UNIVERSITÀ DEGLI STUDI DI CATANIA
Dipartimento di Matematica e Informatica
Dottorato di Ricerca in Matematica e Informatica XXXI Ciclo

Alessandro Ortis

Methods for Sentiment Analysis and Social Media Popularity of Crowdsourced Visual Contents

Tesi di Dottorato di Ricerca

Sebastiano Battiato

Anno Accademico 2017 - 2018

"*Traveller, there are no paths. Paths are made by walking.*"

Antonio Machado

Abstract

This thesis collects all the research work done by the PhD candidate in the Joint Open Lab for Wireless Applications in multi-deVice Ecosystems (JOL WAVE Catania) of TIM Telecom Italia, which granted his doctoral fellowship.

The crowdsourcing paradigm opens new opportunities to understand various aspects of people's interactions, preferences and behaviors. In this thesis we investigated methods aimed to infer people's reaction toward visual contents, under different headings. We first focus on the task of understanding how people visit a place (e.g., a cultural heritage site) and infer what catch most their attention and interest, by means of the analysis of the shared photos taken by the users themselves. Then, addressing the issues related to the noisy text associated to images, we defined a method for Image Popularity Prediction, considering an alternative source of text automatically extracted from the visual content. We first highlight the drawbacks of the text used in most of the state of the art methods, and then experimentally compared the two sources of text. Starting from the analysis of the state of the art in image popularity prediction, we observed that a time-aware approach is needed, as the temporal normalization commonly employed in literature makes two contents published at different times incomparable. For this reason we introduced a new task, named Image Popularity Dynamics Prediction, which aims to predict the evolution of the engagement scores of a photo over a period of 30 days from its upload. To challenge the problem, we introduce a large scale dataset of ∼20K photos whose engagement scores have been tracked for 30 days. Moreover, we presented an approach that is able to perform the prediction at time zero. Furthermore, we investigated methods for scene popularity estimation, from a set of videos taken by people attending a public event. This involved the definition of methods for unsupervised video segmentation and scene clustering, able to work in both mobile and wearable domains. The methods have been developed considering unconstrained scenarios without any prior on the input videos. In appendix, we also report some additional results and the pseudocode of the developed algorithms.

Acknowledgements

I would like to express my gratitude to my advisor, Prof. Sebastiano Battiato, for guiding and supporting me over the years. Prof. Battiato has been supportive and has given me the ability to pursue the work autonomously and improve my skills as researcher, by encouraging professionalism and independence.

I would like to thank Prof. Giovanni Maria Farinella for his scientific advices, the insightful discussions about research topics and the continuous support and contribute to my work during these past three years. Giovanni is someone you will never forget once you meet him. I hope that I could be as enthusiastic and energetic as Giovanni and to be able to command an audience as well as he can.

Thanks to the colleagues of the Joint Open Lab WAVE. A special thank goes to Valeria D'Amico, Giovanni Torrisi and Luca Addesso, who closely followed my work. Their guidance helped me in understanding how research should be performed in the context of a big enterprise like TIM Telecom Italia, which sponsored this doctoral fellowship.

In these years I've had the privilege of attending several editions of ICVSS (International Computer Vision Summer School) and two editions of MISS (Medical Imaging Summer School), prestigious scientific schools, directed by Giovanni and Prof. Battiato. This allowed me to learn from the best experts of this fascinating scientific field from all over the world. These schools have been a continuous stimulus and a great resource of knowledge from which I tried to learn as much as possible. For this opportunity I also thank professors Cipolla (Cambridge University), Stanco (University of Catania) and Schnabel (King's College, London).

I'd like to thank also Prof. Catarina Sismeiro, who invited me as *Visiting Researcher* at the prestigious Imperial College, London, UK. She provided me an opportunity to join her team for 3 months as visiting scholar, as well as to know a city that I'll always love.

Thanks to Francesca, a fundamental presence in my life. Thanks for your patience, for understanding the importance that the research has to me, and the sacrifices that this involved for both of us. Special thanks to my parents, two extraordinary people, for their sacrifices and teachings that helped me to achieve many goals, and for having always supported and guided me in every choice of my life. Thanks to Francesco and Mirella, who always believed in me.

Contents

Chapter 1

Introduction

In 2012 Telecom Italia, one of the major telecommunication company in Italy, created the Joint Open Labs (JOLs), aimed to promote and take advantage from the of the Open Innovation paradigm [1]. Indeed, the JOLs are placed within specific Italian university campuses. In the Open Innovation paradigm, companies and universities research groups collaborate in an innovation process which combines the experiences and high level specific skills of industry and academic research. In such a mutual contamination environment, new assets, products, services ideas are developed employing the most advanced technologies, reducing time and cost of the research and development process (fast prototyping). This dissertation collects all the research work done by the PhD candidate in the Joint Open Lab for Wireless Applications in multi-deVice Ecosystems (JOL WAVE) of TIM Telecom Italia, which is located within the University of Catania campus and sponsored this doctoral fellowship. In this laboratory, novel service applications based on connected smart devices (e.g., smartphones, tablets, cameras, wearable devices, sensors) are designed and developed. The increasing diffusion of mobile and wearable devices equipped with interconnected sensors allows the development of First Person View applications which take into account the user's point of view as a source of information. Future networks will handle high definition multimedia contents such as real-time multi-source video streaming applications. In such a scenario, the Long Term Evolution (LTE-4G) [2] represents a valuable asset for the growing request of multimedia services that requires high performance mobile communication technologies. In this dissertation real use-cases multimedia services are defined and presented.

Nowadays, the amount of public available information encourages the study and development of algorithms that analyse huge amount of users' data with the aim

to infer reactions about topics, opinions, trends and to understand the mood of the users whose produce and share information through the web. Sentiment Analysis is the research field aimed to extract the attitude of people toward a topic or the intended emotional affect the author wishes to have on the readers. The tasks of this research field are challenging as well as very useful in practice. Sentiment analysis finds several practical applications, since opinions influence many human decisions either in business and social activities. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviours. Although NLP (Natural Language Processing) offers several approaches to address the problem of understanding users' preferences and behaviours, the social media context offers some additional challenges. Beside the huge amounts of available data, typically the textual communications on social networks consist of short and colloquial messages. Moreover, people tend to use also images and videos, in addition to the textual messages, to express their experiences through the most common social platforms. The information contained in such visual contents are not only related to semantic information such as objects or actions about the acquired picture, but also cues about affect and sentiment conveyed by the depicted scene. Such information is hence useful to understand the emotional impact (i.e., the evoked sentiment) beyond the semantic. For these reasons images and videos have become one of the most popular media by which people express their emotions and share their experiences in the social networks, which have assumed a crucial role in collecting data about people's opinions and feelings. Images and videos produced by users and shared in social media platforms reflect visual aspects of users' daily activities and interests. Such growing user generated images represent a recent and powerful source of information useful to analyse users' interests. In this context, uploading images and videos to a social media platform is the new way by which people share their opinions and experiences. This provides a strong motivation for research on this field, and offers many challenging research problems. In this dissertation we present several scientific works that analyses images and videos produced by users with a common social context (i.e., a social platform, a social event or a public site), with the aim to infer user's interests and behaviours. The basic task in Visual Sentiment Analysis is the prediction of the sentiment evoked by a visual content (i.e., images

and videos) in terms of sentiment polarity (i.e., positive, negative or neutral) or by using a set of emotion classes (i.e., angry, joy, sad, etc.). However, in this dissertation we further extend the task of Sentiment Analysis applied to visual contents by considering the contribution of crowdsourced media. Indeed, beside the basic task of predicting the polarity of an image, the works presented in this thesis aim to perform inferences based on the analysis of sets of pictures/videos collected from specific groups of people with common interests. Thus, we defined several inference tasks, depending on the source media (photo or video) and the parameter to be predicted (sentiment polarity or popularity). Chapter 2 introduces the paradigm of Crowdsourced Media Analysis, focused on the exploitation of huge amount of visual content publicly available. In Section 2.1 we present a framework that collects huge amount of photos taken from users during a specific period and place, with the aim to infer the behaviour of people visiting a cultural heritage site, or attending a specific event. All the inferences are based on the photos taken by visitors, therefore this work highlights how shared pictures can reflect the users' behaviour and preferences. The analysis performed on large collections of images related to the same place allows the digitalization of a cultural heritage site with very low costs, or the automatic assessing of an event directly through the observation of the multimedia information created and shared by users. Chapter 3 introduces the research field of Visual Sentiment Analysis, analyses the related problems, provides an in-depth overview of current research progress, discusses the major issues and outlines the new opportunities and challenges in this area. In Section 3.2 an overview of the most significant works in the field of Visual Sentiment Analysis, published between 2010 and 2018 is presented. The literature is presented in a chronological order, highlighting similarities and differences between the presented works, with the aim to drive the reader along the evolution of the developed methods. Section 3.3 provides a complete overview of the system design choices, with the aim to provide a depth debate about each specific issue related to the design of a Visual Sentiment Analysis system: emotions representation schemes, existing datasets, features. Each aspect is discussed, and the different possibilities are compared one each other, with related references to the state of the art. Section 3.4 provides a complete formalization of the problem, by abstracting all the components that could affect the sentiment associated to an image, including the sentiment holder and the time factor, often

ignored by the existing methods. References to the state of the art addressing each component are reported, as well as the different alternative solutions are proposed. Section 3.5 introduces some additional challenges and techniques that could be investigated, proposing suggestions for new methods, features and datasets. In Section 3.6 we present our approach to the task of sentiment polarity prediction. After a deep revision of the state of the art, we address the challenge of image sentiment polarity prediction by proposing a novel source of text for this task, dealing with the issue related to the use of text associated to images provided by users, which is commonly used in most of the previous works and is often noisy due its subjective nature. Starting from the task of image popularity prediction, in Section 3.7 we define and present an even more challenging task, named popularity dynamics prediction. In this work we provide a description of the classic problem of predicting the popularity of an image and extend this task by adding the temporal axis. Then we present the first dataset related to this task and propose a solution to the temporal challenge. In Chapter 4 we present our works on videos. In particular, Section 4.2 presents a system that takes a set of videos recorded by different people in the same time and produces a unique video as output, by considering the most popular scenes over time, based on the number of people that are simultaneously paying attention to the same scene. This system is applied in public event contexts, such as concerts or public exhibitions, and implements an automatic selection of the scenes based on the preferences inferred from the users that are attending the event. In Section 4.3 we extend this approach in the context of personal context by proposing a system for the daily living activity monitoring for lifelogging.

1.1 Dissertation Structure

In this dissertation, titled "Methods for Sentiment Analysis and Social Media Popularity of Crowdsourced Visual Contents", we mainly treated image and video contents produced by groups of users (i.e., crowdsourced). For this reason, the dissertation is properly divided into three main chapters: Crowdsourced Media Analysis, Image Sentiment Analysis and Video Segmentation and Clustering for Scene Popularity Detection. The dissertation structure is shown in Figure 1.1. In Chapter 2 we start our discussion by an introduction on Crowdsourced Media Analysis, which

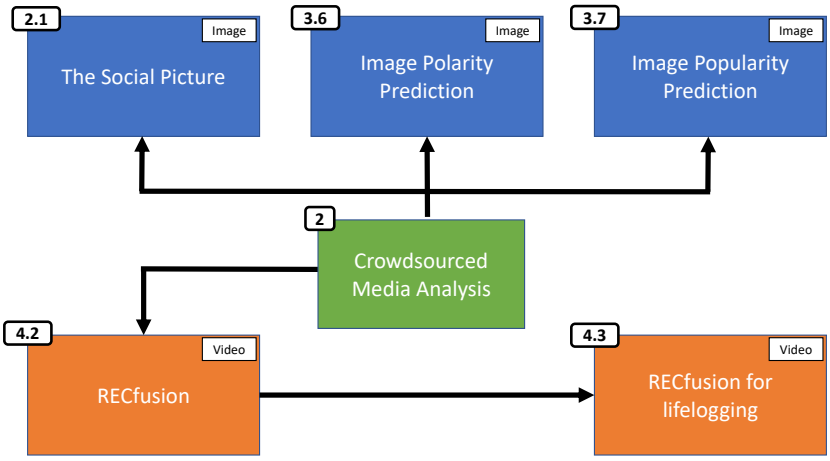


Figure 1.1: Overall structure of the current dissertation. Numbers depicts the Sections which describe the proposed work. Blue blocks represent algorithms that work on images, whereas orange blocks represent algorithms that work on videos. All the media involved in the works presented in this dissertation are produced by users or group of people with common interests (i.e., crowds).

brings together all the works presented in this dissertation. In this Chapter we also present a work for users behaviour analysis from the analysis of crwodsourced collection of photos. In Chapter 3, we present two methods for the tasks of image polarity prediction and image popularity prediction. In Chapter 4, we present our works related to the analysis of video contents, in the context of scene popularity in public events and First Person View for lifelogging. In this research work, the sentiment associated to images has been studied under several meanings. Hence, different research questions have been addressed. One is to understand the most popular subjects related to a place or an event, by the analysis of the images produced by users visiting that place or attending the event. This study produced two main framework: *The Social Picture* (see Section 2.1) and *RECfusion* (see Section 4.2). The first framework is aimed to understand user preferences based on the pictures taken from users themselves in the context of a specific event or place. The output of *The Social Picture* is a set of statistical insights about the collected images, as well as some exploration tools which allow to understand the most important subjects. The second is aimed to understand what is the scene recorded simultaneously by the most number of users attending the same event, based on the videos taken from the users at the same time. The output of *RECfusion* is a

video depicting the most popular scene over time composed by segments of videos selected from the users' ones. The approach defined in *RECfusion* has been further improved and extended to the First Person View (PFV) video domain, for the task of daily monitoring for assistive lifelogging (see Section 4.3). Then, we aimed to design systems able to predict how people react to photos shared on social media. First, we designed a system for image sentiment polarity prediction, which takes an image as input and predicts its polarity (i.e., poitive/negative). Starting from the analysis of the state of the art, we observed that most of the existing works make use of the text accompanying the pictures in the social platform, which is provided by users. Such a text content is often noisy, since the users aim to maximize the diffusion of their contents. Therefore, substantial efforts have been spent to address the issues related to such "Subjective Text". In the work presented in Section 3.6 we propose an alternative source of text, and demonstrate that it provides better results compared to the classic user provided text. Finally, we addressed the problem of image popularity prediction. When an image is shared through a social media, is important to understand, and preferably predict, the capability of such content to reach as much people as possible. This can be measured by a set of engagement values defined in the platform, such as the number of views, number of comments, etc. Given an image posted on a social media, the aim of a popularity prediction system is to predict a popularity score, which is based on a function of one engagement value (i.e., number of views, comments, favorites, shares, likes, etc.). In Section 3.7 we present and address an even more challenging task, which adds the temporal dimension to the predicted value. Thus, the proposed system is able to predict the daily popularity values of a given image for a period of 30 days, at time zero (i.e., before the image is posted).

1.2 Contributions

The main contributions of this thesis are the following:

- the definition and the development of a framework to collect, organize and explore huge collections of public photos, exploiting the crowdsourcing paradigm;
- provide a complete and thorough survey on sentiment analysis applied on images, taking into account methods, existing datasets and future challenges;

- the formulation of a method for Image Polarity Prediction, by tackling the issues related to the subjective text associated to images, which affect the most of the existing works in the field;
- the definition of a new challenging task related to the prediction of image popularity dynamics, as well as the building of a large scale dataset related to the task and publicly available;
- the formulation of a method for image popularity dynamics prediction at time zero;
- the definition and investigations of methods for scene popularity estimation, considering both mobile and wearable domains;
- the definition of a public video dataset for the scene popularity estimation task, related to indoor scenes and public events;
- a system for temporal segmentation and organization of first person videos, able to work in an unconstrained scenario without any prior in the input videos.

In the following the full list of publications of related papers is reported just taking care of group them with respect to the main involved work.

- The Social Picture:
  - Battiato, S., Farinella, G. M., Milotta, F. L., Ortis, A., Addesso, L., Casella, A., D'amico, V., Torrisi, G. (2016, June). The Social Picture. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (pp. 397-400). ACM.
- Image Polarity Prediction:
  - Ortis, A., Farinella, G. M., Torrisi, G., Battiato, S., (2018, September). Visual Sentiment Analysis Based on Objective Text Description of Images. In Proceedings of Content Based Multimedia Indexing 2018 (Accepted).
  - Ortis, A., Farinella, G. M., Torrisi, G., Battiato, S., Exploiting Objective Text Description of Images for Visual Sentiment Analysis. *Submitted to* Multimedia Tools and Applications Journal.

  - Ortis, A., Farinella, G. M., Battiato, S., A Survey on Visual Sentiment Analysis. *Submitted to* ACM Computing Surveys.

- Image Popularity Prediction

  - Ortis, A., Farinella, G. M, Battiato, S., On the Prediction of Social Image Popularity Dynamics. *Submitted to* 10th ACM Multimedia System Conference (MMSys 2019).
  - Ortis, A., Farinella, G. M., Battiato S., Predicting Social Image Popularity Dynamics at Time Zero. *Submitted to* IEEE Transactions on Multimedia.

- RECfusion:

  - Ortis, A., Farinella, G. M., D'amico, V., Addesso, L., Torrisi, G., Battiato, S. (2015, October). RECfusion: Automatic video curation driven by visual content popularity. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 1179-1182). ACM.
  - Battiato, S., Farinella, G. M., Milotta, F. L., Ortis, A., Stanco, F., D'Amico, V., Addesso L., Torrisi, G. (2017, September). Organizing Videos Streams for Clustering and Estimation of Popular Scenes. In International Conference on Image Analysis and Processing (pp. 51-61). Springer, Cham.

- RECfusion for lifelogging:

  - Ortis, A., Farinella, G. M., D'Amico, V., Addesso, L., Torrisi, G., Battiato, S. (2016, October). Organizing egocentric videos for daily living monitoring. In Proceedings of the first Workshop on Lifelogging Tools and Applications (pp. 45-54). ACM.
  - Ortis, A., Farinella, G. M., D'Amico, V., Addesso, L., Torrisi, G., Battiato, S. (2017). Organizing egocentric videos of daily living activities. Pattern Recognition, 72, 207-218.

Other unrelated publications are listed in Appendix D for sake of completeness.

Chapter 2

Crowdsourced Media Analysis

With the rapid growth in communication technology, both companies and research institutes have been given the opportunity to perform large scale analysis on a multitude of real user-generated data, with a huge variety of application contexts. Crowdsourcing provides the opportunity for input from a number of sources, with different degrees of granularity, and allows to find new ways to reach audiences on a broader scale. Social media, blogs, forums, and comment sections in online websites allow the opportunity for people to give suggestions or concerns. There are three main assets that supported the rise of the "crowdsourcing era":

- **Social Platforms**: the diffusion of social networks plays a crucial role in collecting information about people opinion, trends and behaviour. There are general social networks in which people chat, read news, and share their experiences (e.g., Facebook). Furthermore, there are also very specific social platforms aimed to bring together people with common interests. There are platforms by which computer engineers share code and advices, or professional photographers can share their photos, etc. What happens now is that people love sharing their information, tell friends what they are doing and how they feel. And what is very important for the scientific community is that most of these information are public and immediately available.

- **High Bandwidth Connection**: the number of people with an Internet connection is increasing, as well as the bandwidth and the available connection speed. With the 5G connection, it's possible to download an high quality two hour long movie in less than 4 seconds. The connectivity improvements allowed the development of new services based on the transmission of huge amount of data, and real-time services. This allowed, for instance, web-based

services like Netflix and the IP television, with the possibility to watch movies or live events with very high quality and low latency, or to perform a video of the event the user is attending, allowing him to share the live streaming through a social network.

- **Personal Devices**: the diffusion of personal devices like smartphones allows people to be connected in every second of their lives, wherever they are in that moment. This allows the users to access on-line services in any moment of their daytime. By the other hand, the amount of personal data that can be acquired by personal devices allow these services to be more pervasive and user centric.

Companies have been attempting innovative ways to get their customers involved both in production and promotion processes of their products and services. Crowdsourcing brings people together through a web-based platform, generally by means of social media, so businesses can obtain insights about what topics consumers are talking about or are interested in [1]. Asking what people like before offering a new product on the market helps reduce the risk of a product or service failure, while also generating hype around a new offering.

In the last decade, several companies exploited the crowdsourcing paradigm to offer innovative services. For example, crowdsourcing has changed the way people travel. The rise of services like AirBnB, Uber, and what has been termed the "sharing economy", transformed what had been primarily a mass-produced experience into a peer-to-peer economic network.

Companies like AirBnB and Uber have driven down prices by increasing the marketplace offer. Customers also benefit from the increased variety and personalization in their travel options. The traveller's issues and habits has remained rather the same, what have changed are only the service providers, and often times the service provided. Although the low prices can be attractive, the most of users trust the deals of such kind of companies due to the feedbacks of previous customers. Indeed, they do not actually trust the companies, but the opinions of other users of the community (preferably a large amount of them, especially if they are expert

[1] In Appendix A we describe a statistical analysis performed on the social posts published by specific brands and users with large audiences (i.e., social influencers) on Facebook, and the related produced engagement.

users of the platform who already provided useful and fair feedbacks in the past). On the other hand, these companies push users to public comments, express their opinions and tell their experiences by exploiting the "gamification" approach: the more you contribute, the more you earn (in terms of discounts, reputation, platform tools).

Besides new emerging companies, also the main IT companies have sought out innovative ideas to exploit crowdsourcing. Google exploits its users' contributions to improve the quality of Google Translate results, and the GPS locations transmitted by a large number of users' smartphones to infer traffic conditions in real time on major roads and highways. In 2008, Facebook has exploited crowdsourcing to create different language versions of its website [3].

The amount of public available and large-scale information supports the study and development of systems able to translate crowdsourced data into clear actionable insights. In the following, real use-case services based on the exploitation of user gathered visual contents are presented. From a set of crowdsourced videos or images, the presented systems are able to infer information about the sentiment polarity (i.e., positive/negative evoked emotion) and the popularity (i.e., "visual consensus" among large groups of users) of the users viewing or recording the depicted scenes. In Section 2.1, we present *The Social Picture* [4], a framework to collect and explore huge amount of crowdsourced social images about public events, cultural heritage sites and other customized private events, with the aim to extract insights about the behaviour of people attending the same event or visiting the same place. Through *The Social Picture*, users contributes to the creation of image collections about common interests. The collections can be explored through a number of advanced Computer Vision and Machine Learning algorithms, able to capture the visual content of images in order to organize them in a semantic way. The interfaces of *The Social Picture* allow the users to create customized collections by exploiting semantic filters based on visual features, social network tags, geolocation, and other information related to the images. Although the number of images could be huge, the system provides tools for the summary of the useful collection insights and statistics. It is able to automatically organize the pictures in semantic groups, according to several and customizable criteria (also in live mode). *The Social Picture* can be used as a tool for analysing the multimedia activity of the audience of an organized

event, or the activity of people visiting a cultural heritage site, performing inferences on the attitude of the participating people. The obtained information can be then exploited by the event organizers for the event evaluation and further planning or marketing strategies.

2.1 The Social Picture

2.1.1 Introduction

Images and videos have become one of the most popular media by which users express their emotions and share their experiences in the social networks. The redundancy in these data can be exploited to infer social information about the attitude of the attending people. In the context of big social data, Machine Learning and Computer Vision algorithms can be used to develop new advanced analysis systems to automatically infer knowledge from large scale visual data [5], and other multimedia information gathered by multiple sources.

In this Section we introduce a framework called *The Social Picture* (TSP) to collect, analyze and organize huge flows of visual data, and to allow users the navigation of image collections generated by the community. We designed the system to be applied on three main scenarios: public events, cultural heritage sites, private events. TSP is a social framework populated by images uploaded by users or collected from other social media. The social peculiarities of such collections can be exploited not only by the people who partecipate to an event, in fact each scenario distinguishes two kinds of users: the event organizer and the event partecipant. Imagine an art-gallery manager who leases a famous Picasso's painting with the aim to include it in a event exhibition, together with other famous and expensive artworks. How does he know he did a good investment? Which was the more attractive artwork? From which position of the hall have people taken the most number of pictures?

These information can be inferred by analysing the multimedia audience activity (i.e., uploaded images) of the organized event in *The Social Picture*. The collection of the uploaded images for an event, gives the sources analysed in TSP to answer the aforementioned questions. The obtained information can be then exploited by the event organizers for their event evaluation and further planning. On the other hand, from the user point of view, the collection of an event can be exploited trough a set

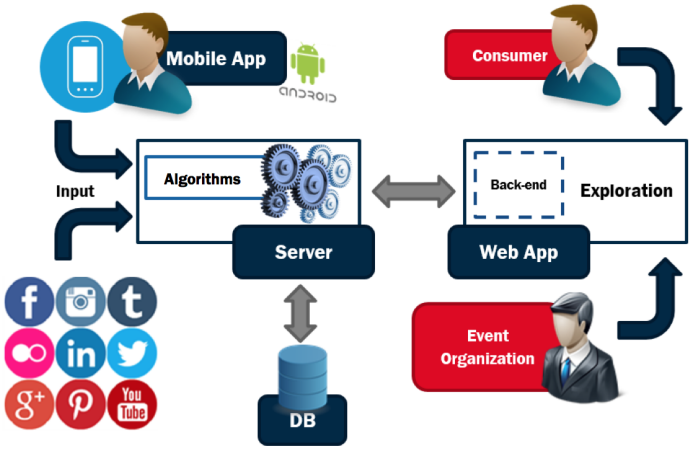


Figure 2.1: The Social Picture's architecture.

of visualization tools which exploits Computer Vision algorithms to organize images by visual content. In this way, the "social picture" of the event can be captured and shared among users.

