Gonçalo Barreto Ferreira Marcelino

Master of Science

A computational approach to the art of visual storytelling

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Computer Science and Informatics Engineering

Adviser: Dr. João Miguel da Costa Magalhães, Assistant Professor, NOVA University of Lisbon

Examination Committee

Chairperson: Dr. João Leite
Raporteur: Dr. Bruno Martins
Member: Dr. João Magalhães

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

December, 2018

A computational approach to the art of visual storytelling

This document was created using the (pdf)LaTeX processor, based in the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Ao meu eu futuro.
Que não te arrependas das minhas escolhas,
que o esforço que estou a fazer agora te ajude a ser
um moço feliz.

Um bem haja.

Acknowledgements

I would like to thank,

my Advisor, João Magalhães, for all the help, the recommendations and indispensable encouragement that made this project possible. It was a pleasure to work under the your guidance;

my colleagues and friends David Semedo, Flávio Martins, André Mourão, Gustavo Gonçalves, not only for the technical help you provided me during the course of the year that passed (I learned much from you) but also for... everything else;

All the friends I made during these past 5 years, even those those to whom I eventually stopped speaking to or who stopped speaking to me;

my family, Ana Baru, Vanessa Sofia, André Pontes and my mother Luisa, without you I wouldn't be myself;

FCT/UNL, for 5 years worth of knowledge, friendships and new experiences.

CMU Portugal research project GoLocal Ref. CMUP-ERI/TIC/0033/2014; the H2020 ICT project COGNITUS with the grant agreement No687605 and the NOVA LINCS project Ref. UID/CEC/04516/2013 for funding this work.

Abstract

For millennia, humanity as been using images to tell stories. In modern society, these visual narratives take the center stage in many different contexts, from illustrated children's books to news media and comic books. They leverage the power of compounding various images in sequence to present compelling and informative narratives, in an immediate and impactful manner. In order to create them, many criteria are taken into account, from the quality of the individual images to how they synergize with one another.

With the rise of the Internet, visual content with which to create these visual storylines is now in abundance. In areas such as news media, where visual storylines are regularly used to depict news stories, this has both advantages and disadvantages. Although content might be available online to create a visual storyline, filtering the massive amounts of existing images for high quality, relevant ones is a hard and time consuming task. Furthermore, combining these images into visually and semantically cohesive narratives is a highly skillful process and one that takes time.

As a first step to help solve this problem, this thesis brings state of the art computational methodologies to the age old tradition of creating visual storylines. Leveraging these methodologies, we define a three part architecture to help with the creation of visual storylines in the context of news media, using social media content. To ensure the quality of the storylines from a human perception point of view, we deploy methods for filtering and raking images according to news quality standards, we resort to multimedia retrieval techniques to find relevant content and we propose a machine learning based approach to organize visual content into cohesive and appealing visual narratives.

Keywords: News media, Social media, Illustration, Storylines

Resumo

Desde os tempos primórdios que a humanidade tem feito uso da imagem como meio de transmitir histórias. No entanto, na sociedade actual, estas narrativas visuais ganharam uma nova importância. Desde ilustrar livros infantis, até informar no contexto de peças jornalísticas, este medium é frequentemente usado pela sua habilidade de apresentar informações de maneira interessante e imediata.

Com o crescimento da Internet, o conteúdo visual através do qual é possível criar estas narrativas tornou-se abundante. Para a imprensa, que faz uso frequente de narrativas visuais para ilustrar noticias, esta mudança trouxe ambas vantagens e desvantagens. Embora possa existir conteúdo de qualidade, online, para ilustrar uma notícia, o processo de encontrar esse conteúdo e organiza-lo de uma forma coesa e apelativa, é uma tarefa demorada e difícil.

Como primeiro passo para resolver este problema, nesta tese propomos trazer o uso de metodologias que são o estado da arte na area das ciências da computação, para auxiliar jornalistas e editores neste processo criativo. Fazendo uso destas metodologias, definimos uma arquitectura composta por três módulos que permite a criação de narrativas visuais através de conteúdos retirados das redes sociais.

Palavras-chave: Imprensa, Redes sociais, Ilustração, Histórias

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context, motivation and challenges

For thousands of years humanity has been telling stories through images. From carvings in prehistoric caves retelling hunts, to classical paintings of bible scenes and modern children's books illustrations, the image has always been a key tool in the process of storytelling. It provides an immediate way of sharing narratives that transcends language barriers while having the capacity to be immensely descriptive.

This unique quality of the medium allows us to still be able to understand the stories the ancient Egyptians carved in stone 4000 years ago. Figure 1.1 presents such a carving. By analyzing it, one can get to know their fishing methods, even without prior knowledge

Figure 1.1: Examples of how our ancestors used images to tell stories. On the left, carvings of acient Egyptians fishing, present in the tomb of Kagemni, a vizier of ancient Egypt. Source: [51]. On the right the interior of the Scrovegni chapel situated in Padua, Italy, with murals depicting the life of Christ. Source: [45].

of the ancient Egyptian civilization and culture. Similarly, in the middle ages, it was through the visual depiction of religious scenes that Christianity came to be understood and worshiped by the general populace. In 1305, Giotto di Bondone finished his famous series of frescos in the Scrovegni chapel, depicting the life of Christ. These murals work in tandem to tell a narrative that can be understood by anyone just by observing them, proving that imagery is able to express complex tales and depict the passage of time. Although, much as changed since the these murals where painted, the inherent qualities of visual stories are still highly valuable in modern society. In fact, since the 19th century we saw the rise of many new mediums that focus on telling stories through images, such as comic books, graphic novels, manga, photography and cinema.

Today, with the large amounts of information available in the most varied contexts, the importance of being able to present compelling and informative narratives in an immediate and impactful fashion has revitalized the importance of the image. As a prime example of this, news media is focusing more and more on the usage of images to tell news stories, providing news in formats such as BBC's *In Picture*[1], where news pieces are presented to the audience through selections of images accompanied by small captions.

However, the high amounts of information and images available, specially online, has also brought new challenges to the task of creating visual narratives. Rooted in the need to help solve these challenges and introduce new technologies to the millenia-old human tradition of telling stories through images, this thesis approaches visual storylines from a computational perspective, a problem yet to be solved by the research community.

In particular, we focus on the problem of creating visual storylines to illustrate news pieces using social media content.

1.1.1 Social media in the newsroom

In the context of news media, visual storylines are consistently used as way to present information to the reader in a concise yet interesting manner. Not only are news pieces normally illustrated by a carefully selected sequence of high quality images, but the image becomes the central focus of the news piece in slide shows as the ones that populate most news websites (BBC In Pictures[1], Reuters Pictures[2] and Euronews NoComment[3] are just some examples). Hence, in the news room, it is the job of the journalist and the news editor to select news worthy images and to organize them in a semantically, visually coherent and appealing fashion.

However, with the advent of large scale social media platforms like Twitter, Facebook and Flickr, interesting and appealing images that can illustrate a piece of news are no longer only present in the portfolio of news photographers. User generated content has become a great source of news images as the photographic quality of mobile devices

[1] www.bbc.com/news/in_pictures
[2] www.reuters.com/news/pictures
[3] www.euronews.com/nocomment

continues to rise. Additionally, social media users document the events they participate in themselves, taking photos that cover many more places and perspectives, than a group of photo journalists could ever be able to create. Consequently news editors are now using user generated content found on social media when creating visual storylines.

1.1.1.1 Social media challenges

Despite the previously described advantages of using social media images to create news media storylines, there are various downsides to the approach:

- Not all social media content presents the same aesthetic quality;
- Fake and altered media are prominent on social media;
- Sources reputation varies greatly;
- There are massive amounts of content to be considered and analyzed;
- Resorting to social media content means dealing with its heterogeneous style and characteristics;
- Finding relevant content to a particular topic can be hard and time consuming.

Consequently, using social media content to create news media visual storylines is highly impractical if the full extent of the task is to be performed manually.

Hence, there is a need for the development of new tools to help the news journalists and news editors in their task. These technologies range from intelligent methods of content filtering based on respective quality and characteristics, to automated suggestions for news illustration and visual storyline creation. With these tools the news editor will be able to take full advantage of the benefits of using social media content without having to deal with the aforementioned problems.

Developed in the context of project COGNITUS, an European project that aims to combine news media, broadcasting technologies and user generated content, this thesis is a first step in this direction. It focuses on the study and development of a method to assist news editors in the process of visual storyline creation. This composes a novel research task, one that, to our knowledge, is yet to be tackled in literature.

1.2 Problem statement

In the context of this thesis we define a *visual storyline* as an organized sequence of images illustrating a sequence of text *segments* related to a particular *topic*. Creating a visual storyline from a topic and a set of segments means picking images to illustrate each segment. These images must not only make sense in the context of the text segment they are illustrating individually but must also form a cohesive and appealing whole when organized sequentially into a visual storyline. The objective of this thesis is to automate

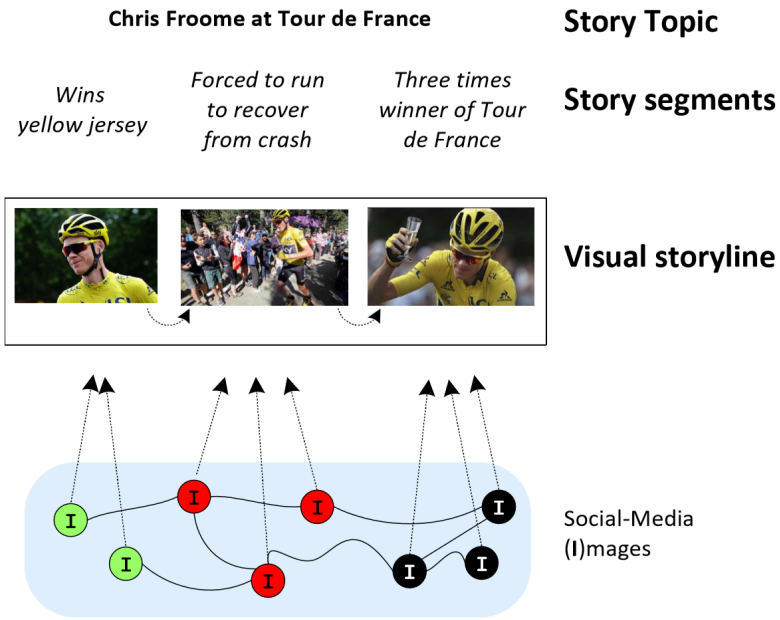


Figure 1.2: *Topics* and *segments* and the process of creating a *visual storyline*. Source of the Tour de France news story being illustrated: https://www.bbc.com/sport/cycling/36879128.

this illustration process resorting to content extracted from social media. Figure 1.2 presents the *topic - segments* hierarchy as well and the illustration process of creating a visual storyline for a real BBC news story with social media content.

Formally we define our query, a story as consisting of N text segments, each denoted by u_i, as:

$$Story_N = (u_1, u_2, ..., u_N) \tag{1.1}$$

The main objective of this thesis is to create a framework, that takes as input a story $Story_N$ and a set of social media posts D, and outputs one or more visual storylines $Storyline_N$ containing N social media images, each denoted by w_i where $w_i \in D$:

$$Storyline_N = (w_1, w_2, ..., w_N) \tag{1.2}$$

Furthermore we aim to understand, from a computational point of view, what are the key characteristics that make particular images more apt to be used in a visual storyline, in the context of news media, and what characteristics make a storyline more appealing and coherent to the viewers consuming them.

1.3 Objectives and proposed solution

Figure 1.3 summarizes the architecture of the framework, composed by three modules, we propose to tackle stated problem. This linear pipeline corresponds directly to part

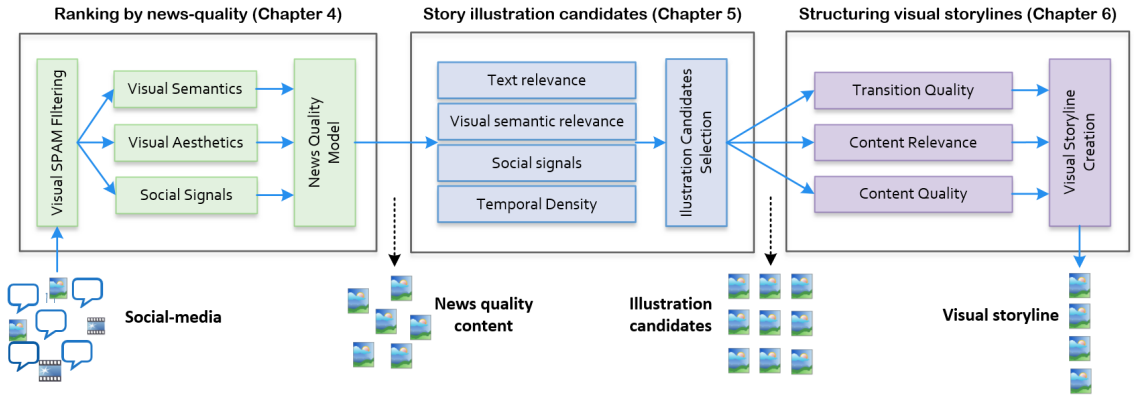


Figure 1.3: Visual storyline creation framework.

of the structure of this thesis, as the first, second and third modules are described and studied in detail in Chapters 4, 5 and 6.

Our research hypothesis is that, by taking advantage of this three part architecture we are able to generate visual storylines that are perceived as having high quality by their viewers. Additionally, by taking advantage of this system we also aim to understand what key criteria can be used to optimize the process of storyline creation from a computational point of view.

The modules that compose the framework are designed as follows.

1.3.1 Ranking by news quality

The first module of the framework is designed to find content inline with news media standards in a pool of social media posts.

This means dealing directly with SPAM and content that is, overall, not suitable for use in a news media context (e.g.: digital adverts). Moreover, after filtering out SPAM one is still left with images of various degrees of quality. This means also taking into account more nuanced news media criteria when picking content. However, understanding and enforcing news media criteria in visual content is, as detailed in Chapter 2 (Related Work), a complex and nuanced task that, as far as we could tell, has not been tackled in literature.

The social media images that pass this filter are then given as input to the second module of the framework.

1.3.2 Retrieving relevant content

Ensuring the storyline is comprised of quality content is not enough, as illustrating a story with quality content that is not relevant to the topic the story describes mutes the purpose of the illustration. Hence, the second module tackles the problem of retrieving relevant social media content to a particular story in an automated way. Taking as input the images filtered by the first model and a story to illustrate, this module finds candidate

images to illustrate each segment of the story and provides them to the third and final module that composes the framework.

1.3.3 Creating the storyline

As already described, a visual storyline is an ordered sequence of images and, as noted in Chapter 2, the way these images are ordered affects how they are perceived: from a human cognition point of view a visual storyline is expected to be semantically and visually coherent while providing an interesting narrative that unfolds over time. Taking the candidate images outputted by the second module as input, this final module is tasked with generating storylines with the candidate images that reflect these characteristics. As discussed in Chapter 2, although there is already research work on tasks such as semi-automated video editing, no works could be found on visual storyline generation in the context of news media. As such this task is also one that is one that is also yet to be tackled in literature.

1.4 Contributions

This thesis resulted in the following contributions:

- A paper published in the proceedings of the 2018 ACM International Conference on Multimedia Retrieval titled Ranking News Quality Multimedia, which was nominated for best paper of the conference.
- Two modules that were integrated in the COGNITUS European project. The first module designed to evaluate images according to news quality standards and the second designed to evaluate the transitions between images in the context of visual storylines.
- A dataset composed of images extracted from social media annotated according to their news quality in the context of news media through crowd sourcing. This is the first dataset of this kind to be publicly available as far as we know.
- A visual storylines dataset created using social media content and annotated according to quality through crowed sourcing. Again, this is also the first dataset of this kind to be publicly available as far as we understand.
- Finally, the work in this thesis contributed to the organization of the first "Social-media video storytelling linking"TRECVid competition, a workshop where world-wide multimedia retrieval and analysis competitions are held. The task focuses on developing a system similar to the one proposed in this thesis. The storyline evalation framework described in Chapter 3 was used to evaluate the competing systems.

1.5 Document structure

The remaining document is organized as follows:

- Chapter 2 discusses related works that either influenced this thesis or that can be used to complement it;
- Chapter 3 proposes an evaluation framework for visual storylines that is the basis for the experiments conducted throughout this thesis;
- Chapter 4 describes the first module of the visual storyline generation framework developed for image quality assessment and filtering, according to news media standards;
- Chapter 5 describes the methods that compose the second module of the framework, used to identify and retrieve candidate images relevant to the stories provided as input to the framework;
- Chapter 6 details the third and final module of the framework, designed to structure candidate images into cohesive and pleasant visual storylines;
- Finally, Chapter 7 concludes the thesis.

Chapter 2

Related work

2.1 Introduction

In this Chapter we perform an analysis on previous research works in order to understand what methodologies already established in literature can be helpful when tackling the problems posed in the context of this thesis.

We start by review works related to various methods of qualifying visual content in Section 2.2. These serve as a first basis for the first module of the storyline generation framework. In turn, Sections 2.3 and 2.4 discuss works related to multi-document summarization, storyline creation and editing. They serve as ground work for the remaining two modules of the framework, tasked with finding media to illustrate stories and organizing this media into a cohesive storyline, respectively.

Finally, we provide a critical review of the works analyzed, identifying their most valuable take away messages as well as possible gaps in existing literature.

2.2 Identifying quality content

Image quality is an abstract concept that describes a large set of characteristics and criteria whose value is variable according to both context and personal subjective preference. A photography may be highly suitable to be used in the context of an advert while not presenting the necessary characteristics that would make it a good image to be used to illustrate a news piece. Additionally, different image characteristics provide different effects on the individuals viewing them. Even so, on average some images tend to be considered more aesthetically pleasing, memorable, interestingness or even exotic then others.

Because we face the task of illustrating news stories, we are interested in being able

Figure 2.1: Usage of the Rule of Thirds in a photography by Henri Cartier-Bresson with Rule of Thirds division lines in overlay. The degree to which the rule is applied greatly varies from photography to photography while also being highly open to interpretation. Source: [50].

to distinguish between newsworthy and non-newsworthy images. In order to understand what makes an image newsworthy one can turn to works on photographic technique, both older [1] and more recent [11, 13] as well as works on photojournalism [25]. These detail the importance of visual criteria like exposure quality, composition and the use of color, among many others, while also elaborating on the importance of semantics in photography, all explained through the photographer's point of view.

However, although of high importance, this perspective is not enough, as translating and combining some of the photographic concepts presented in these works into a computational context is a hard and subjective task. An example of such a concept is the Rule of Thirds [11], as its degree of application is highly subjective in some images and the improvement it provides is fully dependent on the remaining characteristics of the image. The application of this rule is presented in Figure 2.1.

Although the study of methods for measuring the quality of textual news pieces, such as the one described in [3], are a popular topic in literature, no research work could be found on the subject of understanding newsworthy images from a computation point of view. As such, we turn to research on other types of image quality and characteristics as a basis for our work.

2.2.1 Aesthetics

In [32], Marchesotti et. al. attempt to predict the aesthetic quality of images by making use of local generic image descriptors. The approach attempts to implicitly find quality images that adhere to photographic rules by training a machine learning model using low

Figure 2.2: High level features used in [10] to predict aesthetic quality of images. Source: [10].

level features. The evaluation of the method was conducted using two distinct datasets, one created for research purposes through crowd sourcing [24] and one leveraging the opinions of the community of Photo.net[1], a photography social network. This approach to experimentation underlines the trending method of using datasets annotated via crowd sourcing for evaluation purposes in this type of task.

Complementing the focus on low-level visual features, other works make use of high-level visual and semantic features, like compositional attributes, semantic content of the images and illumination quality, as a way to tackle the same task. One such work is [10], where the Dhar et. al. try to predict both aesthetic quality and the percieved interestingness of images using high level features, extrapolated from low level ones. Figure 2.2 presents some of these features and Subsection 2.2.3 discusses this work in greater detail.

Interpreting the choice of features used in [10] and [32] highlights the value of both high and low level features when tackling this type of task.

2.2.2 Memorability

Now focusing on other kinds of quality in [20, 21] Isola et. al. tackle task of computationally quantifying the memorability of different images. They propose a memory game and resort to crowd sourcing as a way of understanding the key factors that make an image more or less memorable. To do so, they consider a collection of visual attributes such as the aesthetics of the image, the emotions the images projects, the location the image depicts, if the image contains people, among others. The authors then correlate memorability with these features by analyzing the results of the memory game that was

[1] https://www.photo.net/

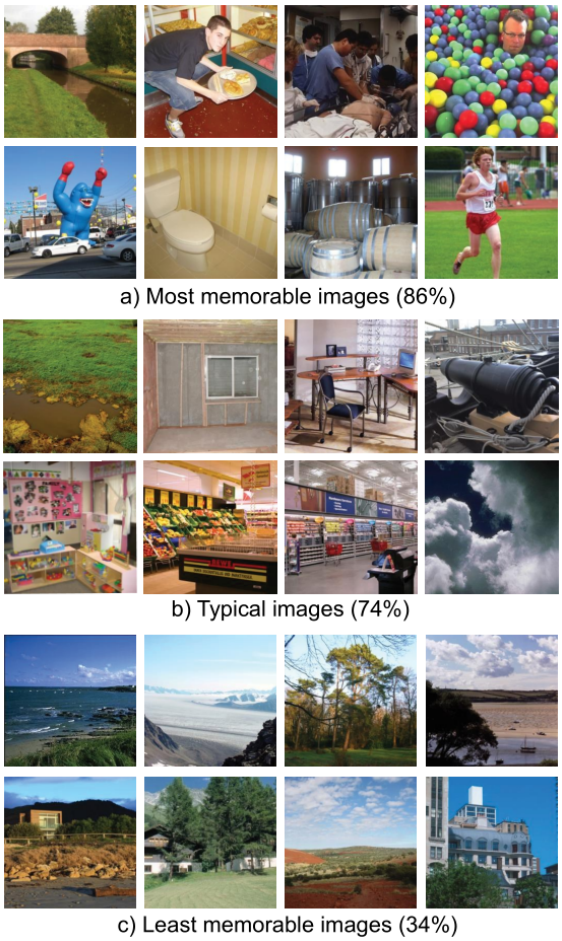


Figure 2.3: Different images and their memorability in the context of a memory game. The percentages presented are the amount of participants in the memory game that remembered the images. Source: [20]

conducted. Figure 2.3 presents different images and their memorability to the users who participated in the memory game.

Following the approaches to quantifying memorability in images proposed in [20, 21], in [31] Mancas et. al. again tackle the same task, however this time by researching the possible relationship between memorability and attention. More specifically the authors research memorability and its connection to a proxy for attention: the eye movements of users (analyzed by an eye-tracking system) when viewing images in the context of a memory game. Additionally, the authors expand on the feature set analyzed in the context of [20, 21] by taking into account two low level image features related to attention: salience map coverage and contrasted structures. Through this novel approach the authors proved the importance of considering attention in the context of image memorability, showing that fixation duration (as measured through an eye-tracking system) is a valuable criteria in the task of predicting image memorability.

