?" In: *Proceedings of the 24th ACM International Conference on Multimedia*. MM '16. Amsterdam, The Netherlands: ACM, 2016, pp. 287–291.

[142] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. "Personalized image aesthetics." In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 638–647.

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

[144] A. Samii, R. Měch, and Z. Lin. "Data-Driven Automatic Cropping Using Semantic Composition Search." In: *Computer Graphics Forum* 34.1 (Feb. 2015), pp. 141–151.

[145] Philipp Sandhaus and Susanne Boll. "Semantic analysis and retrieval in personal and social photo collections." In: *Multimedia Tools and Applications* 51.1 (2011), pp. 5–33.

[146] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. "Gaze-based interaction for semi-automatic photo cropping." In: *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems* 1 (2006), pp. 771–780.

[147] Risto Sarvas and David M Frohlich. *From snapshots to social media-the changing picture of domestic photography*. Springer Science & Business Media, 2011.

[148] Katharina Schwarz, Patrick Wieschollek, and Hendrik PA Lensch. "Will people like your image? learning the aesthetic space." In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 2048–2057.

[149] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. "Attention-based Multi-Patch Aggregation for Image Aesthetic Assessment." In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 879–886.

[150] Gunnar A. Sigurdsson, Xinlei Chen, and Abhinav Gupta. "Learning Visual Storylines with Skipping Recurrent Neural Networks." In: (2016). arXiv: 1604.04279.

[151] Florian Simond, Nikolaos Arvanitopoulos, and Sabine Süsstrunk. "Image aesthetics depends on context." In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2015, pp. 3788–3792.

[152] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *ImageNet Challenge* (2014), pp. 1–10. arXiv: 1409.1556.

[153] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. "Summarization of personal photologs using multidimensional content and context." In: (2011), p. 4.

[154] Pinaki Sinha, Hamed Pirsiavash, and Ramesh Jain. "Personal photo album summarization." In: *Proceedings of the seventeen ACM international conference on Multimedia - MM '09* January 2009 (2009), p. 1131.

[155] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. "To click or not to click: Automatic selection of beautiful thumbnails from videos." In: *Proceedings of the 25$^{th}$ ACM International on Conference on Information and Knowledge Management*. ACM. 2016, pp. 659–668.

[156] Emily L Spratt and Ahmed Elgammal. "Computational beauty: Aesthetic judgment at the intersection of art and science." In: *Workshop at the European Conference on Computer Vision*. Springer. 2014, pp. 35–53.

[157] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H. Hsu, and Shao-Yi Chien. "Scenic Photo Quality Assessment with Bag of Aesthetics-preserving Features." In: *Proceedings of the 19th ACM International Conference on Multimedia*. MM '11. ACM, 2011, pp. 1213–1216.

[158] Hsiao-hang Hang Su, Tse-wei Wei Chen, Chieh-chi Chi Kao, Winston H. Hsu, and Shao-yi Yi Chien. "Preference-Aware View Recommendation System for Scenic Photos Based on Bag-of-Aesthetics-Preserving Features." In: *IEEE Transactions on Multimedia* 14.3 (June 2012), pp. 833–843.

[159] Wei-Tse Sun, Ting-Hsuan Chao, Yin-Hsi Kuo, and Winston H Hsu. "Photo filter recommendation by category-aware aesthetic learning." In: *IEEE Transactions on Multimedia* 19.8 (2017), pp. 1870–1880.

[160] H. Talebi and P. Milanfar. "NIMA: Neural Image Assessment." In: *IEEE Transactions on Image Processing* 27.8 (2018), pp. 3998–4011.

[161] Xiaoou Tang, Wei Luo, and Xiaogang Wang. "Content-Based Photo Quality Assessment." In: *IEEE Transactions on Multimedia* 15.8 (Dec. 2013), pp. 1930–1943.

[162] Christopher Thomas and Adriana Kovashka. "Seeing behind the camera: Identifying the authorship of a photograph." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3494–3502.

[163] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. "YFCC100M: The New Data in Multimedia Research." In: *Communications of the ACM* 59.2 (Jan. 2016), pp. 64–73.

[164] Xinmei Tian, Zhe Dong, Kuiyuan Yang, and Tao Mei. "Query-Dependent Aesthetic Model With Deep Learning for Photo Quality Assessment." In: *IEEE Trans. Multimedia* 17.11 (2015), pp. 2035–2048.