2.1.2 Architecture

The architecture of the developed framework is shown in Figure 2.1. Users can add an image to an event's collection by using a mobile application which gives access to *The Social Picture* repository (TSP). The new images can be uploaded in TSP by using the mobile camera or by selecting images from the most common social networks for images (e.g., Flickr, Panoramio). Once an image is uploaded, it is analysed by a set of algorithms, and then stored in the database together with the extracted features and the inferred high level attributes (e.g., type of scene recognized by the algorithm). These information are exploited in TSP to create smart interfaces for the users, which can be used during the exploration of the images related to an event's collection. The framework collects all the data uploaded by the users of an event, and exploits this crowdsourced multimedia flow of pictures to infer social behavioural information about the event considering the popularity of the uploaded scenes [6].

The collections can be explored with smartphones, tablet or desktop computers via a web application, which exhibits a range of filtering tools to better explore

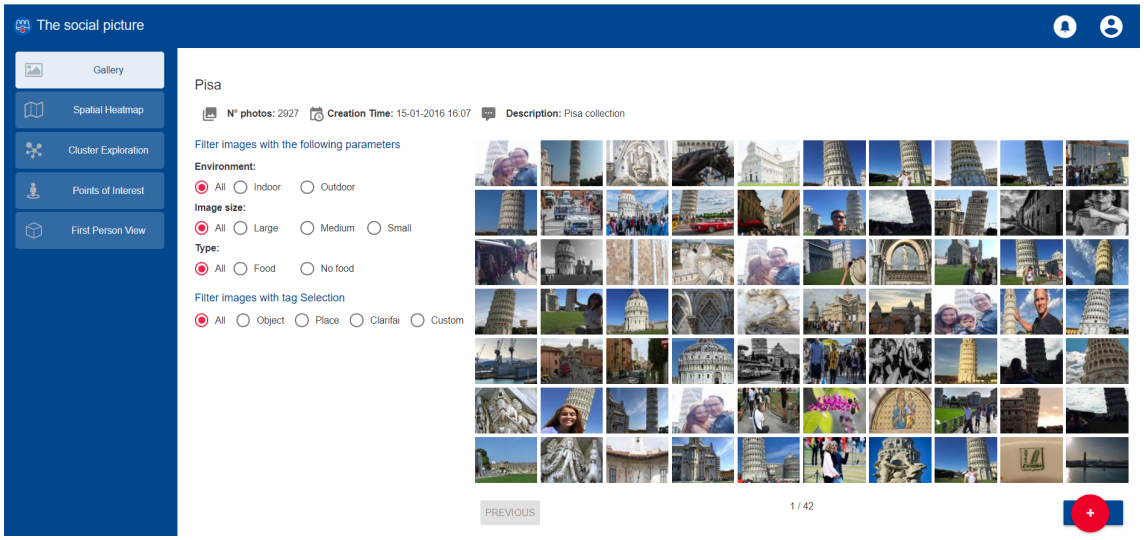


Figure 2.2: Example of exploration interface. It is composed by three main areas: the gallery (right part) shows the collection's pictures according to the selected filters (middle) which allows the users to explore the collection. By selecting an image the system shows all the extracted information and the computed inferences (i.e., objects, places, similar images, if there is food, etc.). The filters allow to customize the set of images shown in the gallery. The menu (left) allows to select the visualization tool of the framework.

the huge amount of data (see Figure 2.2). The web application shows different interfaces depending of the specific user and the event in which he has joined after an invitation from the event manager (the person who created the event). To join an event's collection, the user must upload at least one picture related to that event. Collections can be explored by several data visualization enviroments, which are selected by the event manager. Anyone registered to *The Social Picture* can become event manager and start a social collection: this follows the "prosumer" paradigm, where the users are both producers and consumers of the service. The developed framework is characterized by a modular architecture: new visualization interfaces, as well as new semantic filters can be independently created and futher added to the system. Thus, when an event manager creates a new collection, he is allowed to specify several options to customize the image gathering, the social analysis to be performed and the visualization tools for the users of that collection. The event manager is also allowed to set a range of statistics, which will be available after the analysis of the collected images (see some examples in Figure 2.3 and Figure 2.4).

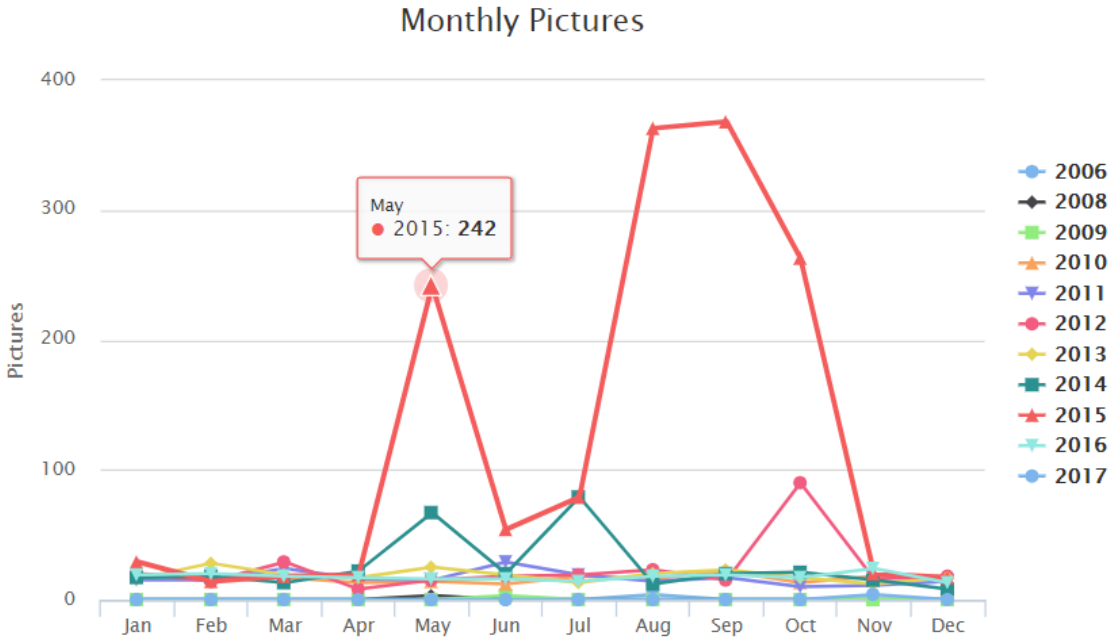


Figure 2.3: Number of pictures monthly uploaded for a set of selected years.

Statistics helps organisers to extract useful social information from the crowdsourced pictures. For example, what is the most popular artwork of a museum? What is the least considered? From which perspective these pictures were taken? These information could be exploited, for example, to perform aimed investments. The system can suggest what is the better subject to use for the advertising campaign of the event, or which of the attractions it worth to mainly reproduce in the souvenir shop products, to support merchandising strategies. Feedback about what is the most interesting part (i.e., the most captured photo) of a landmark building can help on taking decisions about renovating some parts of the building rather than other as first investment, where the connotation of importance is achieved by the crowd who generated "the social picture" for that building by uploading related images.

The several exploration tools are based on both visual and textual data. The system exploits information such as Exif data (camera model, geolocalization, acquisition details, and others) when available, and a number of ad hoc extracted visual features.

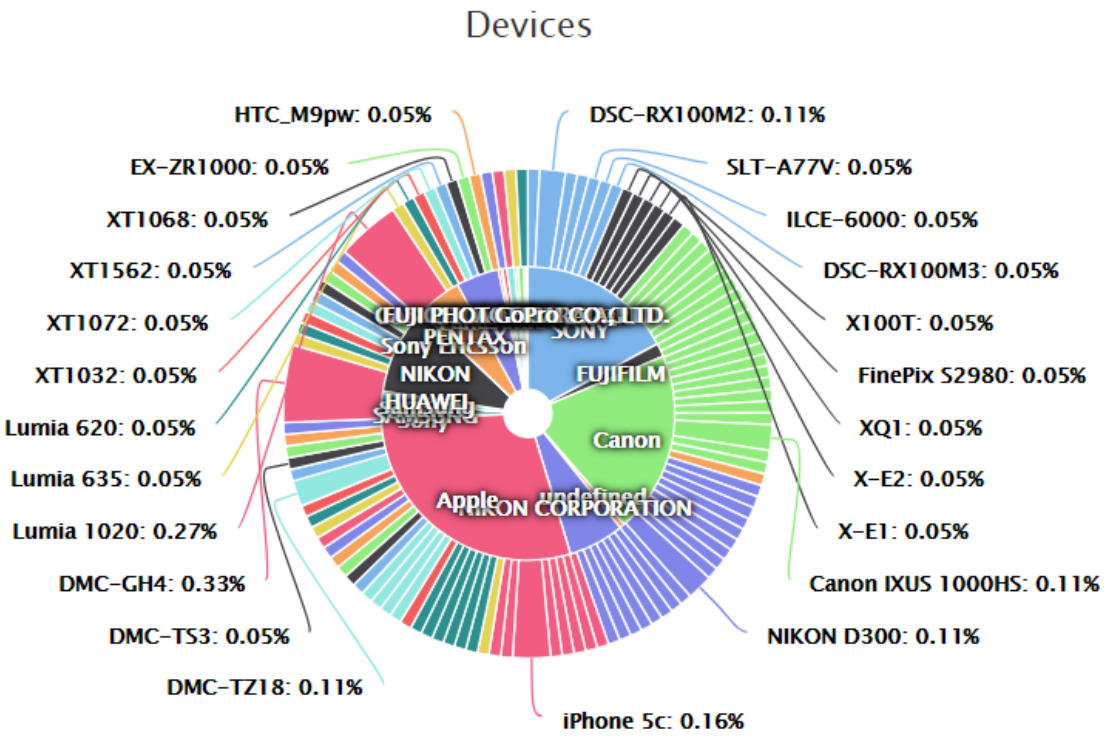


Figure 2.4: Distribution of the devices used to take the pictures included in the collection.

The visual analysis module of the system feds all the images to two different CNNs [7, 8], in order to extract the classification labels and an image representation. To attach semantic labels to the visual content of the images, we used *AlexNet* [7] and *Places205-AlexNet* [8]. The CNN used in [7] consist of seven internal layers with a final 1000-way softmax which produces a distribution over the 1000 predefined classes of the ImageNet dataset [9]. We considered the feature activations induced at the last hidden layer, which consists of 4096 dimensional feature (*fc7* feature), as an image representation to be further used with t-SNE algorithm [10] for visualization purposes. We also fed the images to the *Places205-AlexNet* CNN [8]. This CNN has the same architecture of *AlexNet* CNN, but it is trained on 205 scene categories related to places learned by using the database *Places205-AlexNet* composed by 2.5 million images.

2.1.3 User experience

An event manager (i.e., a user of *The Social Picture* which starts a new collection) creates a new event by selecting among three possible type of event: public event (e.g., a concert), cultural heritage site (e.g., a museum) or private event (e.g., a wedding). The available event categorization can be extended to include other customized categories. We considered these three categories to better focus the aims of the specific analysis, and the inferred information that an organizator wants to extract. The data gathering from users can be performed within a specific time window. The manager is allowed to control the image acquisition by selecting fine-grained criteria such as filtering media by hashtag, associated text or geolocalization distance. After creating the event and its acquisition settings, the manager can select the statistics that the system have to compute by exploiting the collection of multimedia data gathered for that event.

The pictures can be grouped by hierarchical categories depending on the combination of two or more of the extracted visual features. Specific image categorizations help users to better handle huge amount of crowdsourced pictures, this kind of grouping can be exploited as a pre-processing before performing an image based visual search. Given a seed image, the system selects a set of similar pictures. The system provides different exploration tools that can be exploited to better navigate any huge image collection. These exploration tools together with other advanced tools are described in the next subsections. A demonstration video of the framework is available at the following URL: http://iplab.dmi.unict.it/TSP.

Heatmap exploration

In a cultural heritage site, people usually take pictures from different points of view and considering different details of parts related to famous and appreciated attractions and artworks. The heatmap exploration tool of *The Social Picture* aims to infer the "interest" of people with respect to the different parts of a site. An example of heatmap generated from data in *The Social Picture* is shown in Figure 2.5. Through this visualization tool, an organizer of a collection will be able to know which parts of the site captures people's interests. On the other hand, users can explore the collection related to a site in a very simple and intuitive way. So, to

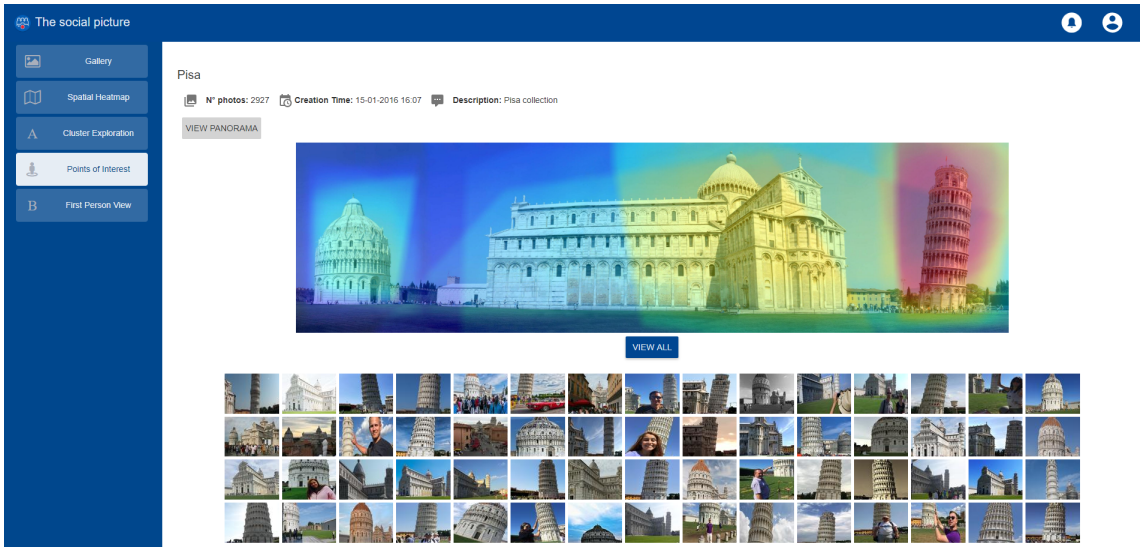


Figure 2.5: Heatmap exploration tool. By selecting a point in the heatmap, the system visualizes all the photos that contributed to that point in the bottom part of the interface.

highlight the "interest" of people related to parts of a site, the proposed system creates an heatmap by aligning images in *The Social Picture* with respect to panoramic images of the site of interest [11].

The heatmap is a visualization used to depict the intensity of images at spatial points. The heatmap consists of a colored overlay applied to the original image. Areas of higher intensity will be colored red, and areas of lower intensity will appear blue. The intensity of the heatmap is given by the number of collected pictures that contain that visual area. By clicking on a point of the heatmap, the user can visualize the subject of images that contribuited to generate the map intensity at that point. This set of pictures can be further refined by selecting one of the images and asking the system to search similar pictures, or use the image subset as a starting point for further analysis. In other words, the heatmap visualization gives the possibility to understand the behaviour of the people, especially if it is combined with the information coming from the geolocation of the devices in the instant of the photos creation. Also it can be considered as a powerful and intuitive image retrieval tool for the collections related to cultural heritage sites.

3D Reconstruction Starting from VSFM (Visual Structure From Motion) [12], we are able to compute a 3D sparse reconstruction of large photos collections. The

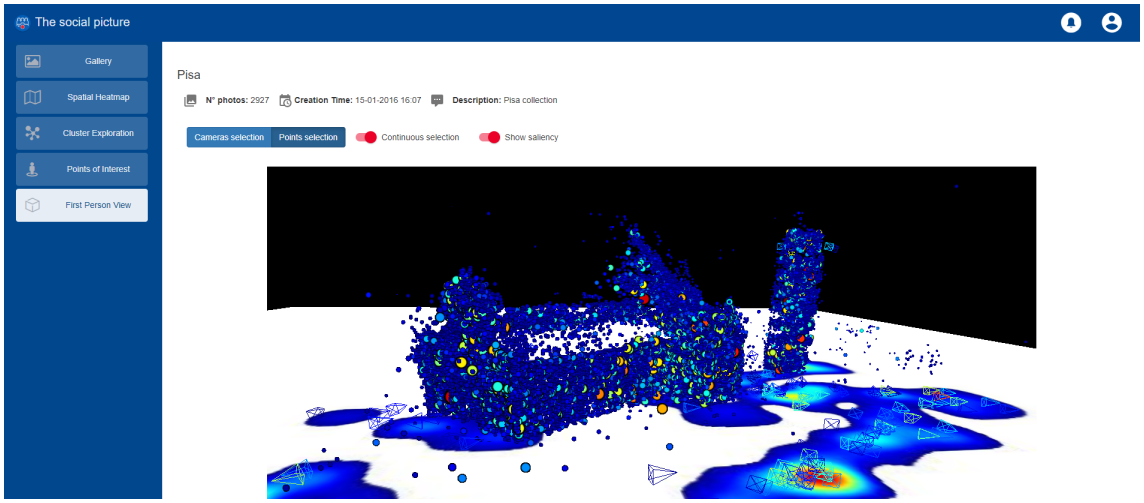


Figure 2.6: 3D sparse reconstruction of a cultural heritage site based on the photos of the collecction. Each point in the 3D space is estimated by considering the projections of the images depicting that part of the scene.

models are augmented with colors for vertices, related to the frequency of been acquired in a photo, colors for cameras, related to the number of visual features acquired by each photo, and with a plane which show the spatial density of contributing users. We embedded in TSP the models through a 3D web viewer allowing the users to browse the 3D sparse reconstructed models gaining a cue about what are the points of view and the subjects preferred by users when take photos.

Embedding Exploration

We exploit the *fc7* feature extracted with the *AlexNet* architecture [7] for each image and use the *t-SNE* embedding algorithm [10] to compute a 2D embedding that respects the pairwise distances between visual features. The *t-SNE* (t-Distribuited Stochastic Neighbour Embedding) is a technique for feature space dimensionality reduction that is particularly well suited for the visualization of high dimensional image datasets. In Figure 2.7 the images are first assigned to a 2D position in the embedding space by means of the *t-SNE* algorithm, then we forced the images to fit a grid layout for a better visualization. Note that images with the same subject are automatically arranged nearby (see Figure 2.7). Moreover, the system arranges very close those images which are not the same but have a similar visual content. It is also important to highlight that the employed CNN [7] has been trained using

Figure 2.7: t-SNE visualization, the images are forced to fit a grid layout. Images of an event are automatically organized by visual content. Images close in the 2D space of the visualization tool are also close in terms of visual content.

a different dataset concerning 1000 classes of objects, but the *fc7* features resulted expressive and representative enough to be applied successfully to a generic event collection.

Another exploration tool allows users to visualize the result of t-SNE embedding without using the grid layout. Images related to similar subjects will be clustered implicitly, without any semantic hints. This allow the exploration of the different sets of images composing the collection. For instance, photos of the same building but different lighting or weather conditions are arranged nearby, depending on their pairwise similarity. Photos of the same building but from different point of views are arranged in further sub-groups. Indeed, depending on the analysed photos, this automatic grouping can have different level of granularity. These automatically generated groups can be exploited to select the "best" photo among the cluster of photos related to the same scene, or remove duplicates in the collection as instance. When we display a set of images using their embedded locations computed by t-SNE, they may overlap one each other. Especially if there are many similar images. For this reason, this interface provides a set of tools to help the user's navigation. With this exploration tool, the user can apply a translation or a zooming to all the viewed images, just clicking and dragging the mouse along the desired direction and by using the scroll wheel respectively. This helps the user to better explore the image distribution in a custom level of detail.

Hierarchical t-SNE The first implementation of the t-SNE exploration tool in [4] was unable to scale with the number of the collections' images. We further extended this tool by implementing an hierarchical version of the t-SNE embedding which allows to explore picture collections without limits on the amount of processed pictures. This helps the user to better explore the image distribution in a custom level

Figure 2.8: photos embedding based on the t-SNE coordinates.

of detail. Furthermore, the user can choose a subset of images and compute the t-SNE embedding of them directly on the browser.

As the number of pictures of a collection is unpredictable, the computation of the t-SNE coordinates could be very expensive. Besides the t-SNE computation, which needs to be executed only one time per dataset, a huge number of pictures can affect the browser efficiency for the visualization of the 2D embedding. We organize the entire collection of pictures in a hierarchical structure. After the collection is analysed (i.e., the *fc7* features have been computed for all the images) the system performs a hierarchical k-means clustering of the image features. The algorithm divides the dataset recursively into k clusters, for each computation the k centroids are used as elements of a *k-tree* and removed from the set. When this new version of the t-SNE tool (hierarchical t-SNE) is executed, it shows to the user the t-SNE embedding computed only for the elements in the root of the *k-tree* (i.e., the picture centroids of the first *k-means* computation). When the user selects one of these pictures, the system computes the t-SNE of the pictures included in the child node corresponding to the selected picture element. This hierarchical exploration

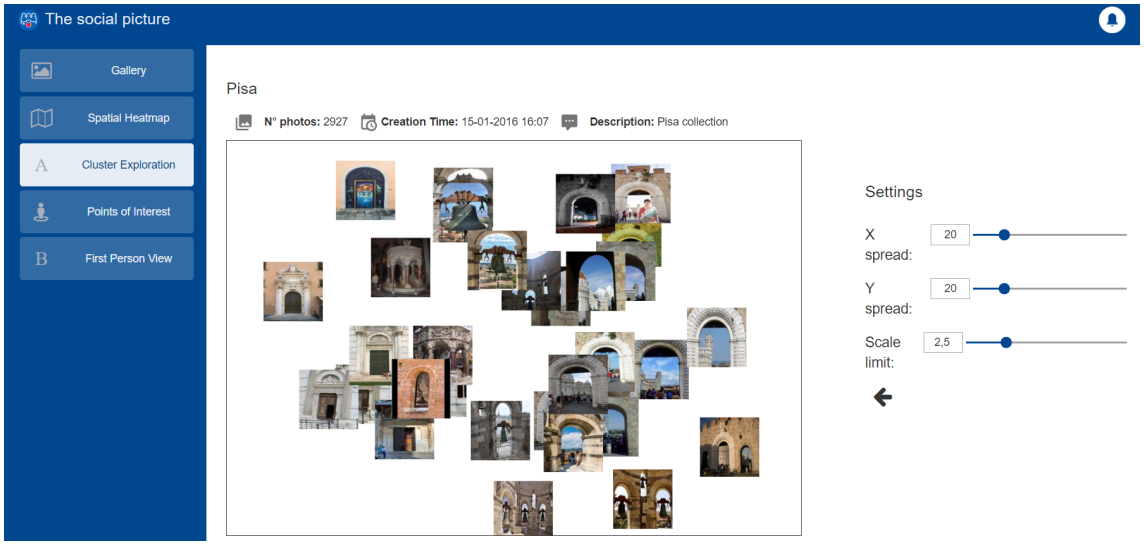


Figure 2.9: visualization interface for t-SNE based image embedding. In this example, a subset of images is shown. All the images are related to photos depicting "archs", however the group if images has been automatically created without taking into account the semantic of the images.

can be continued by selecting one of the shown pictures and computing the t-SNE embedding for its sub-elements in the hierarchy. Figure 2.9 shows an example of t-SNE visualization related to the group of images included in a child node of the tree. All these images are related to "archs", however it worth to highlight that this group has been automatically created without taking into account the semantic in the images, but only considering the pairwise distances between images of the entire collection and the implemented hierarchical organization.

Other Advanced Tools

Among the tools included in *The Social Picture* there is the one useful to generate automatic subsets of images from a specific photo collection. This tool allows the user to set the number of images to obtain as output for a collection in TSP, and automatically generates the subset of images taking into account visual features as well as EXIF information related to the images composing the photo collection (e.g., GPS location, TAGS, day, time, etc). In this way, the user can have some representative image prototypes related to the collection to be used for different purposes

Figure 2.10: visualization of the locations from which users taken the photos of the collection. This tool allow to understand how people visit the site and what are the places with the most popular points of view. By providing the positions of the users during the event it gives some hints about the most interesting parts of the site.

(e.g., printing the most significative pictures of paintings of a museum for a specific social group).

For each photo analysed by *The Social Picture*, the system exploits three CNNs to extract information about object and places depicted, as well as determine if there is food in the picture (i.e., food vs. no-food classifier). Moreover, the automatic image captioning as described in [13] is also exploited to extract a textual description from images. The descriptions of images can be used for text based query performed by the user.

2.1.4 Future Works

The Social Picture is a framework able to provide feedback about the behaviour and preferences of users visiting a place, or attending an event. However, all the performed inferences represent insights about what the users captured and shared in the social platform. An experimental evaluation of the above described system can't be performed by comparing the obtained results with respect to specific sets of Ground-Truth pictures. For this reason, future works will be devoted to the study and the development of evaluation approaches of the proposed system by

comparing the results obtained considering different set of photos related to the same place taken during the same period but in different years.

2.2 Conclusions

In this Chapter we discussed the Crowdsourced Media Analysis paradigm, as well as presenting a framework aimed to infer the interest of people attending an event or visiting a cultural heritage site based on the analysis of the taken photos. Feedback about what is the most interesting part (i.e., the most captured) of a landmark building can help on taking decisions about renovating some parts rather than others as first investment. The t-SNE exploration tool exploits a technique for feature space dimensionality reduction that is particularly well suited for the visualization of high dimensional image datasets. This tool allows the visualization of huge amount of pictures, and the hierarchical implementation allows to scale the number of analysed pictures. Very large collections can be explored, and the pictures are automatically arranged in semantic groups. The system provides different exploration tools (e.g., heatmap, t-SNE exploration), automatic tagging engines (i.e., object classification, places, food vs. no-food, concept tags) and statistics. The extracted information can be exploited to define custom filters for images and to automatically infer how users act when visiting a cultural heritage site or attending to a public event, and what captured their interest.