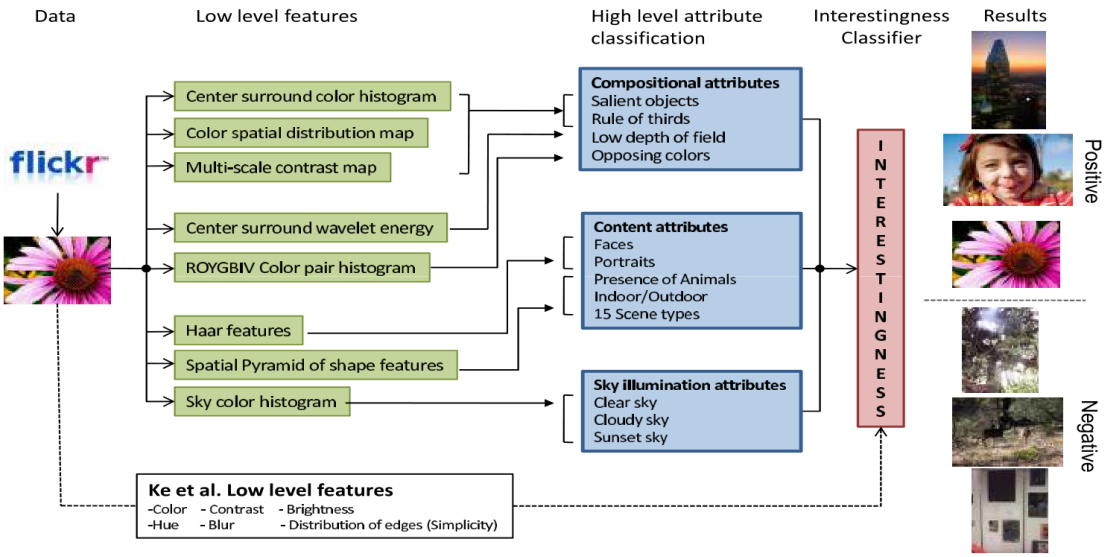


Figure 2.4: Pipepile for image interestingness prediction with low level features extracted from [24]. Source: [10].

2.2.3 Interestingness

Also leveraging a similar approach as the one proposed by Isolda et. al.[20, 21], but towards a different goal, Dhar et. al. [10], research the characteristics of images that make them generally more interesting to viewers. In order to tackle the task the authors propose a pipeline where low-level features are first extracted from visual content and are then used to infer a set of high level characteristics related to the images under scrutiny. These higher level features are then given to a classifier tasked with inferring interestingness. Figure 2.4 presents a diagram of the full pipeline specifying all low and high level features used.

Of note is that, both the authors of [10] and [21] make use of a greedy feature selection method as a way of finding the features that best correlate to their respective goals, establishing the method as a solid approach to the task, even if works such as [23] show that other methods could also have been employed.

2.2.4 Exoticism

Tackling a novel problem in research, in [6] Ceroni et. al. approach the problem of automatically identifying exotic images using deep learning techniques. Although the authors acknowledge that an image can be seen as less or more exotic in a continuous scale, in [6] the task is simplified and tackled from a binary classification perspective. Additionally, the use of deep learning techniques means losing the ability to interpret the learned models. Consequently, although the authors were able to achieve a high precision at the task they were unable to provide insight into what makes an image more or less exotic.

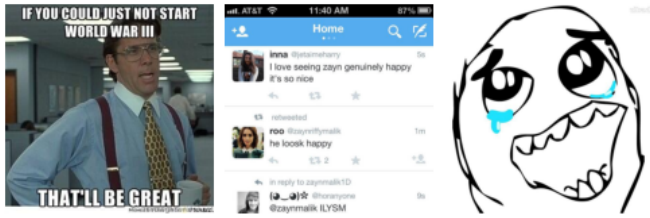


Figure 2.5: Images unsuitable for summarization. Source: [37].

2.2.5 SPAM detection

Although identifying high quality content is important, we are also interested in identifying very low quality content to ensure such content is never used in visual storylines. Consequently, besides optimizing for image quality, we also have to consider and filter out SPAM content, as it accounts for a large portion of the content found on social media. In a context akin to this, McParlane et al. [37] and Schinas et al. [43] address the problem of automatically detecting images unsuitable for visual summarization, stating the importance of dealing with images such as "memes"and captioned images like the ones presented in Figure 2.5. Particularly in [37], the authors also deal with the problem of identifying duplicated and near-duplicated images resorting to techniques such as pHash, as a way to deal with the large amount of duplicate content found on social-media.

2.2.6 Other

Previously discussed research works approach the task of evaluating images by taking only into account their visual characteristics. However, content posted to social media tends to be attached to other types of informative characteristics, as social signals and general metadata. Consequently, we now review works that use such information to infer the quality of content under evaluation.

In this context, Agichtein et. al. [2] propose an architecture for qualifying content in Q&A dedicated forums like Yahoo! Answers. The approach of the authors takes into account the social signals attached to the content under scrutiny, the intrinsic quality of the content calculated by examining it independently and the metadata associated with the content. Although the framework proposed in this research work is designed specifically to evaluate answers posted to Q&A forums, the authors make clear that, with the appropriate modifications, it could be used to qualify any type of user generated content posted to social media. More specifically, in the particular case of this study, intrinsic content quality is evaluated by examining text characteristics like punctuation, orthography, and grammatical correction. In the context of image evaluation these characteristics could be replaced by visual ones, like exposure quality or focus quality, as made explicit by the authors.

Finally, the authors of [37] also take into account social signals to rank images while tackling the problem of visual summarization of events using social media content. In

this context, an image attains a higher rank in function of the quantity of retweets it is associated with and of how many near-duplicates exist of it. Higher ranked images are then chosen as candidates to be part of the summary. A more thorough analysis of this work is detailed in the next Section which regards literature on summarization tasks.

2.3 Multi-document summarization

In the context of this thesis, creating a storyline means organizing content in a way that forms a cohesive, short and pleasing narrative. Consequently, we analyze approaches to summarization tasks. Although, for some of these tasks, the pleasantness or interestingness of the summary may not be a priority, by definition they focus on ways to automatically compile an abridged version of a large pool of information.

2.3.1 Social media

Summarization of social media content has been the focus of various research works throughout the years. This has happened as a result of the increasing need to take advantage of the large amounts of content being posted every day on social media.

In [40], Nichols et. al. propose a methodology to perform automated text summarization of sporting events using data collected from Twitter. In this case, the authors find the most important moments of an event by identifying spikes in the volume of tweets over time, an approach also applied in TwitInfo [34]. A visualization of the result of this approach can be observed in Figure 2.6. After acquiring these tweets and applying spam removal techniques to the set, a phrase graph is created from the text present in the tweets. This graph details the chance of a word appearing next to a previously established sequence of words in the context of a phrase. The phrase graph is used to generate possible sentences that summarize the event. These sentences are also scored through the phrase graph and the best scoring ones are outputted. As a possible alternative to this graph based algorithm, the authors propose Sharifi's modified TF.IDF [46] as a method for generating a text summary of the event. Although not directly related to visual storyline generation, these works highlight the importance of considering the amount of content being posted to social media, at any particular time, as metric for finding if interesting and important events are taking place.

Three years after Nichols et. al. published [40], the authors of [43], Schinas et. al., tackled a similar the problem, by rooting their research in a similar approach. In this work the task of visual event summarization using social media content, again from Twitter, is tackled through the use of topic modeling and graph based algorithms. The authors start by filtering a stream of tweets from a specific event in order to obtain only the most informative ones. This filtering process takes into account image size, image type, text size, text morphology through part-of-speech tagging, among other criteria. A multigraph

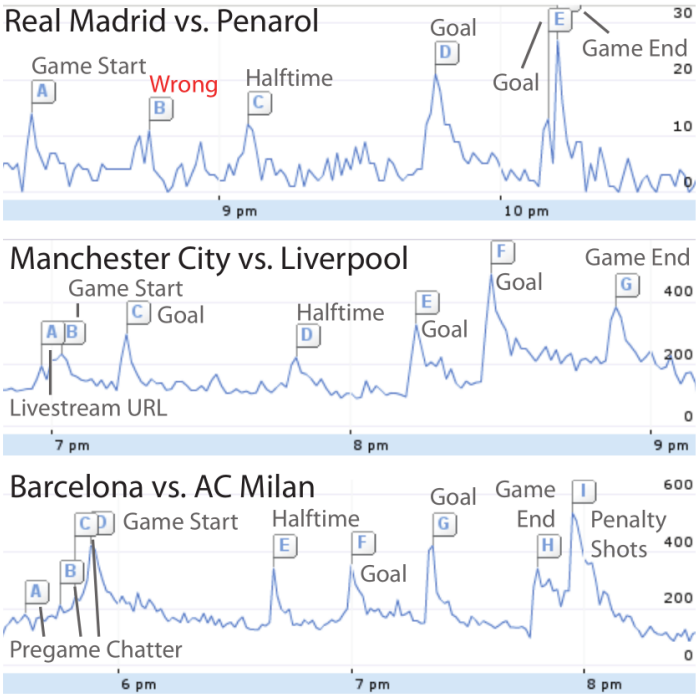


Figure 2.6: Spikes in the volume of data published to Twitter relative to three different football matchs. As shown the spikes tend to align with the important parts of the match. Source: [34].

is then created. Each node representing a tweet and being connected by an edge to other nodes as many times as the following criteria are applicable:

- Text similarity between tweets below a specific threshold.
- Image similarity between tweets below a specific threshold.
- Temporal proximity between tweets below a specific threshold.
- One tweet is a reply to another.

Re-posts are then discarded and tweets with duplicated images are clustered with the help of the graph. The tweets in these clusters are removed from the graph and are replaced by a single node that encompasses the information of the removed tweets. The authors then intent on discovering which tweets are part of the targeted event. In order to do this, they apply the SCAN [55] algorithm to the graph, which returns a set of subgraphs, each corresponding to a different topic. Finally, for each subgraph, the authors extract the respective images and rank them according to their popularity, relevance to the topic and the amount of information they provide.

These approaches emphasize the trend that is the usage of graphs in summarization tasks. In most cases they are used to represent the connections between pieces of content. This then allows for the application of already established graph algorithms, in order to find content highlights and appropriately structuring them in chronological fashion.

In [37] the authors also tackle the problem of visual summarization, although in this case the summarization is done with images not only from Twitter but also from other websites found in the URLs of tweets. After filtering the obtained images in a manner similar to the one proposed in [43], the authors rank the remaining images using social signals. Finally, besides taking into account the popularity of an image the authors also factor into its rank the diversity of said image when compared to the remaining available ones. They achieve this by grouping the images through semantic clustering and then giving priority to the ones with the highest TF.IDF.

Of note is that both authors of [43] and [37] tackle the problem of filtering duplicated content, although through different methodologies. In order to reconcile both approaches one could use pHash as proposed by [37] and cluster the results through a graph based approach in a way similar to what is described in [43].

Additionally, works such as [4] and [5] were also studied. In this particular case the authors discuss the importance of taking into account the media content present in microblog posts during the process of summarization. Consequently they tackle the task of filtering irrelevant or noisy content, as using this type of media may severely degrade the quality of the generated summaries. This problem is approached trough the use of a spectral filtering model. Having filtered the content, the authors propose and define a variation of LDA, CMLDA (or Cross-Media-LDA) designed to simultaneously deal with the textual and visual aspects of social media content.

2.3.2 Personal photo stream

With the advent of the smartphone cameras the act photographing and self documenting events as become a common practice. As such, paralell to social media summarization, contemporary research has focused on applying and developing summarization techniques in the context of personal photo streams. This field of research attempts to help with the process of filtering, cataloging and organizing this kind of content.

As an example of research done on this field we analyze the work of Yang et. al. [56]. In this work the authors tackle the task of creating a temporally organized summary of an event from a set of photo streams extracted from different sources. The authors give as an example events like weddings and family vacations where multiple cameras are used to capture different perspectives and subjects at different moments in time. In this situations organizing the available images into a chronological summary can be complicated: image files may not have a correct timestamp associated with them and the set of available images may contain similar and redundant content. To tackle this problem the authors align photo streams in a common timeline using a bipartite kernel sparse representation graph. Finally, a master stream, corresponding to the summary, is obtained by removing redundant photos and leveraging the informations obtained from the graph.

2.3.3 Other

In [28], Li et. al. aproach the problem of summarization from a different perspective. Although the authors are interested in generating summaries from multimedia content, the generated summaries are composed only of textual information. In this context the authors propose a framework to summarize text, images, video and audio into text. To do so, the media content is first characterized in textual form. A speech recognition method is applied to audio and a graph approach is used to measure the importance of each piece of content. Videos are separated into key-frames and image semantic extraction is performed on both video key-frames and still images. Although the task tackled by the authors is fairly different from the one approached in this thesis, the work draws attention to the value of considering different types of media in summarization tasks. In the context of visual storyline generation, this can mean analyzing not only the images extracted from social media, but also the text of the posts they were extracted from. Additionally, the approach of decomposing videos into key-frames provides a possible method to easily extend the work developed in the context of this thesis regarding images to video content as well.

2.4 Storyline creation and editing

A visual storyline must be succinct and cohesive but also pleasing to the viewer. Overall it must present a sequence of images as an interesting and informative narrative. Whether in the news room or in the context of cinema, the processes of ordering and cutting content to fit these criteria is the job of the editor. Consequently, emulating this process in an automated way means understanding what makes a sequence of images cohesive, interesting and appealing, or not, to the viewer.

2.4.1 Professional

To do so, we could turn to literature on the topic, approaching it first from a non-computational point of view. However, analyzing works on film and video editing such as [42] or [38] yields less interesting insights then those extracted from the previously described literature on photography. This is the case as techniques and rules in editing are generally highly subjective, context driven and, consequently, hard to mimic algorithmically. Regardless, these works elaborate on the importance of ensuring the cohesion and pleasantness in the content as a whole, but also of the transitions between its individual pieces of media that compose it. More specifically, as per [38], the quality of individual transition between two pieces of content is a result of both the visual and semantic characteristics of the pair.

2.4.2 Automatic

Since no works could be found that specifically tackle the task of visual storyline generation, we turn to works on various forms of content editing, in order to understand what technologies and methods where proposed by previous authors as basis and support for the work developed in the context of this thesis.

In [15] the authors approach the task of semi-automated (not fully automated) home video editing as opposed to professional video editing. Additionally, they focus on optimizing only for simple and few editing techniques and shot characteristics, like brightness and length. Although works like this approached the task of editing in a simplified manner, they were the basis for the more complex approached that proceeded them.

In [30], Lino et. al. also approach the task of automated video editing, this time taking into account more complex rules and editing concepts. Here, the goal of the authors is to automate the decision process of which cameras to use, when to cut a shot and which camera to cut to, in the context of animated computer generated videos. All of this while taking into account notions of shot composition, continuity and pacing. To do this, the authors propose an approach in which they try to minimize the total cost of a full video edit. This is calculated by taking into account the cost per shot and the cost per transition between each pair of shots in a final video. In turn, these costs are calculated as the the weighted sum of the violations the shots or transitions incur in. As a whole, this allows the authors to approach the difficult problem that is the subjective evaluating an edit of a video, in a algorithmic way. To optimize for individual shot quality both shot composition and shot duration are taken into account. Here shot composition is used as a metric of the pleasantness of the shot, while shot duration is used to ensure all shots in the video have a similar duration. Regarding transition quality the authors consider the following rules:

- **Screen continuity:** the eyes of the actors should remain in similar position on screen after a cut. This is a simplification of an editing technique that has as an objective preventing the viewers to be forced to search the images for their main subjects after each cut. This creates a more pleasant and balanced viewing experience. Although, in editing, this technique applies to the general subjects of the shots, the authors of this work took only into account the eyes of the actors as a way of simplifying the complex problem that is identifying the subject or subjects of a shot.

- **Gaze continuity:** preserving the actor's gaze direction. This technique is used in editing to help assert and preserve both the position and direction the actors occupy in relation to each other and their surroundings, in the viewers perspective. Figure 2.7 gives an example of a transition that follows this rule and one that does not.

- **Motion duration:** preserving the actor's movement and motion directions. This is done for the same reason and the rule presented above.

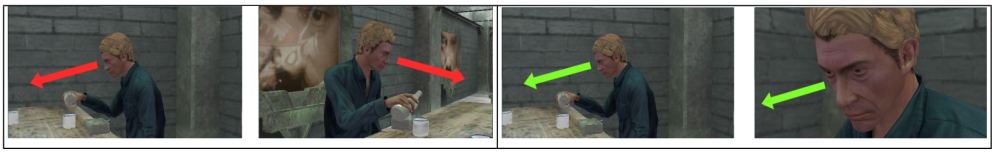


Figure 2.7: The pair of scenes on the left does not follow the gaze continuity rule, while the pair of scenes on the right does. Source: [30].

Although interesting, it is much easier to take advantage of these criteria in the context of animated computer generated videos, then in the context of live action ones. As an example, in a computer generated video finding the orientation and direction of the gaze of an actor requires simply the model of the actor to be tracked, while in a live action scenario the only automated approach would require taking advantage of face recognition techniques.

Hence, we move away from editing in the context computer generated videos to editing in the context of live action ones, analyzing the work of Smith et al., [48]. In this work the authors tackle the task of creating a trailer for a full length feature film in a automated way. As opposed to earlier works the edit provided by the system was actually used in a real life situation, although with manual alterations, serving as a movie trailer for Morgan, a 2016 horror movie. The system leveraged deep learning, audio, visual aesthetics and shot semantics, sentiment analysis and statistics associated with the content of horror movies, to decide which scenes to include in the trailer. Figure 2.8 shows the scene selection and organization for both the fully automated version of the trailer as well as the final, manually edited one.

Finally, not many research works could be found directly regarding news media editing. In this context, one of the few works that provide insight into the task is [9]. In it, the authors observe that presenting repeated images/videos in the context of news related storylines makes viewers perceive said storylines as having less quality, even if relevant information is being shown. A simple but very relevant insight regarding the task tackled by this thesis.

2.5 Critical summary

Regarding literature on identifying quality images, as debated throughout this Chapter, multiple research works have already been conducted regarding various forms of image quality, although none was found that tackled the task of identifying news quality content.

This novel computer vision task, following computational approaches to aesthetics, interestingness, memorability and more recently exoticism, is of less abstract character than previous ones. In itself, this poses new challenges, as we are working with a more constrained and defined set of quality criteria then, for instance, works on general image

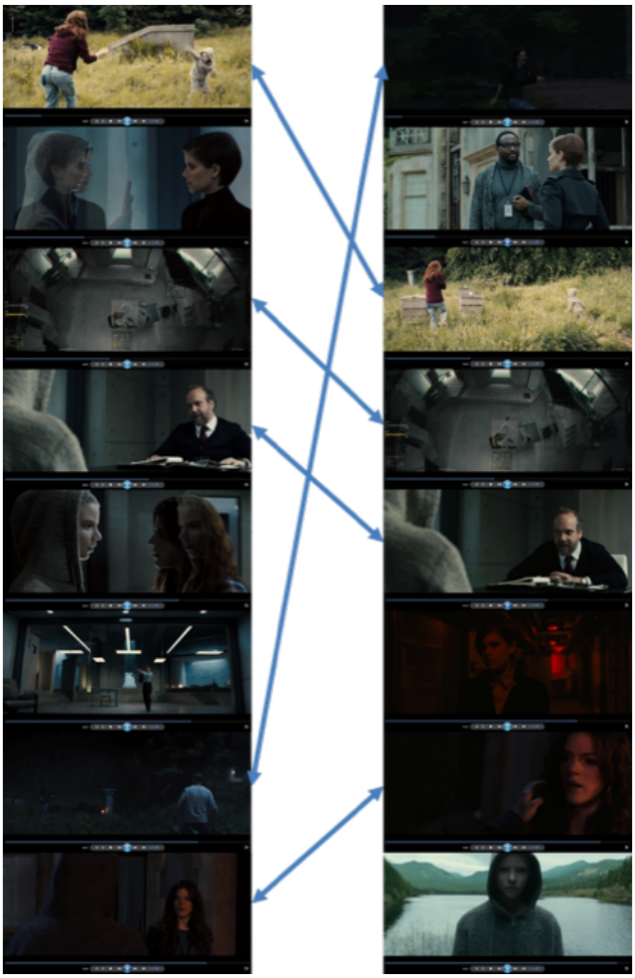


Figure 2.8: The column on the left shows the scenes selected in an automated way by the system proposed in [48] for the trailer of the movie Morgan. The column on the right shows the final trailer's scenes, after being manually improved by professional editors. The blue arrows show the chance in order applied to the scenes in the final edit that where present of the automated one. Source: [48]

aesthetics. Additionally, because this is a novel task, no datasets and ground truth are available to evaluate possible approaches to the problem. As such, we propose to follow an approach similar to the one documented in [24], crowd sourcing the creation of a ground truth for a dataset containing images extracted from social media. Additionally, inspired by the works on image aesthetics, interestingness and memorability, we aim to leverage machine learning methodologies, supplementing them with a large set of low and high level features, in order to tackle the task. Since we aim to, not only correctly identify news quality content, but also explain why a piece of content was deemed as having news quality, models like the one presented in [6] are not valid approaches to the problem. This because, deep learning models suffer from low interpretability, regardless of overall performance.

Regarding works on summarization methods, as already discussed, this is a research

field that has received a lot of attention in recent years and as such, many summarization methodologies have been proposed, discussed and evalutated in literature, be it in the context of social media or otherwise. However, these methods tend to focus solely on the task of creating abridged versions of a large amounts of data, without tackling the problem of optimizing the resulting summaries for aesthetic quality and pleasantness. This makes sense in the context of problems where the objective is only to provide a useful synthesis of the information available. This is not, however, true in the context of this thesis. Although it is our aim to develop a method for automatically creating visual storylines able to summarize a targeted news story, we want to do so while ensuring the aesthetic quality and cohesion of said storylines. As such, we aim to build on the approaches of previous authors while also providing the necessary modifications in order make them fit the specifications of the problem at hand.

Finally, with respect to storyline creation and editing, no research work could be found that tackles the exact challenges and problems posed in Chapter 1. However, there is research on automated video and film editing, even if mostly from a semi-automated perspective. These approaches base themselves on simplified versions of techniques and criteria put forward in the context of professional manual editing. These simplifications are required as editing techniques are highly subjective, context dependent and require the understanding of the visual and semantic characteristics of the content available, all of which is extremely hard to achieve from a computational stand point. Of these works, [48] presents both a particularly interesting problem and approach, leveraging deep learning methodologies to create the first semi-automatically generated trailer for a full length feature film. Inspired by this approach, we intend to leverage machine learning methodologies to identifying and understand what semantic and visual criteria result in a visual storyline being perceived as cohesive and pleasing to the viewer.

Summarizing, various gaps can be found in literature regarding computational approaches to visual storyline evaluation, identifying and understanding news quality visual content, editing in visual storylines and overall automated storyline generation. By tackling the novel research task purposed in this thesis we aim to research and evaluate possible solutions to the aforementioned problems, providing a stepping stone for future works on related subjects.

Chapter 3

An evaluation framework for visual storylines

3.1 Introduction

Achieving *quality* is only possible after understanding what characteristics *quality* derives from. In the context of this thesis, this means understanding by what characteristics should we evaluate a tool designed for visual storyline creation, as by doing so we gain insight into the process of optimizing such a tool. In this context we face two main challenges. The first relative to the high degree of subjectivity associated with the process of visual storyline creation and evaluation. The second associated with the fact that, to our knowledge, no evaluation framework has been proposed in literature for the task at hand.

Hence, in this Chapter we propose a novel framework to evaluate the various approaches to storyline creation that are proposed in this thesis and that may be proposed in future research work. Additionally, this evalutation framework has already been chosen to evaluate the systems submitted to the "Social-media video storytelling linking" competition (https://www-nlpir.nist.gov/projects/tv2018/Tasks/lnk/) that took place during TRECVid 2018, an annual workshop where worldwide multimedia retrieval and analysis competitions take place.