[165] Hanghang Tong, Mingjing Li, Hongjiang Zhang, and Changshui Zhang. "Blur detection for digital images using wavelet transform." In: *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*. Vol. 1. IEEE. 2004, pp. 17–20.

[166] Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. "Classification of digital photos taken by photographers or home users." In: *Pacific-Rim Conference on Multimedia*. Springer. 2004, pp. 198–205.

[167] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. "Learning mixtures of submodular functions for image collection summarization." In: (2014), pp. 1413–1421.

[168] Nancy A Van House and Marc Davis. "The social life of cameraphone images." In: *Proceedings of the Pervasive Image Capture and Sharing: New Social Practices and Implications for Technology Workshop (PICS 2005) at the Seventh International Conference on Ubiquitous Computing (UbiComp 2005)*. Citeseer. 2005.

[169] Marynel Vázquez and Aaron Steinfeld. "An Assisted Photography Framework to Help Visually Impaired Users Properly Aim a Camera." In: *ACM Trans. Comput.-Hum. Interact.* 21.5 (2014), 25:1–25:29.

[170] Tina Caroline Walber, Ansgar Scherp, and Steffen Staab. "Smart Photo Selection: Interpret Gaze As Personal Interest." In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM. 2014, pp. 2065–2074.

[171] Baoyuan Wang, Noranart Vesdapunt, Utkarsh Sinha, and Lei Zhang. "Real-time Burst Photo Selection Using a Light-Head Adversarial Network." In: *arXiv preprint arXiv:1803.07212* (2018).

[172] Weining Wang, Mingquan Zhao, Li Wang, Jiexiong Huang, Chengjia Cai, and Xiangmin Xu. "A multi-scene deep learning model for image aesthetic evaluation." In: *Signal Processing: Image Communication* 47 (2016), pp. 511–518.

[173] Wenguan Wang and Jianbing Shen. "Deep cropping via attention box prediction and aesthetics assessment." In: *IEEE International Conference on Computer Vision*. 2017.

[174] Yinting Wang, Mingli Song, Dacheng Tao, Yong Rui, Jiajun Bu, Ah Chung Tsoi, Shaojie Zhuo, and Ping Tan. "Where2stand: A human position recommendation system for souvenir photography." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.1 (2015), p. 9.

[175] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell. "Event-specific image importance." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4810–4819.

[176] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomír Mech, Gavin Miller, and Garrison W Cottrell. "Recognizing and Curating Photo Albums via Event-Specific Image Importance." In: *arXiv preprint arXiv:1707.05911* (2017).

[177] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomır Mech, Minh Hoai, and Dimitris Samaras. "Good View Hunting: Learning Photo Composition from Dense View Pairs." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5437–5446.

[178] Tom White. "Sampling generative networks." In: *arXiv preprint arXiv:1609.04468* (2016).

[179] Steve Whittaker, Ofer Bergman, and Paul Clough. *Easy on that trigger dad: a study of long term family photo retrieval*. Vol. 14. 1. Springer-Verlag, 2010, pp. 31–43.

[180] Maria K. Wolters, Elaine Niven, and Robert H. Logie. "The Art of Deleting Snapshots." In: *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '14. ACM. ACM, 2014, pp. 2521–2526.

[181] Jun Xiao, Xuemei Zhang, Phil Cheatle, Yuli Gao, and C. Brian Atkins. "Mixed-initiative photo collage authoring." In: *Proceeding of the 16th ACM international conference on Multimedia - MM '08* (2008), p. 509.

[182] Pengfei Xu, Hongxun Yao, Rongrong Ji, Xian-Ming Liu, and Xiaoshuai Sun. "Where should I stand? Learning based human position recommendation for mobile photographing." In: *Multimedia tools and applications* 69.1 (2014), pp. 3–29.

[183] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. "Learning the change for automatic image cropping." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2013), pp. 971–978.

[184] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. "A learning-to-rank approach for image color enhancement." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2987–2994.

[185] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. "Automatic photo adjustment using deep neural networks." In: *ACM Transactions on Graphics (TOG)* 35.2 (2016), p. 11.

[186] Lei Yao, Poonam Suryanarayan, Mu Qiao, James Z Wang, and Jia Li. "Oscar: On-site composition and aesthetics feedback through exemplars for photographers." In: *International Journal of Computer Vision* 96.3 (2012), pp. 353–383.