Chapter 3

Image Sentiment Analysis

3.1 Introduction

With the growth of social media (i.e., reviews, forums, blogs and social networks), individuals and organizations are increasingly using public opinions for their decision making [14]. As instance, companies are interested in monitoring people opinions toward their products or services, as well as customers rely on feedbacks of other users to evaluate a product before they purchase it.

The basic task in Sentiment Analysis is the polarity classification of an input text (e.g., taken from a review, a comment or a social post) in terms of positive, negative or neutral polarity. This analysis can be performed at document, sentence or feature level. The methods of this area are useful to capture public opinion about products, services, marketing, political preferences and social events. For example the analysis of the activity of Twitter's users can help to predict the popularity of parties or coalitions. The achieved results in Sentiment Analysis within micro-blogging have shown that Twitter posts reasonably reflect the political landscape [15]. Historically, Sentiment Analysis techniques have been developed for the analysis of text [16], whereas limited efforts have been employed to extract (i.e., infer) sentiments from visual contents (e.g., images and videos).

Even though the scientific research has already achieved notable results in the field of textual Sentiment Analysis in different contexts (e.g., social network posts analysis, product reviews, political preferences, etc.), the task to understand the mood from a text has several difficulties given by the inherent ambiguity of the various languages (e.g., ironic sentences), cultural factors, linguistic nuances and the difficulty of generalize any text analysis solution to different language vocabularies.

The different solutions in the field of text Sentiment Analysis have not yet achieved a level of reliability good enough to be implemented without enclosing the related context. For example, despite the existence of natural language processing tools for the English language, the same tools cannot be used directly to analyse text written in other languages.

This Chapter is organized as follows. Section 3.2 reviews relevant publications and presents a complete overview of the state of the art in the field. After a description of the task and the related applications, the subject is tackled under different main headings. Then, principles of design of general Visual Sentiment Analysis systems are described in Section 3.3 and discussed under three main points of view: emotional models, dataset definition, feature design. A formalization of the problem is presented in Section 3.4, considering different levels of granularity, as well as the components that can affect the sentiment toward an image in different ways. To this aim, Section 3.3 considers a structured formalization of the problem which is usually used for the analysis of text, and discusses it's suitability in the context of Visual Sentiment Analysis. The Chapter also includes a description of new challenges in Section 3.5 as well as the evaluation from the viewpoint of progress toward more sophisticated systems and related practical applications.

Two main inference tasks related to the image sentiment analysis that have been investigated in this research work are further presented in this Chapter: sentiment polarity (Section 3.6) and sentiment popularity (Section 3.7).

Given an image, the proposed method described in Section 3.6 properly combines visual and textual features to define an embedding space, then a classifier is trained on the embedded features. The novelty of the proposed method consists on the fact that we don't lean on the text provided by users, which is often noisy. Indeed we propose an alternative subjective source of text, directly extracted from images. Starting from an embedding approach which exploits both visual and textual features, we attempt to boost the contribute of each input view. We propose to extract and employ an *Objective Text* description of images rather than the classic *Subjective Text* provided by the users (i.e., title, tags and image description) which is extensively exploited in the state of the art to infer the sentiment associated to social images. During the evaluation, we compared an extensive number of text and visual features combinations and baselines obtained by considering the state of the

art methods. Experiments performed on a representative dataset of 47235 labelled samples demonstrate that the exploitation of *Objective Text* helps to outperform state-of-the-art for sentiment polarity estimation.

The work presented in Section 3.7 addresses the task of image popularity prediction. In particular, we introduce the new challenge of forecasting the engagement score reached by social images over time. We call this task "Popularity Dynamic Prediction". The work is motivated by the fact that the popularity of social images, which is usually estimated at a precise instant of the post lifecycle, could be affected by the period of the post (i.e., how old is the post). The task is hence the estimation, in advance, of the engagement score dynamic over a period of time (e.g., 30 days) by exploiting visual and social features. To this aim, we propose a benchmark dataset that consists of $\sim 20K$ Flickr images labelled with their engagement scores (i.e., views, comments and favorites) in a period of 30 days from the upload in the social platform. For each image, the dataset also includes user's and photo's social features that have been proven to have an influence on the image popularity on Flickr (e.g., number of user's contacts, number of user's groups, mean views of the user's images, photo tags, etc.). The proposed dataset is publicly available for research purposes. We also present a method to address the aforementioned problem. The proposed approach models the problem as the combination of two prediction tasks, which are addressed individually. Then, the two outputs are properly combined to obtain the prediction of the whole engagement sequence. Our approach is able to forecast the daily number of views reached by a photo posted on Flickr for a period of 30 days, by exploiting features extracted from the post. This means that the prediction can be performed before posting the photo. The proposed method is compared with respect to different baselines. In Section 3.8 a summary of the insights resulting from this study is presented.

3.2 State of the Art

Visual Sentiment Analysis is a recent research area. Most of the works in this new research field rely on previous studies on emotional semantic image retrieval [17, 18, 19, 20], which make connections between low-level image features and emotions, with the aim to perform automatic image retrieval and categorization. These works

have been also influenced by empirical studies from psychology and art theory [21, 22, 23, 24, 25, 26]. Other research fields close to Visual Sentiment Analysis are those considering the analysis of the image aesthetic [27, 28, 29, 30], interestingness [31], affect [32] and popularity [33, 34, 35, 36].

The first paper on Visual Sentiment Analysis aims to classify images as "positive" or "negative" and dates back on 2010 [37]. In this work the authors studied the correlations between the sentiment of images and their visual content. They assigned numerical sentiment scores to each picture based on their accompanying text (i.e., meta-data). To this aim, the authors used the SentiWordNet [38] lexicon to extract sentiment score values from the text associated to images. This work revealed that there are strong correlations between sentiment scores extracted from Flickr meta-data (e.g., image title, description and tags provided by the user) and visual features (i.e., SIFT based bag-of-visual words, and local/global RGB histograms).

In [39] a study on the features useful to the task of affective classification of images is presented. The insights from the experimental observation of emotional responses with respect to colors and art have been exploited to empirically select the image features. To perform the emotional image classification, the authors considered 8 emotional output categories as defined in [40] (i.e., Awe, Anger, Amusement, Contentment, Excitement, Disgust, Sad, and Fear).

In [41] the authors built a large scale Visual Sentiment Ontology (VSO) of semantic concepts based on psychological theories and web mining (SentiBank). A concept is expressed as an adjective-noun combination called Adjective Noun Pair (ANP) such as "beautiful flowers" or "sad eyes". After building the ontology consisting of 1.200 ANP, they trained a set of 1.200 visual concept detectors which responses can be exploited as a sentiment representation for a given image. Indeed, the 1.200 dimension ANP outputs (i.e., the outputs of the ANP detectors) can be exploited as features to train a sentiment classifier. To perform this work the authors extracted adjectives and nouns from videos and images tags retrieved from YouTube and Flickr respectively. These images and videos have been searched using the words corresponding to the 24 emotions defined in the Plutchik Wheel of Emotion [42], a well known psychological model of human emotions. The authors released a large labelled image dataset composed by half million Flickr images regarding to 1.200 ANPs (see

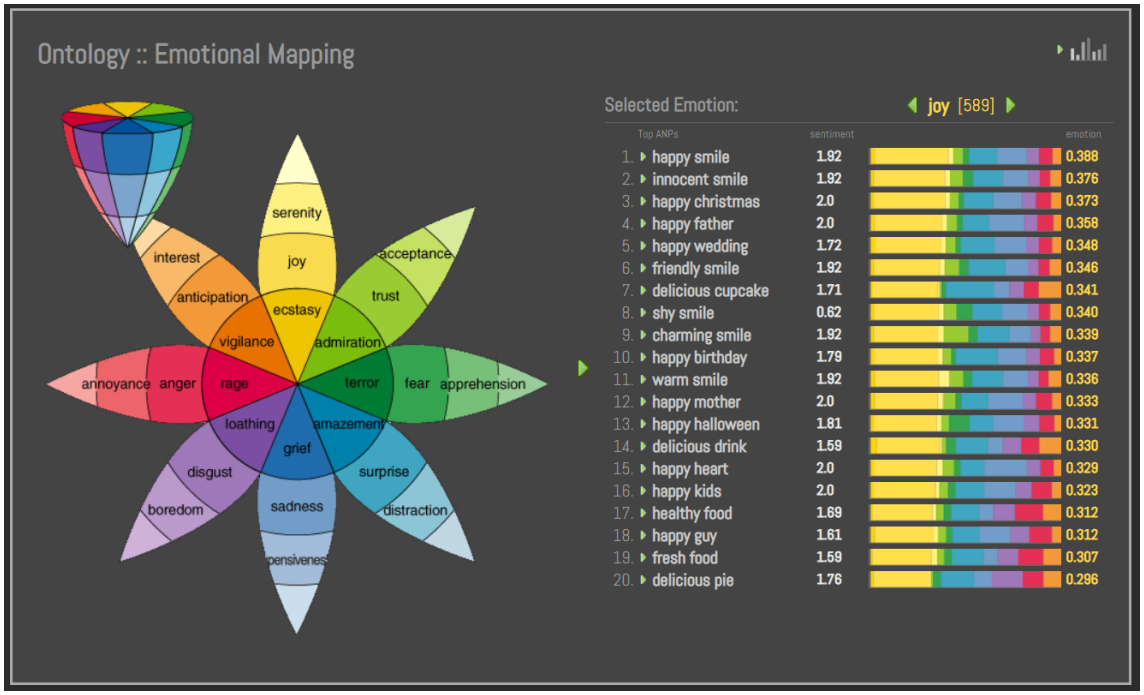


Figure 3.1: VSO (Visual Sentiment Ontology) web interface by which the dataset defined in [41] can be explored. Image borrowed from the project webpage of [41].

Figure 3.1). Results show that the approach based on SentiBank concepts outperforms text based method in tweet sentiment prediction experiments. Furthermore, the authors compared the SentiBank representation with shallow features (colour histogram, GIST, LBP, BoW) to predict the sentiment reflected in images. To this purpose, they used two different classification models (LinearSVM and Logistic Regression) achieving significant performance improvements when using the SentiBank representation. The proposed mid-level representation has been further evaluated in the emotion classification task considered in [39], obtaining better results.

In 2013, Yuan et al. [43] employed scene-based attributes to define mid-level features, and built a binary sentiment classifier on top of them. Furthermore, their experiments demonstrated that adding a facial expression recognition step helps the sentiment prediction task when applied to images with faces.

In 2014, Yang et al. [44] proposed a Sentiment Analysis approach based on a graphical model which is used to represent the connections between visual features and friends interactions (i.e., comments) related to the shared images. The exploited

visual features include saturation, saturation contrast, bright contrast, cool color ratio, figure-ground color difference, figure-ground area difference, background texture complexity, and foreground texture complexity. In this work the authors considered the Ekman's emotion model [45].

Chen et al. [46] introduced a CNN (Convolutional Neural Network) based approach, also known as "SentiBank 2.0" or "DeepSentiBank". They performed a fine-tuning training on a CNN model previously trained for the task of object classification to classify images in one of a 2.096 ANP category (obtained by extending the previous SentiBank ontology [41]). This approach significantly improved the ANP detection with respect to [41]. Similarly to [41], this approach provides a sentiment feature (i.e., a representation) of an image that can be exploited by further systems.

In contrast to the common task of infer the affective concepts intended by the media content publisher (i.e., by analysing the text associated to the image by the publisher), the method proposed in [47] tries to predict what concepts will be evoked to the image viewers.

In [48] a pre-trained CNN is used as a provider of high-level attribute descriptors in order to train two sentiment classifiers based on Logistic Regression. Two types of activations are used as visual features, namely the *fc7* and *fc8* features (i.e., the activations of the seventh and eighth fully connected layers of the CNN respectively). The authors propose a fine-grained sentiment categorization, classifying the polarity of a given image through a 5-scale labelling scheme: *"strong negative", "weak negative", "neutral", "weak positive", and "strong positive"*. The evaluation of this approach considers two baseline methods taken from the state of the art, namely low-level visual features and SentiBank, both introduced in [41], in comparison with their approaches (the *fc7* and *fc8* based classifiers). The experimental setting evaluates all the considered methods on two real-world dataset related to Twitter and Tumblr, whose images have been manually labelled considering the above described 5-scale score scheme. The results suggest that the methods proposed in [48] outperform the baseline methods in visual sentiment prediction.

The authors of [49] proposed to use a progressive approach for training a CNN (called Progressive CNN or PCNN) in order to perform visual Sentiment Analysis in terms of "positive" or "negative" polarity. They first trained a CNN architecture with a dataset of half million Flickr images introduced in [41]. At training time, the

method selects a subset of training images which achieve high prediction scores. Then, this subset is used to further fine-tune the obtained CNN. In the architecture design they considered a last fully connected layer with 24 neurons. This design decision has been taken with the aim to let the CNN learn the respones of the 24 Plutchik's emotions [42]. An implementation of the architecture proposed in [49] is publicly available. The results of experiments performed on a set of manually labelled Twitter images show that the progressive CNN approach obtain better results with respect to other previous algorithms, such as [41] and [43].

Considering that the emotional response of a person viewing an image may include multiple emotions, the authors of [50] aimed to predict a distribution representation of the emotions rather than a single dominant emotion from (see Figure 3.2). The authors compared three methods to predict such emotion distributions: a Supor Vector Regressor (based on hand crafted features related to edge, color, texture, shape and saliency), a CNN for both classification and regression. They also proposed a method to change the evoked emotion distribution of an image by editing its texture and colors. Given a source image and a target one, the proposed method transforms the color tone and textures of the source image to those of the target one. The result is that the edited image evokes emotions closer to the target image than the original one. This approach has been quantitatively evaluated by using four similarity measures between distributions. For the experiments, the authors consider a set of 7 emotion categories, corresponding to the 6 basic emotions defined by Ekman in [45] and the neutral emotion. Furthermore, the authors proposed a sentiment database called Emotion6. The experiments on evoked emotion transfer suggest that holistic features such as the color tone can influence the evoked emotion, albeit the emotion related to images with high level semantics are difficult to be shaped according to an arbitrary target image.

In [51] the textual data, such as comments and captions, related to the images are considered as contextual information. Differently from the previous approaches, which exploit low-level features [32], mid-level features [41, 43] and Deep Learning architectures [49, 50], the framework in [51] implements an unsupervised approach (USEA - Unsupervised SEntiment Analysis). In [52] a CNN pre-trained for the task of Object Classification is fine-tuned to accomplish the task of visual sentiment prediction. Then, with the aim to understand the contribution of each CNN

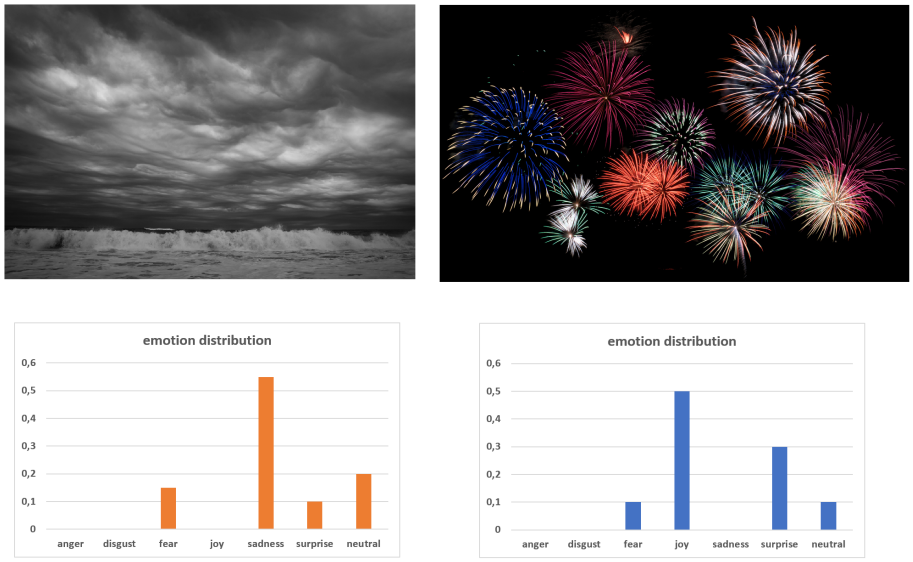


Figure 3.2: Examples of image emotion distributions.

layer for the task, the authors performed an exhaustive layer-per-layer analysis of the fine-tuned model. Indeed, the traditional approach consists in initializing the weights obtained by training a CNN for a specific task, and replacing the last layer with a new one containing a number of units corresponding to the number of classes of the new target dataset. The experiments performed in this paper explored the possibility to use each layer as a feature extractor and training individual classifiers. This layer by layer study allows measuring the performance of the different layers which is useful to understand how the layers affect the whole CNN performances. Based on the layer by layer analysis, the authors proposed several CNN architectures obtained by either removing or adding layers from the original CNN.

Even though the conceptual meaning of an image is the same for all cultures, each culture may have a different sentimental expression of a given concept. Motivated by this observation, Jou et al. [53] extended the ANP ontology defined in [41] for a multi-lingual context. Specifically, the method provides a multilingual sentiment driven visual concept detector in 12 languages. The resulting Multilingual Visual Sentiment Ontology (MVSO) provides a rich information source for the analysis of cultural connections and the study of the visual sentiment across languages.

Starting by the fact that either global features and dominant objects bring massive sentiment cues, Sun et al. [54] proposed an algorithm that extracts and combines

features from either the whole image and "salient" regions. These regions have been selected by considering proper objectness and sentiment scores aimed to discover affective local regions. The proposed method obtained good results compared with [41] and [49] on three widely used datasets presented in [41] and [49]. In [55] Katsurai and Satoh exploited visual, textual and sentiment features to build a latent embedding space where the correlation between the projected features from different views is maximized. This work implements the CCA (Canonical Correlation Analysis) technique to build a 3-view embedding which provides a tool to encode inputs from different sources (i.e., a text and an image with similar meaning/sentiment are projected nearby in the embedding space) and a method to obtain a sentiment representation of images (by simply projeting an input feature to the latent embedding space). This representation is exploited to train a linear SVM classifier to infer positive or negative polarity. The authors used a composition of RGB histograms, GIST, SIFT based Bag of Words and two mid-level features defined in [56] and [41] as visual features. The textual feature is obtained using a Bag of Words approach form text associated to the image, crawled from Flickr and Instagram. The sentiment features are obtained starting from the input text and exploiting an external knowledge base, called SentiWordNet [38], a well-known lexical resource used in opinion mining to assign sentiment scores to words.

The works in [57] and in [58] perform emotional image classification of images considering multiple emotional labels. As previously proposed in 2015 by Peng [50], instead of training a model to predict only one sentiment label, the authors considered a distribution over a set of pre-defined emotional labels. To this aim, they proposed a multi-task system which optimizes the classification and the distribution prediction simultaneously. In [57] the authors proposed two Conditional Probability Neural Networks (CPNN), called Binary CPNN (BCPNN) and Augmented CPNN (ACPNN). A CPNN is a neural network with one hidden layer which takes either features and labels as input and outputs the label distribution. Indeed, the aim of a CPNN is to predict the probability distribution over a set of considered labels. The authors of [58] changed the dimension of the last layer of a CNN pre-trained for Object Classification in order to extract a probability distribution with respect to the considered emotional labels, and replaced the original loss layer with a function that integrates the classification loss and sentiment distribution loss through a weighted

combination. Then the modified CNN has been fine-tuned to predict sentiment distributions. Since the majority of the existing datasets are built to assign a single emotion ground truth to each image, the authors of [58] proposed two approaches to convert the single labels to emotional distribution vectors, which elements represent the degree to which each emotion category is associated to the considered image. This is obtained considering the similarities between the pairwise emotion categories [42]. The experimental results show that the approach proposed in [58] outperforms eleven baseline Label Distribution Learning (LDL) methods, including BCPNN and ACPNN proposed in [57].

In [59] the authors extended their previous work [52] in which they first trained a CNN for sentiment analysis and then empirically studied the contribute of each layer. In particular, they used the activations in each layer to train different linear classifiers. In this work the authors also studied the effect of weight initialization for fine-tuning by changing the task (i.e., the output domain) for which the fine-tuned CNN has been originally trained. Then, the authors propose an improved CNN architecture based on the empirical insights.

The authors of [60] crawled $\sim 3M$ tweets within a period of 6 months. Then, the text contained in the tweets has been labelled using a sentiment polarity classifier. The selected images have been used to build a dataset named Twitter for Sentiment Analysis (T4SA). The authors exploited this dataset to finetune existing CNN previously trained for objects and places classification (VGG19 [61] and HybridNet [8]). The proposed system has been compared with the CNN and PCNN presented in [49], DeepSentiBank [46] and MVSO [53] obtaining better results on the built dataset.

The system proposed in [62] represents the sentiment of an image by extracting a set of ANPs describing the image. Then, the weighted sum of the extracted textual sentiment values is computed, by using the related ANP responses as weights. The approach proposed in this paper takes advantage of the sentiment of the text composing the ANPs extracted from images, instead of only considering the ANPs responses defined in SentiBank [41] as mid-level representations. In particular, the sentiment value of an extracted ANP is defined by summing the sentiment scores defined in SentiWordNet [38] and SentiStrength [63] for the pair of adjective and the noun words of th ANP. A logistic regressor is used to infer the sentiment orientation

by exploiting the scores extracted from the textual information, and a logistic classifier is trained for polarity prediction by exploiting the traditional ANP responses as representations. Then, the two schemes are combined by employing a late fusion approach. The authors compared their method with respect to three baselines: a logistic regression model based on the SentiBank mid-level representation, the CNN and PCNN methods proposed in [49]. Experiments show that the proposed late fusion method outperforms the method based only on the mid-level representation defined in SentiBank, demonstrating the contribute given by the sentiment coefficients associated to the text composing the extracted ANPs. However, the CNN and PCNN approaches proposed in [49] exhibit better performances than the late fusion method.

The works described in this section have led to significant improvements in the field of Visual Sentiment Analysis. However these works address the problem considering different emotion models, datasets and evaluation methods. So far, researchers formulated this task as a classification problem among a number of polarity levels or emotional categories, but the number and the type of the emotional outputs adopted for the classification are arbitrary (see Table 3.1). The difference in the adopted emotion categories makes result comparison difficult. Moreover, there is not a strong agreement in the research community about the use of an universal benchmark dataset. Indeed, several works evaluated their methods on their own datasets. Many of the mentioned works present at least one of the said issues.

Year	Paper	Input	Output
2010	Siersdorfer et al. [37]	Hand crafted visual features	Sentiment Polarity
2010	Machajdik et al. [39]	Hand crafted visual features	Emotional Classification [40]
2013	Borth et al. [41]	ANP output responses [41]	Sentiment Polarity and Emotional Classification [40]
2013	Yuan et al. [43]	Hand crafted visual features	Sentiment Polarity
2014	Yang et al. [44]	Hand crafted visual features	Emotional Classification [40]
2014	Chen et al. [46]	Raw image	ANP annotation [41]
2014	Xu et al. [48]	CNN activations	5-scale Sentiment Score
2015	You et al. [49]	Raw image	Sentiment Polarity
2015	Peng et al. [50]	Hand crafted visual features	Distribution of Emotions
2015	Wang et al. [51]	Textual metadata	Sentiment Polarity
2015	Campos et al. [52]	Raw image	Sentiment Polarity
2016	Sun et al. [54]	Image salient regions	Sentiment Polarity
2016	Katsurai et al. [55]	Hand crafted visual features & textual metadata	Sentiment Polarity
2017	Yang et al. [57]	Raw image	Distribution of Emotions
2017	Yang et al. [58]	Raw image	Distribution of Emotions
2017	Campos et al. [59]	Raw image	Sentiment Polarity
2017	Vadicamo et al. [60]	Raw image	Sentiment Polarity
2018	Li et al. [62]	Text from the inferred ANPs [41]	Sentiment Polarity

Table 3.1: Summary of the most relevant publications on Visual Sentiment Analysis. While the early methods were mostly based on hand crafted visual features, more recent approaches exploit textual metadata and features learned directly from the raw image (i.e., CNN based representations).

3.3 Visual Sentiment Analysis Systems

This section provides a complete overview of the system design choices, with the aim to provide a comprehensive debate about each specific issue, with proper references to the state of the art.

3.3.1 How to represent the emotions?

Basically, the goal of a Visual Sentiment Analysis system is to determine the sentiment polarity of an input image (i.e., positive or negative). Several works aims to classify the sentiment conveyed by images into 2 (positive, negative) or 3 polarity levels (positive, neutral, negative) [37, 41, 49]. However, there are also systems that adopt more then 3 levels, such as the 5-level sentiment scheme used by Xu et al. [48] or the 35 "impression words" used by Hayashi et al. [64]. Beside the polarity estimation, there are systems that perform the sentiment classification by using a set of emotional categories, according to an established emotion model based on previous psychological studies. However, each emotional category usually corresponds to a positive or negative polarity [39]. Thus, these systems can be evaluated also for the task of polarity estimation. In general, there are two main approaches for emotion modelling:

- ***Dimensional:*** this model represents emotions as points in a 2 or 3 dimensional space. Indeed, as discussed in several studies [21, 23, 24, 25], emotions have three basic underlying dimensions: valence, arousal and control (or dominance). The valence dimension ranges from extreme pain or unhappiness to extreme happiness, and represents the attractiveness toward an item. Arousal is defined as a mental activity which values range from sleep to frantic excitement, and refers to the intensity of emotional activation (i.e., being mentally reactive and awake to stimuli). It can be considered as the degree of activity that is generated by a particular stimulus. Dominance is related to feelings of control with respect to a situation. The VAC (Valence Arousal control) 3D emotion space is shown in Figure 3.3. However, as can be seen from Figure 3.3, the control dimension has a small effect. Therefore, a 2D emotion space is often considered. This space is obtained by considering only the arousal and the

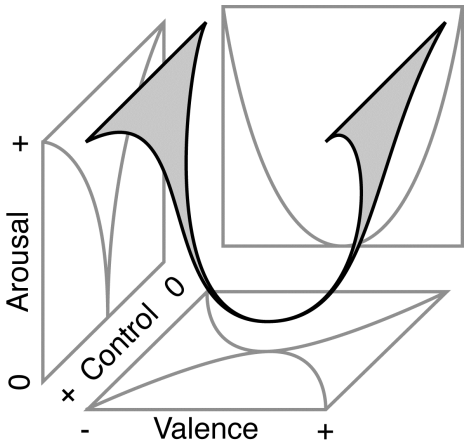


Figure 3.3: Illustration of the 3D emotion space (VAC space), image borrowed from [66].