To create this framework we based ourselves in the Cranfield Experiments [52], a literature standard for evaluating the performance of information retrieval systems. In this context the systems are evaluated by receiving as input a set of queries and being tasked to retrieve documents from a dataset. The quality of the retrieved documents to a particular query is evaluated according to ground truth.

As such, we first establish and detail four distinct datasets composed of images collected from Twitter, leveraging a pool of social media content with heterogeneous characteristics. Complementing these datasets, we also define the queries, these being a manually curated set of stories extracted from news media, with which it is possible to create quality visual storylines, by using the images from the datasets. Additionally, we define the method by which the ground truth was created through crowd sourcing. In this context there is the need for an interface in order to present visual storylines to the viewers (the annotators). Furthermore, the quality of the interface may interfere directly with the viewers perception of a storyline. Consequently, we establish the interface developed by Marcucci et. al. in the context of [33] as the one used to present visual storylines to the viewers during evaluation processes, elaborating on the characteristics that make it a solid choice for the task.

Finally, we derive a metric for visual storyline evaluation composed of two distinct dimensions that are easier to evaluate then overall quality. This metric provides an empirical method to qualify visual storylines that is consistently used in the process of attaining ground truth throughout this thesis.

3.2 Dataset and queries

Image datasets and curated stories (the queries) are necessary to evaluate storyline generation methods. These datasets must contain images from which it is possible to create quality visual storylines that illustrate the stories. Since no datasets with these characteristics could be found, our approach was to pick a set of datasets for which no ground truth or stories existed, create the stories and, as required, proceed to the creation of ground truth through crowd sourcing.

Due to the nature of the task we pursued datasets containing only posts from social media related to individual events. We chose Twitter as the source of social media images. This choice is supported by a proven correlation between what is posted on the social network and news media. As an example, [27] shows that over 85% of the topics trending on Twitter are also covered by the news. As criteria for picking the events we focused on those that span over multiple days and that gather a lot of social media and news media traction like music festivals and sports competitions. This allows us to have large amounts of content to analyze and use for experimentation, with varied quality and characteristics. Finally, the focus on events covered by news media means that we can look directly at news pieces in order to find stories to illustrate. In total four datasets where chosen. The targeted events were:

The Edinburgh Festival (EdFest) a celebration of the performing arts, including dance, opera, music and theatre performers from all over the world. The event takes place in Scotland and has a duration of 3 weeks.

Event	Stories	Docs	Docs w/images
EdFest 2016	20	82348	15439
EdFest 2017	13	102227	34282
TDF 2016	20	325074	34865
TDF 2017	15	381529	67022

Table 3.1: Summary of the characteristics of the 4 datasets.

Le Tour de France (TDF) a world famous road cycling race competition. The event takes place mainly in France and has a total duration of 16 days.

More specifically, we considered EdFest 2016 and 2017 and TDF 2016 and 2017. For each event we then created stories with 3 to 4 segments each based on news media content we researched, related the events. A sample of these stories for TDF 2016 and EdFest 2016 can be found in Tables 3.4 and 3.3. Throughout this thesis we will evaluate methods of storyline generation by their ability to pick content from the EdFest and TDF datasets to ilustrate these stories.

Table 3.1 summarizes the content of the datasets and the amounts of stories created for each dataset.

3.3 Ground truth generation

In order to present a visual storyline to a viewer (or crowd source annotator) an interface is required. For this task, we carefully choose an interface that was purposefully designed for visual storyline visualization, which provides us with some guaranty that bias in the annotation process is reduced. We opted for the interface proposed in [33] by Marcucci et. al., an interface designed specifically to present visual storylines composed of social media content, carefully build and tested through an iterative development process in which testers were constantly being interview to provide quality feedback. Hence, all experiments elaborated throughout this thesis where viewers were shown visual storylines were conducted using this interface. Figure 3.1 presents the interface.

3.4 Visual storyline quality metric

As detailed in Chapter 2 media editors are constantly judging the quality of news material to decide if it deserves being published. The process is highly skillful and deriving a methodology from such a process is not straightforward. The motivation for why some content may be used to illustrate specific segments can derive from a variety of factors. While subjective preference obviously plays a part in this process (which cannot be replicated by an automated process), other factors are also important which come from common practice and general guidelines, and which can be mimicked by objective quality assessment metrics.

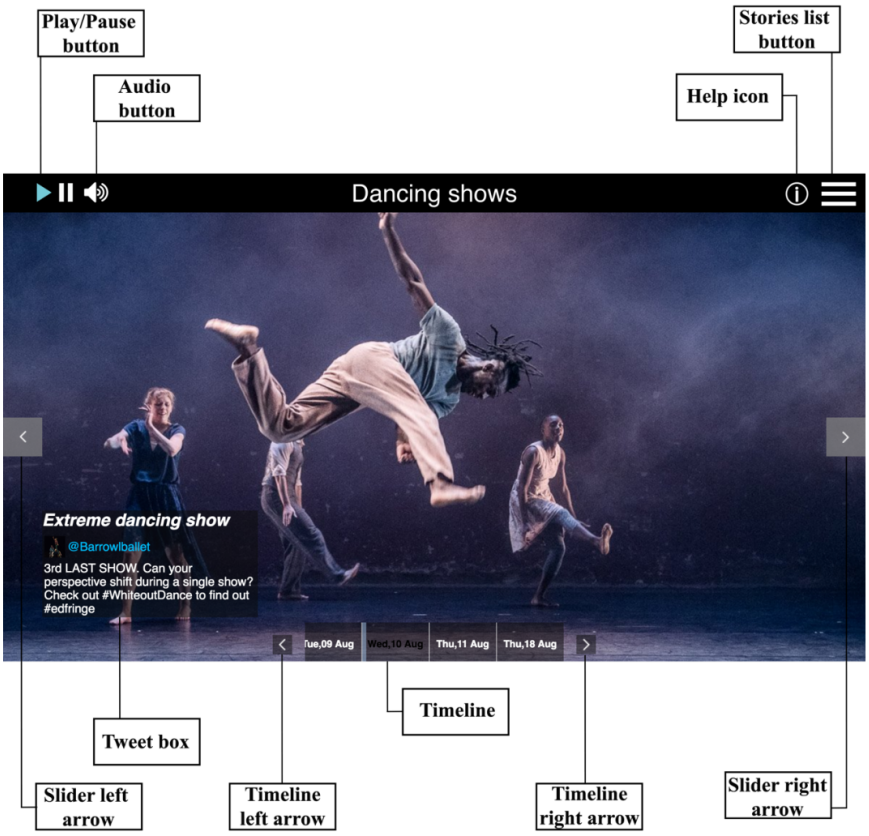


Figure 3.1: Visual storyline interface used in the experiments conducted throughout this thesis. In this particular case presenting a storyline created using content extracted from Twitter. Source: [33]

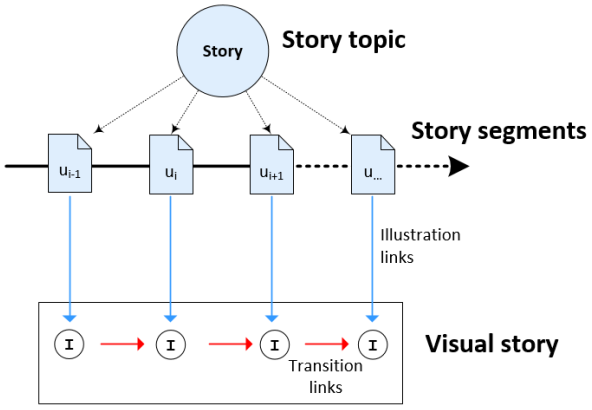
Therefore we propose a metric, inspired by [30], aimed at qualifying storylines by means of judging specific, more objective characteristics – Figure 3.2 illustrates the visual storyline quality assessment framework. In particular, visual storylines are assessed in terms of *relevance of illustrations* (blue links in Figure 3.2) and *transition quality* (red links in Figure 3.2). Formally, given

$$Story_N = (u_1, u_2, ...u_N) \tag{3.1}$$

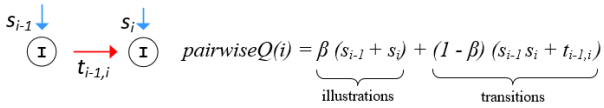
and

$$VisualStoryline_N = (w_1, w_2, ...w_N) \tag{3.2}$$

we consider s_i, the *relevance of illustration* w_i to the segment u_i. Similarly with respect to *transition quality* of the pairs of images in a visual storyline we consider $t_{i,k}$, representative of the visually and semantic coherence of a transition between illustrations w_i and w_k. Both s_i and $t_{i,k}$ are values ranging between 0 and 1. These two dimensions are then used to obtain overall expression of the *"quality"* of a given visual storyline for a story of N



(a) Social story evaluation framework.



(b) Social story quality assessment metric.

Figure 3.2: Methodology for evaluating visual storyline illustration.

segments. This is formalized by the expression:

$$Quality = \alpha \cdot s_1 + (1-\alpha) \cdot \frac{1}{2(N-1)} \sum_{i=2}^{N} pairwiseQ(i) \tag{3.3}$$

$$pairwiseQ(i) = \underbrace{\beta \cdot (s_i + s_{i-1})}_{\text{segments illustration}} + \underbrace{(1-\beta) \cdot (s_{i-1} \cdot s_i + t_{i-1,i})}_{\text{transition}} \tag{3.4}$$

where the function $pairwiseQ(i)$ defines quantitatively the perceived quality of two neighbouring segment illustrations based on their relevance and transition, on a 0 to 2 continuous scale. In turn, α weights the importance of the first segment, and β weights the trade-off between *relevance of segment illustrations* and *coherence of transitions* towards the overall quality of the story.

Given the underlying subjectivity of the task, the values of α or β that optimally represents the human perception of visual storylines, are in fact average values. Nevertheless, we posit the following two reasonable criteria: (i) illustrating with non-relevant elements ($s_i = 0$) completely breaks the story perception and should be penalised. Thus, we consider values of $\beta > 0.5$; and (ii) the first image perceived is assumed to be more important, as it should grab the attention towards consuming the rest of the story. Thus, α is a boost to the first story segment s_1. Finally, because of the proved negative impact repeated images have on the quality of visual storylines, as shown in [9], we add that visual storylines that present the same image more then once are rated by the Quality metric with a score of 0.

Annotator rating	Avg. *Quality* Score	RMSE
1	1.42	0.546
2	2.22	0.523
3	3.07	0.574
4	3.82	0.596
5	4.84	0.353

Table 3.2: Performance of the metric proposed in 3.4, when compared against the judgement of the annotators.

3.5 Metric evaluation

To test the accuracy of the proposed quality metric at emulating the human's perception of visual storyline quality we resorted to crowd sourcing. To do so, were illustrated 40 Edinburgh Festival 2016 stories using images from the EdFest 2016 dataset and the *BM25* baseline described in Chapter 5. Afterwards, 5 annotators were asked to (i) rate each story segment according to relevance as 0 ("not relevante") or 1 ("relevante"), (ii) rate each transition according to quality as 0 ("bad") and 1 ("good"), and (iii) rate overall quality on a scale of 1 to 5. Using these judgments we fine tuned the parameters of the metric, setting α and β to 0.1 and 0.6 respectively, values which were used in the remaining experiments elaborated throughout this thesis.

Table 3.2 shows the average *Quality* score predicted by the metric for the stories annotated with the 5 possible ratings and the RMSE of the quality score against the actual ratings. These values show that linear increments in the ratings provided by the annotators were matched by the metric with an average RMSE of 0.552. Thus, these results show that the metric *Quality* effectively emulates the human perception of visual storyline quality.

3.6 Visual storylines guidelines

After the evaluation process of the *Quality* metric, the annotators where asked to provide written commentary on what factors impacted their perception of quality of the visual storylines they where presented. By analyzing the aforementioned commentaries we gained new insights into what characteristics impact visual storyline quality. Bellow we present a summarized version of the four main conclusions that resulted from this analysis.

- Storylines with repeated images tend to be perceived as bad by viewers.
- It is more important for overall storyline quality that the images in the storyline are relevant to the segments they are illustrating then that the storyline contains quality transitions between images.

- Storylines that start with images that are not relevant are perceived as bad by viewers.
- Although the annotators presented diverging opinions on what results in good transitions between images, in general annotators commented they took into account, amoung other factors, the semantic and color similarity in image pairs, when rating transitions.

These commentaries are a crucial first step to understand how to optimize visual storyline generation methods. Additionally, they confirm the validity of some of the decisions made when creating the quality metric, as well as enforce the importance of some concepts already proposed in literature. Namely that color and semantic similarity are of high importance in visual storyline transitions and that repeated images in storylines should be avoided.

3.7 Conclusion

In this Chapter, we established a framework for the evaluation visual storyline creation methods. Basing ourselves on the Cranfield Experiments we define datasets and respective queries (which in our countext, are stories). Furthermore we proposed a method for attaining ground truth through crowd sourcing, defining a quality metrics for general visual storyline evaluation and establishing a graphical a interface to be used in the annotation process.

Finally, we tested the evaluation framework through crowd sourcing, proving that the proposed visual storyline quality metric does effectively emulate the human perception of visual storyline quality while also gaining, in the process, insight into some of the characteristics that impact the quality of visual storylines.

With an evaluation framework established we are now ready to begin tackling the problems related to visual storyline generation.

Table 3.3: Edinburgh Festival 2016 stories and segments.

Story Title	Segment (1)	Segment (2)	Segment (3)	Segment (4)
Edinburgh comedy awards 2016: the nominees in full	Eight comedy shows are in the running for the prestigious prize at the Edinburgh festival	James Acaster receiving a fifth consecutive nomination	The most eye-catching nominee is the Fife comic Richard Gadd, overlooked for his extraordinary stunt-comedy show Waiting for Gaddot last year	Completing a trio of Australians on the shortlist, Zoe Coombs Marr is the only female act nominated – and even she dressed as a man to get there
My first fringe: the Edinburgh baby shows getting gurgles of applause	Full of costumes and musical magic, the fringe can cast a spell over most adults – but for the very young, it's a creative introduction to a whole new world	The range and quality of shows for babies at the fringe is delightful	The Royal Botanic Gardens and its dreamy musical delights can be enjoyed by all families	Babies and kids at Edinburgh Festival
Comedy at Edinburgh Festival	Edinburgh Festival is Full of Comedy Shows	A comedian and a microphone: Several Stand-up shows take place across festival's stages	Comedy crowds sometimes have the size of a football team	The best joke award goes to Masai Graham's
Theatre at Edinburgh Festival	Actors performing on Stage - Play	Edinburgh Festival has plenty of theatre shows	There's a bevy of Shakespeare-related shows at Edinburgh in the playwright's quadricentenary year	The Glass Menagerie – triumphant take on Tennessee Williams
Edinburgh Festival locations	Edinburgh Castle	Streets	Stages	Parks and woods
Edinburgh Festival attractions	Music shows	Theater and Comedy	Circus	Street Performances
Gastronomy at Edinburgh Festival	Pizzas	Hamburgers	Deserts	Drinks
Scottish Elements	Bagpipes	Food and Drink	Outfits	Military Parade
Edinburgh Castle is one of the main attractions	Deep time Show	Fireworks	Beautiful streets of Edinburgh with its castle on the background	People enjoying Edinburgh Castle clear blue sky
Street Performances	The Edinburgh Festival is home to one of the most unique celebrations of arts	Street circus is a popular attraction at Edinburgh Festival with several artists such as unicycle jugglers	Street circus is full of colorful artists	Bagpipes
Music shows	Audiences at Edinburgh Festival music shows	Guitar on stage	Band on stage	Singer close-up

Table 3.4: Tour de France 2016 stories and segments.

Story Title	Segment (1)	Segment (2)	Segment (3)	Segment (4)
Chris Froome's path on TDF2016	Chris Froome pedaling	Chris Froome as Yellow Jersey	Chris Froome forced to run to recover from crash, at Mont Ventoux	Chris Froome became Britain's first three-time winner of Tour de France
Nice Attack repercussions on the Tour	Lorry attack on France	Minute's silence held before time trial on stage 13 of Tour	Several Cyclists paid tribute to the Nice Attack	Security stepped up after the incident
Tour de France Highlights	Thermal cameras will be used at the Tour de France to detect motors in bikes	Sprinting to the finish line	Mountain stages are the harder ones	Riders close-ups
"Out-of-control"Spectators	Large number of Spectators on TDF 2016	"Out-of-control" spectators who jam the path of Tour de France riders	Fans take their selfies and run along riders	Pile-up on the 12th stage caused by spectators
Happy moments at TDF2016	Cyclists celebrating	Cyclists at the podiums	Camaraderie is a big part of the sport	Crowd cheering for the athletes
Adversities at TDF2016	Sometimes cyclists crash	Bad weather during TDF2016	Cyclists getting back to the race after a crash	Animals interfering with the race
TDF: not only about racing	Multiple interviews take place during TDF2016	Beautiful landscapes and views surrounding the tracks	People taking selfies during the event	Some people dress up in costumes during the event
Popular cyclists	Highlights of Chris Froome	Highlights of Mark Cavendish	Highlights of Peter Sagan	Highlights of Adam Yates
Cycling Visual Semantic Patterns	Group of cyclists	Single Cyclist	Getting Assistance	Close-up
TDF2016 Popularity Highlights	TDF2016 Starts	Stage 9	Stage 12 and Froome crash	TDF2016 Ends
TDF2016 Monuments	Mont Saint-Michel	Eiffel Tower	Triumphal Arch	Louvre

CHAPTER 4

RANKING NEWS-QUALITY MULTIMEDIA

4.1 Introduction

Picking images to be used in the context of news media is a difficult and nuanced task, one normally attributed to news editors. The process itself is complex and takes into account many variables: the visual quality of the image, how it relates to the news it is supposed to illustrate and how much of the story it conveys by itself, are just some of them [25].

In the context of creating a framework for news media visual storyline generation we aim to automate this process, ensuring the quality of the content used to create storylines matches what is expected by news media professionals. Hence, as depicted in Figure 4.1, this first module of our storyline generation framework is tasked with filtering and ranking social media content according to news media standards. To do so, we specify a machine learning based approach for selecting high quality media, that can be used to

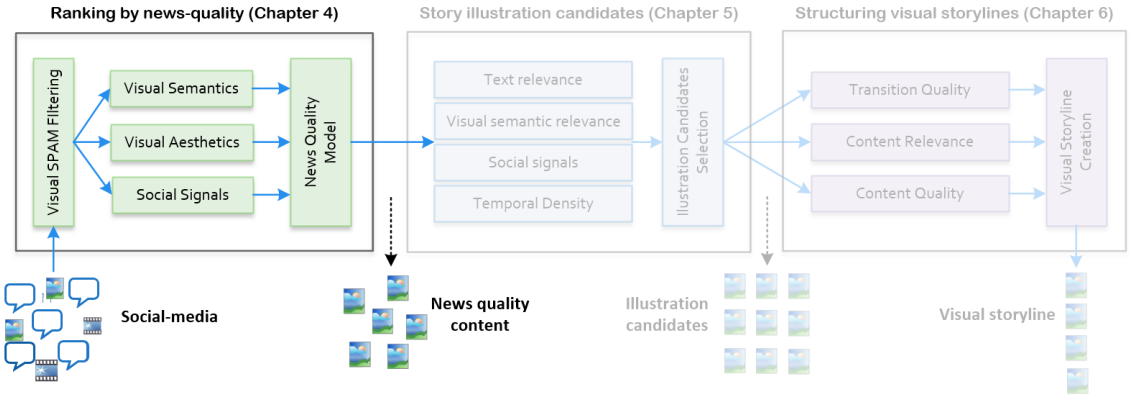


Figure 4.1: Highlight of the first module of the visual storyline generation framework.

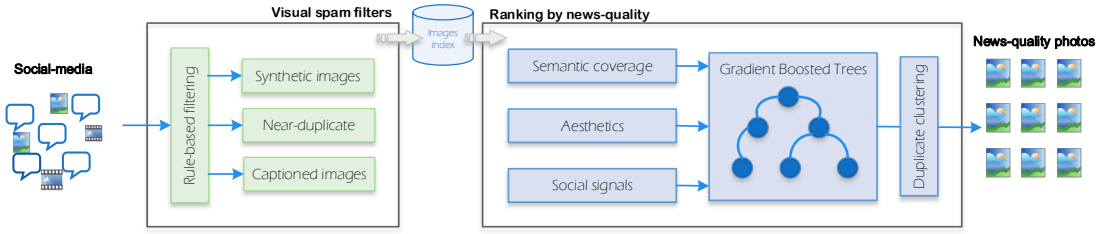


Figure 4.2: The framework for ranking news-quality pictures in social-media is leveraged by a machine learning algorithm that merges social, visual, semantic and aesthetic evidence.

illustrate a piece of news.

To model and quantify the photographic quality that news editors are looking for, our hypothesis is that one needs to consider the problem across three fundamental dimensions: aesthetic, semantic and social. The argument is that ranking by visual aesthetics alone, is not enough – the sharpness and colorfulness of pictures needs to be complemented by strong and clear semantic content. Also, getting some preliminary human feedback is crucial, hence, social features are also an important element.

Furthermore SPAM is a big part of the content present in social media. As such, we explicitly tackle the task of SPAM detection, to ensure that low-quality photos such as memes and adverts are not even considered for analysis. We do this by reviewing the textual and visual components of the content under scrutiny. Enforcing this specialized SPAM detection methodology allows us to simplify the task of the filtering and ranking methods, as these can be designed to solely work with content that is beyond a basic threshold of quality. Finally, we remove redundant duplicated content from the list of photos ranked and filtered ensuring the non-redundancy of the images outputed.

4.2 Finding news-quality pictures

Figure 4.2 illustrates the architecture of the proposed framework designed to filter and rank news worthy photos. Its main components are:

- **Visual SPAM filter.** The social media posts are first processed and filtered by a spam detection module. This way, images such as memes, adverts, and images of extremely low resolution, are immediately removed from the pipeline and are never considered in the ranking and filtering processes.
- **Visual redundancy.** The content that is not discarded by the Visual SPAM filter is then processed by the duplicate and near-duplicate image detection algorithms.
- **News-quality ranking.** Finally, we have a component responsible for filtering and ranking photos by their news-quality, i.e., a machine learning model, Gradient Boosted Trees [7] (GBT), that combines aesthetic, semantic and social criteria to infer how news worthy an image is.

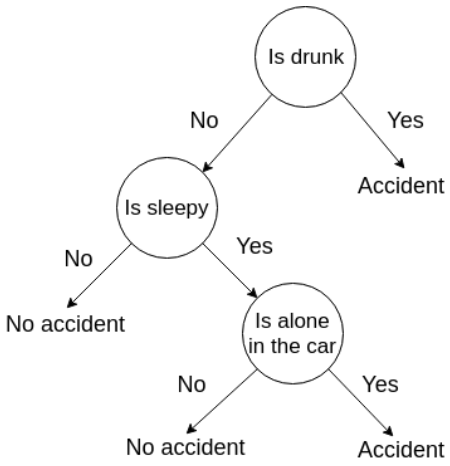


Figure 4.3: Example of a Decision Tree for binary classification using binary features, after training. Here the problem is that of predicting if a traffic accident will occur given the state of the driver as input.

In the following sections, these components are presented in detail.

4.3 Ranking by news-quality

Determining if a picture has news-quality is a complex task that cannot be solved by taking only into account its visual appeal. The picture can, for instance, be visual appealing but severely lacking in interesting content and information. To solve this, problem we consider not only the *visual aesthetics* of pictures, but also the *semantic content* and the the *social signals* associated with them. Moreover, we argue that there are non-linear interactions among these distinct sets of features. Due to this, and inspired by the work of [16], we propose to solve the present problem with Gradient Boosted Trees (GBT) – a tree based machine learning model designed for supervised learning.