[187] Che-Hua Yeh, Yuan-Chen Ho, Brian A Barsky, and Ming Ouhyoung. "Personalized Photograph Ranking and Selection System." In: *Proceedings of the 18$^{th}$ ACM international conference on Multimedia*. ACM. 2010, pp. 211–220.

[188] Hsin-Ho Yeh, Chun-Yu Yang, Ming-Sui Lee, and Chu-Song Chen. "Video aesthetic quality assessment by temporal integration of photo-and motion-based features." In: *IEEE Transactions on Multimedia* 15.8 (2013), pp. 1944–1957.

[189] Mei-Chen Yeh and Yu-Chen Cheng. "Relative features for photo quality assessment." In: *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE. 2012, pp. 2861–2864.

[190] Wenyuan Yin, Tao Mei, Chang Wen Chen, and Shipeng Li. "Socialized mobile photography: Learning to photograph with social context via mobile devices." In: *IEEE Transactions on Multimedia* 16.1 (2014), pp. 184–200.

[191] Fang-Lue Zhang, Miao Wang, and Shi-Min Hu. "Aesthetic image enhancement by dependence-aware object recomposition." In: *IEEE Transactions on Multimedia* 15.7 (2013), pp. 1480–1490.

[192] Luming Zhang, Yue Gao, Chao Zhang, Hanwang Zhang, Qi Tian, and Roger Zimmermann. "Perception-Guided Multimodal Feature Fusion for Photo Aesthetics Assessment." In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. ACM, 2014, pp. 237–246.

[193] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

**Titre :** Évaluation de photos dans des albums basée sur l'esthétique et le contexte

**Mots clés :** Évaluation d'image, Sélection de photos, Clustering, Organisation d'albums photo

**Résumé :** Le processus de sélection de photos dans des albums peut être considérablement amélioré à l'aide d'un critère d'évaluation automatique des qualités d'une photo. Cependant, les méthodes existantes abordent ce problème de manière indépendante, c'est à dire en évaluant chaque image séparément des autres images d'un album.

Dans cette thèse, nous explorons la modélisation du contexte d'une photo via une approche de *clustering* de collections de photos et la possibilité d'appliquer l'information de contexte à l'évaluation d'une photo.

Nous avons effectué des études subjectives permettant d'étudier la manière dont les utilisateurs regroupent et sélectionnent des photos dans un album. Ces études ont permis une estimation du niveau de l'accord entre les différents utilisateurs. Nous avons aussi etudié la manière dont le contexte influence leurs decisions.

Après avoir étudié la nature des décisions des utilisateurs, nous proposons une approche informatique pour modéliser leur comportement. Tout d'abord, nous introduisons une méthode de clustering hiérarchique, qui permet de regrouper des photos similaires selon une structure de similarité à plusieurs niveaux, basée sur des descripteurs visuels. Ensuite, les informations de contexte de la photo sont utilisées pour adapter le score de la photo pré-calculé indépendamment, en utilisant les données basées sur des statistiques et une approche d'apprentissage automatique.

De plus, comme la majorité des méthodes récentes d'évaluation de la photo sont basées sur des réseaux de neurones convolutionnels, nous avons exploré et visualisé les caractéristiques esthétiques apprises par ces méthodes.

**Title :** Assessment of photos in albums based on aesthetics and context

**Keywords :** Image assessment, Photo selection, Clustering, Photo collection organization

**Abstract :** An automatic photo assessment can significantly aid the process of photo selection within photo collections. However, existing computational methods approach this problem in an independent manner, by evaluating each image apart from other images in a photo album.

In this thesis, we explore the modeling of photo context via a clustering approach for photo collections and the possibility of applying such context information in photo assessment.

To better understand user actions within photo albums, we conduct experimental user studies, where we study how users cluster and select photos in photo collections. We estimate the level of agreement between users and investigate how the context, defined by similar photos in corresponding clusters, influences users' decisions.

After studying the nature of user decisions, we propose a computational approach to model user behavior. First, we introduce a hierarchical clustering method, which allows to group similar photos according to a multi-level similarity structure, based on visual descriptors. Then, the photo context information is extracted from the obtained cluster data and used to adapt a pre-computed independent photo score, using the statistics-based data and a machine learning approach.

In addition, as the majority of recent methods for photo assessment are based on convolutional neural networks, we explore and visualize the aesthetic characteristics learned by such methods.