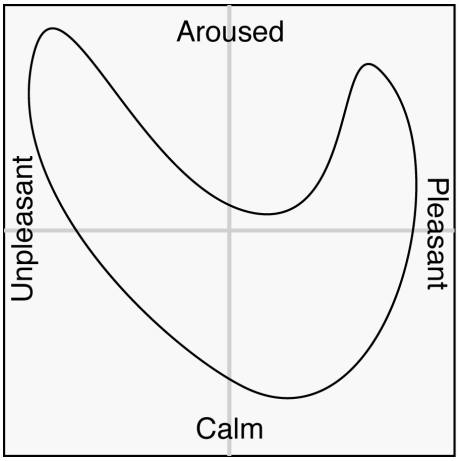


Figure 3.4: Illustration of the 2D emotion space (VA space) obtained considering just the arousal and the valence axis. Image borrowed from [66].

valence axis (Figure 3.4). Indeed, for example, Hanjalic et al. [65] considered the VA (Valence-Arousal) space to model the affective video content.

- ***Categorical approach:*** according to this model there are a number of basic emotions. Having just a few emotional categories is convenient for tasks such as indexing and classification. This set of descriptive words can be assigned to regions of the VAC (Valence-Arousal-Control) space. Thus it can be considered a quantized version of the dimensional approach. Considering such approach, there are several emotion models that can be defined. The choice of the emotional categories in not an easy task. Since emotion belongs to the psychology domain, the insights and achievements from cognitive science can be beneficial for this problem.

What are the basic emotions? There are several works that aim to ask this question. As observed in Section 3.2, the most adopted model is the Plutchnik's wheel of emotions [42] that defines 8 basic emotions with 3 valences each (see Figure 3.5). Thus it defines a total of 24 emotions:

- "ecstasy" → "joy" → "serenity"
- "admiration" → "trust" → "acceptance"

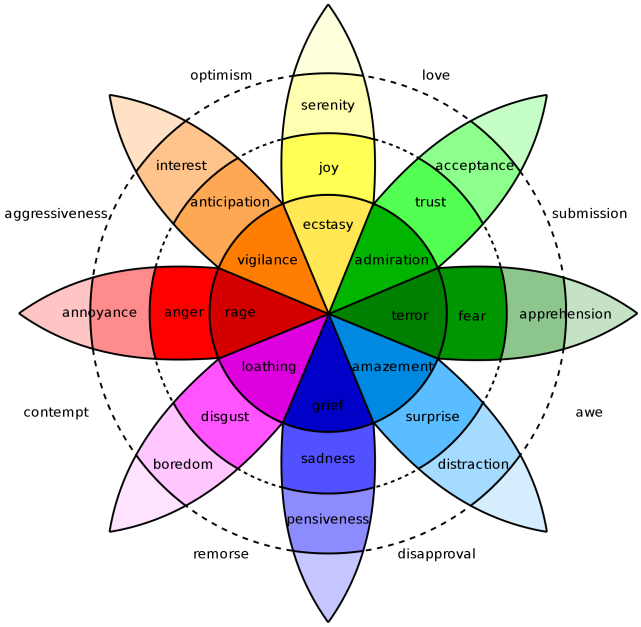


Figure 3.5: Plutchik's wheel of emotions [42].

- "terror" → "fear" → "apprehension"
- "amazement" → "surprise" → "distraction"
- "grief" → "sadness" → "pensiveness"
- "loathing" → "disgust" → "boredom"
- "rage" → "anger" → "annoyance"
- "vigilance" → "anticipation" → "interest"

According to Ekman's theory [45] there are just five basic emotions ("anger", "fear", "disgust", "surprise" and "sadness"). Another emotions categorization is the one defined in a psychological study by Mikell et al. [40] where the authors perform an intensive study on the *International Affective Picture System* (IAPS) in order to extract a categorical structure of such dataset [67]. As result, a subset of IAPS have been categorized in eight distinct emotions: "amusement", "awe", "anger", "contentment", "disgust", "excitement", "fear" and "sad". A deeper list of emotion

is described in Shaver et al. [68], where emotion concepts are organized in a hierarchical structure.

As discussed in this Section, there is a wide range of research on identification of basic emotions. By way of conclusion, the 24 emotions model defined in Plutchnik's theory [42] is a well established and used model. It is inspired by chromatics in which emotions are organized along a wheel scheme where bipolar elements are placed opposite one to each other. Moreover the three intensities provides a richer set of emotional valences. For these reasons it can be considered the reference model for the identification of the emotional categories.

3.3.2 Existing datasets

There are several sources that can be exploited to build Sentiment Analysis datasets. The general procedure to obtain a human labelled set of data is to perform surveys over a large number of people, but in the context of Sentiment Analysis the collection of huge opinion data can be alternatively obtained by exploiting the most common social platforms (Instagram, Flickr, Twitter, Facebook, etc.), as well as websites for collecting business and products reviews (Amazon, Tripadvisor, Ebay, etc.). Indeed, nowadays people are used to express their opinions and share their daily experiences through the Internet Social Platforms.

In the context of Visual Sentiment Analysis, one of the first published dataset is the International Affective Picture System (IAPS) [67]. Such dataset has been developed with the aim to produce a set of evocative color images that includes contents from a wide range of semantic categories. This work provides a set of standardized stimuli for the study of human emotional process. The dataset is composed by hundreds of pictures related to several scenes including images of insects, children, portraits, poverty, puppies and diseases, which have been manually rated by humans by means of affective words. The dataset has been used in [39] in combination with other two datasets built by the authors: a set of artistic photos (to investigate whether the choice of colors and textures chosen by the photographer affects the classification) and a set of abstract paintings which consist only of combinations of color and texture without any recognisable object. These datasets are publicly

Figure 3.6: Example images from the GAPED dataset [70].

available[1]. In [69] the authors considered a subset of IAPS extended with subject annotations to obtain a training set categorized in distinct emotions according to the emotional model described in [40] (see Section 3.3.1). However, the number of images of this dataset is very low. In [39], the authors presented the Affective Image Classification Dataset. It consists of two image sets: one containing 228 abstract painting and the other containing 807 artistic photos. These images have been labelled by using the 8 emotions defined in [40]. The authors of the dataset presented in [37] considered the top 1.000 positive and negative words in SentiWordNet [38] as keywords to search and crawl over 586.000 images from Flickr. The list of image URLs as well as the collected including title, image resolution, description and the list of the associated tags is available for comparisons [2].

The Geneva Affective Picture Database (GAPED) [70] dataset includes 730 pictures labelled considering negative (e.g., images depicting human rights violation scenes), positive (e.g., human and puppies) as well as neutral pictures which show static objects. All dataset images have been rated considering the valence, arousal, and the coherence of the scene also in this case. The dataset is available for research purposes [3]. Figure 3.6 shows two example photos of the GAPED dataset, depicting human rights violation and animal mistreatment.

In 2013 Borth et al. [41] proposed a very large dataset ($\sim$ 0.5 million) of pictures gathered from social media and labelled with ANP (Adjective Noun Pair) concepts. Furthermore, they proposed Twitter benchmark dataset which includes 603 tweets

[1] www.imageemotion.org

[2] http://www.l3s.de/ minack/flickr-sentiment

[3] http://www.affective-sciences.org/home/research/materials-and-online-research/research-material/

with photos. It is intended for evaluating the performance of automatic sentiment prediction using features of different modalities (text only, image only, and text-image combined). Such dataset has been used by most of the state-of-the-art works as evaluation benchmark for Visual Sentiment Analysis, especially when the designed approaches involve the use of Machine Learning methods such as in [43] and in [49] for instance, due to the large scale of this dataset. The Emotion6 dataset, presented and used in [50], has been built considering the Elkman's 6 basic emotion categories [45]. The number of images is balanced over the considered categories and the emotions associated with each image is expressed as a probability distribution instead of as a single dominant emotion. In [49] You et al. proposed a dataset with 1,269 Twitter images labelled into positive or negative by 5 different annotators. Given the subjective nature of sentiment, this dataset has the advantage to be manually labelled by human annotators, differently than other datasets that have been created collecting images by automatic systems based on textual tags or predefined concepts such as the VSO dataset used in [41]. In [55] two large sets of social pictures from Instagram and Flickr (CrossSentiment) have been crawled. The list of labelled Instagram and Flickr image URLs is available on the Web [4].

Vadicamo et al. [60] crawled $\sim 3M$ tweets from July to December 2016. The collected tweets have been filtered considering only the ones written in English and including at least an image. The sentiment of the text extracted from the tweets has been classified using a polarity classifier based on a paired LSTM-SVM architecture. The data with the most confident prediction have been used to determine the sentiment labels of the images in terms of positive, negative and neutral. The resulting Twitter for Sentiment Analysis dataset (T4SA) consists of $\sim 1M$ tweets and related $\sim 1.5M$ images.

Datasets such as GAPED and IAPS rely on emotion induction. This kind of datasets are very difficult to be built in large scale and maintained over time. The Machine Learning techniques and the recent Deep Learning methods are able to obtain impressive results as long as these systems are trained with very large scale datasets (e.g., VSO [41]). Such datasets can be easily obtained by exploiting the social network platforms by which people share their pictures every day. These datasets allowed the extensive use of Machine Learning systems that requires large

[4] http://mm.doshisha.ac.jp/senti/CrossSentiment.html

Table 3.2: Main benchmark datasets for Visual Sentiment Analysis. Some datasets contains several additional information and annotations.

Year	Dataset	Size	Labelling	Social Media	Polarity	Additional Metadata
1999	IAPS [67]	716 photos	Pleasure, arousal and dominance	✗	✗	✗
2005	Mikels et al. [40]	369 photos	Awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
2010	Affective Image Classification Dataset [39]	228 paintings 807 photos	Awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
2010	Flickr-sentiment [37]	586.000 Flickr photos	Positive, negative.	✓	✓	✓
2011	GAPED [70]	730 pictures	Positive, negative, neutral.	✗	✓	✗
2013	VSO [41]	0,5 M Flickr Photos 603 Twitter Images	- Adjective-Noun Pairs - Positive or negative	✓	✓	✓
2015	Emotion6 [50]	1.980 Flickr photos	- Valence-Arousal score - 7 emotions distribution	✓	✗	✗
2015	You et al. [49]	1.269 Twitter images	Positive, negative.	✓	✓	✗
2016	CrossSentiment [55]	90.139 Flickr photos 65.439 Instagram images	Positive, negative, neutral.	✓	✓	✗
2017	T4SA [60]	1,5 M Twitter images	Positive, negative, neutral.	✓	✓	✗

scale datasets. This furthered the building of very large datasets such as T4SA in the last few years. Table 3.2 summarizes the main dataset just reported with details about the number of images, the source (e.g., Social Platform, paintings, etc.) and the labelling options.

3.3.3 Features

One of the most difficult step driving the design of a Visual Sentiment Analysis system, and in general for the design of a data analysis approach is the selection of the data features that better encode the information that the system is aimed to infer. Image features for Visual Sentiment Analysis can be categorized within three levels:

- ***Low-level*** **-** These features describe distinct visual phenomena in an image mainly related in some way to the color values of the image pixels. They usually includes generic features such as color histograms, HOG, GIST. In the context of Visual Sentiment Analysis, previous works can be exploited to extract particular low-level features derived from proper studies on art and perception theory. These studies suggest that some low-level features, such as colors and texture can be used to express the emotional effect of an image [39].

- ***Mid-level -*** This group of features bring more semantic, thus they are more interpretable and have stronger associations with emotions [71]. One example is given by the scene-based 102-dimensional feature defined in [43]. Furthermore, many of the aforementioned works on Visual Sentiment Analysis exploit the 1200-dimensional mid-level representation given by the 1200 Adjective-Noun Pairs (ANP) classifiers defined by Borth et al. [41].

- ***High-level -*** These features describe the semantic concepts shown in the images. Such a feature representation can be obtained by using pre-trained classification methods or semantic embeddings [55].

In 2010 Machajdik and Hanbury [39] performed an intensive study on image emotion classification by properly combining the use of several low and high visual features. These features have been obtained by exploiting concepts from and art theory [22, 26], or exploited in image retrieval [72] and image classification [27, 19] tasks. They selected 17 visual features, categorized in 4 groups:

- **color:** mean saturation and brightness, 3-dimensional emotion representation by Valdez et al. [26], hue statistics, colorfulness measure according to [27], number of pixels of each of the 11 basic colors [73], Itten contrast [22], color histogram designed by Wang Wei-ning et al. [19];

- **texture:** wavelet textures for each HSB channel, features by Tamura et al. [74], and features based on GLCM (i.e., correlation, contrast, homogeneity, and energy for the HSB channels);

- **composition:** the number of resulting segments obtained after the application of a waterfall segmentation (denoted as "level of detail" in [39]), depth of field (DOF) [27], statistics on the line slopes by using the Hough transform (denoted as "dynamics"), rule of thirds;

- **content:** number of detected front faces, number of the biggest face pixels, count of skin pixels, ratio of the skin pixels over the face size.

Most of the mentioned works combine huge number of hand-crafted visual features. Although all the exploited features have been proven to have a direct influence on the perceived emotion by previous studies, there is not agreement about which

of them give the most of the contribution on the aimed task. Besides the selection of proper hand-crafted features, designed with the aim to encode the sentiment content conveyed by images, there are other kind of approaches that lean on representation learning techniques based on Deep Learning [46, 48, 49]. By employing such representation methods, image features are learned from the data. This avoid the designing of a proper set of feature, because the system automatically learns how to extract the needed information from the input data. These methods requires huge amounts of labelled training data, and an intensive learning phase, but obtain better performances in general.

Another approach, borrowed from the image retrieval methods, consists on combining textual and visual information through multimodal embedding systems [55]. In this case, features taken from different modalities (e.g., visual, textual, etc.) are combined to create a common vector space in which the correlations between projections of the different modalities are maximized (i.e., an embedding space).

So far, there is not an established strategy to select of visual features that allows to address the problem. Most of the previous exploited features demonstrated to be useful, but recent results on Visual Sentiment Analysis suggest that it's worth investigating the use of representation learning approaches such as Convolutional Neural Networks and multimodal embedding.

3.4 Problem analysis

In this section we propose a formulation of the problem, which highlights the related issues and the key tasks of Visual Sentiment Analysis. This allows to better focus the related sub-issues which form the Visual Sentiment Analysis problem and support the designing of more robust approaches. Moreover, to address the overall structure of the problem is useful to suggest a common framework helping researchers to design more robust approaches. Starting from the definition of the Sentiment Analysis problem applied to the natural language text given by Liu [75], we propose to generalize the definition in the context of Visual Sentiment Analysis.

Text based Sentiment Analysis can be performed considering different levels of detail:

- at the **document level** the task is to classify whether a whole document (i.e., the whole input) expresses a positive or negative sentiment. This model works on the underling assumption that the whole input discusses only one topic;
- at the **sentence level** the task is to find each phrase within the input document and determine if each sentence express a positive or negative (or neutral) sentiment;
- the **entity and aspect level** performs finer-grained analysis by considering all the opinions expressed in the input document and defining a sentiment score (positive or negative) for each detected target.

Similarly, if the subject of the analysis is an image, we can:

- consider a Sentiment Analysis evaluation for the whole image. These systems work with global image features (e.g., color histograms, saturation, brightness, colorfulness, color harmony, etc.);
- consider an image as a composition of several sub-images according to its specific content. A number of sub-images is extracted and the sentiment analysis is performed on each sub-image obtained by exploiting methods such as multi-object detection, image segmentation, objectness extraction [76];
- define a set of image aspects, in terms of low level features, each one associated to a sentiment polarity based on previous studies [39, 69]. This is essentially the most fine-grained analysis to be considered.

When a system aims to perform Sentiment Analysis on some textual content, basically it is looking for the opinions in the content and extracting the associated sentiment. An opinion consists of two main components: a target (or topic), and a sentiment. The opinions can be taken from more than one person, this means that the system has to take into account also the opinion holder. Furthermore, opinions can change over time, thus also the time an opinion is expressed has to be taken into account. According to Liu [75], an opinion (or sentiment) is a quintuple

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l) \tag{3.1}$$

where e_i is the name of an entity i, a_{ij} is an aspect j related to e_i, h_k is the k-th opinion holder, and t_l is the time l when the opinion is expressed by the opinion holder h_k. Finally, s_{ijkl} is the sentiment score with respect to the aspect a_{ij}, expressed by the k-th holder at time t_l. In the above definition the subscripts are used to highlight that all the components must correspond. That is, if one of the five component changes it defines a new quintuple. The sentiment score s_{ijkl} can be expressed in terms of polarity, considering positive, negative or neutral polarity; or with different levels of intensity. The special aspect "GENERAL" is used when the sentiment is expressed for the whole entity. In this case, either the entity e_i and the aspect a_{ij} represent the opinion target.

This definition is given in the context of opinion analysis applied on textual contents which express positive or negative sentiments. In the case of Sentiment Analysis applied on visual contents there are some differences. Indeed, when the input is a text, Sentiment Analysis can easily lean on context and semantic information extracted directly from the text. Thus the problem is to be considered into the NLP (Natural Language Processing) domain. When the input is an image, because of the *affective gap* between visual content representations and semantic concepts such as human sentiments, the task to associate the visual features with sentiment labels or polarity scores results challenging. Such *affective gap* can be defined as:

"the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal", A. Hanjalic [77].

In the following paragraphs each of the sentiment components previously defined (i.e., entity, aspect, holder and time) are exploited in the context of Visual Sentiment Analysis.

3.4.1 Entity and Aspects

The entity is the subject (or target) of the analysis. In the case of Visual Sentiment Analysis the entity is the input image. In general, an entity can be viewed as a set of "parts" and "attributes". The set of the entity's parts, its attributes, plus the special aspect "GENERAL" forms the set of the aspects (see Figure 3.7).

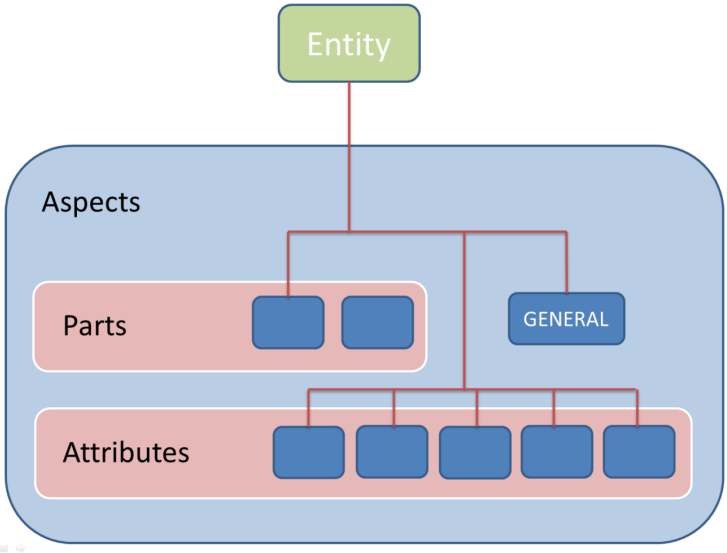


Figure 3.7: Relationship between an entity and its aspects.

This structure can be transferred to the visual domain considering different levels of visual features. Indeed, as mentioned above, in the case of visual contents, Sentiment Analysis can be performed considering different level of visual detail. The most general approach performs Sentiment Analysis considering the whole image, this corresponds to apply a Visual Sentiment Analysis method on the "GENERAL" aspect. The parts of an image can be defined by considering a set of sub-images. This set can be obtained by exploiting several Computer Vision techniques, such as background/foreground extraction, image segmentation, multi object recognition or dense captioning [78, 79]. The attributes of an image regards its aesthetic quality features, often obtained by extracting low-level features. Exploiting this structured image hierarchy, a sentiment score can be inferred for each aspect. Finally, the partial scores can be properly combined to obtain the sentiment classification (e.g., data can be used as input features of a regression model). As an example, the Figure 3.8 shows an image related to a dish with pancakes and some fruit. The sentiment associated to this image could be inferred by considering the input from different perspectives. Considering the whole image (i.e., the GENERAL aspect), the inherent context expresses the concept of "breakfast", or "food" in general. From this perspective one can consider the concept associated to the image context. For this purpose, several works about personal contexts [80, 81] and scene recognition can

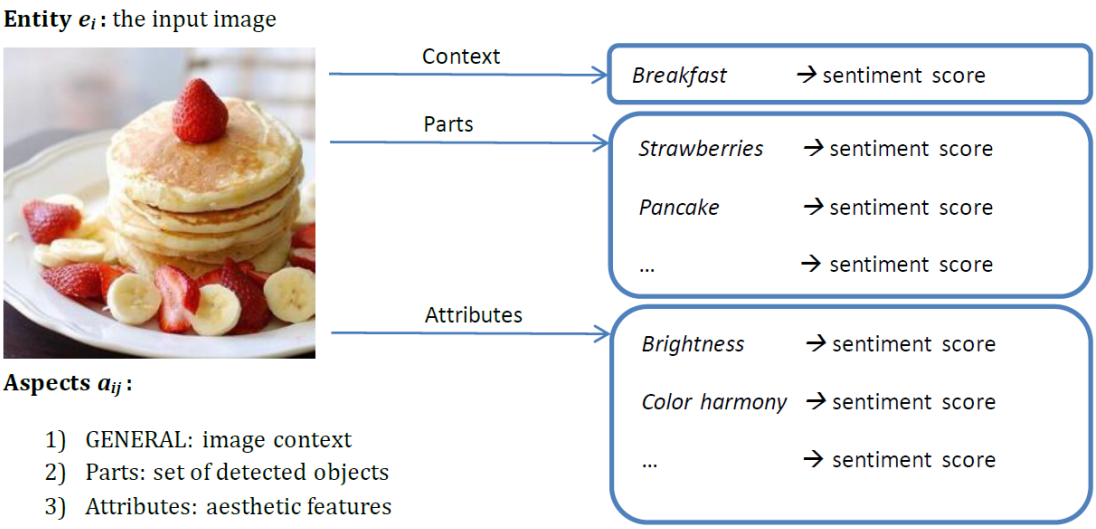


Figure 3.8: Example of the different scores that can be extracted from an image. The entity e_i is the input image, the context (i.e., breakfast), the parts (i.e., objects in the scene) and the attributes (i.e., aesthetic features) represent the different aspects. Each aspect a_{ij} is associated to a sentiment score s_{ijkh}. The sentiment scores can be expressed in terms of polarity (e.g., positive or negative) or by considering different levels of strength (e.g., a score from 1 to 5).

be exploited from the visual view, and the inferred concepts can be used to extract the associated sentiment. Moreover, sentiment scores can be further extracted from image parts and attributes, according to the model described above.

Instead of representing the image parts as a set of sub-images, an alternative approach can rely on a textual description of the depicted scene. The description of a photo can be focused on a specific task of image understanding. By changing the task, we can obtain multiple descriptions of the same image from different points of view. Then, these complementary concepts can be combined to obtain the above described structure. Most of the existing works in analysing social media exploit textual information manually associated to images by performing textual Sentiment Analysis. Although the text associated to social images is widely exploited in the state-of-the-art to improve the semantics inferred from images, it can be a very noisy source because it is provided by the users; the reliability of such input is often based on the capability and the intent of the users to provide textual data that are coherent with respect to the visual content of the image. There is no guarantee that the subjective text accompanying an image is useful. Moreover, the tags associated

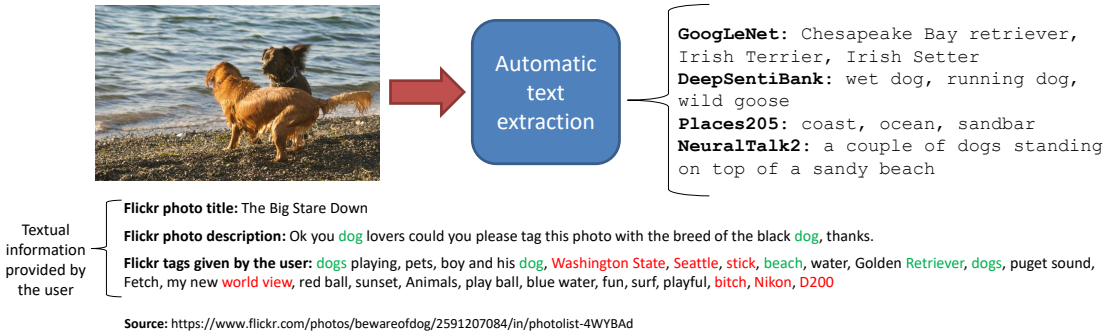


Figure 3.9: Given an image, the text describing the visual content can be extracted by exploiting four different deep learning architectures. The considered architectures are used to extract text related to objects, scene and image description. The figure shows also the text associated to the image by the user (i.e., title, description and tags) at top left. The subjective text presents very noisy words which are highlighted in red. The words that appears in both sources of text are highlighted in green.

to social images are often selected by users with the clear goal to maximize the visibility. For a deeper analysis, a comprehensive treatise of image tag assignment is presented in [82].