In a similar fashion to other boosting methods, GBT leverages combinations of weak learners (simple machine learning predictors whose performance is only slightly better then chance at the task they where intended to perform) to create a strong learner. In this case GBT's weak learners are Decision Tree models (Figure 4.3). Used individually, these models have the advantage of being highly interpretable. However they also present serious performance disadvantages, most notably they do not generalize well, being highly prone to overfitting. GBT retains some of this interpretability while presenting a much better performance. The model combines the weak learners through the following iterative learning process: 1) a weak learner is trained on the input training data 2) the error of this learner is calculated 3) a new weak learner is trained to predict this error 4) using this error prediction the original learner's predictions are modified to improve it's performance, hence a new more robust learner is attained 5) steps 2 to 4 are repeated until convergence or until a pre-specified stop condition is met [7].

Besides being robust to outliers, GBT are known to work well with categorical and

Feature	Description
#Edges	the number of vertical, horizontal and diagonal edges present in an image.
Rule of 1/3	real value representing how much an image complies with the commonly used photography composition rule.
Focus	real value describing how focused an image is.
Entropy	real value measuring an image's entropy.
Faces	the number of human faces present in an image.
Luminance	real value describing an image's brightness.
Simplicity	real value representing how simple an image is in therms of the distribution of its colors.
Area	the width × height of an image in pixels.
Aspect	the height of an image divided by its width.
Orientation	if an image is square or in a portrait or landscape orientation.
Colorfulness	real value describing an image colorfulness.

Table 4.1: Visual features and respective descriptions. Figure 4.4 presents a visual representation of each of these features.

continuous data, which is a critical advantage to solve our problem, where both types of feature co-exist. Additionally, GBT also works well with both larger and smaller datasets, which means the ability to train the model with sets of news related media of different sizes. Further benefits of GBT also include the fact that it performs implicit feature selection, the ability to deal with non-linear relationships in the data as well as capture high-order interactions between features, making it, overall, a very versatile model. These, and other advantages are discussed in greater detail in [14] and [39].

Finally, GBT exist for regression and classification. We take advantage of both, using the classification variant of the model to allow for filtering by news quality while, the regression variant is used for ranking according to the same criteria. The benefits of each approach for finding news quality content are discussed in Section 4.5.3.1.

While we also tested other models, such as Linear and Ridge regression models, SVM^{rank} (an instance of SVM^{struct} [22]), Naive Bayes and Logistic regression. In the end, the model that yielded the best performance was GBT.

4.3.1 Visual quality

Deciding if a photo is news worthy is a very subjective task. Nevertheless, when approaching news media one expects a certain set of characteristics to be present in its

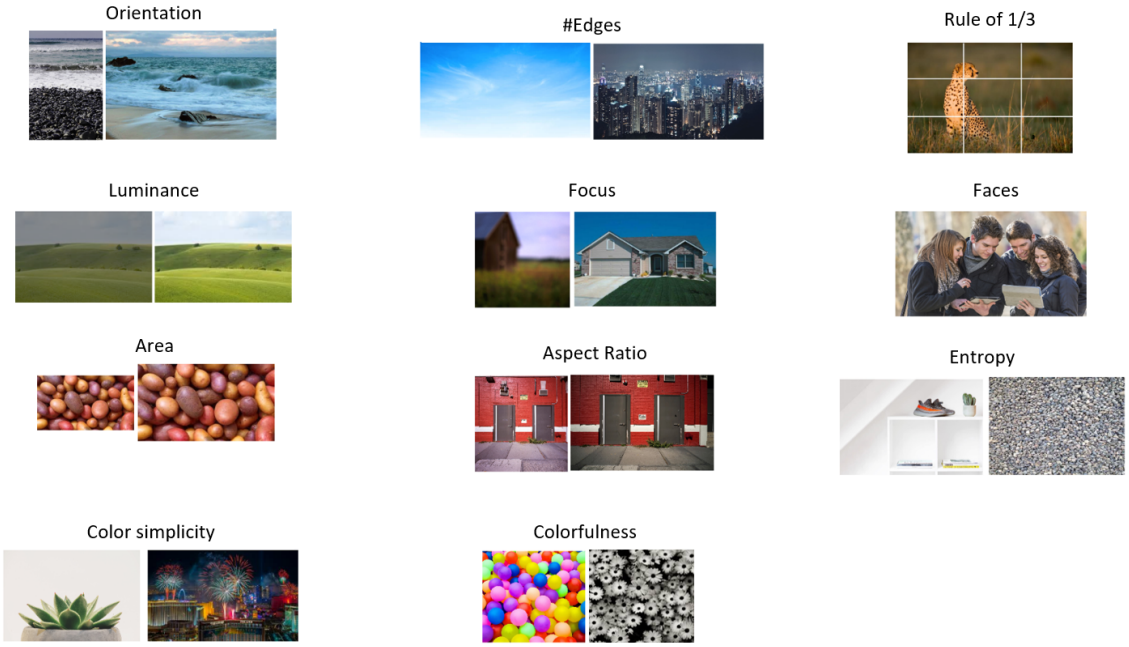


Figure 4.4: A visual representation of the features presented in Table 4.1.

visual content, even if only subconsciously. In order to make use of these latent characteristics we extracted a large set of visual features with which to quantify the visual quality of photos. This large set of features allow the Gradient Boosted Trees to perform implicit feature selection and capture complex feature interactions. These features are presented and described in Table 4.1, being that the first seven were extracted using the image feature extractor made available[1] by [35] and Colorfulness was extracted through the method proposed in [17]. Figure 4.4 presents a visual representation of these features.

The feature regarding the number of faces present in a photo was chosen as works such as the one presented in [21] show the positive visual impact of the presence of faces in photographies. Additionally, aspect ratio was added as a feature because certain photography equipment is directly associated with a specific image aspect ratio. As an example, DSLR's, cameras normally used in a professional or semi professional setting, normally output images with an aspect ratio of 3:2 [13]. Finally, the remaining features were chosen as proxies for various photography criteria as debate in [13] such as exposure correctness, composition quality, image sharpness, among others.

4.3.2 Visual concepts

Our initial intuition, was that images that are used to illustrate news pieces have a particular distribution of visual concepts associated with them. As Table 4.2 shows, visual concepts such as *selfie* are expected to be less frequent in news media images then a

[1] https://github.com/pcpmartins/extractor

Low quality concepts	Prob.	High quality concepts	Prob.
performance	1.980	product	0.282
performing arts	0.189	advertising	0.126
entertainment	0.153	font	0.118
performance art	0.126	selfie	0.115
dancer	0.108	facial hair	0.099
crowd	0.108	recreation	0.099
singer	0.105	dog	0.090

Table 4.2: Most common visual concepts associated with news-worthy and non-news-worthy images present in the *news quality photos* dataset described in Section 4.5.1, ordered by decreasing probability of appearance.

concept like *performing arts*. Using this knowledge, we propose a way to calculate two additional features that take advantage of these trends to improve our filtering and ranking methodologies.

Given two sets, Y containing images known to have news-quality, and N containing images known to not have news-quality, we first calculate Py_i and Pn_i the probability of concept i appearing in the images present in Y and N, respectively. We do this for all concepts extracted from the images in Y and N, that appear in more then one image. When a new image containing the set of concepts C is given as input to the framework, both $\sum_{x \in C} Py_x$ and $\sum_{x \in C} Pn_x$ are calculated so that they can be used by the Gradient Boosted Trees model. These values are the sum of probabilities of each concept belonging to an image that is news worthy and not news worthy, respectively.

Although multiple visual concept extraction methodologies are currently available, we choose to use the Google Cloud Vision API for the purpose. We considered a total of 850 unique concepts and, on average, each image was annotated with 7.7 concepts. Table 4.2 shows the most common concepts associated with the news worthy and non news worthy images present in the *social-media photos datset* described in Section 6.1.

4.3.3 Social signals

Given the subjectivity associated with the task of identifying news worthy images, it is important to take into account not only the data extracted directly from the images but also the social signals generated by the users who interacted with the associated social media posts, the images were extracted from. In practice, we consider the following social signals:

- **#RT:** number of retweets associated with the post the image was extracted from;
- **#FL:** the number of followers associated with the user who posted the tweet containing the image;
- **#UN:** the number of times an image is featured in the individual available posts;

Figure 4.5: Examples of unwanted images that can be immediately discarded (i.e., logos, adverts and memes).

- **#DD:** the number of times a visually near-duplicated image is featured in the individual available posts.

This information is used as a proxy for the users opinions regarding an image's importance and its entertainment and informative value.

4.4 Visual SPAM and redundancy

Images of adverts, captioned images, memes and similar visual content are big portion of the content posted by social media users which must be filtered by the framework. To solve this problem we propose a method to filter low-quality and redundant visual information, in order to prevent content like the one presented in Figure 4.5 from being indexed together with valid photos.

We propose a filtering pipeline, extending what was already developed in the context of [41], composed of four distinct parts. The first is the application of a set of simple thresholds well established in literature [36] to features extracted from both the images and posts they were taken from. The second is the usage of a linear regression model trained to detect synthetic images, to filter images such as digital adverts. The third is the application of Optical Character Recognition (OCR) technology to subsequently filter out captioned images, such as memes. Finally, the forth part deals with the large number of duplicated images found among social-media content. The pipeline is detailed in the following subsections.

Figure 4.6: Example of near-duplicate images. The first is the original image. The second is a cropped version of the first with different contrast.

4.4.1 Coarse filtering

To filter thumbnails, banners and adverts, we follow the same approach taken in [41], based on [37] and [36], and exclude images extracted from posts that contain more than 3 hashtags, more than 3 mentions or more than 2 URLs. Additionally, we also discard small images that, due to their size, are not useful in an illustration context (i.e., images with less than 200 pixels width or height).

4.4.2 Synthetic images detection

In order to filter synthetic images we made use of the logistic regression model trained, tuned and tested in the context of [41], that uses some of the features proposed by [29, 53]: *number of corners*, *number of vertical and horizontal lines*, *number of dominant colors*, *most common color*, and 3 additional features derived from the color transitions (the measure of color distance between two neighbor pixels).

4.4.3 Visual redundancy

Since a lot of images present in social-media are slightly altered versions of their respective originals, we take advantage of the methodologies used in [41] to find not only duplicated but also near-duplicate images. This is important as it means the ability to filter redundant content. Additionally, we propose a method for clustering the previously found near-duplicate images, enabling the grouping of different versions of the same original image through the use of a clustering algorithm already established in literature, DBSCAN [12].

4.4.3.1 Duplicate detection

To access whether two images are exact duplicates of each other we make use of the MD5 hash algorithm applied to the pixel values of the image. More specifically, we consider two images to be exact duplicates if their respective MD5 hash is the same. Before presenting the results of the ranking and filtering models, all duplicated images are removed.

4.4.3.2 Near-duplicate detection

To detect near-duplicated images, we employ perceptual hash (pHash)[2]. Previous work already proved it presents a high performance in the this task [49], regardless if the image is rotated, resized, cropped, exposure compensated or even if small elements are added to it (like a logo or signature).

As a method for assessing if two images are near-duplicates, we calculate the Hamming distance between their pHash codes, which corresponds to the amount of bit positions where those codes differ [26]. We consider two images to be near-duplicates if the Hamming distance between their pHash values is below 8 as proposed in [37]. An example of two near-duplicate images can be found in Figure 4.6.

4.4.3.3 Forming clusters of near-duplicates

In order to find near-duplicates we must take into account that one image might have multiple near duplicates and these near duplicates might themselves have near-duplicates, being important then to consider the transitive property of the concept. As an example, a cluster of near duplicated images can be created by successively cropping small amounts of an original image.

As a result, centroid based clustering algorithms are a bad choice for this task, as well as those that do not deal well with noise (images that do not belong to a cluster), such as KMeans. Due to these peculiarities DBSCAN [12], with parameters ϵ equal to 8 and $MinPts$ equal to 2, was chosen, as the algorithm clusters points according to their spacial proximity to the borders of existing clusters, as opposed to considering the clusters centroids. Additionally, the algorithm deals well with noise, allowing images for which there are no near-duplicates, to be left without cluster.

Since multiple near-duplicate versions of the same image might have been given as input to the framework, the framework hides them before presenting the results of the ranking and filtering models. The framework then outputs only the best ranked near-duplicate present in the input image set.

4.5 Evaluation

4.5.1 Datasets

To evaluate the different components of the proposed framework, we used two datasets: (i) newswire photos, used to train the high-quality photos models and (ii) social media images from which we need to retrieve high-quality photos.

News-quality photos. To create a robust model that is able to qualify photos according to their news-quality, we obtained newswire photos from The New York Times and

[2] http://www.phash.org/

Agreem.	Images	High quality	LQ/HQ ratio
57%	124	58	1.14
71%	129	55	1.35
86%	144	39	2.69
100%	103	17	5.06
78%	500	169	1.96

Table 4.3: Results of the annotations performed on the news-quality dataset according to the question "Could this image have appeared in the New York Times?".

the BBC web sites. We collected a total of 100 newswire photos and added 400 social media photos, sampled from the EdFest 2016 dataset. This new dataset comprises a total of 500 images that were annotated by 7 annotators with respect to their "*news-quality provenance*", as described in the following section. Moreover, the annotation effort allowed us to better understand the specific characteristics of news-quality photos.

Social-media photos. To create a dataset comprised only of social media content, we again resorted to sampling the EdFest 2016 dataset documented in Chapter 3. In order to evaluate our ranking method, we created a small sample of 1,500 photos for results pooling. Ground-truth was obtained through crowdsourcing by resorting to 7 annotators that judged the top-*k* best ranked photos photos of each approach tested.

4.5.2 News-quality photos ground truth

The *News-quality photos* datset was used to train the classification (for filtering) and regression (for ranking) models. All 500 images in this dataset were annotated via crowdsourcing by 7 annotators. The annotators were presented with the images and asked the question "*Could this image have appeared in the New York Times?*". Table 4.3 presents, in an abbreviated manner, the results of the annotation process. Through their answers we can infer the ambiguity of the task, as the 7 annotators only fully agreed on 103 images. As ground truth for the ranking task, 7 quality levels were attributed to each image according to the number of annotators that agreed that the image might have appeared in the New York Times. For this task all 500 images were considered and the regression models used in it were trained to predict these quality levels. As a ground truth for the filtering task only images where 71% or more of the annotators agreed, were considered. In this case, the image was regarded as having news quality if the majority of the crowd answered yes to the already mentioned question. The classification models used in the filtering task were trained to predict this binary judgment.

As Table 4.3 shows, out of all 500 images, only 17 of them were annoated as possibly having appeared in the New York Times, by all 7 annotators. Of these 17 images, 14 belong to the set of images extracted from news sources, which shows the ability of the annotators to distinguish news-quality images.

Figure 4.7: True positives: examples of images the annotators correctly assessed as being extracted from news media.

4.5.3 Results and discussion

4.5.3.1 Analysis of the crowd sourcing results

By analyzing the *news quality photos* dataset ground truth we concluded that images that were easily and correclty identified as having news-media provenance by the annotators have at least one of three characteristics:

1. They have high visual quality, being sharply focused, correctly exposed and adequately framed. Prime examples of this are the images depicting the athlete and the otter present in Figure 4.7.

2. They depict a situation where the elements involved are popular news subjects or events. Images depicting President Trump, Queen Elizabeth II and renown athletes are examples of this trend.

3. They depict interesting situations and perspectives that are difficult to photograph without the clearance levels and resources available to professional news photographers. Both the images featuring President Trump and the fire in an African village, present in Figure 4.7, are examples of this. To some extent, images of exotic animals also fit this criteria.

Here, characteristic number one illustrates the need to have a visual quality assessment incorporated into the framework, while numbers two and three illustrate the importance of evaluating social signals and visual concepts. Visual concepts easily allow the differentiation between a *selfie* and a photography of an animal and social signals are useful to distinguish more popular and interesting images from more common ones.

Figure 4.8: False negatives: examples of images the annotators incorrectly assessed as not being extracted from news media.

Conversely, it is also interesting to identify the main characteristics of the images that the majority of the annotators were mislead into labeling as not having news provenance, when in fact they do. Since in only one case all 7 annotators were mislead to annotate an image extracted from news medias as not belonging to news media, the sample of images present in Figure 4.8 also features false negatives annotated by 6 annotators. This set of images has the following characteristics:

1. They are images of seemingly low quality that are needed to illustrate a piece of news and can not be re-shot. In particular, simple snapshots of deceased people are common in this set since news editors resort to images previously taken by the subject's family and friends, to illustrate the piece of news regarding the deceased. An example of this case is the lesftmost picture in Figure 4.8 featuring Rogelio Martinez, a deceased border patrol officer.

2. They are images of low quality featuring common subjects, and are used to illustrate news that are of less importance, or that have no direct visual representation. The rightmost image found in Figure 4.8, depicting a parking ticket, is an example of this.

These characteristics show the need for a ranking method of image selection as opposed to using a binary filtering method only. Specifically, when trying to find images that depict an event for which there are only available a small set of low quality pictures, raking is a much more sound approach than binary filtering. This because filtering could end up discarding images with potentially important and rare information.

4.5.3.2 Evaluating the filtering approach

The classification and regression models were trained using 70% of the *news-quality photos* dataset, while the results presented next regarding the classifier's performance were measured using the remaining 30%. We tested classification models where visual ($GBTC_V$), social ($GBTC_S$) and semantic ($GBTC_C$) features were used separately and combined ($GBTC_F$) to understand the impact of the different feature sets. Table 4.4 shows the results of these tests. The advantage of joining multiple groups of features, to tackle the proposed task, is being able to attain clearly higher precision and accuracy values in

Features	Prec.	Acc.
$GBTC_V$	0.672	0.787
$GBTC_C$	0.555	0.742
$GBTC_S$	0.639	0.834
$GBTC_F$	0.701	0.854

Table 4.4: News-quality assessment results on the filtering task. Models were tested on 30% of the news-quality images dataset.

Features	Prec@30	nDCG@50	MAP
GBT_V	0.833	0.837	0.448
GBT_C	0.833	0.859	0.532
GBT_S	0.733	0.836	0.454
GBT_F	0.967	0.906	0.645

Table 4.5: Results of the performance tests done on the various ranking models.

comparison to the models where only one feature group is used. Consequently, using only the visual quality, semantics or social signals associated with an image as criteria for deciding if it has news-quality, equates to having a worst performance in the task overall.

4.5.3.3 Evaluating the ranking approach

We trained 4 distinct regression models using the *news quality photos* dataset. Again, the first three taking only advantage of visual (GBT_V), social (GBT_S) and semantic (GBT_C) features individually and the forth using all of the three feature sets simultaneously (GBT_F). Then, we resorted to results pooling to perform the evaluation: each model was applied to the *social media photos* dataset and the k better ranked images were extracted. These images were, in turn, annotated by 7 annotators, again according to the question "*Could this image have appeared in the New York Times?*". Finally, images were labeled as news worthy if the majority of the annotators answered yes to the question.

Starting with a numeric interpretation of the results, Table 4.5 shows the precision@30 and nDCG@50 values of the various models tested, while Figure 4.9 presents their precision-recall curve. By analyzing both these metrics we discover that, overall, the models that performed worst were GBT_V and GBT_S. In turn, the model trained only with semantic features, GBT_C, was marginally more successful, specially when retrieving the first half of the relevant images. This shows the importance of semantics in the context of news media. Finally, the complete model (GBT_F) was the one that performed better as it was able to take advantage of the combined strengths of the feature groups used.

Turning to a qualitative interpretation of the results, in Table 4.6 we exemplify this tendency by examining specific examples of images ranked by each model while identifying, in a broad way, the features that influenced the model's choices. The images GBT_V ranked higher (shown on the left side of the table) are of high visual quality, but

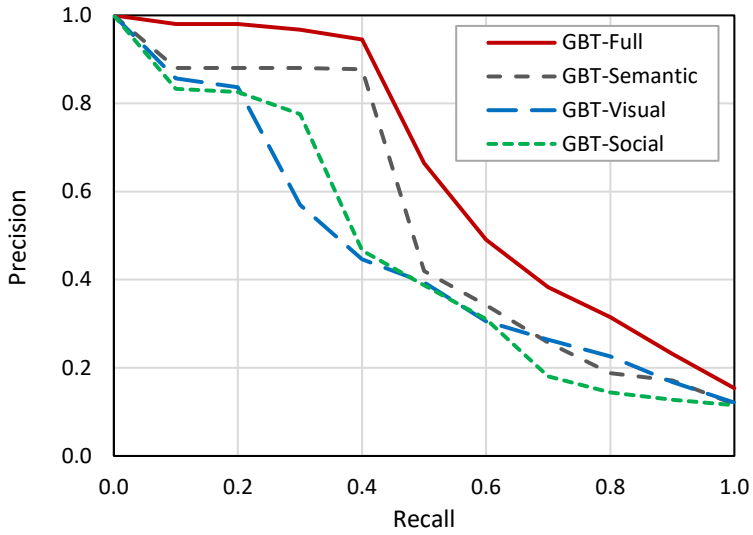


Figure 4.9: Precision recall curves of the various ranking models.

GBT_V	Luminance$^{\uparrow}$, Focus$^{\uparrow}$, Color$^{\uparrow}$	Luminance$^{\uparrow}$, Focus$^{\uparrow}$,	Luminance$^{\downarrow}$, Focus$^{\downarrow}$	Aspect$^{\downarrow}$, Faces$^{\uparrow}$ Focus$^{\downarrow}$ Entropy$^{\downarrow}$
				
GBT_C	Performing Arts$^{\uparrow}$, Event$^{\uparrow}$, Stage$^{\uparrow}$	Event$^{\uparrow}$, Festival$^{\uparrow}$	(No interesting concepts)	Girl$^{\downarrow}$, Selfie$^{\downarrow}$
				
GBT_S	#Duplicates$^{\uparrow}$, #Retweets$^{\uparrow}$	#Duplicates$^{\uparrow}$, #Retweets$^{\downarrow}$	#Retweets$^{\uparrow}$ #Duplicates$^{\downarrow}$	#Duplicates$^{\downarrow}$, #Retweets$^{\downarrow}$
				
GBT_F	Visual$^{\uparrow}$, Social$^{\uparrow}$, Semantic$^{\uparrow}$	Semantic$^{\uparrow}$, Visual$^{\uparrow}$	Social$^{\downarrow}$, Semantic$^{\uparrow}$	Visual$^{\downarrow}$, Social$^{\downarrow}$, Semantic$^{\downarrow}$
				

Table 4.6: Examples of images ranked by four distinct models with increasing ranks from left to right.