As discussed in [83], the semantic of an image can be expressed by means of an object category (i.e., a class). However the tags provided by users usually include several additional terms, related to the object class, coming from a larger vocabulary. As an alternative, the semantic could be expressed by using multiple keywords corresponding to scenes, object categories, or attributes.

Figure 3.9 shows an example image taken from Flickr. The textual information below the image is the text provided by the Flickr's user. Namely the photo title, the description and the tags are usually the text that can be exploited to make inferences on the image. This example shows how the text can be very noisy with respect to any task aimed to understand the sentiment that can be evoked by the picture. Another drawback of the text associated to social images is that two users can provide rather different information about the same picture, either in quality and in quantity. Finally, there is not guarantee that such text is present; this is an intrinsic limit of all Visual Sentiment Analysis approaches exploiting subjective text.

Starting from the aforementioned observations about the user provided text associated to social images, one can exploit an objective aspect of the textual source

that comes directly from the understanding of the visual content. This text can be achieved by employing a set of deep learning models trained to accomplish different visual inference tasks on the input image. At the top right part of Figure 3.9 the text automatically extracted with different scene understanding methods is shown. In this case, the inferred text is very descriptive and each model provides distinctive information related to objects, scene, context, etc. The objective text extracted by the three different scene understanding methods has a pre-defined structure, therefore all the images have the same quantity of textual objective information. In Section 3.6 we present our work on Image Polarity Prediction exploiting Objective Text extracted directly from images, and experimentally compare such text with respect to the Subjective (i.e., user provided) text information usually used in previous works. Such approach provides an alternative user-independent source of text which describes the semantic of images, useful to address the issues related to the inherent subjectivity of the text associated to images. Indeed, several papers faced the issues related to the subjective text associated to images, such as tag refinement and completion [82, 84, 85, 86], which aims at alleviating the number of noisy tags and enhancing the number of informative tags by modelling the relationship between visual content and tags.

3.4.2 Holder

Emotions are subjective, they are affected by several factors such as gender, individual background, age, environments, etc. However, emotions also have the property of stability [87]. This means that the average emotion response of a statistically large set of observers is stable and reproducible. The stability of emotion response enables researchers to generalize their results, when obtained on large datasets.

Almost all the works in Visual Sentiment Analysis ignore the sentiment holder, or implicitly consider only the sentiment of the image publisher. In this context at least two holders can be defined: the image owner and the image viewer. Considering the example of an advertising campaign, where the owner is the advertising company and the viewers are the potential customers, it's crucial the study and analysis the connection between the sentiment intended by the owner and the actual sentiment induced to the viewers.

These days, the social media platforms provide a very powerful mean to retrieve real-time and large scale information of people reactions toward topics, events and advertising campaigns. The work in [47] is the first that distinguishes publisher affect (i.e., intent) and viewer affect (i.e., reaction) related to the visual content. This branch of research can be useful to understand the relation between the affect concepts of the image owner and the evoked viewer ones, allowing new user centric applications. User profiling helps personalization, which is very important in the field of recommendation systems. The insights that could be obtained from such a research branch can be useful for several business fields, such as advertisement and User Interface design (UI). And the community of user interface designers started to take into account the emotional effect of the user interfaces toward users who are interacting with a website, product or brand. The work in [88] discusses about methods to measure user's emotion during an interface interaction experience, with the aim to assess the interface design emotional effect. Progresses in this field promote the definition of new design approaches such as Emotional UI [89], aimed to exploit the emotions conveyed by visual contents. Indeed, emotions have been traditionally considered to be something that the design evoked, now they represent something that drives the design process. While so far, designers focused on "user friendly" design (i.e., interfaces easy to use), now they need to focus on design that stimulates and connects the product with users deeply.

Although the interesting cues in this paragraph, currently the development of Visual Sentiment Analysis algorithms that concern the sentiment holder find difficulties due the lack of specific datasets. In this context, the huge data shared on the social media platforms can be exploited to better understand the relationships between the sentiment of the two main holders (i.e., owner/publisher and viewer/user) through their interactions.

3.4.3 Time

Although almost all the aforementioned works ignore this aspect, the emotion evoked by an image can change depending on the time. This sentiment component can be ignored the most of times, but in specific cases is determinant. Moreover, is very difficult to collect a dataset related to the changes in the emotion evoked by images

Figure 3.10: Picture of the World Trade Center, taken in 1990.

over time. For example, the sentiment evoked by an image depicting the World Trade Center (Figure 3.10) is presumably different if the image is shown before or after 9/11.

Although there are not works on Visual Sentiment Analysis that analyse the changes of image sentiments over time, due to the specificity of the task and the lack of image datasets, there are several works that exploits the analysis of images over time focused on specific cognitive and psychology applications. As an example, the work in [90] employed a statistical framework to detect depression by analysing the sequence of photos posted on Instagram. The findings of this paper suggest the idea that variations in individual psychology reflect in the use social media by the users, hence they can be computationally detected by the analysis of the user's posting history.

In [91] the authors studied which objects and regions of an image are positively or negatively correlated with memorability, allowing to create memorability maps for each image. This work provides a method to estimate the memorability of images from many different classes. To collect human memory scores, the adopted experimental procedure consists of showing several occurrences of the same images at variable time intervals. The employed image dataset has been created by sampling images from a number of existing dataset, including images evoking emotions [39].

3.5 Challenges

So far we discussed on the current state of the art in Visual Sentiment Analysis, describing the related issues, as well as the different employed approaches and features. This section aims to introduce some additional challenges and techniques that can be investigated.

3.5.1 Popularity

One of the most common application field of Visual Sentiment Analysis is related to social marketing campaigns. In the context of social media communication, several companies are interested to analyse the level of people engagement with respect to social posts related to their products. This can be measured as the number of post's views, likes, shares or by the analysis of the comments. These information can be further combined with web search engine and companies website visits statistics, to find correlations between social advertising campaigns and their aimed outcomes (e.g., brand reputation, website/store visits, product dissemination and sale, etc.) [92, 93, 94].

The level of engagement of an image posted on a social network is usually referred as "Image Popularity". It is a difficult index to measure or predict, however human beings are able to predict what visual contents other people will like in specific contexts (e.g., marketing campaigns, professional photography). This suggests that there are some common appealing factors in images. So far, researches have been trying to gain insights into what features make an image popular.

As mentioned in the previous sections, the text associated to images is often pre-processed in order to avoid noisy text. On the other hand, the users who want to increase the reach of their published contents are used to associate popular tags to their images, regardless their relevance with the image content. This is motivated by the fact that image associated tags are used by the image search engines in these platforms. Thus, the use of popular tags rises the reach of the pictures. Therefore, in this context, the tags associated to images become a crucial factor to understand the popularity of images in social platforms. For instance, Yamasaki et al. [95] proposed an algorithm to estimate the social popularity of images uploaded on Flickr by using only text tags. Other features should be also taken into account such as:

number of user followers and groups, which represent the reach capability of the user. These factors make the task of popularity prediction very different from the task of sentiment polarity classification in the selection of features, methods and measures of evaluation.

In 2014, Khosla et al. [34] proposed a log-normalized popularity score that has been then commonly used in the community. Let c_i be a measure of the engagement achieved by a social media item (e.g., number of likes, number of views, number of shares, etc.), also known as popularity measure. The popularity score of the i^{th} item is defined as follows:

$$score_i = \log\left(\frac{c_i}{T_i} + 1\right) \tag{3.2}$$

where T_i is the number of days since the uploading of the image on the Social Platform.

Although the task of image popularity prediction is rather new, there are interesting datasets available for the development of massive learning systems (i.e., deep neural networks). The Micro-Blog Images 1 Million (MBI-1M) dataset is a collection of 1M images from Twitter, along with accompanying tweets and metadata. The dataset was introduced by the work in [96]. A subset of the the Trec 2013 micro-blog track tweets collection [97] has been selected.

The MIR-1M dataset [98] is a collection of 1M photos from Flickr. These images have been selected considering the interestingness score used by Flickr to rank images.

The Social Media Prediction (SMP) dataset is a large-scale collection of social posts, recently collected for the ACM Multimedia 2017 SMP Challenge [5]. This dataset consists of over 850K posts and 80K users, including photos from VSO [41] as well as photos collected from personal users' albums [99, 100, 101]. In particular, the authors aimed to record the dynamic variance of social media data. Indeed, the social media posts in the dataset are obtained with temporal information (i.e., posts sequentiality) to preserve the continuity of post sequences. Two challenges have been proposed:

[5]Challenge webpage: https://social-media-prediction.github.io/MM17PredictionChallenge

- **Popularity Prediction:** the task is to predict a popularity measure defined for the specific social platform (e.g., number of photo's views on Flickr) of a given image posted by a specific user;
- **Tomorrow's Top Prediction:** given a set of photos and the data related to the past photo sharing history, the task is to predict the top-n popular posts (i.e., ranking problem over a set of social posts) on the social media platform in the next day.

The SMP dataset includes features such as unique picture id (pid) and associated user id (uid). From these information one can extract almost all the user and photo related data available in Flickr. Some metadata of the picture and user-centered information are also included in the dataset. Moreover, the popularity scores (as defined in Equation 3.2) are provided. The SMP dataset furthered the development of time aware popularity prediction methods, which exploit time information to define new image representation spaces used to infer the image popularity score at a precise time or at pre-defined time scales.

Popularity Dynamics

Equation 3.2 normalizes the number of interactions reached by an image by dividing the engagement measure by the time. However, the measures c_i related to social posts are cumulative values as they continuously collect the interactions between users and the social posts during their time on-line. Therefore, this normalization will penalize social media contents published in the past with respect to more recent contents, especially when the difference between the dates of posting is high. Indeed, the most of the engagement obtained by a social media item is achieved in the first period, then the engagement measures become more stable. For example, the study presented in [102] shows that photos obtain most of their engagement within the first 7 days since the date of upload. However, this study is focused on Flickr, and each social platform has its own mechanisms to show contents to users. There are very few works which takes into account the evolution of the image popularity over time (i.e., the dynamics of the image popularity). In Section 3.7 we present our work on predicting the popularity dynamics of photos published on Flickr, as well as a detailed description of the state of the art related to the task of Image Popularity

Prediction. The proposed method is able to predict the popularity score evolution with a daily granularity over time. Furthermore, we introduce a new large dataset obtained by tracking the engagement scores of $\sim$ 20K photos shared on Flickr that allows to investigate and tackle this new challenging task.

3.5.2 Image Virality

A recent emerging task, closely related to image popularity, is the prediction of the level of virality of an image. The image virality is defined as the quality of a visual content (i.e., images or videos) to be rapidly and widely spread on social networks [103]. Differently than popularity, the virality score takes into account also the number of resubmission of an image by different users. Therefore, images that became popular when they are posted, but not reposted, are not considered to be viral [104]. These often involve images which content itself is less relevant, but are related to current events that drawn attention to the image in a specific period such as a flash news, or a tragedy. The work in [103] focused on understanding the influence of image parts on its virality. In particular, the authors presented a method for the task of simultaneously detection and localicazion of virality in images. The detection consists on the prediction of the virality score on an image. The localization aims to detect which areas in an image are responsable for making the image viral, this allows to produce an heatmap which highlights the relevant areas of the input image.

3.5.3 Relative Attributes

As discussed in previous sections, several Visual Sentiment Analysis works aim to associate an image one sentiment label over a set of emotional categories or attributes. However, given a set of images that have been assigned to the same emotional category (e.g., joy), it would be interesting to determine their ranking with respect the specific attribute (see Figure 3.11). Such a technique could suggest, for example, if a given image A conveys more "joy" than another image B. For this purpose, several works on relative attributes can be exploited [105, 106, 107, 108]. Furthermore, a ground truth dataset can be built by exploiting human annotators. Given a pair of

Figure 3.11: Example of images ranking based on the emotional category "joy".

images, the annotator is requested to indicate which image is closer to the attribute. In this way it's possible to obtain a proper ranking for each sentiment attribute.

3.5.4 Common Sense

With the aim to reduce the affective and cognitive gap between images and sentiments conveyed by them, we further need to encode the "affective common-sense". An Halloween picture can be classified as a negative image by an automatic system which considers the image semantics, however the knowledge of the context (i.e., Halloween) should affect the semantic concepts conveyed by the picture, hence its interpretation. This corresponds to the "common-sense knowledge problem" in the field of knowledge representation, which is a sub-field of Artificial Intelligence. Clearly, besides inferential capabilities, such an intelligent program needs a representation of the knowledge. By observing that is very difficult to build a Sentiment Analysis system that may be used in any context with accurate classification prediction, Agrawal et al. [109] considered contextual information to determine the sentiment of text. Indeed, in this paper is proposed a model based on common-sense knowledge extracted from ConceptNet [110] ontology and context information. Although this work addresses the problem of Sentiment Analysis applied on textual data, as discussed above, the knowledge of the context related to what an image is depicting should affect its interpretation. Moreover, such results on textual analysis can be transferred to the visual counterpart. Furthermore, emerging approaches based on the Attention mechanism could be exploited to add such a context. The Attention mechanism is a recent trend in Deep Learning, it can be viewed as a method for making the Artificial Neural Network work better by letting the network know where to look as it is performing its task. For example, in the task of

image captioning, the attention mechanism tells the network roughly which pixels to pay attention to when generating the text [111, 112].

3.5.5 Emoticon/Emoji

In this section we discuss about the possibility to exploit text ideograms, such emoticons and emoji, in the task of Sentiment Analysis on both visual and textual contents. An emoticon is a textual shorthand that represents a facial expression. The emoticons have been introduced to allow the writer to express feelings and emotions with respect to a textual message. It helps to express the correct intent of a text sentence, improving the understanding of the message. The emoticons are used to emulate visual cues in textual communications with the aim to express or explicitly clarify the writer's sentiment. Indeed, in real conversations the sentiment can be inferred from visual cues such as facial expressions, pose and gestures. However, in textual based conversations, the visual cues are not present.

The authors of [113] tried to understand if emoticons could be useful as well on the textual Sentiment Analysis task. In particular, they investigated the role that emoticons play in conveying sentiment and how they can be exploited in the field of Sentiment Analysis. The authors manually labelled 574 emoticons as positive or negative, and combined this emoticon-lexicon with the text based Sentiment Analysis to perform document polarity classification considering both sentence and paragraph levels.

A step further the emoticon, is represented by the emoji. An emoji is an ideogram representing concepts such as weather, celebration, food, animals, emotions, feelings, and activities, besides a large set of facial expressions. They have been developed with the aim to allow more expressive messages. Emojis have become extremely popular in social media platforms and instant messaging systems. For example, in March 2015, Instagram reported that almost half of the texts on its platform contain emojis [114].

In [115], the authors exploited the expressiveness carried by emoji, to develop a system able to generate an image content description in terms of a set of emoji. The focus of this system is to use emoji as a means for image retrieval and exploration. Indeed, it allows to perform an image search by means of a emoji-based query. This approach exploits the expressiveness conveyed by emoji, by leaning on the textual

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name	Unicode block
		0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY	Emoticons
		0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART	Dingbats
		0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT	Miscellaneous Symbols
		0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES	Emoticons
		0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
		0x1f618	3648	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS	Emoticons
		0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES	Emoticons
		0x1f44c	2925	0.805	0.094	0.249	0.657	0.563		OK HAND SIGN	Miscellaneous Symbols and Pictographs
		0x1f495	2400	0.766	0.042	0.285	0.674	0.632		TWO HEARTS	Miscellaneous Symbols and Pictographs

Figure 3.12: Some examples of the *Emoji Sentiment Ranking* scores and statistics obtained by the study conducted in [117]. The sentiment bar (10th column) shows the proportion of negativity, neutrality and positivity of the associated emoji.

description of these ideograms (see the eleventh column in Figure 3.12). The work in [116] studied the ways in which emoji can be related to other common modalities such as text and images, in the context of multimedia research. This work also presents a new dataset that contains examples of both text-emoji and image-emoji relationships.

Most of them contains also strong sentiment properties. In [117] the authors presented a sentiment emoji lexicon named Emoji Sentiment Ranking. In this paper, the sentiment properties of the emojis have been deeply analyzed, and some interesting conclusions have been highlighted. For each emoji, the *Emoji Sentiment Ranking* provides its associated positive, negative and neutral scores. These scores are represented by decimal values between -1 and +1. The authors also proposed a visual tool, named sentiment bar, to better visualize the sentiment properties associated to each emoji (see Figure 3.12). The data considered in this analysis consists of 1.6 million labelled tweets. This collection includes text written in 13 different languages. The authors found that the sentiment scores and ranking associated to emojis remain stable among different languages. This property is very useful to overcome the difficulties addressed in multilingual contexts. This lexicon represents

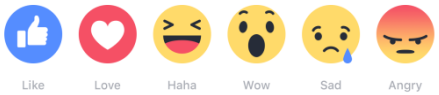


Figure 3.13: Facebook emoji reactions: Like, Love, Haha, Wow, Sad, and Angry.

a precious resource for many useful applications [6].

The results and the insights obtained in [115, 116] and [117] could be combined to exploit the sentiment conveyed by emoji on the task of Visual Sentiment Analysis. Indeed, as the most common systems lean on the text associated to images to obtain the corresponding sentiments, it worth to investigate if the sentiment conveyed by emoji can improve the performances of such systems. To this aim, an image dataset with emoji annotation can be defined, by asking people to select a set of meaningful emojis to express the sentiment evoked by a given image. This dataset, combined with the sentiment insights obtained in [117], can be exploited to build systems able to better predict the sentiment evoked by images. For instance, an image could be represented by considering the distribution of the associate emojis as a sentiment feature, taking a cue from the approach presented in [50].

By a few years, Facebook has released a new "reactions" feature, which allows users to interact with a Facebook post by using one of six emotional reactions (Like, Love, Haha, Wow, Sad, and Angry) , instead of just having the option of "liking" a post. These reactions corresponds to a meaningful subset of emoji (see Figure 3.13).

[6]The Emoji Sentiment Ranking scores computed by [117] can be visualized at the following URL: http://kt.ijs.si/data/Emoji_sentiment_ranking/.

3.6 Image Polarity Prediction

3.6.1 Introduction

The rise of social media has opened new opportunities to better understand people's interests towards topics, brands or products. Social media users continuously post images, opinions and share their emotions. This trend has supported the growing of new Machine Learning application areas, such as semantic-based image selection from crowdsourced collections [4, 118], Social Event Analysis [119] and Sentiment Analysis on Visual Contents [46]. As well as the definition of new approaches to address classic tasks such as products rating prediction [120] and election forecasting [121], based on the Web contents publicly shared by users. Early methods in Visual Sentiment Analysis focused only on visual features [39, 43] (ignoring the text associated to the images) or have employed text to define a sentiment ground truth [37, 41]. More recent approaches exploit a combination of visual and text features in different ways. Most of them, consider SentiWordNet [38] and the WordNet [122] lexicons as external knowledge to extract useful semantic information from textual data. In particular, SentiWordNet provides three types of sentiment polarity scores for each word defined in WordNet [122], and describes quantitatively score.

In this Section we present a method which exploits the text (automatically extracted from images) to build an embedding space where the correlation among visual and textual features is maximized. Some previous works in literature define models which learn a joint representation over multimodal inputs (i.e., text, image, video, and audio) to perform Image Classification [123], Visual Sentiment Analysis [55], image retrieval [83, 124, 125, 126, 127, 128, 129, 130], and event classification [131] by exploiting social media contents. The text associated to images is typically obtained by considering the meta-data provided by the user (e.g., image title, tags and description). Differently than previous approaches, our framework describes images in an "objective" way through state-of-the-art scene understanding methods [8, 79, 132]. Since the text describing the images is automatically inferred, in our approach, we denote it as "objective" emphasizing the fact that it is different to the "subjective" text written by the user for describing and/or commenting a visual content (i.e., image) of a social media post.

In [55] two different datasets are considered, by crawling public images from

Instagram and Flickr respectively. To represent contents for sentiment analysis estimation, the authors proposed three different type of features extracted considering pairs of images and the related subjective texts: a visual feature defined by combining different visual descriptors usually used for visual classification [83, 127, 133], a feature obtained by using the traditional Bag of Words approach on the subjective text, and a sentiment feature obtained by selecting the words of the subjective text whose sentiment scores (positive or negative) reported in SentiWordNet [38] are larger than a threshold, and applying the Bag of Words on this restricted vocabulary. These three types of features, called views, are then combined to form an embedding space by using multi-view Canonical Correlation Analysis (CCA) [134]. The aforementioned features projected to the computed embedding space are then exploited to train a binary classifier which is used to infer the final positive or negative sentiment (i.e., sentiment polarity).

Although the subjective text associated to social images can be exploited as additional source of input to infer the sentiment polarity of an image, two different users could associate very different texts to the same image. This makes the features extracted from the subjective text prone to be noisy. Indeed, in the definition of the dataset used in [55] the authors observed that the tags associated to social images can be very noisy, and for this reason they avoided to exploit the textual data for the definition of the sentiment ground truth of their experimental dataset (i.e., they decided to build the ground truth by manual labelling). The authors compared a pool of representations usually used for the task of Visual Sentiment Analysis [37, 41, 51, 135] mainly based on hand-crafted visual features and textual information provided by users (i.e., subjective), using the same evaluation protocol and the same dataset. Differently than [55], we built the textual and sentiment views by exploiting the objective text as input instead of the subjective text provided by users, with the aim to assess the benefits of using the proposed source of text in lieu of the text commonly used in previous approaches. To this aim, we exploited four state of the art deep learning architectures to automatically extract the objective text from the input images. To further assess the effectiveness of our approach, we have also considered different combinations of subjective, objective text and visual features for the definition of the embedding space to be used for sentiment polarity estimation. Since the visual representations used in the state of the art are based

on the combination of hand crafted features (e.g., GIST, color histograms, etc.), we consider the possibility to further boost the performances of the system by exploiting a deep visual representation. To this aim, for each considered deep architecture we extracted an internal representation of the input image and trained an SVM for the task of polarity prediction (i.e., binary classification). The results of this experiment provide another strong baseline for the performance evaluation. Deep visual features have been combined with objective text features for comparative evaluation with respect to the baseline (i.e., deep feature only). We selected the best performing deep visual feature and performed the evaluation pipeline of the proposed approach considering this stronger visual feature.

Differently than common methods, in the proposed system the input text is automatically extracted from the images. In particular, the objective text associated to a given image is obtained by considering the output text labels of *GoogLeNet* [132] (Object Classification), *Places205* [8] (Scene Recognition), *DeepSentiBank* [7] [46] (Adjective-Noun Pair) and the image description generated by *NeuralTalk2* [79] (Image Captioning). Based on the extracted text, we built three textual features that are combined in different ways with the visual and subjective text descriptors to obtain different embedding spaces. The overall contributions are the following:

- the weakness of the subjective text associated to images usually provided by users for the sentiment prediction task are discussed and experimentally demonstrated;
- an alternative source of text, which is user-independent, has been proposed. For this reason we refer to it as *Objective Text*. Experimental results demonstrate the effectiveness of such textual data, which allows to obtain the best results compared with several baseline and state-of-the-art approaches;
- considering the proposed source of text, several number of combinations of textual and visual features have been evaluated. Furthermore, for each experimental setting the possibility to reduce the dimensionality of the exploited features has been investigated, by employing a truncation strategy which keeps the 99% of the original information;

[7] Our implementation exploits the *MVSO English* model provided by [53], that corresponds to the *DeepSentiBank* CNN fine-tuned to predict 4342 English Adjective Noun Pairs.

- in the second part of the evaluation, we attempt to further improve the performances of the proposed system by employing a deep based visual feature together with objective text. To properly select the deep representation, a comparative evaluation of three state-of-the-art deep architectures has been performed.

The feature evaluation performed in this study focuses on the task of Visual Sentiment Analysis, however the observations and the achieved insights result useful also to other systems which exploit the text associated to social images. The presentation is organized as follows. In Section 3.6.2 a brief review of the state of the art in this context is presented. Section 3.6.3 describes the features we have used to infer the sentiment polarity. Section 3.6.4 details the experimental settings and discusses the results. Finally, Section 3.6.5 concludes the work presentation.

3.6.2 Related Works

The aim of Visual Sentiment Analysis is to infer the sentiment polarity associated to images in terms of positive or negative engagement. Recently, the rise of social media provides huge amount of pictures with user generated accompanying text, such as title, description, tags and comments. This allows the analysis, by Machine Learning approaches, of huge amount of real-word images published on social media by the combination of proper visual and text features. As Section 3.2 provides a detailed presentation of the state of the art, in this Section we only focus on the methods that combine user generated (i.e., noisy) text and images for the task of Visual Sentiment Analysis. Several papers investigated the problem of joint modelling the representation of Internet images and associated text or tags for different tasks, such as image retrieval [82, 83, 128], social images understanding [4], image annotation [136] and visual sentiment analysis [37, 41, 55, 137]. A model that combines textual and visual information is presented in [51]. The subjective textual data such as comments and captions on the images are considered as contextual information. In [137] the authors presented different learning architectures for sentiment analysis of social posts containing short text messages and an image (i.e., Tweets). They exploited a representation learning architecture that combines the input text with the polarity ground truth. This model is further extended with a Denoising Autoencoder when the visual information is present. The approach proposed in [55]

combines visual features with text-based features extracted from the text subjectively associated to images (i.e., descriptions and tags). Specifically, the authors exploited a feature obtained by using the traditional Bag of Words (BoW) [138] approach on the subjective text, and a sentiment feature obtained by selecting the words of the subjective text whose sentiment scores (positive or negative) reported in SentiWordNet [38] are larger than a threshold, and applying the Bag of Words on this revised vocabulary. The considered features are exploited to define an embedding space in which the correlation among the projected features is maximized. Then a sentiment classifier is trained on the features projected in the embedding space. This approach outperformed other state-of-the-art methods [37, 41, 51, 135].

In previous approaches, the authors face different issues related to the subjective text associated to images. For instance, the framework presented in [51] implements an unsupervised approach aimed to address the lack of proper annotations/labels in the majority of social media images. In [133], the authors tried to learn an efficient image-sentence embedding by combining a large amount of weakly annotated images (where the text is obtained by considering title, descriptions and tags) with a smaller amount of fully annotated ones. In [139] the authors exploit large noisily annotated image collections to improve image classification.

Is important to notice that the text sources associated to images exploited in the aforementioned works can be very noisy due the subjectivity of such text. Different users can describe and tag the same image in different ways, including also text which is not related to the content. We investigated the use of an objective text source. In our framework Objective Text is automatically extracted from the visual content of images for the task of Visual Sentiment Analysis.

To the best of our knowledge, this is the first study that propose the exploitation of objective text automatically extracted from images to deal with the issues related to the subjectivity nature of the text provided by users for Visual Sentiment Analysis purposes.

3.6.3 Proposed Approach

In this Section we highlight the main differences between subjective and objective text, present the features extraction process and detail how to build the embedding space in order to exploit jointly different kind of features (views).

Subjective vs Objective Text

Analysing social pictures for Sentiment Analysis brings several advantages. Indeed, pictures published through social platforms are usually accompanied by additional information that can be considered. Several meta-data are available, depending on the specific platform, but in general all the pictures published through a social platform have at least a title, a description and a number of "significant" tags. Most of the existing works in the field exploit social subjective textual information associated to images either to define the ground truth [41] (i.e., by performing textual Sentiment Analysis on the text) or as an additional data modality (i.e., views) [55, 137]. In the latter case, both the visual and the textual information are used as input to establish the sentiment polarity of a post.

Although the text associated to social images is widely exploited in the state-of-the-art to address different tasks and to improve the semantics inferred from images, it can be a very noisy source because it is provided by the users; the reliability of such input is often based on the capability and the intent of the users to provide textual data that are coherent with respect to the visual content of the image. There is no guarantee that the subjective text accompanying an image is useful for the sentiment analysis task. It is usually related to a specific purpose or intention of the user that published the picture on the platform. Often, the subjective user description and tags are related to the semantic of the images or to the context of acquisition rather than sentiment. In addition, the tags associated to social images are often selected by users with the purpose to maximize the retrieval and/or the visibility of such images by the platform search engine. In Flickr, for instance, a good selection of tags helps to augment the number of views of an image, hence its popularity in the social platform. These information are hence not always useful for sentiment analysis.

As proposed in [83], the semantic of an image can be defined by a single object category, while the user-provided tags may include a number of additional terms correlated with the object coming from a larger vocabulary. Alternatively, the semantic might be given by multiple keywords corresponding to objects, scene types, or attributes. In the context of image retrieval, the authors of [83] exploited three views to build the embedding space with a Canonical Correlation Analysis approach (CCA). The first and the second views were related to visual and textual features

respectively, whereas the third view was obtained considering the ground truth annotations (i.e., category) and the search keywords used to download the images. When these information were missing, the authors obtained the third view by clustering the tags, aiming to reduce the overall noise.

To better explain the problem, is useful to reason on a real case. Figure 3.14 shows an example image taken from the Flickr dataset used in [55]. The textual information below the image is the subjective text provided by the Flickr's user. Namely the photo title, the description and the tags are usually the text that can be exploited to make inferences on the image. As shown by this example, the text can be very noisy with respect to any task aimed to understand the sentiment that can be evoked by the picture. Indeed the title is used to describe the tension between the depicted dogs, whereas the photo description is used to ask a question to the community. Furthermore, most of the provided tags include misleading text such as geographical information (i.e., Washington State, Seattle), information related to the camera (i.e., Nikon, D200), objects that are not present in the picture (i.e., boy, red ball, stick) or personal considerations of the user (i.e., my new word view). Moreover, in the subjective text there are many redundant terms (e.g., dog). Another drawback of the text associated to social images is that two users can provide rather different information about the same picture, either in quality and in quantity. Finally, there is not guarantee that such text is present; this is an intrinsic limit of all Visual Sentiment Analysis approaches exploiting subjective text.

Starting from the aforementioned observations about the subjective text associated to social images, we propose to exploit an objective aspect of the textual source that comes directly from the understanding of the visual content of the images. This text is achieved by employing four deep learning models trained to accomplish different visual inference tasks on the input image. At the top right part of Figure 3.14 the objective text automatically extracted with different scene understanding methods is shown. In this case, the inferred text is very descriptive and each model provides distinctive information related to objects, scene, context, etc. The objective text extracted by the three different scene understanding methods has a pre-defined structure, therefore all the images have the same quantity of textual objective information. For each considered scene understanding method (i.e., GoogLeNet [132], DeepSentiBank [46] and Places205 [8]) the classification results

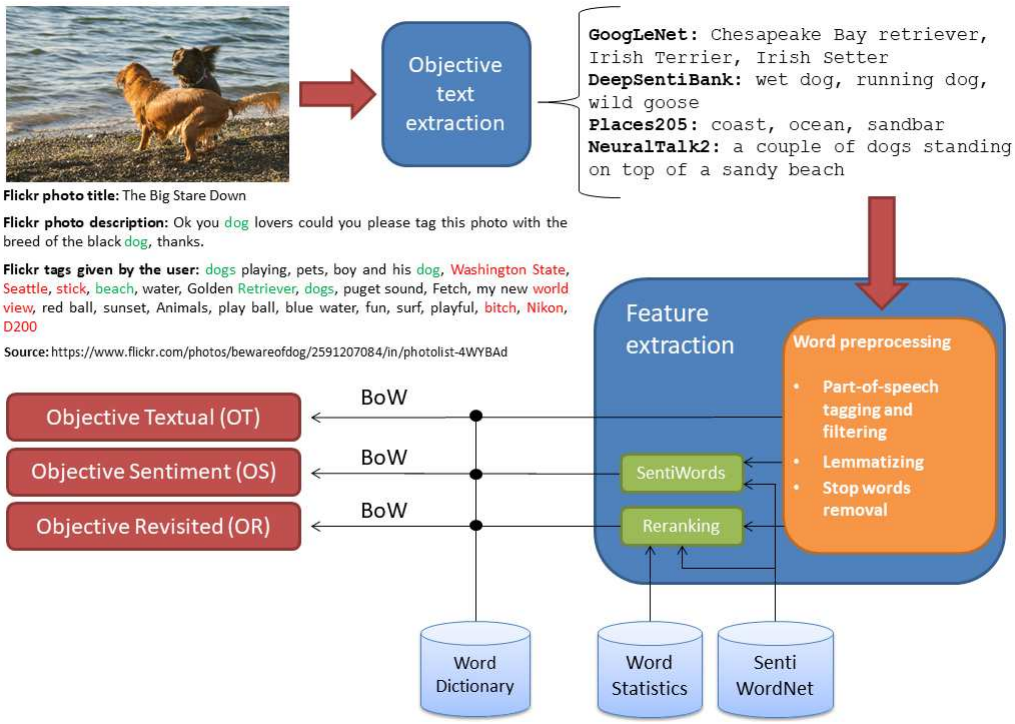


Figure 3.14: Given an image, the proposed pipeline extracts Objective Text by exploiting four different deep learning architectures. The considered architectures are used to extract text related to objects, scene and image description. The obtained Objective Text is processed to produce three different features: the Objective Textual feature (OT) which is the BoW representation of the extracted text based on the whole dictionary, the Objective Sentiment feature (OS) which is the BoW representation obtained considering only the words with strong sentiment scores according to SentiWordNet, and the Objective Revisited feature (OR) which is the weighted BoW representation of the extracted text, in which the weight of each word is given by its statistics and sentiment scores according to the SentiWordNet lexicon. The figure shows also the subjective text associated to the image by the user (i.e., title, description and tags) at top left. The subjective text presents very noisy words which are highlighted in red. The words that appears either in the subjective and objective texts are highlighted in green.

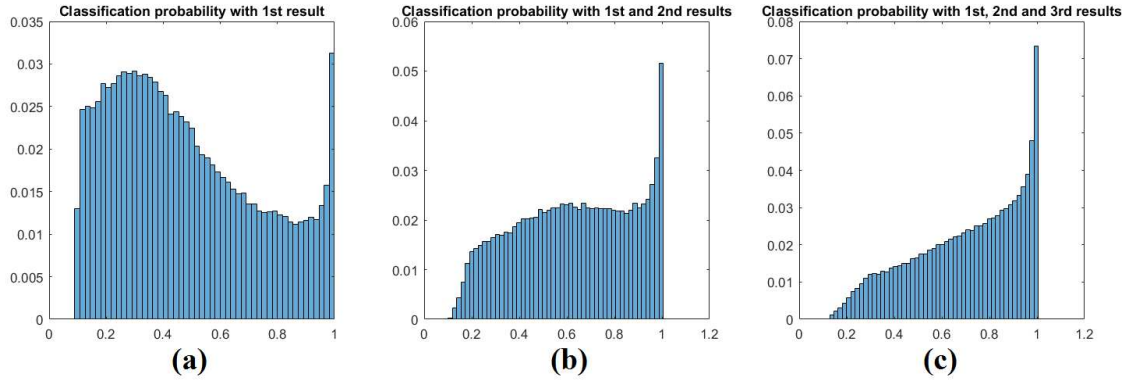


Figure 3.15: In order to choose the number of classification labels to be taken into account for the objective text, we performed a statistical analysis of the output probabilities. We extracted the out probability of the first three labels for each CNN, then we evaluated the probability distributions considering only the first, the sum of the first two or the sum of the first three output probabilities. The histograms show the probability distributions computed over a set of classification probability outputs of the exploited Deep Learning Architectures. The first histogram (a) is the distribution of the output probabilities obtained by considering only the first label in the classification ranking (i.e., corresponding to the highest classification output probability), whereas the second and the third histograms (b) and (c) show the distributions of the cumulative probability obtained by summing the first two and the first three probabilities of the obtained labels respectively. The distributions (a) and (b) have a strong tendency around the values 0.3 and 0.6. The histogram (c) instead shows a monotonic increasing shaped distribution with only one peak close to the value 1; indicating that the correct prediction is mostly within the first three labels.

are ranked by the output probability of the classifier and only the first three labels related to the classification results are considered in our framework. Augmenting the number of the classification results leads to the inclusion of wrong categories. Therefore, we considered the minimum number of labels that guarantee (in a probabilistic sense) a total classification probability close to 1 (i.e., the minimum number of outputs which probabilities sum distribution has a tendency near the value 1). To this aim, we analysed the distribution over the output classification probabilities (see Figure 3.15) to understand the number of labels to be considered to describe the images. In our experiments, we observed that considering only the first three labels is a reliable approach to achieve a total classification probability very close to 1, avoiding to include noisy labels with respect to the visual content.

Finally, we used also a method able to produce an image caption (NeuralTalk2 [79]). That provides one more objective description (i.e., one more view) which we consider

as objective text feature for sentiment purposes.

Features Extraction

The proposed approach exploits one hand-crafted feature and three deep visual representations as visual views, and three text features to represent the objective text extracted from the images, namely Objective Textual (OT) and Objective Sentiment features (OS) and Objective Revisited (OR). As mentioned above, we use scene understanding approaches to extract objective text for the images.

Classic Visual View: As in [55] we considered five image descriptors used in various Computer Vision tasks, such as object and scene classification. In particular, the extracted visual features are a 3×256 RGB histogram, a 512 dimensional GIST descriptor, a Bag of Words image descriptor using a dictionary with 1000 words with a 2-layer spatial pyramid and max-pooling, the 2000 dimensional attribute features presented in [56] and the SentiBank 1200 mid-level visual representation presented in [41]. Each image descriptor has been mapped by using the random Fourier feature mapping [140] or the Bhattacharyya kernel mapping [141]. Then, all the obtained representations have been reduced to 500 dimensions using Principal Component Analysis (PCA). The final visual feature vector associated to each image has a length of 2500 and is obtained as concatenation of all the PCA projected visual features.

Deep Visual View: In the last few years Convolutional Neural Networks (CNNs) have been showing outstanding performances in many Computer Vision challenges. Furthermore, CNNs have proved to be very effective for transfer learning problems. In order to improve the contribution given by the visual view in the computed embedding space, in our experiments we exploited three state-of-the-art deep architectures to extract their inner visual representations (i.e., GoogLeNet [132], DeepSentiBank [46] and Places205 [8]). We first performed a set of baseline experiments based only on the deep visual representation. Then, we evaluated the contribute of the deep visual view in the embedding space, combined with the other extracted views.

Text Views: Five text-based features are used in our experiments. Two of them are the same textual (T) and sentiment (S) views used in [55]. These features reflect the subjective text information provided by the users. Moreover, we built three textual features based on the Objective Text obtained through deep learning architectures. The overall pipeline for Objective Textual based features extraction is sketched in Figure 3.14. Each exploited deep learning architecture provides a description, in some sense objective, of the input image from a different point of view, as each architecture has been trained for a different task (e.g., object recognition, place recognition, etc.). For instance, the deep architecture specialized for object classification (i.e., GoogLeNet [132]) finds the principal objects within the picture (e.g., dog), providing information about dog breeds in Figure 3.14. The Adjective-Noun Pair classifier (i.e., DeepSentiBank [46]) agrees with the previous result that the main object is a dog and provides other information in form of Adjectives-Noun Pairs (i.e., "wet dog" and "running dog"). The network devoted to place classification (i.e., Places205 [8]) gives further information about the location and the depicted environment (i.e., "coast", "ocean" and "sandbar"). Furthermore, the caption generated by NeuralTalk2 [79] provides a confirmation of all the previous inferences putting them in context through a description. The use of different architectures allows to obtain a wide objective description of the image content which consider different semantic aspect of the visual content. Although the exploited deep learning architectures are different, they all describe the same image, and it implies the generation of some redundant terms. This has not been considered as a drawback, indeed the presence of more occurrences of similar or related terms (e.g., dog, dogs, retriever, terrier, setter, etc.) enhance the weight of these correct terms in the representation extracted by our framework. On the other hand, this redundancy reduces the effect of noisy results such as the third result extracted with DeepSentiBank in Figure 3.14 (i.e., "wild goose"). For these reasons, in the Bag of Words text representation exploited in the proposed paper, we considered the number of occurrences of each word of the vocabulary in the text associated to the image, instead of considering a binary vector representation which encodes the presence or the absence of each word [55, 83, 142].

To further compare the considered Objective Textual representation with other state of the art solutions, we implemented the feature extraction process described

in [143]. According to this approach, a given text is represented as a feature vector which elements are obtained by multiplying the sentiment scores of the contained words by their frequencies. The sentiment scores are taken from SentiWordNet [38], and a re-ranking of such scores is performed for the words whose neutral score is higher than either the negative and the positive ones. In our experiments, we implemented both the re-ranking procedure and the feature extraction process of [143] for comparison purposes.

In this work all the text-based features are obtained through a Bag of Words (BoW) representation of the objective text extracted from the input picture. These representations share the same pre-processing stage of the text extracted with the deep learning architectures. This includes the procedures commonly applied in text mining:

- **Speech tagging and filtering:** this step choose a proper part of speech tag of each word, to solve its ambiguity. This step is needed since in SentiWordNet a word with a different part of speech tag might have a different sentiment value and, hence, dominant polarity. Considering our input source, we already know that the two words resulting from DeepSentiBank corresponds to an adjective-noun pair, and the most of the Places205 and GoogLeNet outputs are nouns. Therefore, this preprocessing mainly contributes on the text obtained with NeuralTalk2;

- **Lemmatizing:** since only base form of words are stored in SentiWordNet, we performed a lemmatizing step;

- **Stop words removal:** this step removes words that contain no semantic concepts, such as articles and prepositions.

The above pre-processing steps allow to obtain co-occurrences of the words describing the image from different semantic aspects of the visual content. Indeed, the proposed approach benefits from the inferences coming from architectures trained for different tasks: object classification, places classification, Adjective-Noun Pair classification and image description. Starting from the pre-processed Objective Text, we propose to extract the following text-based features:

- **Objective Text (OT):** we obtained this feature by computing a classic Bag of Words representation followed by a SVD dimensionality reduction. The final feature has dimension 1500.

- **Objective Sentiment (OS):** we computed the Bag of Words representation by using a reduced dictionary of sentiment related words (called sentiment vocabulary), followed by a SVD (Single Value Decomposition) feature dimensionality reduction to obtain 20 dimensional vectors. We considered only the words which either positive or negative sentiment score in SentiWordNet, is higher than 0.15.

- **Objective Revisited (OR):** the paper described in [143] proposed an interesting text representation for the task of sentiment analysis. Furthermore, it highlights an issue related to the use of SentiWordNet scores for sentiment analysis. Indeed, most of the existing sentiment feature extraction methods (including [55]) ignore words which neutral sentiment is higher than either positive and negative ones, albeit they comprise the 93.75% of SentiWordNet entries. The authors of [143] proposed a revisiting procedure of the sentiment scores associated to the neutral words that modules the sentiment scores according to the probability of a word to appear in a positive or a negative sentence. Then, the representation of a given text is a weighted BoW vector which elements are obtained by weighting the word counts with the predominant sentiment score (positive, negative or zero if the neutral score remains the higher even after the scores revisiting). We use this process on the proposed Objective Text. The OR feature we compute is hence a vector W in which each W_i element is defined as follows:

$$W_i = \begin{cases} TF_i \times posW_i, & where\ W_i \in [pos\ words] \\ TF_i \times negW_i, & where\ W_i \in [neg\ words] \\ 0 & otherwise \end{cases} \tag{3.3}$$

where $posW_i$ and $negW_i$ denote the positive and negative sentiment scores of the *i-th* word, and TF_i is the number of occurrences of the *i-th* word in the Objective Text extracted from the considered image.

All the process described above for word dictionaries definitions, SVD computation and OT, OS and OR parameter settings have been done considering only the Objective Text associated to the training set images of the dataset used for our experiments. The methods are then evaluated on a different test set.

Embedding Different Views

Recently, several papers for jointly modelling images and associated text with Canonical Correlation Analysis (CCA) have been proposed [55, 83, 124, 125, 126, 127]. CCA is a technique that maps two or more views into a common embedding space. The CCA is used to find the projections of multivariate input data such that the correlation between the projected features is maximized. This space is cross-modal, therefore the embedded vectors representing the projections of the original views are treated as the same type of data. Thus, in the CCA embedding space, projections of different views are directly comparable by a similarity function defined over the elements of the embedding space [83].

Let ϕ^i be the data representation matrix in the *i-th* view. The n_v projection matrices W^i are learned solving the following minimization problem:

$$\begin{aligned} \min_{\{W^i\}_i^{n_v}} &= \sum_{i,j=1}^{n_v} Trace(W^i \Sigma_{ij} W^j) \\ &= \sum_{i,j=1}^{n_v} \left\| \phi^i W^i - \phi^j W^j \right\|_F^2 \\ s.t. \left[W^i\right]^T \Sigma_{ii} W^i &= I \quad \left[w_k^i\right]^T \Sigma_{ij} w_l^j = 0 \\ i \neq j, k \neq l \quad i,j &= 1,\ldots,n_v \quad k,l = 1,\ldots,n \end{aligned} \tag{3.4}$$

where W^i is the projection matrix which maps the *i-th* view matrix $\phi^i \in \Re^{n \times m_i}$ into the embedding space, w_k^i is the *k-th* column of W^i and Σ_{ij} is the covariance matrix between ϕ^i and ϕ^j. The dimensionality of the embedding space m_e is the sum of the input view dimensions $m_e = \sum_i^{n_v} m_i$. Therefore $W^i \in \Re^{m_i \times m_e}$ transforms the m_i dimensional vectors of the *i-th* view into the embedding space with dimension m_e. As demonstrated in [134], this optimization problem can be formulated as a standard eigenproblem. In the proposed work we exploited the multi-view CCA implementation provided by [127].

Table 3.3: Number of positive and negative images in our dataset and in the original dataset used in [55].

	Positive	Negative
Dataset in [55]	48139	12606
Our Dataset	37622	9613

In Section 3.6.4, we describe how to use the embedding space learned from multiple views to obtain the features used in the proposed approach.

3.6.4 Experimental Settings and Results

Dataset

In [55] the authors performed the experiments with two different datasets crawled from Instagram and Flickr. For each image in the dataset, the image description and tags have been taken into account to obtain the text on which build the text based features. The images are not available for download, but the authors published the list of images' ids to allow performing comparison with other approaches. Due to the recently changes in Instagram policies, we were unable to download images from this platform. Therefore we used only the dataset obtained downloading Flickr images. Some of the pictures were missing at the moment of crawling (e.g., removed by the users). Only 69893 Flickr images were available at the time of our analysis. Following the experimental protocol, we considered the images with positive or negative ground truth, discarding the images labelled as neutral. The final dataset used in the experiments has a total of 47235 images. Although the dataset used in our experiments is a subset of the Flickr images used in [55] due to aforementioned reasons, the number of either positive and negative images is comparable with the number of positive and negative images of the original dataset (see Table 3.3).

To evaluate the performances of the different compared sentiment classification approaches, we considered the original sentiment labels which have been obtained via crowdsourcing [55]. During the labelling process, each image has been presented to three people who were asked to provide a five-point scale sentiment score. The final ground truth has been defined by considering the majority votes of polarity for each image. The images labelled as neutral, as well as the images resulting in disagreement among people, have been discarded.

Table 3.4: Performance Evaluation of the proposed method with respect to the baseline method presented in [55]. The best result is highlighted in bold, whereas the second best result is underlined. See text for details.

	Experiment ID	Embedded Views	Full Feature	Truncated Features (99%)
Subjective Features Proposed in [55]	K1	V+T+S	66.56 ±0.43 %	66.11 ±0.45 %
	K2	V+T	71.67 ±0.36 %	71.55 ±0.57 %
	K3	V+S	62.19 ±0.63 %	62.89 ±0.45
Considering Subjective and/or Objective Features	O1	V+T+OS	68.88 ±0.49 %	69.23 ±0.38 %
	O2	V+OT+S	66.97 ±0.57 %	66.34 ±0.68 %
	O3	V+OT	<u>73.48</u> ±0.54%	<u>72.54</u> ±0.65 %
	O4	V+OS	66.58 ±0.70 %	66.41 ±0.53 %
	O5	V+OT+OS	69.83 ±0.58 %	69.62 ±0.53 %
	O6	V+T+S OT+OS	68.04 ±0.55 %	67.39 ±0.19 %
	O7	V+T+OR	66.04 ±0.54 %	66.74 ±0.45 %
	O8	V+OT+OR	68.29 ±0.54 %	67.84 ±0.68 %
	O9	V+OR	64.60 ±0.70 %	63.08 ±0.82 %
	O10	V+T+OT	**73.96** ±0.39 %	**72.66** ±0.70 %

Embedded Vectors

In Section 3.6.3 we described the CCA technique, and defined how to obtain the projection matrices W_i, related to each view i, by solving an optimization problem.

We exploited a weighted embedding transformation which emphasize the most significant projection dimensions [127]. The final representation of the data from the *i-th* view into the weighted embedding space is defined as:

$$\Psi^i = \phi^i W^i \left[D^i\right]^\lambda = \phi^i W^i \tilde{D}^i \tag{3.5}$$

where D^i is a diagonal matrix which diagonal elements are the eigenvalues in the embedding space, λ is a power weighting parameter, which is set to 4 as suggested in [127].

In our experiments we further considered a reduced projection obtained by taking only the first components of W^i encoding the 99% of the original information. The number of components to keep is obtained by considering the minimum number of eigenvalues (i.e., the diagonal elements of D) which normalized sum is greater or equal than 0.99. We call these representations *Truncated Features* in our experiments.

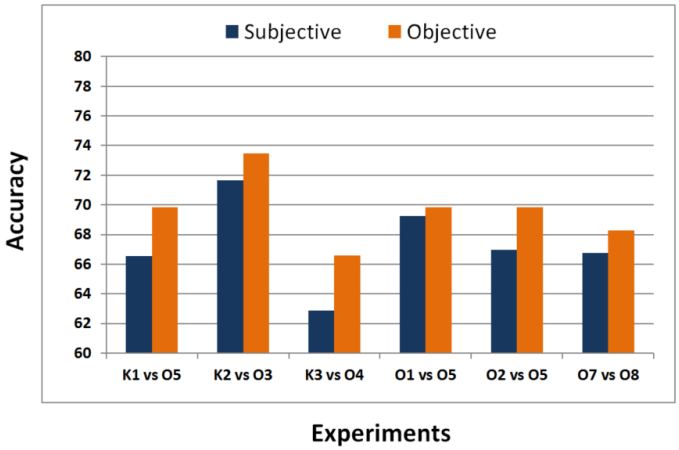


Figure 3.16: Explicit comparison between the experimental settings that differ by the use of one or more particular objective/subjective feature (e.g., O5 differs from K1 by the exploitation of OT and OS instead of T and S, respectively). In all the experiments, the objective features (orange bins) perform better than the corresponding subjective ones (blue bins). For each experimental setting, the accuracy value reported in this histogram is the best achieved result between the experiment performed by using the full feature and the truncated one (see Table 3.4).

Performance Evaluation

The dataset has been randomly separated into a training set and test set, considering a proportion of 1:9 between the number of test and training images, and including a balanced number of positive and negative examples. As a performance evaluation metric, we computed the average and standard deviation of test classification accuracy over 10 runs, repeating the data shuffling at each run [8]. A linear SVM has been used to establish the sentiment polarity over the different compared representations. For each experimental setting we used LibLinear [9] to train a linear SVM classifier. The parameter C of the linear SVM was determined by 10-fold cross validation.