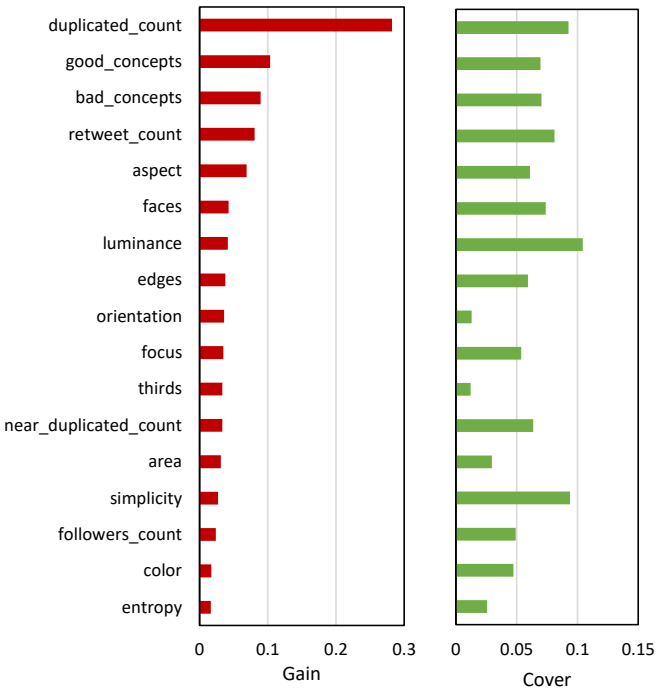


Figure 4.10: The importance of each feature measured through its gain and cover in the Gradient Boosted Trees regression model.

the model is unable to ensure the interestingness of the images selected, the image with the mobile phone being a good example of this problem. GBT_C is able to correctly assert that a photo of a concert is more likely to be used in news media then a selfie. However, the model ends up ranking an extremely blurry image as one of the best in the set, when possibly better suited alternatives were available, like the one displayed to its right. In turn, GBT_S ranks images according to social signals, consequently discarding good images that did not gain social traction. The model ranks correctly images that have a lot of social traction but, when this ceases to be the case, the existing social signals stop being enough to distinguish between images. This tendency is not only observable in Table 4.6 but also in the precision-recall curve, as the model is the worst for recall values higher then 0.6. Lastly, GBT_F leverages the benefits of the other models to correct, to a degree, their individual faults. The GBT_F model is still able to distinguish a *selfie* from a photo of a concert while also being able to assure the visual quality of the better ranked images. Additionally, the model does not focus singularly on social signals meaning that, although these are considered, an unpopular but visual appealing image, semantically tied to news media, is still ranked high by the model.

Finally, Figure 4.10 presents, for each visual, social and semantic feature, its associated gain and cover in the context of the GBT_F model. The higher the gain, the more important a feature is in improving the accuracy of the model. Similarly, cover equates to the amount of coverage of a feature when used in the trees. Here, the gain table shows that, although most visual features have a small gain individually, the model comprised only of visual

features still retains a decent performance due to the high number of different and distinct visual features used. Additionally, we can find visual, social and semantic features in the top 5 features with more gain, confirming that all feature groups increase, by themselves, the performance of the model.

4.6 Conclusions

In this Chapter we detailed the first of the three modules that compose the visual storyline framework proposed in this thesis. The module proposed in this Chapter receives as input social media images, filters them according to news quality standards and provides the remaining images to the module presented in the next Chapter, tasked with finding candidate images to illustrate a particular story.

To do so, we take advantage of visual, social and semantic feature groups. Through our experiments we prove the importance of leveraging these feature groups to tackle the task in a successful manner. Hence, the take away lessons are:

- Social features can be used as proxies to measure the interestingness and quality of an image but the lack of strong social signals does not directly imply the image is not news worthy.
- Semantic features can be used to discard images that are generally not employed in the context of news media, such as *selfies*, while giving priority to topics covered more often in the news. However, semantic features not only do not ensure the visual quality of images but also may not be of great help with images that have rare concepts associated with them, that the model was not able to interact with in the training phase.
- Finally, visual features can be used to ensure the visual quality of an image but are not enough to ensure the interestingness and quality of the information it provides.

Consequently, the machine learning model that performed systematically better during evaluation was the one that leverages simultaneously these three feature groups.

The above results were only possible to achieve in real world social media data because we deployed a thorough visual SPAM and redundancy filtering process. SPAM is a big part of the social media, thus, we combined synthetic image detectors, captioned image filters, near-duplicate removal and other heuristics to clean low quality data. This allows the ranking and filtering methods to work with cleaner data.

CHAPTER 5

STORY ILLUSTRATION CANDIDATES

5.1 Introduction

It is only possible to tell a story through images if the images are semantically relevant to the story they try to illustrate.

Hence, in this Chapter we tackle the problem of finding relevant content to illustrate a particular story. More specifically, we propose a module, as depicted in Figure 5.1, that given a set of social media posts and a story composed of various text segments, outputs relevant candidate images to illustrate each segment. Figure 5.2 illustrates this task.

Approaching this problem as an ad hoc information retrieval one, we make use of text retrieval methodologies augmenting them through multi-model retrieval techniques. Formally, given $Story_N = (u_1, u_2, ..., u_N)$ a story composed of N segments, we define the story segments we intent on illustrating (u_i) as queries and the available social media posts as a set of documents, D. Hence, we wish to find a retrieval method that given

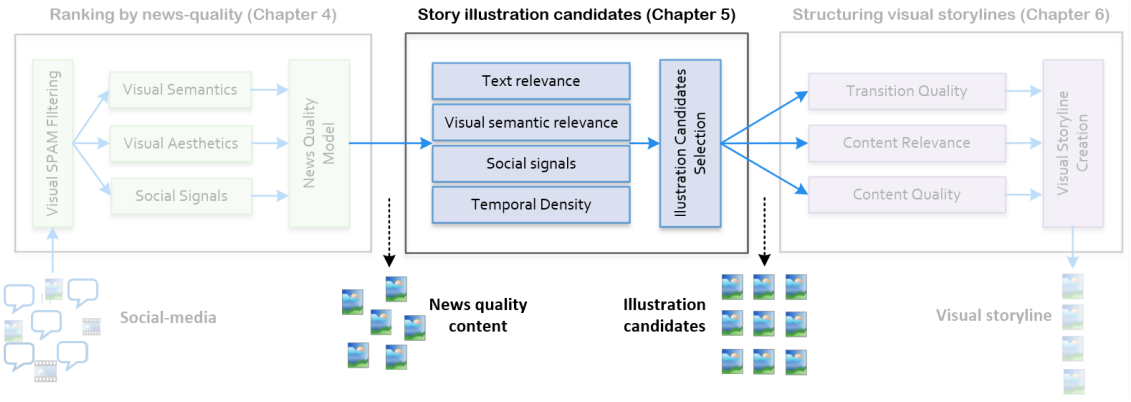


Figure 5.1: Highlight of the second module of the visual storyline generation framework.

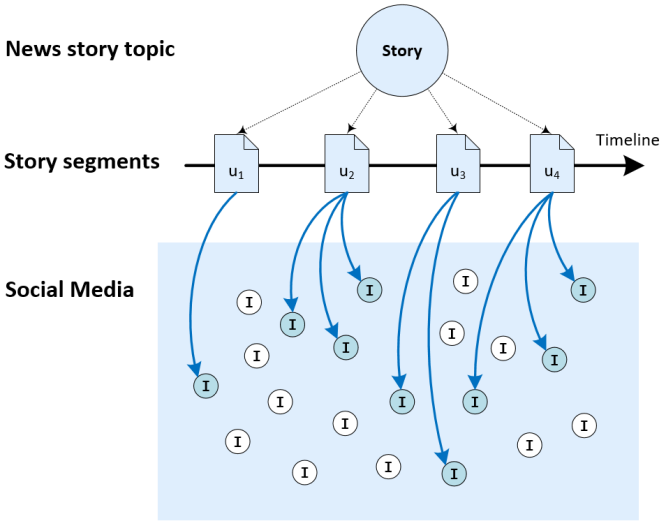


Figure 5.2: Finding relevant candidate images to illustrate each segement of a story.

$(u_1, u_2, ..., u_N)$ and D produces $(C_1, C_2, ..., C_N)$ a list of sets of candidate images where C_i contains relevant candidate images to illustrate u_i.

To do so, we leverage social signals related to the available social media posts, visual concepts of the images present in the publications, as well as analyze publication dates, in order to provide different approaches for selecting candidate images to illustrate story segments.

Finally, in order to evaluate the proposed methods, we ran a human relevance judgment task gaining insight into the advantages and disadvantages of each approach, exploring possible changes that could improve their performance.

5.2 Retrieving relevant content

5.2.1 Text retrieval

Given a set of social media posts containing text and images, and a segment to be illustrated, we want to find the publication with the text that better matches the text of an individual segment. Consequently, we approach the problem from a text retrieval perspective. Making use of a text retrieval engine, we index the publications by their text and then score them through a scoring function according to the story segment.

In this context, we are first tasked with choosing a scoring function. To do so we tested multiple alternatives already well established in literature such as BM25, TF.IDF, Frequency (the score of a document is equal to the number of words in a query found in a document) and Binary (the score of a document is proportional to the number of words in a query found in a document).

We also preprocessed both the text of the segments and the text of the social media posts in several different ways including removing stopwords (very common words that

have no value for finding relevant documents; e.g.: "the", "a", "and"), word lemmatization, word stemming (removing morphological affixes from words, generalizing the retrivel process; e.g.: transforming "generously"into "generous") and transforming words into ngrams. Afterwards tests were performed to understand what combination of these methodologies would be preferred. The tests were focused on sampling the 10 best scoring documents for random segments and calculating overall precision. By analyzing the best performing combination of methodologies we choose BM25 as the ranking function and stop word filtering and stemming as the text preprocessing methods to be used.

Regarding the methodologies chosen, the version of BM25 used is formally defined as

$$score(d,q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,d) \cdot (k_1+1)}{f(q_i,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \tag{5.1}$$

where $f(q_i,d)$ is the frequency of the query term i in document d, $|d|$ is the number of words in document d, and $avgdl$ is the average document length. Additionally, k_1 and b are free parameters, that in this particular case where set to 1.5 and 0.75 respectively by following [8]. Finally $IDF(q_i)$ is the inverse document frequency of the term q_i which is calculated as

$$IDF(q_i) = log(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}) \tag{5.2}$$

where N is the total number of documents, in this case, total number of social media posts available, and $n(q_i)$ is the number of documents in which the query term i is present.

In practice we apply stemming and a stop word filter to both the text of the social media posts and the text describing the segment. Afterwards, BM25 is used to score and rank the publications containing images, according to their textual relevance to the segment. This first part of the approach is also used as a basis for the remaining baselines. As candidates to illustrate the input segment, this approach outputs the images of the top 10 ranked publications. We refer to this as the *BM25* baseline.

5.2.2 Reranking with social signals

Social signals provide a direct measure for popularity as well as a proxy for the quality, informativeness and interestingness of social media content. Aiming to leverage this in order to pick illustration candidates that are more inline with these positive characteristics, we propose two baselines making use of social signals.

First, for both baselines, the social media publications are ranked in the same fashion as for the *BM25* baseline.

Then, in the first baseline, referred to as *#Retweets*, the 20 best ranked documents by *BM25* are reranked by the amount of times they where shared (e.g. "re-tweeted" in the case of Twitter). Alternatively, in the second baseline, referred to as *#Duplicates*, they are reranked by the number of times the image present in the post appears in all the available

social media posts. In both cases, after the rerank, the images of the top 10 ranked posts are chosen as candidates to illustrate the segment

5.2.3 Reranking with visual concepts

Frequently, images present in a social media publications do not directly match their textual content. This means that, by taking only into account the text present in the publications, we may be missing interesting and valuable images that could illustrate a segment well. Hence, we propose two additional baselines with the objective of exploiting the visual concepts associated with the images in each post. In both cases these visual concepts where extracted from the images using VGG-16 [47], a deep Convolutional Neural Network (CNN) specifically designed for large-scale image recognition and pre-trained on the ImageNET Large Scale Visual Recognition Challenge. Also, for both baselines, the posts with images are first ranked in the same manner as for the *BM25* baseline.

In the first baseline, *Concept Pool*, the visual concepts associated with the images of the top 10 ranked posts are considered. These concepts are pooled together. Finally, the 10 images, of the top 20 previously ranked posts, containing the most visual concepts present in the aforementioned pool, are picked as candidates to illustrate the segment.

The second baseline, *Concept Query*, is based on pseudo-relevance feedback. Visual concepts are extracted from the images in the top 5 ranked posts. These concepts are concatenated to form a new query, which is then used to rank all available posts a second time, this time according to the visual concepts present in the posts images (and not according to the posts text, as previously). At this stage, Frequency is used as the ranking function, as we are trying to simple matche two sets of words. Finally, the ranks created by the BM25 and Frequency ranking functions are fused using Reciprocal Rank Fusion, parameterized with $k = 60$. The images from the top 10 posts of the rank that results from this fusion are chosen as candidates to illustrate the segment. An unsupervised rank fusion method, Recriprocal Rank Fusion was chosen due to the lack of training data for this task. It works by attributing and ranking documents by a score calculated by the following expression that leverages the previous ranks:

$$RRFscore(d) = \sum_{i}^{nr} \frac{1}{k + r_i(d))} \tag{5.3}$$

Here, d is a document for which we want a fused score, nr is the number of input ranks (in this case 2) and $r_i(d)$ is the rank of document d in rank i.

5.2.4 Reranking with temporal signals

Some events occur only in a specific moments in time. By finding publications posted in or near those moments we are more likely to retrieve content from those specific events. Taking this insight into account a final baseline was proposed, referred to as *Temp*.

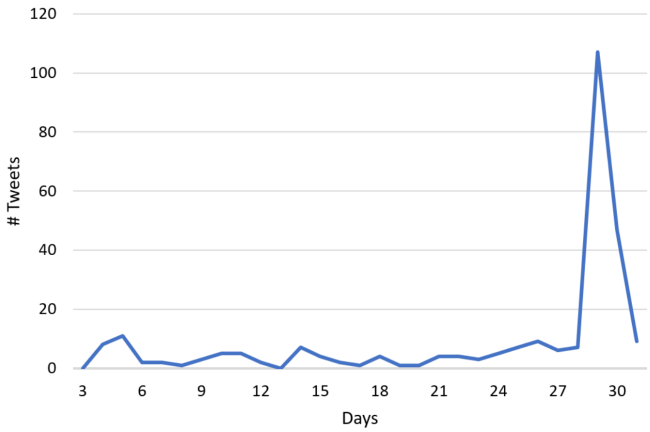


Figure 5.3: Amount of tweets containing the word *fireworks* in the 2016 Edinburgh Festival dataset published per day of the event.

Modelling, that prioritizes content published closer (in time) to peak publication dates related to the segment being illustrated.

To exemplify this reasoning, Figure 5.3 shows the amount of tweets containing the word *fireworks* in the 2016 Edinburgh Festival dataset per publication date. The peak in the number of tweets published during the last days of the event correctly marks the time at which a fireworks show took place. Consequently, when illustrating a story segment such as *fireworks at Edinburgh Festival 2016* one would want to use images from tweets published during that peak, ensuring the relevance of the content to the topic.

For this baseline, posts are first ranked in the same way as for the *BM25* baseline and at this point only posts containing any of the words also present in the segment to illustrate are considered. Considering this new set of posts, we then calculate the number of publications per day, achieving a distribution like the one shown in Figure 5.3. Following this, a Kernel Density Estimator (KDE) with a Gaussian Kernel is applied to the distribution. KDE is defined by:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right) \tag{5.4}$$

With h being the bandwidth, n the total number of data points in the original destribution and K being the kernel, in this case the Gaussian Kernel:

$$K(m) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}m^2) \tag{5.5}$$

In practice the KDE is used to smooth out the original distribution. Hence, after applying the KDE, we achieve a probability distribution of posts, with a text containing words also present in the segment to illustrate, being published in different days. The kernel bandwidth was fixed according to the method defined in [44]. Finally, the top 20 ranked posts are reranked according to the probabilities associated with the dates they where

Baseline	EdFest 2016		TDF 2016	
	Relevance	Transition	Relevance	Transition
Text Retrieval	0.55	0.35	0.58	0.54
#Retweets	0.50	0.39	0.47	0.46
#Duplicates	0.53	0.41	0.52	0.60
Concept Pool	0.51	0.31	0.55	0.53
Concept Query	0.53	0.36	0.46	0.26
Temp. Modeling	0.44	0.23	0.58	0.53

Table 5.1: Performance of the baselines described in the task of illustrating the EdFest 2016 and TDF 2016 stories, measured by the average relevance and transition scores provided by the annotators.

posted in. After this, the images of the top 10 ranked posts are selected as candidates to illustrate the segment.

5.3 Evaluation

5.3.1 Protocol

The goal of this experiment is to evaluate the baselines proposed in the previous Section. To do so, we used the baselines to illustrate the EdFest 2016 and TDF 2016 stories by selecting the best candidate image proposed by each baseline to illustrate each story segment.

Hence, we illustrated a total of 40 storylines (20 for each event). Ground truths for both relevance of illustrations and transition quality were obtained as described in the following section.

5.3.2 Ground truth

In order to evaluate the performance of the proposed baselines, we resorted to crowd sourcing. Three annotators were presented with each story and respective visual storyline, and asked to rate each segment illustration as 1 ("relevant") or 0 ("non-relevant"), as well as rate the transitions between each of the segments as 1 ("good") or 0 ("bad"). Finally, using the subjective assessment of the annotators, the quality metric proposed in Chapter 3, Section 3 was calculated for each story.

5.3.3 Results

Figure 5.4 present the performance of the proposed baselines in the task of illustrating EdFest and TDF storylines evaluated through the quality metric proposed in Chapter 3. In turn, Table 5.1 presents the average relevance and transition scores as provided by the annotators.

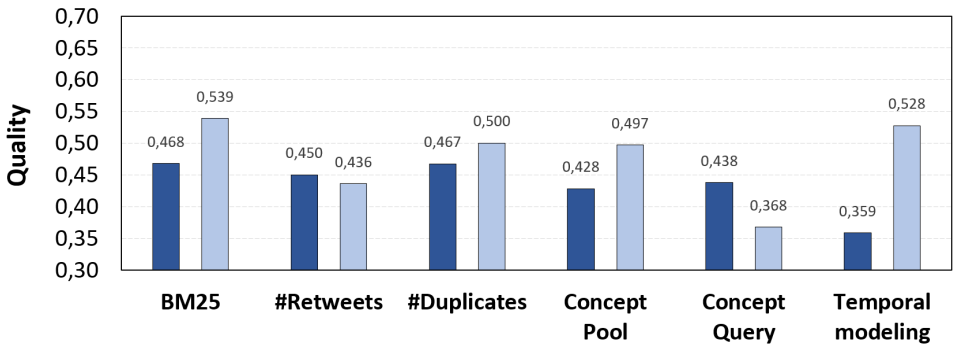


Figure 5.4: Average performance of the baselines in the task of illustrating the EdFest 2016 (dark blue) and TDF 2016 (light blue) stories, according to the annotators, measured by the quality metric proposed in Chapter 3.

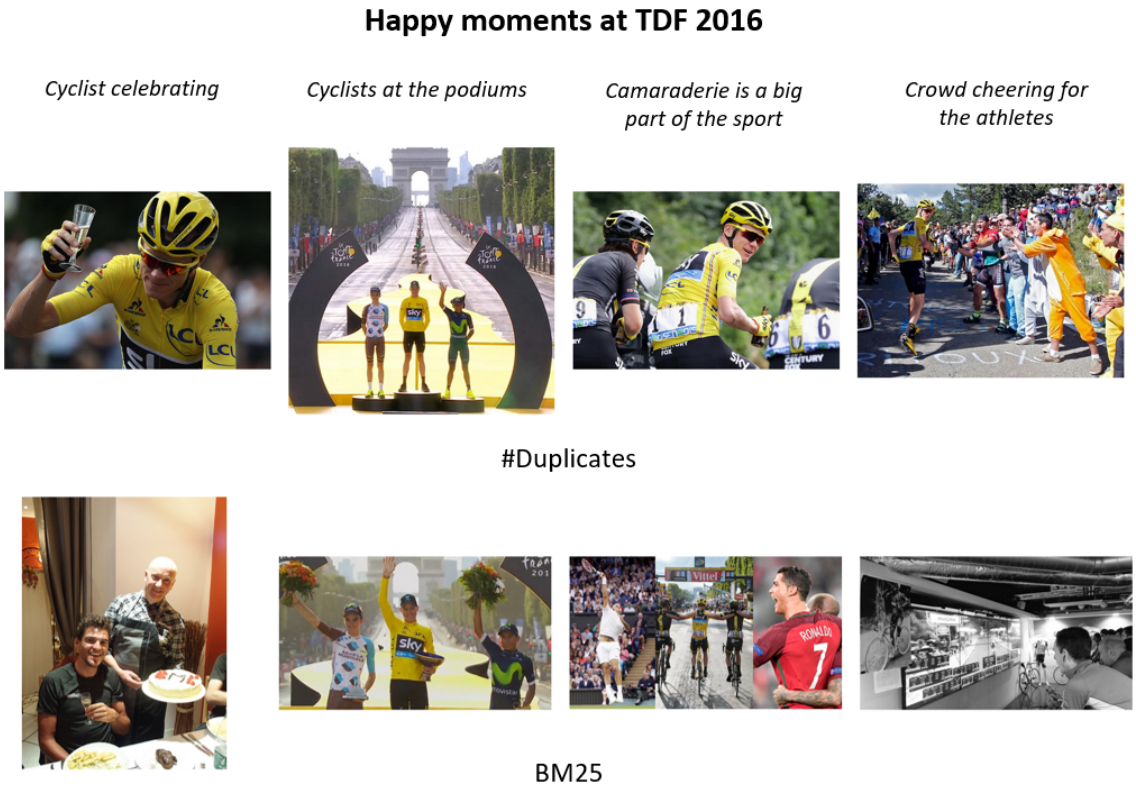


Figure 5.5: Illustrations of the "Happy moments at Tour de France 2016" story achieved by resorting to the *BM25* and *#Duplicates* baselines. Although all images of both storylines were considered relevant by the the annotators to the segments they illustrate, the transitions of the storyline created by the *#Duplicates* baeselines were consistently annotated as having higher quality then those of the storyline created by the *BM25* baseline.

Regarding the relevance metric, the *BM25* baseline performed better in relation to the remaining ones. This shows the importance of regarding the text of the social media publication when choosing the images to illustrate the segments. Analyzing the baselines that leverage social signals, *#Duplicates* was the best performing one, both in terms of relevance and transition quality. Particularly, when inspecting the storylines that result from using this baseline, an increase in aesthetic quality of the images selected for illustration can be noticed. Not only that, the *#Duplicates* baseline was the one that achieved the stories with the best transitions (Figure 5.5). This happens because the number of times an image is published on social media is a good indicator of it's quality. Furthermore, high quality images related to the same event seem to share similarities in terms of their visual and semantic content. However, after analyzing the stories generated by this baseline individually we verified that in scenarios where there are not many images to choose from the approach is hindered by noise. This is specially problematic in cases where there are not strong social signals associated with the few publications available. This problem could be softened by setting a threshold for a minimum number of shares or duplicates needed for the content to be considered for analysis. The threshold can be selected manually, but this means running the risk of filtering out to much content and leaving the segment without illustration candidates, in some cases.

Regarding the baselines that make use of image concepts, it is important to note that the VGG16 model sometimes failed to correctly attribute concepts to the images of both datasets. As an example, for the images of the Tour de France 2016 dataset featuring cyclists, racing concepts such as *"bicycle-built-for-two"*, *"unicycle"*, *"bathing_cap"* and *"ballplayer"* appeared very frequently. However, even though the extracted concepts lack precision, the errors are consistent: VGG16 may identify an unicycle instead of a normal bicycle in an image, but it is consistently doing so in photos where bicycles are found. This means both the proposed baselines are not affected by this issue, since they work by searching images with concepts similar to the ones already deemed relevant by the *BM25* baseline. Overall, both baselines that leverage visual concepts underperformed in situations where text retrieval alone presented good results, while outperforming the other baselines in situations where the text retrieval was not enough to pick relevant content.

Finally, the *Temp. Modeling* approach brought varying results. In story segments that take place over a large portion of time, the approach is particularly flawed as the probabilities attributed to the each of the candidate tweets are marginally the same. However, in story segments with large variations on the number of tweets posted per day the approach performs as expected. Although this approach performed relatively well in the context of the Tour de France 2016 dataset, it performed worst then every other model in the Edinburgh Festival 2016 dataset.

Figures 5.5 and 5.6 present examples of storylines created by each baseline.

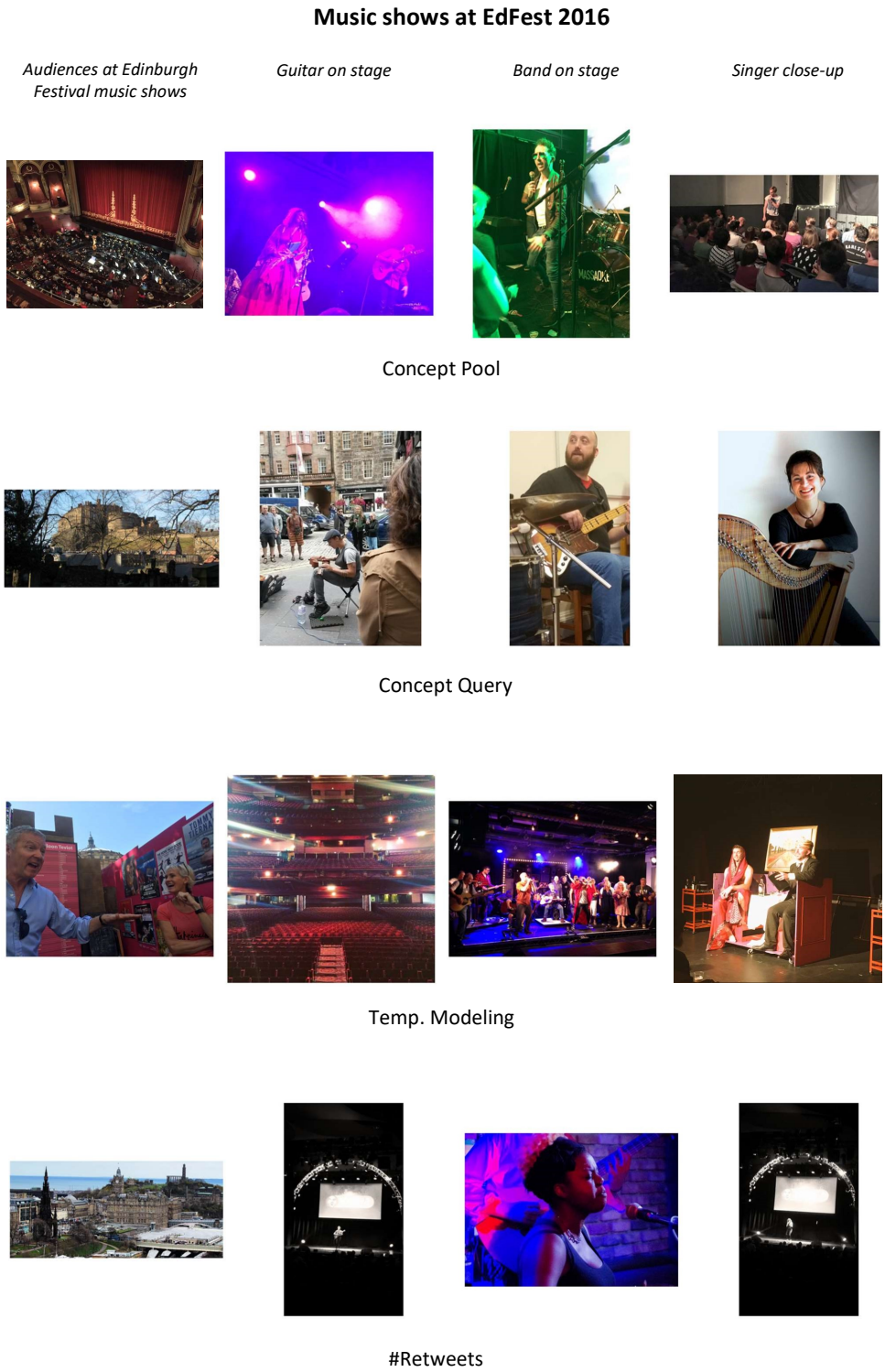


Figure 5.6: Illustrations of the "Music shows at Edinburgh Festival 2016" story achieved by resorting to the *Concept Pool*, *Concept Query*, *Temp. Modeling* and *#Retweets* baselines. From top to bottom, they attained an average score of 0.83, 0.5, 0.25 and 0.17 regarding illustration relevance, respectively, according to the annotators.

5.3.4 Discussion

After analyzing these first storylines it is now possible to provide insight into the difficulty of creating visual storylines for both EdFest and TDF stories using the respective datasets. The images found in the Tour de France 2016 dataset are highly homogeneous both visually and sematically, most of them showing a cyclists pedaling, cyclists on podiums or general closeups of cyclists. Additionally, most appear to have been taken by professional photographers: they are correctly focused, sharp even when the subject is moving, correctly composed and commonly present a depth of field only achievable with high-end photographic equipment. Confirmation of this intuition can be found by analyzing the users who posted the tweets related to the images. As a broad approach we extracted the users that most published tweets related to Tour de France and verified that most of them were attached to highly recognizable news corporations such as The Guardian, BBC or Sky News. The ones that did not fall in this criteria were Twitter accounts solely dedicated to cycling or to the Tour de France itself.

Opposingly, images found in the Edinburgh Festival 2016 dataset vary a lot in visual quality and thematic. They feature several aspects of the event such as fireworks, street performances, theatrical performances among others. Additionally, a lot of them appear to be photographies taken by common festival attendees, using mobile devices or amateur photographic equipment. Again, we verify this tendency by analyzing the accounts that most published tweets related to the event. We found these to be a mix of common users and Edinburgh Festival related accounts.

Hence, overall, inferring visual storylines from the Tour de France 2016 stories is an easier task then doing so for the 2016 Edinburgh Festivals stories. Adding to the fact that it is easier to find quality media related to Tour the France it also appears to be easier to compose visually and semantically cohesive storylines from the available content. The heterogeneous nature of Edinburgh Festival makes it more difficult to illustrate stories in a cohesive manner that entails good transitions between pair of images. Finally, the lower quantity of images present in the Edinburgh Festival 2016 dataset also accentuates this problem as there is less media to choose from when creating storylines. This difficulty trend is confirmed when analyzing the results of the storyline generation methods when applied to both events. As shown in Table 5.1 and Figure 5.4, the scores attained by the baselines on the Tour de France stories are higher overall.

5.4 Conclusions

In this Chapter we propose several approaches designed to retrieve relevant candidate images to illustrate stories. These approaches compose the second module of the storyline generation framework. In the context of the entire framework these candidate images are then provided to the third and final module of the framework, discussed in the next Chapter, designed to create cohesive and appealing visual storylines from these sets of

candidate images.

Having conducted the evaluation of the described retrieval methods through a human relevance judgment task, we concluded that these present individual advantages suited to find candidate images for different types of stories.

- The approach leveraging only text retrieval techniques presents a good overall performance, being outperformed by the remaining approaches only in cases where the publication's text is not enough to determine the relevance of the visual content under scrutiny.
- In situations where the story being illustrated garnered a lot of social traction, the baselines that leverage social signals are good choice for retrieving better quality content.
- In cases where text retrieval is not enough to find relevant content, the approaches leveraging image concepts can be used to find relevant visual content the text retrieval approach would not prioritize.
- Finally, for stories related to events that took place during specific moments in time, the approach based on temporal signals can be applied to ensure the media being used to illustrate the stories was published during or short after the event took place, increasing the chances of it being relevant.

Additionally, by reviewing storylines like the ones present in Figure 5.5, we again confirmed the need for a method of storyline creation that does not take only into account relevance, but also transition quality.

Finally, analyzing the evaluation results, we concluded that the approaches had a good enough performance, although the possibility for improvement is available. These baselines provide only candidate images to each segment, aiming to reduce the computational space of the problem of inferring visual storylines. They do not pick a fixed set of individual images to illustrate a story. Consequently, having some images that are not relevant present in the candidate sets is not a major problem, as the images in each set will be analyzed again during the process of visual storylines creation. As such we leave possible improvements to these baselines for future work.

CHAPTER 6

STRUCTURING VISUAL STORYLINES

6.1 Introduction

A storyline is composed of a set of images organized in a sequence that, together, form a cohesive narrative. As such, tackling the task of creating a storyline means taking into account not only the quality of the individual pieces of content that compose it and their relevance to the story, but also the way they transition from one to the other. Ensuring the quality of these transitions is commonly described as editing and is a process that impacts various forms of content production, from cinema to news media. Consequently, a framework designed for storyline generation must strive to emulate this process in order to provide storylines that are agreeable to their viewers.

Hence, in this Chapter, we present the last module of the storyline generation framework that, as depicted in Figure 6.1, given a set of candidate images to illustrate each segment of a story, is tasked with generating one or more visual storylines optimized for

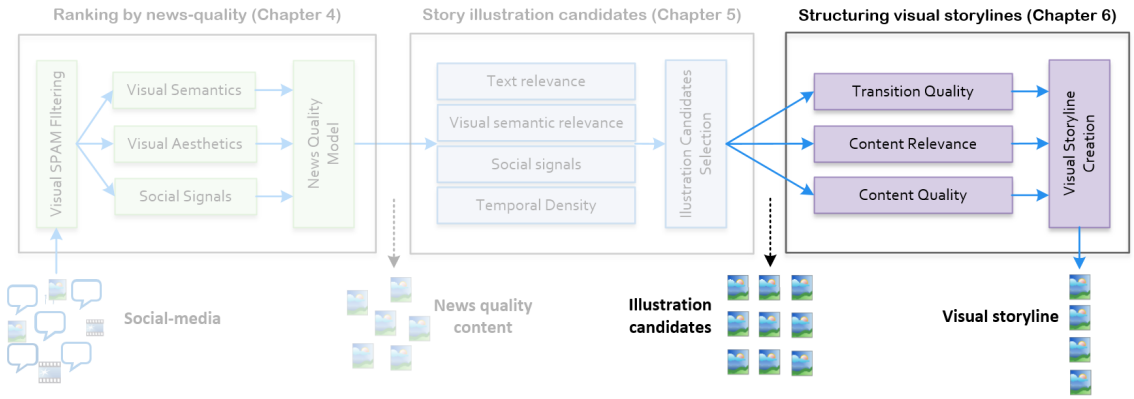


Figure 6.1: Highlight of the third and last module of the visual storyline generation framework.

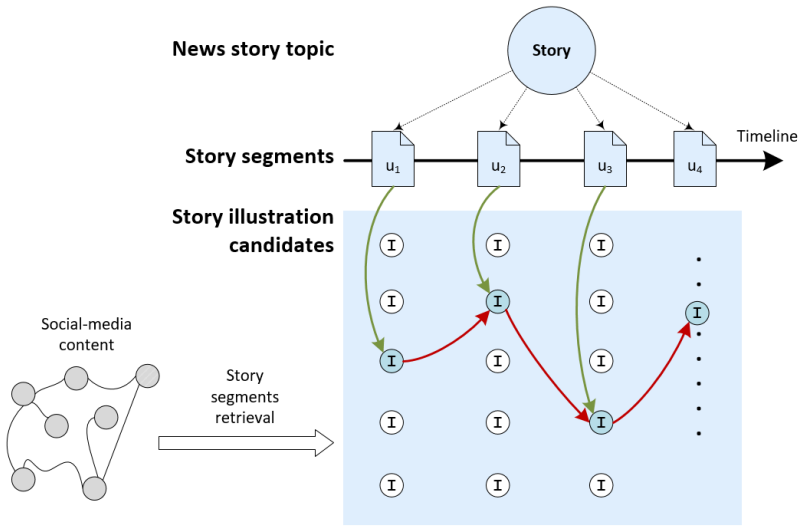


Figure 6.2: Generating visual storylines by taking into account transition quality and the relevance of illustrations to their respective segments. Green arrows represent the need to find relevant content in a pool of candidate images while red arrows represent the need to optimize for transition quality.

both transition quality and relevance. Figure 6.2 further illustrates this challenge.

However, before elaborating on what attributes characterize good and bad transitions and how these can be computed and optimized for, there is the need to define the concept of transition itself, from a computational standpoint. Consequently, in Section 6.3 , we first propose a novel formal definition of transition between two images. Leveraging this definition we are then able to study the impact of a large set of semantic and visual criteria in the quality of image transitions.

Furthermore, because the human perception of transition quality for a particular transition results from various interactions between the images that compose it, there is the need to consider the quality of a transition across multiple dimensions, fusing the various insights that are taken from observing individual transition characteristics. To do so, we resort to machine learning, deploying a model able to predict the perceived transition quality of pairs of images by analyzing their visual and semantic characteristics.

Finally, we tackle the task of generating visual storylines. Formally, we propose a set of graph based methods that, given a story with N segments, $Story_N = (u_1, u_2, ..., u_N)$ and a list of sets of candidate images $(C_1, C_2, ..., C_N)$ to illustrate each segment, produce visual storylines composed of images (w_i), $VisualStoryline_N = (w_1, w_2, ...w_N)$ where $\forall i \in [1, N], w_i \in C_i$

To evaluate these methods we ran a crowd sourcing task where a total of 432 stories were annotated to obtain ground truth.

6.2 Definitions

To tackle the task of understanding transition quality between pairs of images we first need a computationally valid approach to describe the concept of transition. From a non-computational, professional perspective, literature characterizes transitions as the semantic and visual relationships perceived to exist between a pair of images [38]. We emulate this approach, proposing a novel computationally valid formalization of the concept of transition based on distances. More specifically, we define the *transition distance* between two sequential images a_1 and a_2 of a visual storyline as:

$$(distance_p(feature_p(a_1), feature_p(a_2)), \forall p \in P) \tag{6.1}$$

where P is a set of image features under consideration, $feature_p(a)$ is a function that outputs the value of feature p for an image a, and $distance_p(f_1, f_2)$ is a function that outputs the difference between the values f_1 and f_2 of the same feature p for two distinct images. Hence, a transition between two images is formalized as a *transition distance*: a sequence of numeric values representing how distinct two images are in terms of the numeric differences that exist between their features.

Leveraging this definition, we are now able to propose a computational model that takes as input a *transition distance* between two images and qualifies the respective transition.

6.3 Transition quality

Rating a transition between a pair of images, according to its quality, is a non-linear process that results from the interpretation of the features of the individual images and of the manner in which they interact. To tackle the automation of this process, we again resort to the regression version of Gradient Boosted Trees (GBT), defining the problem as one of predicting a quality score, given the *transition distance* of a pair of images. Formally, we propose a function that models ground truth regarding pairwise transition quality:

$$trans(a_1, a_2) \in [0, 1] \tag{6.2}$$

where a_1 and a_2 are images, and the output of the function is a real value scoring the transition between 0 and 1.

To create rich *transitions distances* between the pairs of images analyzed, we propose the use of a large set of visual and semantic features. Through them, we aim to emulate the editing criteria described in literature and elaborated upon in Chapter 2, regarding the importance of maintaining fixed visual and semantic elements when transitioning between different pieces of content. The features and respective distances are presented in Table 6.1 (visual features) and Table 6.2 (semantic features). The next subsections describe these features in detail, while Section 6.5.1 specifies how we aquired the data to train the model and provides detail regarding the training process.

Feature Name (p)	$distance_p(f_1, f_2)$	$feature_p(a)$
Luminance	$abs(f_1 - f_2)$	A positive real value representing the luminance.
Color histogram	$\sum abs(f_1 - f_2)$	A 3D color histogram with 16 bins per RGB channel converted to CIELAB color space.
Color moment	$euclidean(f_1, f_2)$	A vector representing the first color moment of the image in CIELAB color space.
Color correlogram	$\sum abs(f_1 - f_2)$	A 16 bins 3D color correlogram in CIELAB color space.
Entropy	$abs(f_1 - f_2)$	A positve real value representing the entropy of the image.
#Edges	$\sum abs(f_1 - f_2)$	A vector containing the number of horizontal, vertical and diagonal edges.
pHash	$hamming(f_1, f_2)$	A pHash vector.

Table 6.1: Visual features, respective distance functions and descriptions.

Feature Name (p)	$distance_p(f_1, f_2)$	$feature_p(a)$
Concepts	$\#(f_1 \cap f_2)$	A set of image concepts extracted using VGG16.
CNN Dense	$euclidean(f_1, f_2)$	The embeddings extracted from the last layer of the ResNet CNN.
Environment	$f_1 = f_2$	Either "outdoors"or "indoors".
Scene category	$\#(f_1 \cap f_2)$	The location depicted in an image described through labels (e.g.: "bridge", "forest path", "skyscraper", etc.).
Scene attributes	$\#(f_1 \cap f_2)$	The attributes of the location depicted in an image described through labels (e.g.: "man-made", "open area", "natural light", etc.).

Table 6.2: Semantic features, respective distance functions and descriptions.

6.3.1 Visual aesthetics

Visual aesthetics refers to the visual characteristics of the images, such as color signatures and visual entropy. It is expected that images presented in sequence, in a storyline should share some visual traits.

As detailed in Chapter 2, literature underlines the importance of color in transitions, an aspect that was also confirmed in our preliminary experiment detailed in Chapter 3 Section 3.6. We leverage this knowledge by taking into account various color related features each with their unique characteristics. Specifically, we resorted to comparing not only the images *Luminance* values but also their *Color histograms* and their first *Color moment* in the CIELAB color space. CIELAB was chosen because euclidean distances in this color space uniformly match differences in human perception, something that does not occur in the RGB color space [54]. These three features are extracted using the extractor made available[1] in [35]. However, these methods of representing the colors of an image only take into account color distributions in terms of quantity. Nonetheless, two images may have similar quantities of the same colors while presenting these colors in different positions in relation to their boundaries. As such, we make use of an additional feature, the *Color correlogram* as proposed in [19], for its ability to encode both color quantity and its spacial position in the context of the image.

Furthermore, the *Entropy* of each image (also extracted through the aforementioned tool[1]) is used to measure how distinct subsequent images are in terms of the quantity of information they present, while also measuring how simple they are from a human perception point of view. Finally, for its capacity to find similar images, *pHash*[2] is also used.

6.3.2 Semantics

Considering the aesthetic similarity between images is not enough to ensure good transitions. Two images can be very similar in terms of visual aesthetics, but completely different semantically. To tackle this problem we propose the use of several semantic-based methods. For the first, the VGG-16 [47] was again used. We had already taken advantage of the this method of extracting visual concepts from images for the *Concept Pool* and *Concept Query* retrieval baselines detailed in Chapter 5, Section 5.2.3. In this case, each image in the dataset is labled with a set of visual concepts. Then the number of shared semantic labels between the images is compared. We refer to this baseline as *Visual concepts*.

The VGG-16 was, however, trained for multi-class annotation, i.e. to associate a single visual concept to each image. For images with multiple concepts such strategy may not be optimal, in that the concept distribution will be skewed towards the most salient concept, missing other important concepts. To overcome this issue, we propose second

[1] https://github.com/pcpmartins/extractor
[2] http://www.phash.org/

approach we refer to as *CNN Dense*, where embedded representations produced by the penultimate layer of the network are extracted instead of individual concepts. Each image is thus embedded in a 2048-dimensional space. Then the images are compared by this differences in their embeddings. We opt for a ResNet-50 [18] CNN instead of VGG-16 as, according to literature, it is more effective and yields a lower dimensional representation (2048-D vs. 4096-D of the VGG-16).

Finally, we take advantage of methodologies designed to analyze specific semantic characteristics of images. First, we look at the difference in the number of *Faces* present in sequential images in a storyline, using this feature as proxy for the type of scene depicted in said images. Through it, we avoid abrupt transitions from portraits to street scenes to landscapes. Finally, we leverage the environment depicted in the images by resorting to the *Environment*, *Scene Attributes* and *Scene Category* features, extracted through [57], attempting to generate visual storylines where the environments depicted remain consistent in sequential images of the storyline, when possible.

6.4 Story illustration

We now tackle the final task of designing a method for visual storyline creation. This method takes as input sets of candidate images to illustrate each of the segments of a story and is tasked with outputting visually and semantically cohesive storylines composed of the images in said sets. We propose two different graph based approaches to tackle this problem, each with two variants.

6.4.1 Sequence of bipartite graphs - Shortest path

This approach, referred to as *Sequential*, optimizes for storylines with the best possible sequential transitions. In this context, the transition quality is only measured between the candidate images of consecutive segments.

We follow a graph based approach to tackle this problem. Specifically, we define $G = (V, E)$, a *sequence of bipartite weighted directed graphs*, where given a story $Story_N = (u_1, u_2, ..., u_N)$ of N segments, the graph G is constructed as follow:

1. **Vertices:** the graph's vertices V correspond to all the candidate images in the sets $(C_1, C_2, ..., C_N)$, of k images each. Each set $C_i = a_1, ..., a_k$, corresponds to a story segment u_i. Hence, each candidate image a_* of candidate set C_i becomes a vertice in the graph;

2. **Edges:** the graph's edges E, associate all the candidate images from neighboring cadidate sets. In other words, all vertices in set C_i are fully connected and directed to vertices in set C_{i+1}. Hence, the bipartite property of graph;

3. **Edges-weight:** the weight associated with each edge $e \in E$, connecting two vertices v_1 and v_2 is given by a function $pairCost(v_1, v_2)$

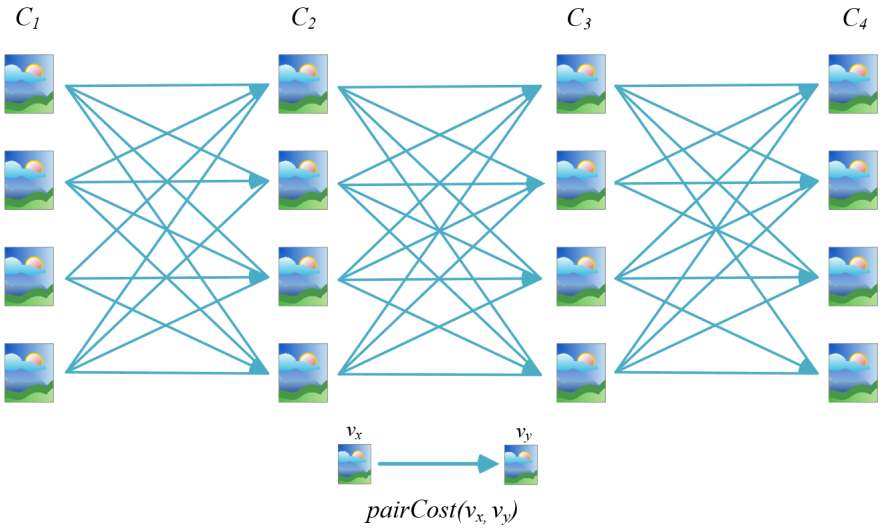


Figure 6.3: Example of a graph for storyline creation using the *Sequential* approach, for a 4 story segment. Images are represented by the vertices of the graph, each vertex belonging to a candidate set C_i. The cost associated with an edge directed from vertices v_x to v_y is given by the $pairCost(v_x, v_y)$ function.

Leveraging this graph structure, exemplified in Figure 6.3 for a 4 segment story, we propose two different methods.

6.4.1.1 Relevance agnostic

The first *Sequential* approach, regarded as *Sequential without relevance* (Seq_T), is optimized for creating storylines with the best possible sequential transitions, regardless of the relevance of the candidate images to the segments they attempt to illustrate. It is designed to present added value in situations where most or all candidate images are already highly relevant to their respective segments or where content relevance is not the highest priority in the context of the story being illustrated.