Table 3.4 shows the obtained results. Each row describes a different experimental setting, corresponding to a specific combination of the input features described in Section 3.6.3 used to build the embedding space. The column "Full Feature" reports the results obtained by considering the full-size representation in the embedding space obtained by applying Equation (3.5), whereas the results of the experiments

[8]The code to repeat the performance evaluation is available at the URL: http://iplab.dmi.unict.it/sentimentembedding/

[9]http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

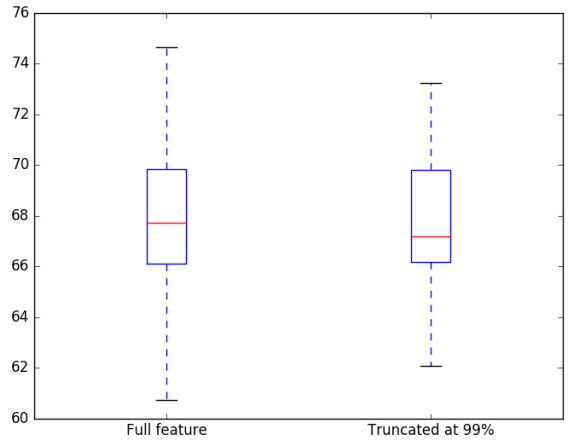


Figure 3.17: Comparison between the distributions of the achieved accuracy values between the experimental settings which exploit the full size features and the ones which exploit the truncated features. The two distributions have been obtained by considering the accuracy and standard deviation values reported in Table 3.4.

performed with the truncated feature representations are reported in the last column (i.e., "Truncated Features"). In Table 3.4 all the tests with prefix "O" (Objective) are related to the exploitation of features extracted with the proposed method, whereas the features V, T and S refer to the features extracted with the method presented in [55] (Visual, Textual and Sentiment respectively). The third column lists the views used for the computation of the embedding space. For instance, V+T refers to the two-view embedding based on Visual and Textual features, V+OT+OS is related to the three-view embedding based on Visual, Objective Textual, and Objective Sentiment features, and so on.

As first, it is simple to note that all the tests where the Objective Text description is used achieve better results with respect to the experimental settings in which the corresponding Subjective Text features are exploited (see Table 3.4 and Figure 3.16). Figure 3.16 compares the experimental settings which differ by the exploitation of one or more subjective or objective features. This allows the comparison between the exploitation of the proposed "Objective Text" with respect to the classic "Subjective Text". Among this subset of experimental settings, the good results are obtained by exploiting Visual feature and Objective Text (O3 in

Table 3.4). In particular, using the feature OT instead of T provides a mean accuracy improvement of 1.81% (compare O3 and K2 in Table 3.4). Adding the view T to the experimental setting O3 yields an increment of 0.48% (O10 in Table 3.4). Note that, adding the proposed OT feature to the experimental setting K2 provides an improvement of 2.29% (compare K2 with respect to O10). These observations highlight the effectiveness of the features extracted from Objective Text with respect to the features extracted from the subjective one. Finally, when the proposed truncated features are employed the classification accuracy has a mean decrease of 0.31%, which is lower than the standard deviation of the accuracy values computed over 10 runs in all the performed experiments. This means that, even if the exploitation of the truncated features causes a decrease of the mean accuracy, the range of the values obtained in the two cases are comparable. To better assess this observation, Figure 3.17 shows the box-and-whisker plots (henceforth, referred to simply as boxplots) obtained from the distributions of the accuracy values reported in Table 3.4.

The fact that the sentiment features (i.e., S and OS) do not achieve good results is probably due to the fact that the number of sentiment words considered to build the Bag of Words representations according to the method proposed in [55] is very limited. Furthermore, the sentiment score of most of the SentiWordNet words is often neutral, indeed the 97.75% of SentiWordNet words are words which neutral score is higher than either the negative and the positive ones. Therefore, most of the words of the sentiment vocabulary are still neutral. In particular, we observed that about 61% of the words in the sentiment vocabulary used in this work are neutral for SentiWordNet, the 24% are negative and 15% are positive. As a result, the feature extraction process for sentiment features produces very sparse and rather uninformative Bag of Words sentiment representations. This observation is confirmed by the fact that the best results are obtained considering Visual, Objective Textual and Subjective Textual features (O10 in Table 3.4). To summarize, from this first set of experiments we observed that:

- the Objective text features (i.e., OT and OS) performs better than the Subjective text features (i.e., T and S);
- the features obtained by exploiting the full vocabulary (i.e., T and OT) performs better than the ones obtained by the usage of sentiment biased textual

features (i.e., S, OS and OR);

- the truncation of the projected features performed by exploiting the principal eigenvalues causes a slightly reduction of the mean accuracy. However, taking into account the variability of the achieved results, we observed that the statistical distribution of the accuracy values of the experiments performed with the truncated features is comparable with the one related to the experiments performed with the full features;

- the best result are obtained by the experimental setting O10 (i.e., V+T+OT, 73.96%). The second best result is obtained by the experimental setting O3 (i.e., V+OT, 73.48%). The comparison between several experimental settings demonstrate that the contribute given by the exploitation of the proposed Objective Textual feature (OT) is significantly higher than the contribute given by the Subjective Textual feature (T).

Improving the Visual Representation

The experiments of the previous section are needed to perform a fair comparison with the method presented in [55]. Indeed, the first part of the experimental evaluation performed in this paper is aimed to compare the exploitation of the Objective Text with the Subjective one, keeping the same Visual View (V) in all the embeddings. However, recent works in Computer Vision provide several stronger deep learning based visual representations. Some of them can be easily extracted from the deep architectures to be exploited in this paper jointly with the Objective Text.

Since, from a computational point of view, the effort to extract the Objective Text and the corresponding deep features is similar (i.e., a single feed forward step), it worth to consider such deep visual representations as alternative features for the addressed semantic classification task. Therefore, we extracted the deep features representations of images by using the considered CNNs (i.e., GoogLeNet [132], DeepSentiBank [46] and Places205 [8]). Instead of focusing on the final output (i.e., classification), we extracted the activations of the earlier layer and trained an SVM classifier, according to the above described evaluation protocol. Since the achieved representations are based on stronger visual features than the one which are included in the visual view (V), the results of this procedure provide a strong and challenging

Table 3.5: Results obtained by training an SVM on the deep features extracted from GoogLeNet [132], DeepSentiBank [46] and Places205 [8] (pool5/7x7_s1, fc7 and fc7 respectively).

Architecture	**Feature Dimension**	**Results**
DeepSentiBank [46]	4096	**75.92** ±0.65
GoogLeNet [132]	1024	75.14 ±0.46
Places205 [8]	4096	73.83 ±0.65

Table 3.6: Performance Evaluation considering Deep Visual Representations. The best result is highlighted in bold, whereas the second best result is underlined. See text for details.

	Experiment ID	**Embedded Views**	**Full Feature**	**Truncated Features (99%)**
Deep Visual and Subjective Features [55]	DK1	DS+T+S	69.19 ±0.52 %	67.36 ±0.64 %
	DK2	DS+T	74.87 ±0.52 %	73.74 ±0.75 %
	DK3	DS+S	64.70 ±0.68 %	64.29 ±0.79
Deep Visual and Objective Features	DO1	DS+T+OS	71.30 ±0.25 %	70.34 ±0.34 %
	DO2	DS+OT+S	69.42 ±0.44 %	68.29 ±0.68 %
	DO3	DS+OT	**76.78** ±0.42 %	<u>74.46</u> ±0.67 %
	DO4	DS+OS	69.01 ±0.88 %	68.90 ±0.49 %
	DO5	DS+OT+OS	72.00 ±0.37 %	71.16 ±0.86 %
	DO6	DS+T+S OT+OS	69.77 ±0.31 %	68.58 ±0.34 %
	DO7	DS+T+OR	69.59 ±0.55 %	68.36 ±0.55 %
	DO8	DS+OT+OR	70.61 ±0.65 %	69.14 ±0.52 %
	DO9	DS+OR	66.43 ±0.61 %	66.58 ±0.73 %
	DO10	DS+T+OT	<u>76.31</u> ±0.55 %	**74.52** ±0.45 %

additional baseline for our evaluation experiments. Table 3.5 shows the classification results obtained by training an SVM for the task of sentiment polarity prediction when only the aforementioned deep visual features are employed. As we expected, the deep feature extracted from DeepSentiBank outperforms the others, as this CNN has been trained for the task of Adjective Noun Pair (ANP) prediction, which is strongly related to the task of sentiment polarity classification. The boxplot diagram shown in Figure 3.18 is useful to compare the results reported in Table 3.5 achieved by the different deep visual representations on the different runs. As described in Section 3.6.3, the Visual View (V) exploited by the proposed approach includes the SentiBank 1200 mid-level visual representation. This feature can be considered an earlier version of the one provided by DeepSentiBank and it takes into account only 1200 ANPs. The DeepSentiBank CNN is trained to predict 4342 different ANPs.

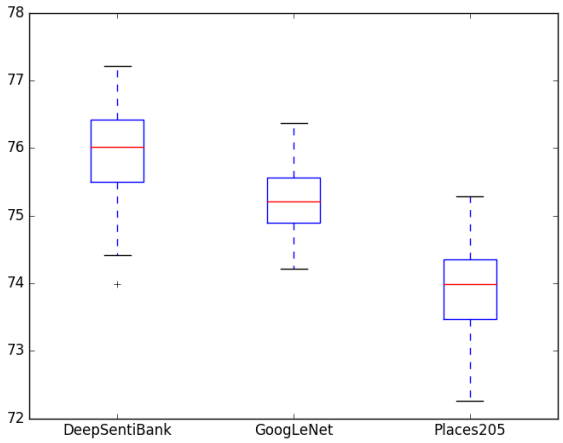


Figure 3.18: Comparison between the distributions of the accuracy values reported in Table 3.5 achieved by exploiting only a deep visual image representation to train the SVM classifier.

Considering that the performances achieved by only using the deep feature representation extracted with DeepSentiBank are better than the two best results obtained by the proposed method (O3 and O10 in Table 3.4), we repeated the performance evaluation of the CCA embedding based representations considering the DeepSentiBank visual feature (DS) instead of the visual feature (V). The results are reported in Table 3.6. The exploitation of the deep visual representation (DS) produced a further improvement of the performances in all the experimental settings (compare Table 3.5 and Table 3.6). In particular, the combination of the DeepSentiBank visual feature (DS) and the Objective Text (OT) feature provides a mean improvement of 2.82% in accuracy with respect to the best two results (DO3 versus O3 and DO10 versus O10). Also, the results obtained exploiting jointly DeepSentiBank features and the Objective Text (i.e., 76.78% of DO3 in Table 3.6) are better than the results obtained when only DeepSentiBank features are used (i.e., 75.92% in Table 3.5). Considering the results detailed in Table 3.6, we observe that in this case the truncation procedure of the projected features produces a more significant decrease of the accuracy score (1.02% in mean). An interpretation of such results is that in the case of the projected representations obtained by considering hand crafted visual features (V), there are some components that can be truncated

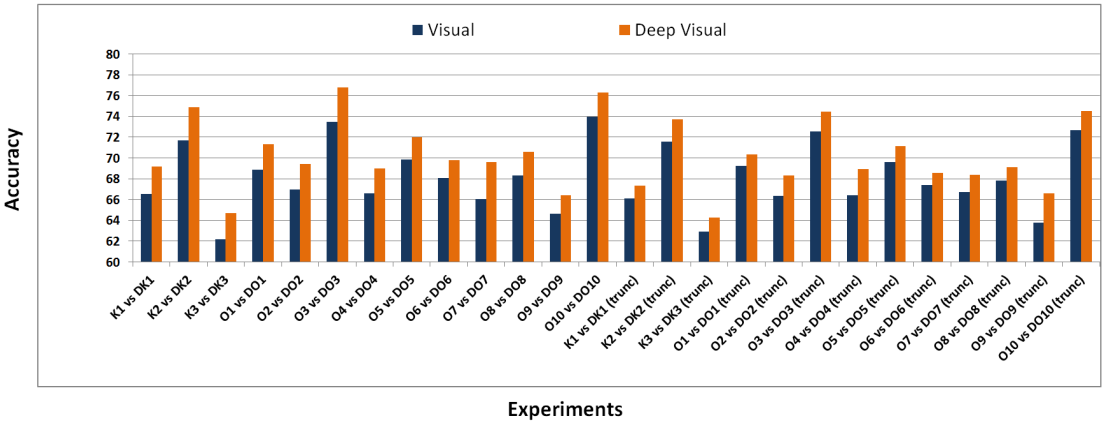


Figure 3.19: Comparison of all the considered experimental settings, directly comparing the settings in which the same textual features are combined with the visual or the deep visual one. In all the experiments, the settings that exploit the deep visual features in the embeddings (orange bins) perform better than the corresponding settings that exploit the visual features (blue bins).

without a significant decrease in performance. Whereas when the projected representations are obtained by considering the deep visual features (DS), almost all the representation components (i.e., including the ones not in the 99% most informative components) provide an non negligible contribute for the classification task. For a better and complete comparison of the achieved results, Figure 3.19 shows the accuracy values obtained with all the different experimental settings detailed in Table 3.4 and Table 3.6, considering either the full feature and the truncated feature experiments.

3.6.5 Final Remarks

The work presented in this Section addresses the challenge of image sentiment polarity classification by exploiting the correlations among visual and textual features associated to images. The considered features are exploited to build an embedding space in which correlations among the different features are maximized with CCA. Then, an SVM classifier is trained on the features of the built embedding space for prediction purposes. The contribute of either visual and textual features in the CCA embedding is deeply investigated and assessed. The first part of the work presents a study in which Objective Text extracted considering the visual content of images is compared with respect to the Subjective Text provided by users. The best results

have been obtained by exploiting the proposed Objective Text based features. This study demonstrates that the exploitation of Objective Text associated to images provides better results than the use of the Subjective Text provided by the user.

On the other hand, we identified several drawbacks brought by the Subjective Text due its intrinsic nature. Indeed, the subjective text associated to images by users presents very noisy terms. It doesn't respect a pre-defined scheme or length constraints, therefore the text sources associated to different images may have very different structures. Moreover, there is no guarantee of the presence of such text for all the considered images. The Objective Text exploited by our approach doesn't present the aforementioned issues, and it is automatically extracted from the image. Two similar images are likely to have very similar Objective Text, whereas we cannot say anything about their subjective text provided by the users. These observations and our experimental results support the use of Objective Text automatically extracted from images for the task of Visual Sentiment Analysis in lieu of the Subjective Text provided by users, and suggest the investigation of the exploitation of such text also for other task related to the association between text and images. Finally, the obtained results show that all the text features based on the SentiWordNet scores do not achieve good results, mainly due to the lack of strong positive or negative terms in the analysed text. An in-depth investigation on this aspect is needed: future works will be devoted to the exploitation of more sentiment oriented information to build features that reflects the emotions evoked by images. Social platforms provides interesting data to infer people reactions toward images (i.e., users social engagement). They includes comments, likes and shares of users that see the image in the platform which can be used to extend the Visual Sentiment Analysis methods to improve the overall accuracy for sentiment polarity estimation.

With the aim to further boost the performances of the proposed system, we considered different visual features based on deep architectures. This evaluation demonstrated that deep based visual representations performs better than the hand crafted visual features proposed in previous works for the task of sentiment polarity classification. Furthermore, these experiments confirmed that the contribution of Objective Text based features is higher than the one provided from Subjective Text ones. Our performance evaluation considers 52 different combinations of features to build CCA embedding spaces, obtained by considering different textual and visual

features, and different strong baselines based on the exploitation of deep based visual features.

3.7 Image Popularity Prediction

3.7.1 Introduction

In Section 3.5.1 we introduced the task of Image Popularity Prediction, that is the prediction of the level of engagement achieved by social contents shared through social media platforms. The media engagement can be measured in terms of number of views, likes or shares. However, the popularity score defined in [34] (i.e., Equation 3.2) is commonly used as index to summarize the popularity of an image shared through a social platform.

Many researches tried to gain insights to understand which features make an image popular [4, 144]. Yamasaki et al. [95] proposed an algorithm to estimate the social popularity of images uploaded on Flickr by using tags. The method is used to recommend tags to users to gain greater attention from other users. The importance of each tag is obtained by combining frequency and model weights. The results show that the popularity can be estimated from low dimensional (but meaningful) features such as image's tags. The exploited features are used either for regression (i.e., estimate the number of views, comments, favorites) and classification (i.e., distinguish between popular and unpopular images) tasks.

The work in [35] considers the effect of 16 features in the prediction of the image popularity. Specifically, they considered image context (i.e., day, time, season and acquisition settings), image content (i.e., image content provided by detectors of scenes, faces and dominant colors), user context and text features (i.e., image tags). The authors cast the problem as a binary classification by splitting the dataset between images with high and low popularity measure. As popularity measures the authors considered the views and the comments counts. Their study highlights that comments are more predictable than views, hence comments are more correlated with the studied features. Accuracy values achieved only considering textual features (i.e., tags) outperform the performances of the classification based on other features and the combinations of them, for both comments and views counts classifications.

The authors of [34] analysed the importance of several cues related to the image or to the social profile of the user that lead to high or low values of popularity. In particular, they considered the relevance of user context features (e.g., mean views, number of photos, number of contacts, number of groups, etc.), the image

context features (e.g., title length, description length, number of tags), as well as visual features extracted from the image (e.g., GIST, LBP, BoW color patches, CNN activation features, etc.). The selected features were used to train a Support Vector Regressor (SVR) to predict the image popularity score defined in Equation 3.2. Cappallo et al. [96] addressed the popularity prediction task as a ranking problem. They used a latent-SVM with an objective function defined to preserve the ranking of the popularity scores between pairs of images. They considered the number of views and the comments for Flickr images, whereas the number of re-tweets and favorites for Twitter posts. The problem of image popularity prediction was also addressed by Gelli et al. [33]. They used sentiment ANPs features (Adjective-Noun Pairs) defined in the Visual Sentiment Ontology (VSO) [41] pointing out that those features have a strong correlation with popularity. Aloufi et al. [145] evaluated several techniques to combine different features and investigated the effect of such combinations to predict different levels of interactions (i.e., number of views, comments and favorites on Flickr).

Most of the works addressing the problem of popularity prediction follow a very similar pipeline. First, a set of interesting features that have been demonstrating correlation with the images sentiment or popularity is selected. Then a model for each distinctive feature is trained to understand the predictive capability of each feature. In the above discussed works, the popularity prediction task is cast as a ranking or a regression problem. Therefore, the exploited algorithms are ranking SVM and latent SVM in the case of ranking, and SVR for regression. Then, the features are combined with the aim to improve the performances of the method. The evaluation is usually quantified through the Spearman's correlation coefficient.

Few works have considered the evolution of the image popularity over time (i.e., popularity dynamics). In [146] the authors considered the number of likes achieved within the first hour of the post lifecycle to forecast the popularity after one day, one week or one month. The problem has been treated as a binary classification tasks: popular vs non popular. To this aim they used a popularity threshold obtained with the Pareto principle (80 % - 20 %). Therefore, the problem is reduced to a binary classification task. Three features for the classification task were evaluated: social context (based on the user's number of followers), image semantics (based on image caption and NLP), and number of likes in the first hour. The binary classification

has been performed by using a Naive Bayes Model with a Gaussian likelihood. The experimental results show that the number of likes of the first hour outperforms other features. Wu et al. [147, 148, 149] explored social media popularity by modelling time-sensitive context and proposed to represent popularity in multi-time scales. The achieved results shown that the time of a post plays an important role for popularity prediction. A large-scale social media dataset, namely Social Media Prediction (SMP) dataset, has been collected to set the ACM Multimedia 2017 SMP Challenge [10]. Differently than the SMP Challenge, in our study we propose a dataset that reports the engagement scores (i.e., number of views, comments and favorites) of all the involved images recorded on a day-to-day basis for 30 consecutive days. This allows to compute the popularity score for an arbitrary day and to try the forecasting of the overall sequence. Moreover, the sequence of daily scores can be exploited to analyse the photo lifecycle.

Li et al. [150] extracted multiple time-scale features from a set of timestamps related to the photo post. The timestamp "postdate" is converted to several features with different time-scales: season of year, month of year, week of year, week of month, day of week, day of month, and moment of day, etc. The framework presented in [151] exploits an ensemble learning method to combine the outputs of SVR and CART models, previously trained to estimate the popularity score. The features used to train the models were extracted from user's information, image meta-data and visual aesthetic features extracted from the image. In particular, the authors used the post duration (i.e., the number of days the image was posted), the upload time, day and month.

The popularity score (Equation 3.2) defined in [34] considers in an explicit way the time of a post. Time aware popularity prediction methods exploit time information to define proper image representations used to infer the image popularity score as in Equation 3.2 at a precise time or at pre-defined time scales [146, 150]. In other words, these systems include the time information in the input, in order to better predict the popularity score defined as a single value.

In the context of modelling and predicting dynamics, several works have been presented. The authors of [152] aimed to predict the popularity evolution of user generated videos. They defined a set of popularity "behaviours" of the contents

[10] Challenge webpage: https://social-media-prediction.github.io/MM17PredictionChallenge

(i.e., patterns). The proposed method compares the popularity achieved by different contents up to the current time step, performing a clustering over the set of defined behaviour patterns. The clustering procedure is performed over time and, for each content, the system is able to predict the next clustering. In other words, given a set of contents (i.e., videos) and the clustering at time t, the system aims to predict the most likely cluster that each element will belong to at time $t+k$.

In [153] the authors aimed to predict the popularity of news (i.e., Digg stories) based on the users' early reaction. In particular, each user is represented as a stochastic process with a small number of states. Group of users with similar behaviours are grouped, this allow to improve the popularity prediction (i.e., number of votes) of news by means of the first users' interactions (e.g., views, votes, etc.). The paper in [154] proposed two models to predict the future popularity of YouTube videos, by exploiting a set of daily samples of the content's popularity measures up to a given reference date, which are properly weighted. The work in [155] analyses the early patterns of YouTube videos and Digg stories to predict the long-term popularity of such contents. For instance, the information related to the users access to a a Digg content (i.e., number of votes) allow the system to the method to predict the popularity 30 days ahead with an error of 10%. To achieve the same error rate for the popularity prediction of YouTube contents, the system needs to know the information related to the first 10 days.

All the mentioned works aim to predict popularity by exploiting features based on the popularity achieved just in the first period (e.g., [146]), often referred as early popularity or early reaction. Bandari et al. [156] propose a method which aims to predict popularity of items prior to their posting, however it focuses on news article (i.e., tweets) and the popularity prediction results (i.e., regression) are not satisfactory, as claimed by the authors themselves. Indeed they achieve good results by quantizing the range of popularity values and performing a classification. Differently than previous works, our approach focuses on the prediction of the whole temporal sequence of image popularity scores (30 days) with a daily granularity. We investigate the correlation between social image popularity, social features (including user's features and photo's features), and visual features extracted from images. The set of features have been selected taking into account the insights achieved by the state of the art.

The main contributions of our study are the following:

- it poses new questions around on-line behaviour, popularity, and social media content lifecycle;
- it proposes a new challenging task which finds very practical applications in recommendation systems and advertisement placement;
- a new benchmark dataset is released. The proposed dataset includes $\sim$ 20K Flickr images, with related user and photo meta-data and three engagement scores tracked for 30 days (the number of views, number of comments and number of favorites);
- the proposed framework empowers the development of systems that support the publication and promotion of social contents, by providing a forecast of the engagement dynamics. This can suggest when old contents should be replaced by new ones before they become obsolete.
- an in-deep investigation of social features that characterize the influence of a user and the level of diffusion of a photo is provided. In particular, in addition to the main user and photo features, we considered statistics related to groups in which the user is enrolled or in which the photo is shared;
- baseline approaches that can be considered as reference to compare other approaches are presented;

We provide a study of the time effect on the problem of popularity prediction in Section 3.7.2. Section 3.7.3 provides a description of the proposed dataset, by detailing the crawling procedure and the analysis of the crawled data, as well as the data preprocessing. Section 3.7.4 gives the details of the proposed approach. The evaluation of the proposed approach is reported in Section 3.7.5. The obtained results are commented in Section 3.7.6, which also provide insights and cues for future works.

3.7.2 Motivations

In order to estimate the popularity with Equation 3.2, the cumulative engagement obtained by an image post until a specific day is normalized with respect to the total

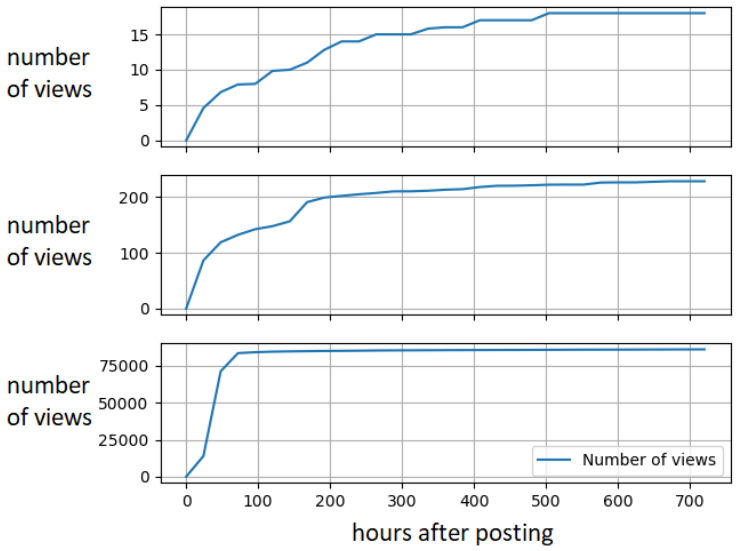


Figure 3.20: Examples of dynamics of the overall number of views achieved by three photos shared on Flickr related to 30 days. The shape of the three sequences is very similar, however they have very scales.

number of days of the post. However, the formulation of the popularity score does not take into account the dynamic of the engagement of the image post over time. We observe that the relative increment of the engagements over time is not constant and tends to decrease with the passing of the days. Therefore, the Equation 3.2 tends to penalize social media contents published in the past with respect to more recent contents, especially when the difference between the dates of the upload of different posts is large. Most of the engagements (e.g., views) obtained by a social media item are achieved in the first period. The study presented in [102] shows that photos shared on Flickr gain the majority of their engagements within the first week (early period).