Formally, this first proposed method computes the shortest path of size N in the graph (i.e. the path will contain a number of vertices equal to the number of segments in the story being illustrated). Hence, we aim to minimize the following expression:

$$\min_{v_1 \in C_1, v_2 \in C_2, \ldots, v_N \in C_N} \sum_{i=1}^{N-1} pairCost(v_i, v_{i+1}) \tag{6.3}$$

where $pairCost(v_x, v_y) = transC(v_x, v_y)$ and the function $transC(v_x, v_y)$ is defined as

$$transC(v_x, v_y) = 1 - trans(v_x, v_y) \tag{6.4}$$

the function $trans$ being defined as described in Section 6.3;

The resulting storyline from this approach is the one composed by the images represented by the vertices that minimize expression 6.3. In practice we resort to a variation of Dijkstra's minimum cost path algorithm to solve this problem.

6.4.1.2 Relevance weighted edges

The previous alternative considers only transition quality when generating visual storylines. For situations where relevant content is scarce we propose a second approach, *Sequential with relevance* (Seq_{TR}). In such situations, some of the available candidate images might not be relevant to the segment they attempt to illustrate. Hence, we leverage both transition quality and relevance of the candidate images, encouraging the creation of storylines with the most relevant candidate images that also present quality transitions.

Thus, for this approach, we again aim to find a path composed of N vertices in the graph. However, this time, the expression to minimize weighs both the importance of the first segment in the storyline being relevant, as well as the importance of transitions *vs.* relevance, to overall storyline quality. We do so by basing ourselves in the quality metric proposed in Chapter 3. In practice, we attain this path by minimizing the following expression:

$$\min_{v_1 \in C_1, v_2 \in C_2, ..., v_N \in C_N} 0.1 \cdot relC(v_1) + 0.9 \cdot \frac{1}{2(N-1)} \cdot \sum_{i=1}^{N-1} pairCost(v_i, v_{i+1}) \tag{6.5}$$

where the function $pairCost(v_x, v_y)$, that ranges from 0 to 2, is defined as:

$$pairCost(v_x, v_y) = \underbrace{0.6 \cdot (relC(v_x) + relC(v_y))}_{\text{segments illustration}} + \underbrace{0.4 \cdot (relC(v_x) \cdot relC(v_y) + transC(v_x, v_y))}_{\text{transition}} \tag{6.6}$$

Here, $relC(v) = 1 - rel(c)$. In turn, $rel(c)$ is the normalized relevance of the image represented by vertice v to the text segment it is candidate to illustrate, as calculated through the BM25 baseline proposed in Chapter 3.

The resulting storyline is the one composed by the images represented by the vertices that minimize expression 6.5.

6.4.2 Multipartite graph - Minimal clique

The *Sequential* approachs aim to produce storylines with high transition quality for sequential pairs of images. However, a visual storyline is consumed as a whole by it's viewers, not as a disconnected set of pairs. Consequently we posit a second approach, *Fully connected*, designed to ensure quality transitions between all elements of the generated visual storylines, leveraging the possibility that individual transition quality is affected by the remaining elements of the storyline they are part of.

Again, we follow a graph based approach to tackle this problem. We define $G = (V, E)$, *a N-partite weighted graph*, where given a story $Story_N = (u_1, u_2, ..., u_N)$ of N segments, the graph G is constructed as follow:

1. **Vertices:** the graph's vertices V correspond to all the candidate images in the sets $(C_1, C_2, ..., C_N)$, of k images each. Each set $C_i = a_1, ..., a_k$, corresponds to a story

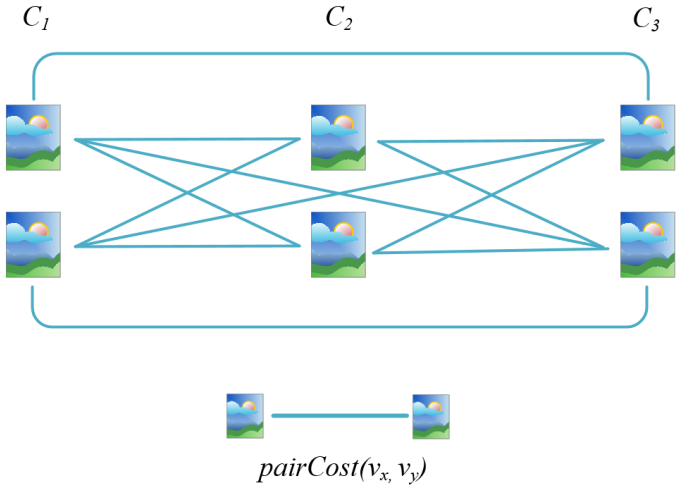


Figure 6.4: Example of 3-partite graph for storyline creation using the *Fully connected* approach, for a 3 story segment. Images are represented by the vertices of the graph, each vertex belonging to a candidate set C_i. The cost associated with an edge connecting vertices v_x to v_y is given by the $pairCost(v_x, v_y)$ function.

segment u_i. Hence, each candidate image a_* of candidate set C_i becomes a vertice in the graph;

2. **Edges:** the graph's edges E, associate all the candidate images from *all other candidate sets*. In other words, all vertices in set C_i are connected to vertices in all the candidate sets except C_i. Hence, the multipartite property of the graph;

3. **Edges-weight:** the weight associated with each edge $e \in E$, connecting two vertices v_1 and v_2 is given by a function $pairCost(v_1, v_2)$

Leveraging this graph construct, exemplified in Figure 6.4 for a 3 segment story, we propose two different methods.

6.4.2.1 Relevance agnostic edges

First, again, we optimize only for transition quality. Hence, for this first *Fully connected* approach, we compute the minimal weighted clique containing N vertices of graph G. (i.e. the clique will contain a number of vertices equal to the number of segments in the story being illustrated). Additionally, we pose the following restriction to the clique: it can only contain one vertex per candidate set. Figure 6.5 provides an example of such a clique.

We attain this clique by minimizing the following expression:

$$\min_{v_1 \in C_1, v_2 \in C_2, \ldots, v_N \in C_N} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} pairCost(v_i, v_k) \tag{6.7}$$

$$pairCost(v_x, v_y) = transC(v_x, v_y) \tag{6.8}$$

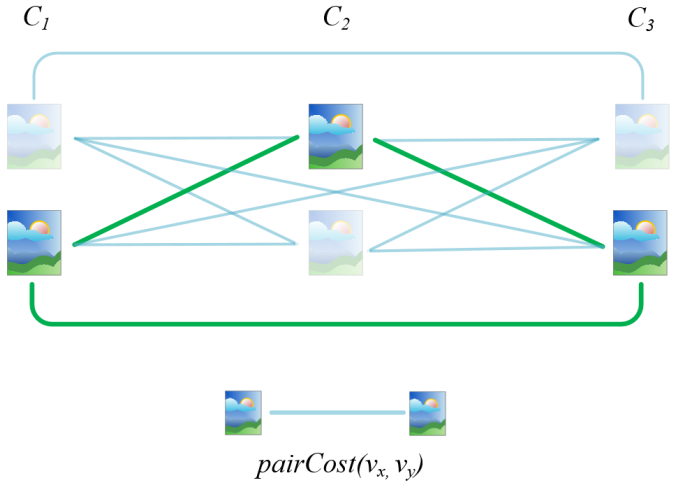


Figure 6.5: Example of a clique containing three vertices, each from a different candidate set, for graph depicted in Figure 6.4. Highlighted vertices (images) and the green edges indicate the parts of the graph that belong to the clique.

The resulting storyline from this approach is the one composed by the images represented by the vertices that minimize expression 6.7. We refer to this approach as *Fully connected without relevance* (Ful_T).

6.4.2.2 Relevance weighted edges

This final alternative builds on the previous one, ensuring high transition quality between all pairs of images in the resulting visual storylines, not just sequential pairs, while optimizing for relevance as with the Seq_{TR} approach.

To do so, we compute a weighted clique containing N vertices of graph G, again with the restriction that it can only contain one vertex per set of candidate images. To find this clique we minimize the following expression.

$$\min_{v_1 \in C_1, v_2 \in C_2, \ldots, v_N \in C_N} 0.1 \cdot relC(v_1) + 0.9 \cdot \frac{1}{N(N-1)} \cdot \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} pairCost(v_i, v_k) \tag{6.9}$$

where the function $pairCost(v_x, v_y)$, a function that ranges from 0 to 2, is again defined as:

$$pairCost(v_x, v_y) = \underbrace{0.6 \cdot (relC(v_x) + relC(v_y))}_{\text{segments illustration}} + \underbrace{0.4 \cdot (relC(v_x) \cdot relC(v_y) + transC(v_x, v_y))}_{\text{transition}} \tag{6.10}$$

The resulting storyline from this approach is the one composed by the images represented by the vertices that minimize equation 6.9. We refer to this final method as *Fully connected with relevance* (Ful_{TR}).

6.5 Evaluation

6.5.1 Crowd sourcing transition quality data

6.5.1.1 Protocol

The goal of this experiment is to crowd source training data to train the machine learning model proposed in Section 6.3. Additionally, we aim to understand which individual features have more impact in transition quality.

In practice, we create storylines by considering the following features individually: *Luminance*, *Color histogram*, *Color moment*, *Entropy*, *#Edges*, *pHash*, *Concepts*, *CNN Dense*, each feature resulting in a different baseline. These features are a subset of those in Tables 6.1 and 6.2.

The protocol is as follows. We first manually selected a total of 1572 relevant images to illustrate the segments of the 40 stories related to the Edfest 2016 and TDF 2016 datasets. This corresponds to an average of 10 relevant image candidates for each story segment. Afterwards, these stories were illustrated using the baselines: for each story, each baseline considers all the 10 relevant images per segment, and chooses the segment's illustration sequence that minimizes the sum of the pairwise *transition distances*, composed only of a single feature, between sequential images. Hence, each baseline focuses on creating storylines where sequential pairs of images have either similar colors, similar semantics, shapes, etc. As a result, 40 distinct storylines were generated by each of the 8 baselines. Thus, in total, 320 storylines were generated.

6.5.1.2 Ground truth

Starting from a story topic and their respective visual storylines (comprising only relevant content), the goal is to assess the quality of the visual storyline as a whole. In practice, the 320 distinct visual storylines were presented to 5 annotators. For each visual storyline, annotators were asked to rate the transitions between each sequential pair of images with a score of 0 (*"bad"*) or 1 (*"good"*).

6.5.1.3 Analysis of crowd sourcing results

Figure 6.6 shows the performance of the proposed transitions baselines at the task of illustrating the 2016 EdFest and TDF stories, using the story quality metric introduced in Chapter 3, calculated based on the judgments of the annotators. Table 6.3 presents the performance of the baselines measured by averaging the sum of the scores given by the annotators to each pairwise transition. As previously noticed, all baselines use the same pool of manually selected relevant visual content.

In this experiment the *CNN Dense* baseline was one of the three best performing baselines, highlighting the importance of taking into account semantics when optimizing the

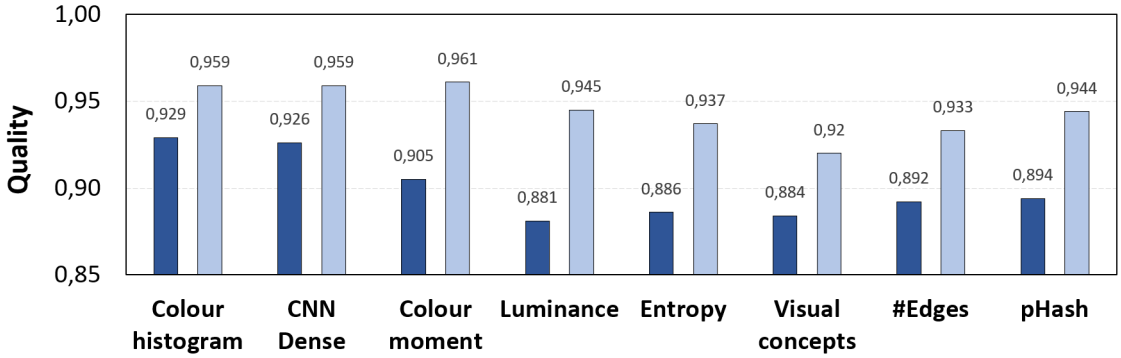


Figure 6.6: Average performance of the baselines described in Section 6.5.1 in the task of illustrating the EdFest 2016 (dark blue) and TDF 2016 (light blue) stories, according to the annotators, measured by the quality metric proposed in Chapter 3.

Baseline	EdFest Transitions	TDF Transitions	Avg.
Color histogram	0.61	0.74	0.68
Color moments	0.53	0.72	0.63
Visual concepts	0.43	0.61	0.52
#Edges	0.41	0.63	0.52
Entropy	0.45	0.68	0.57
CNN Dense	0.58	0.76	0.67
Luminance	0.45	0.67	0.56
pHash	0.42	0.70	0.56

Table 6.3: Performance of the baselines described in Section 6.5.1 in the task of illustrating the EdFest 2016 and TDF 2016 stories, measured by the average transition scores provided by the annotators.

quality of transitions. Semantics may not be enough to evaluate the quality of transitions though. In fact, using single concepts as is the case for the *Visual concepts* baseline provides very poor results, stressing the importance of considering other criteria.

Regarding visual aesthetics, the best performing baselines where the ones that focus on minimizing the color difference between sequential images in a storyline: *Color histograms* and *Color moment*. This supports the assumption that illustrating storylines using content with similar color palettes is a solid way to optimize the quality of visual storylines. Now turning to the *Luminance* and *pHash* baselines, these presented varying results, not always being able to ensure high quality transitions. Regarding luminance, this may be the case because two images can be very distinct while still presenting the same overall lumincance value. Conversely, illustrating storylines by selecting images with similar entropy and number of edges, using the *Entropy* and *#Edges* baselines, provided worst results. This happens because the aesthetic similarities between the sequential images presented in these storylines are, most of the times, not easily perceptible to the naked eye.

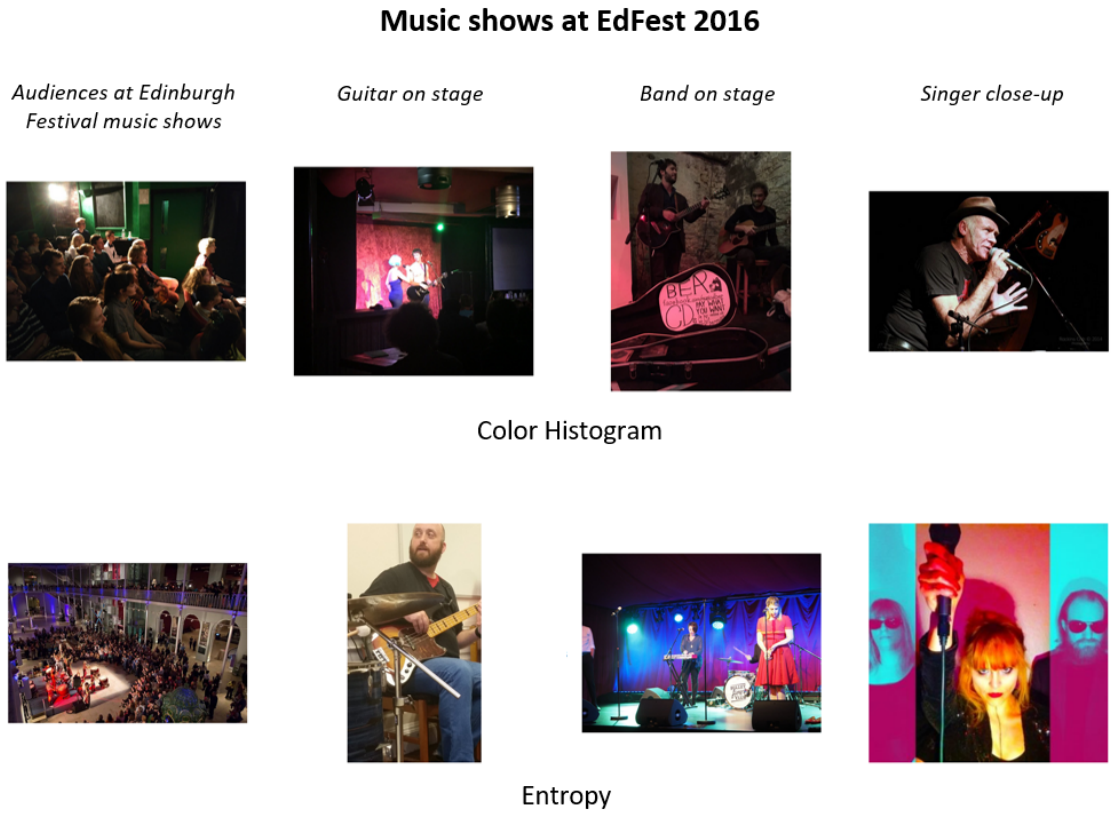


Figure 6.7: Illustrations of the "Music shows at Edinburgh Festival 2016" story achieved by resorting to the *Color histogram* and *Entropy* baselines. The transitions of the storyline created with the *Color histogram* baseline obtained an average score of 1 while the ones in the storyline created by the *Entropy* baseline obtained an average score of 0.6.

Furthermore, by comparing these transitions scores to those of the storylines generated by using the approaches (that optimize only for illustration relevance) proposed in Chapter 5, Section 5.2, we verify a large improvement in terms of transition quality. This proves that the baselines reviewed in this Section are in fact a good first step in the task of generating storylines with high quality transitions.

Finally, and similarly to what was observed in Chapter 5, Figure 6.6 shows that creating storylines with good transitions is easier for the TDF stories than for the EdFest stories. Figures 6.7 and 6.8 show examples of stories illustrated by the the *Color histogram*, *Entropy*, *CCN Dense* and *#Edges* baselines.

6.5.2 Transition quality model

In order to train the Grandient Boosted Trees model to predict the transition quality of a pair of images, we made use of the ground truth that resulted from the crowd sourcing task described in the previous Section. By taking advantage of the annotations made to the 320 storylines (232 composed of 4 segments and the remaining 88 composed of 3 segments) we attained ground truth for a total of 872 pairs of images regarding transition

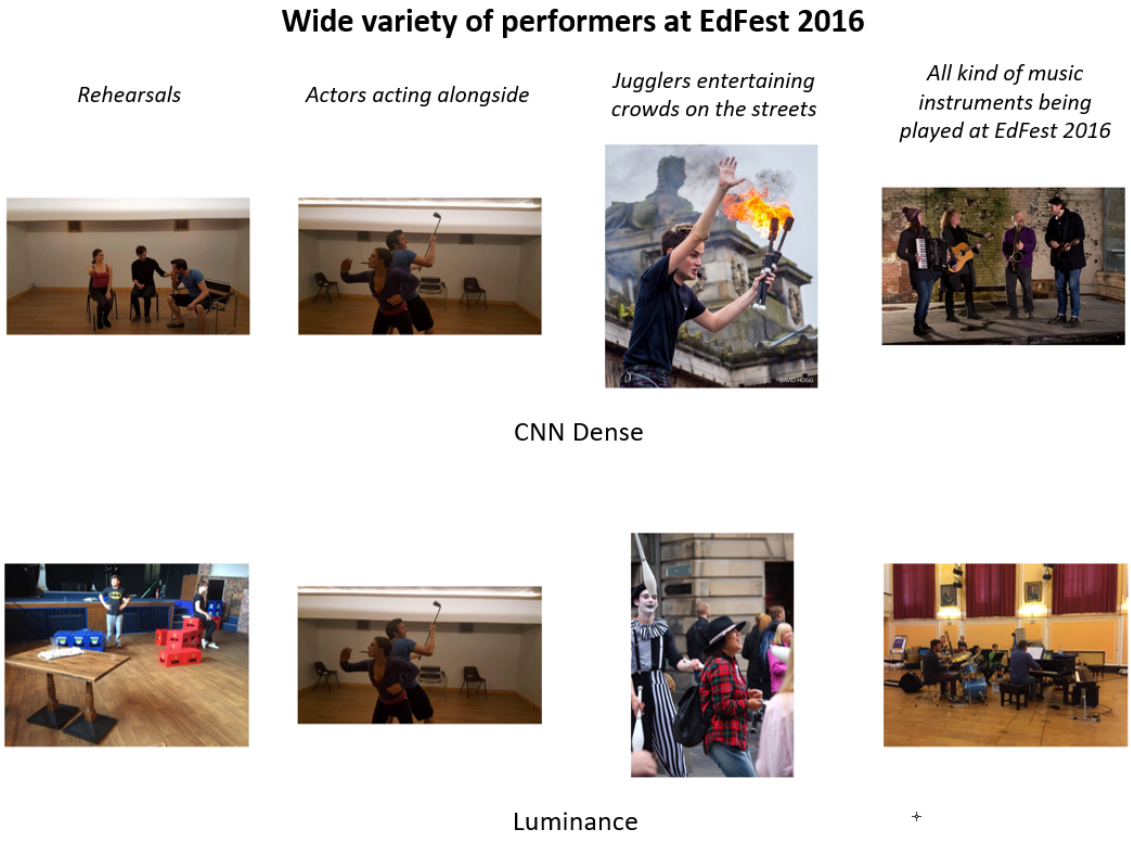


Figure 6.8: Illustrations of the "Wide variety of performers at Edinburgh Festival 2016" story achieved by resorting to the *CNN Dense* and *Luminance* baselines. The transitions of the storyline created with the *CNN Dense* baseline obtained an average score of 0.93 while the ones in the storyline created by the *Luminance* baseline obtained an average score of 0.87.

quality, by averaging the scores (0 or 1) provided by the 5 annotators to each pair.

The *transition distances* for the aforementioned pairs were calculated, using the features defined in Table 6.1 and in Table 6.2. The resulting values where then standardized. Finally, the model was trained to, given a *transition distance* of a pair of images, predict its quality score, according to the ground truth. Dividing the resulting dataset into train and test sets, we first trained the Gradient Boosted Trees model with 70% of the data available, then testing the model using the remaining 30%.

From the tests performed we concluded that the model performs well, presenting a mean average error of 0.245 in the test set.

6.5.3 Creating storylines

6.5.3.1 Protocol

Finally, we test the graph approaches to storyline generation proposed in Section 6.4. To do so, we considered the Edfest 2017 and TDF 2017 datasets, for which 13 and 15 stories are available, respectively. These datasets were yet to be used in any experiment,

Baseline	EdFest 2017			TDF 2017		
	Relevance	Transition	Quality	Relevance	Transition	Quality
Seq_T	0.49	0.72	0.51	0.56	0.81	0.56
Seq_{TR}	0.48	0.71	0.50	0.55	0.78	0.54
Ful_T	0.47	0.77	0.52	0.62	0.91	0.64
Ful_{TR}	0.42	0.61	0.42	0.59	0.72	0.57

Table 6.4: Average performance of the graph based storyline generation methods on the task of illustrating the 2017 Edinburgh Festival and Tour de France stories, measured through the relevance and transition quality scores provided by the annotators, as well as through the quality metric proposed in Chapter 3.

providing a completely new set of images and stories with which to evaluate the storyline generation methods proposed.

The *BM25* baseline described in Chapter 4 was used to select at most 10 candidate image to illustrate each segment of each story. In total, 953 distinct images were retrieved, resulting in an average of 10 candidate images to illustrate each segment. Afterwards the 4 approaches proposed in Section 6.4 where applied resulting in the creation of a total of 112 storylines.

6.5.3.2 Ground truth

We proceeded to assess the quality of each visual storyline related to each story topic. Hence, the 112 distinct visual storylines were presented to 3 annotators. For each visual storyline, the annotators were again asked to rate the relevance of the images to the segment they illustrate and the transition quality between each sequential pair of images with a score of 0 (*"bad"*) or 1 (*"good"*).

6.5.3.3 Results

Table 6.4 shows the performance of the graph based storyline generation methods proposed in Section 6.4 in terms of the average relevance and transition scores given by the annotators, as well as through the quality metric proposed in Chapter 3.

By analyzing them we can verify that, in terms of average transition quality, all approaches performed much better then the simpler baselines presented in Section 6.5.1. Additionally, the *Fully connected without relevance* (Ful_T) approach was the best performing one in terms of average storyline quality, as measured by the quality metric, but also in terms of transition quality. Specifically, the storylines created by this approach were rated as having 91% high quality transitions for the TDF 2017 stories, and 77% high quality transitions for the Edfest 2017 stories. This shows that, in a storyline, the transition quality between a pair of images is not just affected by the images of said pair but also by the remaining images in the storyline. Not only that, but this approach was also the best

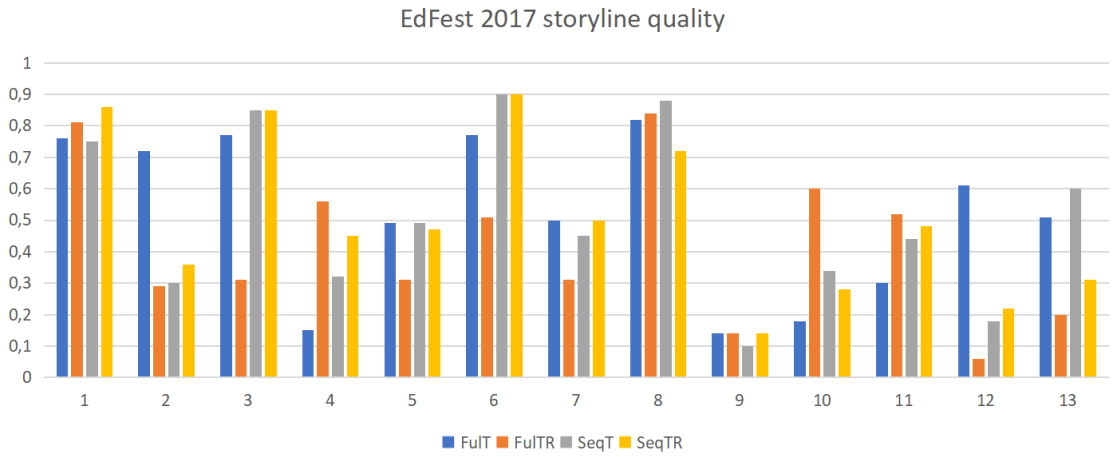


Figure 6.9: The performance of the different graph based approaches at illustrating the EdFest 2017 stories.

performing one according to the quality metric, attaining a score of 0.52 for the EdFest stories and a score of 0.64 for the TDF stories.

Furthermore, by inspecting results we verify that the storylines generated by the approaches that leveraged content relevance, Seq_{TR} and Ful_{TR}, were not scored as containing more relevant images when compared with the storylines generated by the two remaining approaches. By analyzing the storylines individually we see that, the approaches that leverage relevance are, in certain cases, able to pick relevant images to illustrate story segments while the remaining approaches fail to do so. However, in turn, seeking higher quality transitions seems to also improve relevance in particular situations. This happens because, higher transition quality means the images in a storyline tend to present the similar semantics. Because stories are related to a particular topic, semantically similar images to the ones already relevant to a story have a high chance of also being relevant to that same story. Consequently, the approaches that leverage content relevance were not able to outperform the remaining ones in terms of finding relevant content.

Finally, Figures 6.9 and 6.10 present the performance of the graph approaches at illustrating individual stories, according to the quality metric. By analyzing these results we can see that the performance of the approaches is sensitive to the story being illustrated. Although the best average performing approach was Ful_T, in a significant number of cases, the storylines created by the remaining approaches were scored higher by the annotators. This proves the importance of having different methods of storyline creation, as they provide different illustration alternatives that news editors and news editors can work with and build upon. Taking this into account, we can state that in 10 out of 15 TDF stories, at least one of the approaches was able to provide a storyline that was rated with a score near or above 0.7, according to the quality metric. Additionally, in 8 out of 13 EdFest stories, the same can be stated for a score near or above 0.6.

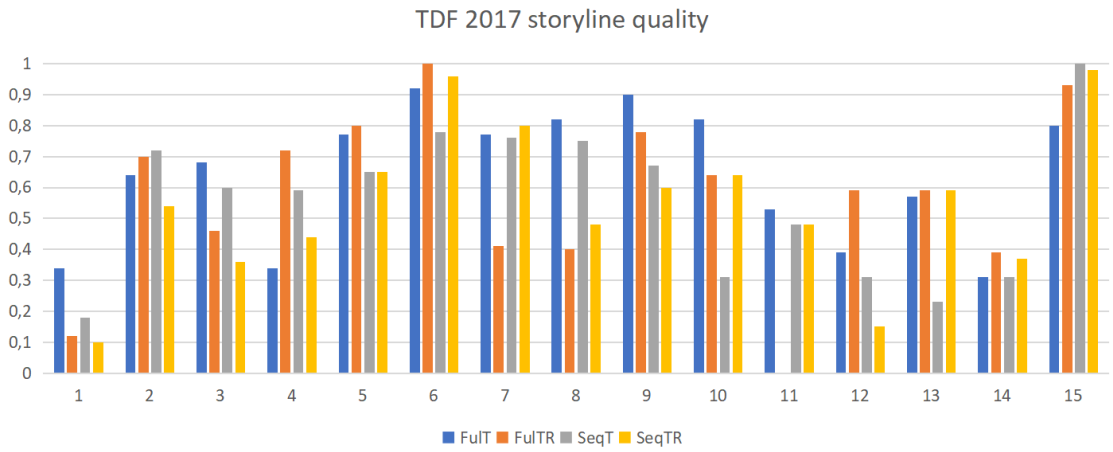


Figure 6.10: The performance of the different graph based approaches at illustrating the TDF 2017 stories.

Presenting a visual example of the performance of the approaches, Figure 6.11 displays the different storylines created by each approach to illustrate the "What is Edinburgh Festival 2017" story. Finally, all storylines created by the graph approaches to illustrate the 2017 EdFest and TDF stories are available for viewing online[3].

6.6 Conclusion

In this Chapter we propose and formalize a novel computational definition of transition between a pair of images. Through it, it is possible to express the visual and semantic relationships present in said pairs, in a non-subjective manner. Consequently, this is a important first step in research regarding visual storyline editing from a computational perspective.

Leveraging this novel definition, we study the impact of semantic and visual characteristics in transition quality and propose and test a method for predicting the quality of transitions using the Gradient Boosted Trees model. This approach presented a good performance at the task, proving that, although transition quality is a subjective topic, it is possible to systematically and accurately predict transition quality in an automated manner. This could only be achieved using the large set of the carfully picked low and high level features with which we trained the model.

Finally, we propose four distinct graph based approaches to visual storyline creation and evaluate them. They comprise the final module of the storyline generation framework depicted in Figure 6.1. These approaches proved to have a high performance at creating high quality visual storylines, while also providing a solid baseline for future research related to this novel task.

[3] http://datasets.novasearch.org/trecvid-storylines/

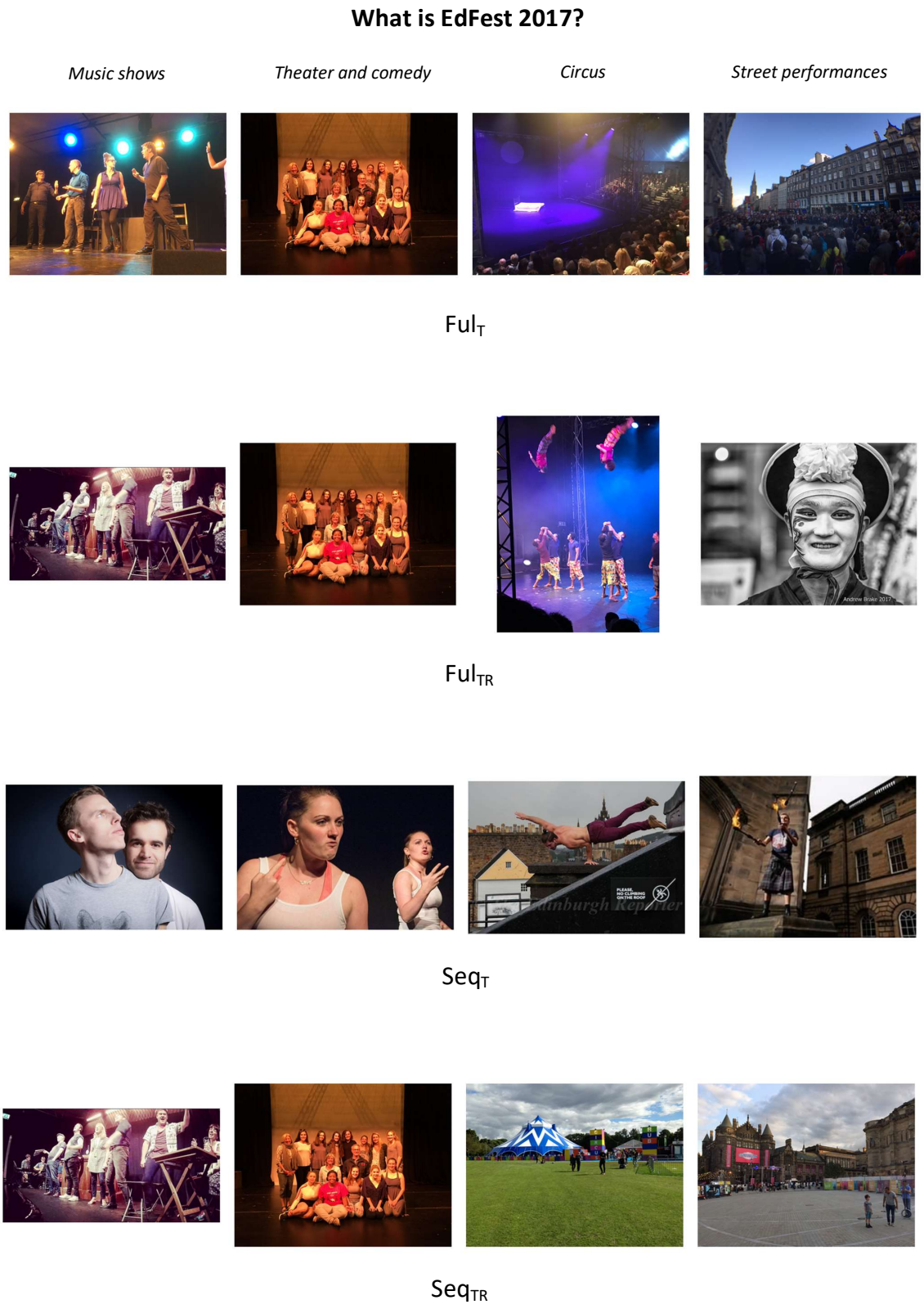


Figure 6.11: Illustrations of the "What is Edinburgh Festival 2017" story achieved by resorting to methods described in Section 6.4. From top to bottom, they attained an average score of 0.76, 0.81, 0.75 and 0.86 regarding the quality metric, respectively.

By inspecting their performance at illustrating stories we were able to gain novel insights into what characteristic impact storyline quality: i) the transition quality between sequential images is affected by the remaining images in a storyline; ii) optimizing for for high quality transitions in the context of a visual storylines positively affects illustration relevance.

Finally, although the four approaches to storyline generation presented different performances during evaluation, in the context of news illustration they can all be taken advantage of. This because they present four alternative storylines with which to illustrate a story that can later be reviewed and polished by a professional journalist or news editor.

Chapter 7

Conclusions and future work

7.1 Conclusion

In this thesis we studied a computational approach to the age-old art of visual storytelling. Aiming to aid news journalists and news editors in the process of news illustration, we proposed a three part framework designed to create news quality visual storylines from a pool of social media content.

Leveraging real world social media data, we tested the framework, proving its ability to select high quality, relevant images, and subsequently organize them, creating semantically and visually cohesive narratives.

From a research standpoint, we tackled a set of problems yet to be solved in literature, including those of news quality assessment, multi-media retrieval for social media content and storyline illustration and editing.

In this context we highlight the following achievements:

1. A novel contribution to the state of the art in news media quality assessment (Chapter 4);
2. A new graph based model for visual storyline creation (Chapter 6);
3. The compilation of a news quality dataset with ground truth (Chapter 4);
4. The creation of a storyline relevance dataset with ground truth (Chapter 5);
5. The creation of a storyline transitions dataset with ground truth (Chapter 6);

In the context of news quality assessment, we resorted to a machine learning based approach, following state of the art literature that tackles the problem of qualifying images according to criteria such as visual aesthetics or memorability. However, we wanted

to not only be able to emulate the content filtering process of news media professionals, but also gain insight into it from a computational point of view. Hence, we resorted to Gradient Boosted Trees, a machine learning model that leverages both precision and interpretability, diverging from the CNN based approach commonly found in literature. Thus, leveraging carefully chosen low and high level social, visual and semantic features, we built an image quality assessment pipeline that proved to have a high performance at the task.

Regarding the task of multimedia retrieval, we approached it from various perspectives, proposing several baselines to tackle the problem. We based our first approach on tried and tested text retrieval methodologies, that are currently an industry standard. Building on it, we incorporated elements of multi-media retrieval, rank fusion methodologies and pseudo relevance feedback. In the end, the proposed baselines proved to have particular advantages and disadvantages, making them useful in different contexts.

Finally, tackling the problem of storyline creation, we started by formalizing, in a non subjective manner, the concept of transition. This novel definition allowed us to work the concept of transition through computational approaches. Leveraging it, we proposed four graph based methods of storyline creation that perform well, as shown in the results of our experimental evaluation. These where created and implemented to take full advantage of existing graph algorithms, ensuring their correctness and high computational performance. They leverage a strong machine learning predictor which was trained to predict transition quality based on both the semantic and visual features present in the pair of images under scrutiny.

7.2 Impact in the newsroom

In the newsroom, the storylines created by the proposed framework can then be reviewed and build upon by news media professionals. This approach aims to expedite the news illustration process that, originally, had news media professionals manually searching social media for high quality relevant content. In practice, the modules that compose the framework can also be used individually, allowing news journalists to easily filter content according to its quality, find relevant content to illustrate news stories or construct storylines from hand picked content.

As such, this work is valuable in a real world, practical context, introducing a novel set of semi-automated methods designed to aid news journalists and news editors perform their everyday tasks.

7.3 Future work

Despite our positive results, the framework proposed in this thesis still has room for improvement. The following list comprises possible approaches to future work in the context of the framework that was developed.

- In situations where no relevant text is associated with a relevant image to illustrate a story segment, the baselines proposed in Chapter 4 will never identify that image as a proper way of illustrating the segment. As an approach to solving this problem, a new baseline could be created, where a search engine (e.g.: Google) is first queried for images related to the terms in a segment. The best ranking images are then retrieved and images similar to these are searched in the available dataset.
- After having attained relevance ground truth for the datasets that compose the experimental framework defined in Chapter 3, it is now possible to fine tune the various parameters associated with the retrieval baselines proposed in Chapter 5. This process might result in improvements in the performance of the baselines.
- Again, making use of the now available relevance ground truth, it is also possible to resort to LETOR, a machine learning approach to rank fusion. Through it, the baselines in Chapter 5 could be combined to achieve a single baseline that presents a better overall performance.
- After having attained ground truth for illustration relevance and transition quality for a large set of storylines, it is now possible to train a machine learning model to create quality visual storylines. This model would leverage the predicted relevance and transition quality values associated with images to predict which images should illustrate which segments. This approach was not tested in the context of this thesis and has the potential to present a good performance.

7.4 Research opportunities

This work is a first approach to a set of problems that where yet to be tackled by literature. Consequently, besides providing a solid baseline for future works on related subjects, it also poses many new and interesting challenges that may be pursued by future research. What follows is a comprehensive description of some of these questions and challenges that we find to be more interesting and of higher importance to pursue in the near future.

- All methods developed in the context of this thesis target still images only. However, visual storylines could contain video content. Hence, there is interest in exploring what changes could be made to the propose storyline generation framework in order to accommodate for both images and videos.
- Leveraging text summarization and event detection methodologies already established in literature, the proposed framework could be expanded to one that automatically detects events covered in social media, by inspecting social media posts, and creates visual storylines narrating those events.
- It is possible that additional dimensions can be added to the storyline quality metric proposed in Chapter 3 to improve its performance at mimicking human perception,

as currently the metric considers only transition quality and illustration relevance. Studying this may help to better understand the human perception of visual storyline quality.

- In chapter 4 we propose a framework designed to filter and rank social media images according to news media standards. The concept of *news media standards* can, however, be further dissected, as different news media outlets present different quality standards. Studying how these standards differ may present a particular interesting window of opportunity to help understand news media content from a computational point of view.

Bibliography

[1] A. Adams. *The Camera, The Ansel Adams Photography Series*. Little, Brown and Company, 1980.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. "Finding high-quality content in social media." In: *Proceedings of the 2008 international conference on web search and data mining*. ACM. 2008.

[3] I. Arapakis, F. Peleja, B. B. Cambazoglu, and J. Magalhaes. "Linguistic Benchmarks of Online News Article Quality." In: *ACL*. 2016.

[4] J. Bian, Y. Yang, and T.-S. Chua. "Multimedia summarization for trending topics in microblogs." In: *Proceedings of the 22nd ACM international conference on Conference on information knowledge management*. ACM. 2013.

[5] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua. "Multimedia summarization for social events in microblog stream." In: *IEEE Transactions on multimedia* (2015).

[6] A. Ceroni, C. Ma, and R. Ewerth. "Mining Exoticism from Visual Content with Fusion-based Deep Neural Networks." In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM. 2018.

[7] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system." In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016.

[8] D. M. Christopher, R. Prabhakar, and S. Hinrich. *Introduction to information retrieval*. Cambridge University Press, 2008.

[9] D. Delgado, J. Magalhaes, and N. Correia. "Assisted news reading with automated illustration." In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010.

[10] S. Dhar, V. Ordonez, and T. L. Berg. "High level describable attributes for predicting aesthetics and interestingness." In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2011.

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

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: AAAI Press, 1996.

[13] M. Freeman et al. *The Photographer's Eye: Composition and Design for Better Digital Photos*. CRC Press, 2007.

[14] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics, 2001.

[15] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox. "A semi-automatic approach to home video editing." In: *Proceedings of the 13th annual ACM symposium on User interface software and technology*. ACM. 2000.

[16] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. "Predictors of answer quality in online Q&A sites." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2008.

[17] D. Hasler and S. E. Suesstrunk. "Measuring Colourfulness in Natural Images." In: *Proceedings of SPIE - The International Society for Optical Engineering* (2003).

[18] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[19] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. "Image Indexing Using Color Correlograms." In: *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*. IEEE Computer Society, 1997.

[20] P. Isola, J. Xiao, A. Torralba, and A. Oliva. "What Makes an Image Memorable?" In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2011.

[21] P. Isola, D. Parikh, A. Torralba, and A. Oliva. "Understanding the intrinsic memorability of images." In: *Advances in Neural Information Processing Systems*. 2011.

[22] T. Joachims. "Training linear SVMs in linear time." In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006.

[23] A. Jović, K. Brkić, and N. Bogunović. "A review of feature selection methods with applications." In: *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE. 2015.

[24] Y. Ke, X. Tang, and F. Jing. "The design of high-level features for photo quality assessment." In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2006.

[25] K. Kobre and B. Brill. *Photojournalism: the professionals' approach*. Vol. 5. Focal Press Burlington, MA, 2004.

[26] N. Krawetz. *Looks Like It*. http://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html. Accessed: 2017-03-23.

[27] H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010.

[28] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong. "Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

[29] R Lienhart and A Hartmann. "Classifying images on the web automatically." In: *Journal of Electronic Imaging* (2002).

[30] C. Lino, M. Chollet, M. Christie, and R. Ronfard. "Computational model of film editing for interactive storytelling." In: *International Conference on Interactive Digital Storytelling*. Springer. 2011.

[31] M. Mancas and O. Le Meur. "Memorability of natural scenes: The role of attention." In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE. 2013.

[32] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. "Assessing the aesthetic quality of photographs using generic image descriptors." In: *2011 IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2011.

[33] A. Marcucci. "UX design for fluid visual storytelling on the web." Master's thesis. Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[34] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. "Twitinfo: aggregating and visualizing microblogs for event exploration." In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2011.

[35] P. Martins and N. Correia. "Semi-automatic Video Assessment System." In: *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM. 2017.

[36] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. "Building a large-scale corpus for evaluating event detection on twitter." In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM. 2013.

[37] P. J. Mcparlane, A. J. Mcminn, and J. M. Jose. ""Picture the scene ...";: Visually Summarising Social Media Events Categories and Subject Descriptors." In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (2014).

[38] W. Murch. *In the blink of an eye: A perspective on film editing*. Silman-James Press, 2001.

[39] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[40] J. Nichols, J. Mahmud, and C. Drews. "Summarizing sporting events using twitter." In: *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM Press, 2012.

[41] R. Pinto. "Summarization of Social-Media Content about Real-World Events." Master's thesis. Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[42] R. Rosenblum and R. Karen. *When the shooting stops... the cutting begins: A film editor's story*. Ingram publisher services US, 1986.

[43] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, and P. A. Mitkas. "Visual event summarization on social media using topic modelling and graph-based ranking algorithms." In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM. 2015.

[44] D. W. Scott. "On optimal and data-based histograms." In: *Biometrika* (1979).

[45] *Scrovegni Chapel*. http://www.dailyartmagazine.com/things-must-know-scrovegni-chapel/.

[46] B. Sharifi, M.-A. Hutton, and J. K. Kalita. "Experiments in microblog summarization." In: *2010 IEEE Second International Conference on Social Computing (SocialCom)*. IEEE. 2010.

[47] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *arXiv* (2014).

[48] J. R. Smith, D. Joshi, B. Huet, W. Hsu, and J. Cota. "Harnessing AI for Augmenting Creativity: Application to Movie Trailer Creation." In: *Proceedings of the 2017 ACM on Multimedia Conference*. ACM. 2017.

[49] Z Tang, Y Dai, and X Zhang. "Perceptual hashing for color images using invariant moments." In: *Appl. Math* (2012).

[50] *The Rules of Photography*. https://akhilphotographyblog.wordpress.com/2012/09/25/the-rules-of-photography.

[51] *Tombs of Ancient Egypt*. https://www.osirisnet.net/mastabas/kagemni/e_kagemni_02.htm.

[52] E. M. Voorhees. "The philosophy of information retrieval evaluation." In: *Workshop of the cross-language evaluation forum for european languages*. Springer. 2001.

[53] F. Wang and M.-y. Kan. "NPIC : Hierarchical Synthetic Image Classification Using Image Search and Generic Features." In: *CIVR* (2006).

[54] G. Wyszecki and W. S. Stiles. *Color science*. Wiley New York, 1982.

[55] X. Xu, N. Yuruk, Z. Feng, and T. A. Schweiger. "Scan: a structural clustering algorithm for networks." In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2007.

[56] J. Yang, J. Luo, J. Yu, and T. S. Huang. "Photo stream alignment and summarization for collaborative photo collection and sharing." In: *IEEE Transactions on Multimedia* (2012).

[57] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. "Places: An image database for deep scene understanding." In: *arXiv* (2016).