Figure 3.20 shows the dynamics of the number of views related to three different photos shared on Flickr. The number of views of each photo is extracted daily for a period of 30 days. We can observe that all three images' dynamics have similar shapes (especially the first and the second plot from the top), but they have different scales (i.e., the number of views after 30 days is different of about an order of magnitude). As claimed in previous works [102], the number of views increases in the first few days, and then remains stable over time. As consequence, the popularity score computed as in Equation 3.2 decreases with the oldness of an image post, but

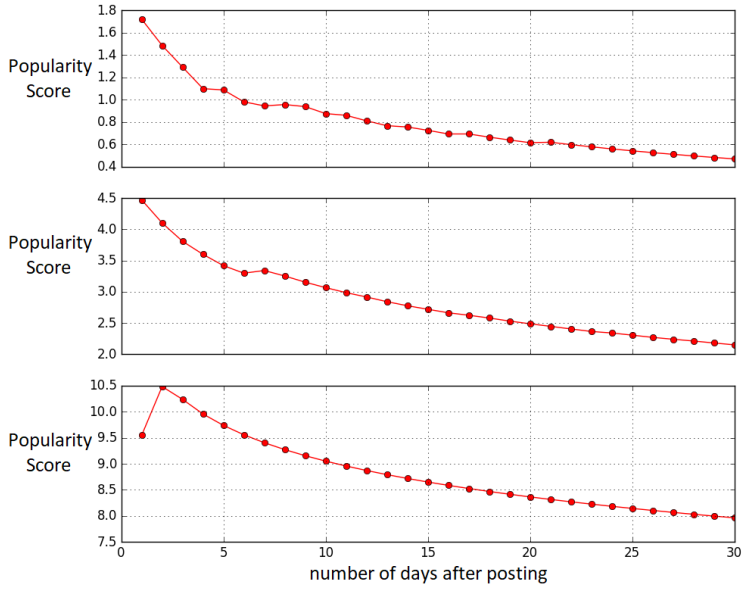


Figure 3.21: Popularity scores over time computed using Equation 3.2. After the first period the popularity score decreases with the post age. This effect is caused by the very low engagement around the photo after the early period.

this fact has not been considered in the popularity score formulation. In other words, two pictures with similar engagement dynamic but very different lifetime, have to be ranked differently. This is simple to understand by considering Figure 3.21, which shows the popularity score of Equation 3.2 computed over time for the three examples depicted in Figure 3.20. Furthermore, the presented study also reveals that several engagement dynamic trends (i.e., engagement shapes) exist. This also affects the comparison of two pictures by means of their popularity scores computed at different time instants.

In our study we address the problem of popularity prediction taking into account the temporal dynamic of the engagement score (i.e., the variation of the engagement value over time).

3.7.3 Proposed Dataset

The dataset has been obtained by exploiting the Flickr API [11], which allows to retrieve images and the related information shared by users which are publicly available on the social platform. With a crawling process, more than 20K images have

[11] https://www.flickr.com/services/api/

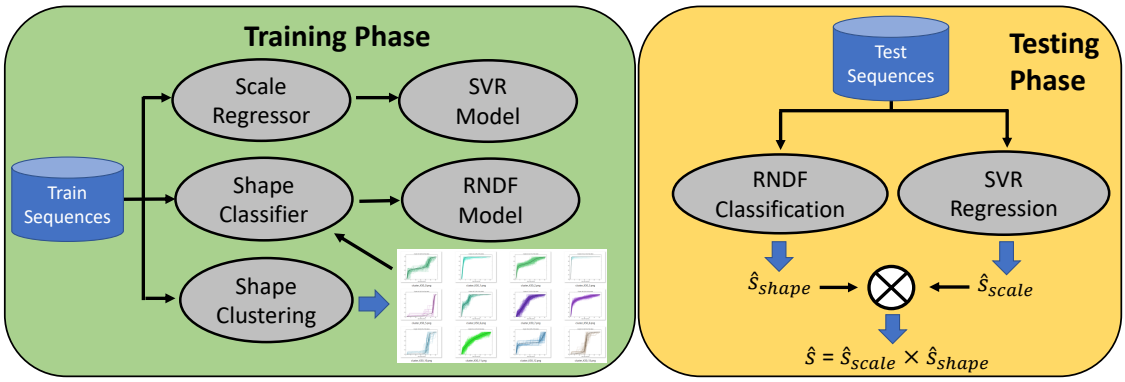


Figure 3.22: The proposed approach models the problem as the estimation of two properties: the sequence engagement shape $\hat{s}_{shape}$ and scale $\hat{s}_{scale}$. The shape property is estimated by exploiting a Random Forest Classifier (RNDF), whereas a Support Vector Regressor (SVR) is employed to produce the scale factor $\hat{s}_{scale}$. The labels used to train the shape classifier are generated by mapping the shapes s_{shape} (Eq. 3.6) in one of the 50 shape prototypes obtained by clustering. This pre-processing step is performed after a clustering procedure over the training dataset (green backgroud). The regressor is trained by considering the value s_{scale} of the training set as Ground Truth.

been downloaded and tracked considering a period of 30 days. In particular, the first day of crawling the procedure downloads a batch of the latest photos uploaded on Flickr by users that day. All the social features related to the users, the photos and groups were collected, as well as a number of information useful to compute the engagement scores are downloaded for the following 30 days (i.e., number of views, number of comments, and number of favorites). In all the experiments, the engagement score is computed using the number of views, as it is the most used score in the state-of-the-art. However, the other two information are available with the dataset for further studies. In order to obtain a fine-grained sampling, the first sample of the social data related to a post is taken within the first 2 hours from the time of the image upload. Then, a daily procedure takes at least 2 samples per day of the social data for each image. We removed posts with a dimension of the image less than 10 Kb, since it usually corresponds to icons or place-holder images used by Flickr to indicate that the picture is no longer available.

The engagement score sequences of 30 days have been pre-processed in order to obtain values with regular time intervals of 24 hours. To this aim, we approximated the score function between two consecutive samples with a linear function. Figure 3.20 shows three examples of the score obtained with this process.

Several social features related to users, photos and groups have been retrieved. In particular, for each user we recorded: the number of the user's contacts, if the user has a pro account, photo counts, mean views of the user's photos, number of groups the user is enrolled in, the average number of members and photos of the user's groups. For each picture we considered: the size of the original image, title, description, number of albums, groups the picture is shared in, the average number of members and photos of the groups in which the picture is shared in, as well as the tags associated by the user. Moreover, for each photo, the dataset includes the geographic coordinates (when available), the date of the upload, the date of crawling, the date of when the picture has been taken, as well as the Flickr IDs of all the considered users, photos and groups. These information can be exploited to extend the crawling with more specific data and analysis. The collected dataset is publicly available [12].

We tracked a total of 21.035 photos for 30 days. Some photos have been removed by authors (or not longer accessible through the APIs) during the period of crawling. As result, there are photos that have been tracked for a longer period than others. In particular the dataset consists of:

- 19.213 photos tracked at least for 10 days;
- 18.838 photos tracked at least for 20 days;
- 17.832 photos tracked at least for 30 days;

In our experiments we considered the set of photos tracked for 30 days, as it represents the most challenging scenario: forecasting the image popularity (i.e., the number of views) over 30 days at posting time. The average number of images per user in the dataset is lower than two ($\sim$ 1.6). Therefore, the dataset represents a setting where different images belong to different users. This is often the case in search engine results. Indeed, search engines can exploit popularity prediction systems in order to better rank the retrieved results.

[12]https://www.visiongarage.altervista.org/popularitydynamics/

3.7.4 Proposed Method

Given a sequence s related to the number of views of a posted image over n days, the proposed framework models it as the combination of two properties, namely the *sequence shape* and the *sequence scale*. We define these properties as in Equation 3.6. In particular, s_{scale} is defined as the maximum value of s, whereas s_{shape} is obtained by dividing each value of s by s_{scale}:

$$\begin{aligned} s &= [v_0, v_1, ..., v_n] \\ s_{scale} &= max\{s\} = v_n \\ s_{shape} &= \left[\frac{v_0}{v_n}, \frac{v_1}{v_n}, ..., \frac{v_n}{v_n}\right] \end{aligned} \tag{3.6}$$

Therefore, given s_{shape} and s_{scale}, we can obtain the sequence s as follows.

$$s = s_{shape} \times s_{scale} \tag{3.7}$$

Note that, since each sequence represents a cumulative function, the last value of the sequence (i.e., v_n) always corresponds to the maximum value of the sequence s. The engagement sequence is hence considered as a pair of shape and scale. The shape describes the general dynamic (i.e., trend) of the sequence in the monitored period, regardless its actual values. The sequence scale represents the degree of popularity reached by the photo in n days. We perform two separate estimation for the shape and the scale of the engagement related to a period of n=30 days. Then, the two information are combined to perform the estimation of the final engagement sequence associated to the photo.

The proposed method assumes the independent relationship between the scale and the shape of a sequence. Indeed, two sequences with the same shape could have very different scales (e.g., as the three examples depicted in Figure 3.20 of the paper), and vice-versa. To motivate this assumption, we analysed the distributions of the s_{scale} values related to sequences grouped by the assigned shape prototype (i.e., s^*_{shape}). Figure 3.23 shows the distributions of the s_{scale} values related to sequences grouped by the assigned shape prototype. (i.e., s^*_{shape}). There is a huge variability of the s_{scale} values within the sequences of the same shape. Note that the s_{scale} axis

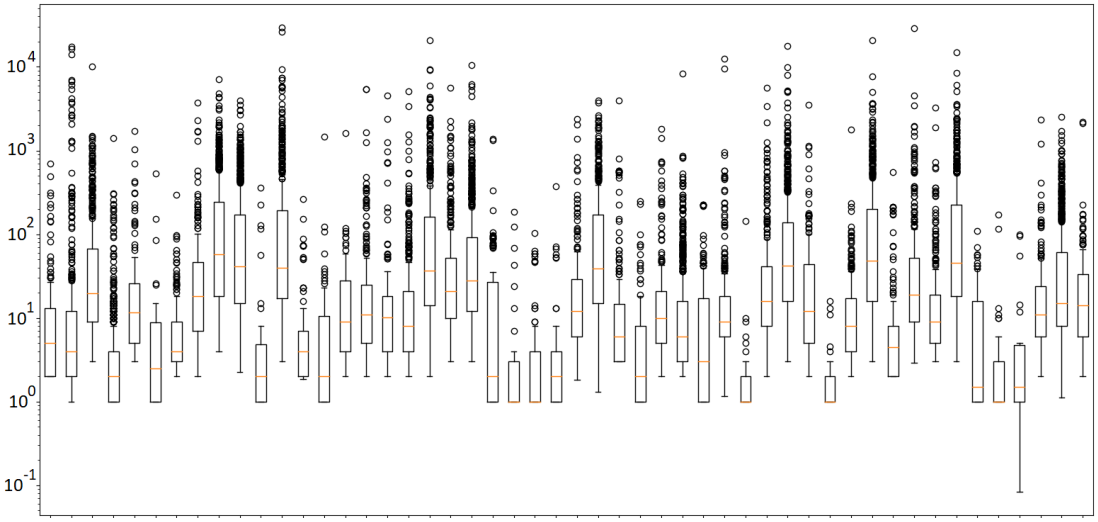


Figure 3.23: Distributions of the s_{scale} values of sequences grouped by the s^*_{shape} prototypes assigned by the clustering procedure.

is represented in logarithmic scale, to better visualize the range of the distributions. This motivates our assumption that the two properties are unrelated.

Figure 3.22 shows the scheme of the proposed approach. The shape of the training sequences are grouped with a clustering procedure during the training phase. All the sequences in the same cluster are represented by a shape prototype denoted by $s^*{}_{shape}$. The obtained shape prototypes $s^*{}_{shape}$ are used as labels to train a classifier in order to predict the shape prototype to which a new sequence has to be assigned, given the set input social features. The predicted shape prototype for a new sequence is denoted by $\hat{s}_{shape}$. A Support Vector Regressor (SVR) is also trained to infer the value of s_{scale} given the social features. The output of the regressor is denoted by $\hat{s}_{scale}$. The final engagement estimation of an image post for the period of n days is obtained as $\hat{s} = \hat{s}_{shape} \times \hat{s}_{scale}$. To measure the performance of the system we use a test dataset and the Root Mean Squared Error (RMSE) between the Ground Truth sequences s of the test sequences and the predicted ones $\hat{s}$. The experimental results reported in this paper have been obtained by averaging the results of 10 random train/test splits with a proportion of 1:9 between the number of test and training images.

In the experiments, we evaluated the proposed approach by exploiting features extracted from the user information, the photo metadata or from the visual content. Although in previous works on popularity prediction the most effective results

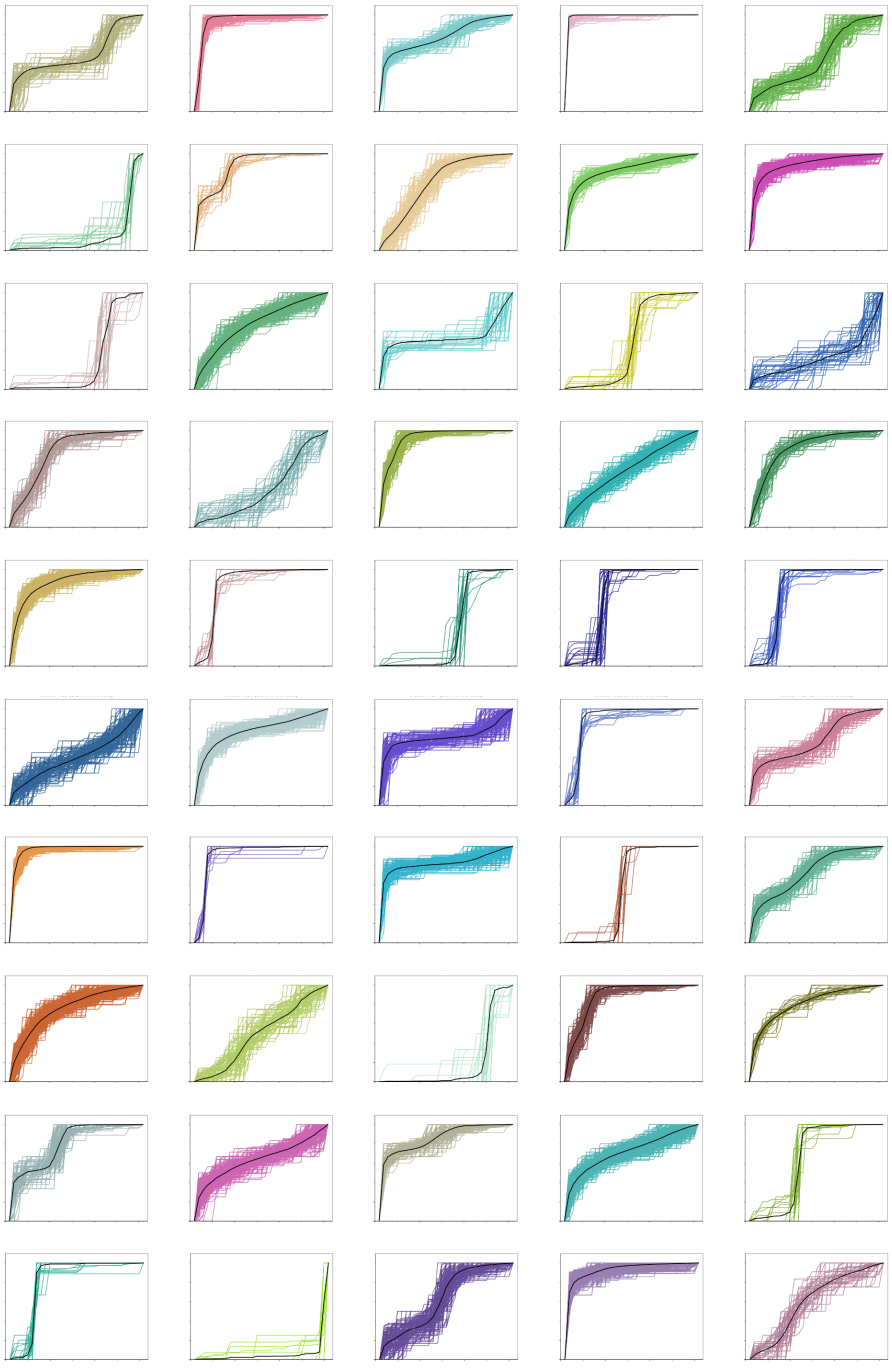


Figure 3.24: The considered shape prototypes obtained by clustering the training sequences. (best seen in color).

have been obtained by considering a combination of user information and photo meta-data, some experiments [34] suggest that the semantic content of the picture also influence the prediction. For this reason we evaluated 6 visual representation extracted from the pictures by exploiting three state of the art Convolutional Neural Networks (CNN). For each architecture we extracted the last two layers of activations before the softmax, referred here as $f1$ and $f2$. Specifically we considered the following architectures:

- **Hybridnet [8]:** specialized to classify images into one of 1183 categories (978 objects and 205 places).
- **DeepSentiBank [46]:** specialized to assign an Adjective-Noun Pair to an input image among 4342 different ANPs [41].
- **GoogleNet [132]:** specialized to perform image classification among 1000 object categories.

The following paragraphs describe the details of each step of the proposed approach.

Shape Prototyping

Considering Equation 3.6, all the s_{shape} sequences have values in the range $[0, 1]$. We consider that all the sequences with the "same" dynamics will have a very similar s_{shape}. Since groups of sequences with the same shape embody instances with a common engagement trend, in the first step of our approach we try to infer a number of popularity shape prototypes representing the different groups of shapes. To this aim, we perform a K-means clustering to group the training sequences. The obtained cluster centroids represent the dynamic models for the sequences within clusters (i.e., each centroid sequence is a shape prototype). In order to select the best value for K, we run the K-means algorithm multiple times considering a large range of values for K and evaluating the quality of the clustering for each run. T hen we selected the optimal K considering the Within cluster Sum of Squares (WSS) and the Between cluster Sum of Squares (BSS) indices, which provide a measure of cluster cohesion and cluster separation respectively. In particular, we want the WSS value to be low and the BSS value to be high. The WSS/BSS analysis has been

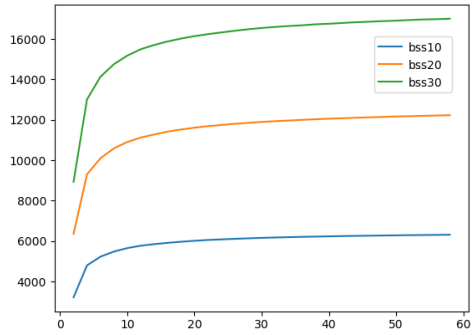


Figure 3.25: BSS analysis.

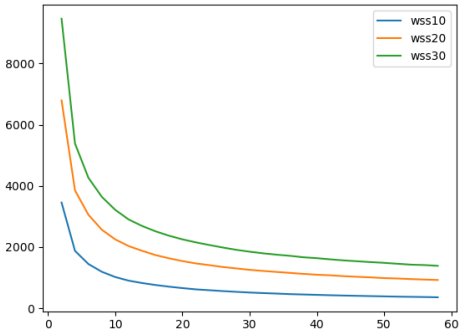


Figure 3.26: WSS analysis.

performed considering the groups of sequences of 10, 20 and 30 days. Figure 3.25 and Figure 3.26 show the values of BSS and WSS respectively. When K increases the cluster elements are closer to the cluster centroid. The improvements will decline, and at some point become more stable. We then tested a few values of K which corresponds to the plateau of the BSS and the WSS functions and selected the best parameter.

Indeed, the experiments based on BSS and WSS analysis suggest more than one value for the optimal number of clusters (i.e., K). The best results have been obtained with K = 45 (10 and 20 days sequences) and K = 50 (30 days sequences).

Figure 3.24 shows the clustering result for the sequences of 30 days, with a set of plots. In particular, for each plot, the black bold line depicts the cluster centroid (i.e., the shape prototype), whereas the coloured lines represent the shapes of the training sequences belonging to the cluster.

Shape Prediction

The result of the shape clustering is a set of shape prototypes. Given this "dictionary" of shapes, any sequence can be assigned to a cluster by comparing its shape with respect to the prototypes. By exploiting the set of shape prototypes we labelled all the sequences in the training and testing dataset by assigning each sequence to a prototype. Then, considering the training sequences, we built a classifier that takes only the social features associated to a post, and predicts the shape (i.e., the prototype) of the corresponding sequence. In order to find the best classifier, we

Algorithm 1: Popularity Sequence Prediction.

Data: Input feature X, Ground Truth sequence s, shape prototype ${s^*}_{shape}$
Result: Inferred sequence $\hat{s}$
$n \leftarrow |s|$;
$\hat{c} \leftarrow RNDF.predict(X)$; // predict the cluster
$\hat{s}_{shape} \leftarrow getPrototype(\hat{c})$; // get the prototype sequence
$\hat{p} \leftarrow SVR.predict(X)$; // predict the popularity score
$\hat{s}_{scale} \leftarrow (e^{\hat{p}} - 1) \times n$;
$\hat{s} \leftarrow \hat{s}_{scale} \times \hat{s}_{shape}$; // predicted sequence

evaluated a pool of algorithms common used for classification tasks, as well as several variations of the social features to be used during classification. In particular, we considered the following classifiers: Random Forest Classifier (RNDF), Decision Tree Classifier (DT), k-Nearest Neighbour (kNN), SVM with RBF or linear kernel and Multi-layer Perceptron Classifier (MLP). The best results have been achieved by using all the social features as input combined with a RNDF Classifier. The DT classifier obtains slightly lower results, whereas kNN and SVM achieved lower performances. The MLP Classifier resulted the worst approach, probably due to the limited number of examples for each class and the unbalanced distribution of the elements in the clusters. To deal with this issue we performed a stratified approach by considering 10 random splits in all our experiments. Considering stratification we ensure that each fold contains roughly the same proportion of the classes. Some of the considered methods required the selection of parameters (e.g., the number of neighbours K for kNN, the parameters C and γ for LSVM and RBFSVM, etc.) which we have established with a grid search method over the training data. In the proposed approach, given a new test image post, the RNDF classifier is hence used to assign a shape prototype starting only from social features.

Scale Estimation

In order to estimate the value s_{scale}, we trained a Support Vector Regressor (SVR) by considering the training set. In particular, we consider the s_{scale} values to compute the popularity score as in Equation 3.2, then the SVR is trained to predict the popularity score. After the prediction, the estimation of the number of views is

obtained by inverting Equation 3.2. Let $\hat{p}$ be the popularity score estimated by the SVR, the number of views is hence computed by the following equation:

$$\hat{s}_{scale} = (e^{\hat{p}} - 1) \times n \tag{3.8}$$

In our experiments, we considered several combinations of social features. In particular, we evaluated the estimation performances to infer the scale by training the SVR with a single social feature as input. For each experimental setting, we computed the Spearman's correlation between the predicted score and the Ground Truth. This provided a measure of correlation between the features and the value of s_{scale} (it represents the common way to evaluate the classic popularity prediction approaches). In the second stage of experiments, we trained the SVR by considering the concatenation of social features as input. The groups of features have been selected in a greedy fashion considering the performance obtained individually. We have also evaluated a set of approaches that exploits as input the results obtained by the SVRs trained with the single features. As these methods perform a fusion of several outputs after the prediction, they are often referred as "late fusion" strategies. In particular we evaluated the following approaches:

- **Late Fusion 1:** the outputs of the SVRs are averaged;
- **Late Fusion 2:** the outputs of the SVRs are concatenated and used to train a new SVR;
- **Borda Count:** the output is obtained by computing a weighted average of the single feature SVR outputs. The m evaluated features are ranked in descending order based on the achieved Spearman's correlation. The weight of the output corresponding to the top ranked feature is set to m, the second ranked feature is weighted with $m - 1$, and so on;
- **Weighted Fusion k:** the features are ranked in descending order considering the achieved Spearman's correlation. Then, the final output is obtained by computing a weighted average of the first k single feature SVR outputs. In this case the weights corresponds to the Spearman's correlations achieved individually.

Table 3.7: Summary of the methods compared in the evaluation.

	Shape	**Scale**	**Note**
Baseline A (shape+scale)	Shape prototype (s^*_{shape})	Max views (s_{scale})	This method achieves the minimum possible error. The measured error is due to the clustering approximation of the sequences.
Baseline B (scale)	Predicted ($\hat{s}_{shape}$)	Max views (s_{scale})	The measured error is due to the clustering approximation of the sequences and the error in the prediction of the shape prototype.
Baseline C (shape)	Shape prototype (s^*_{shape})	Predicted ($\hat{s}_{scale}$)	The measured error is due to the clustering approximation of the sequences and the error in the estimation of the scale.
Proposed Method	Predicted ($\hat{s}_{shape}$)	Predicted ($\hat{s}_{scale}$)	The measured error is due to all the above factors.

Table 3.8: RMSE errors of Baseline A and Baseline B. These measures are not affected by the prediction of the scale and depends only on the clustering and shape prototyping steps.

	RMSE		
	10 days	**20 days**	**30 days**
Baseline A (shape+scale)	2.81	3.85	4.97
Baseline B (scale)	22.99	29.59	7.51

Scale estimation results, in terms of Spearman's correlation, are reported in Table 3.9 and 3.10 (see Spearman column).

3.7.5 Evaluation

Given an image represented by a set of social feature, the proposed approach exploits a Random Forest Classifier to predict the shape prototype $\hat{s}_{shape}$. Then an SVR is used to estimate the popularity score of the image after n days. This value is then transformed by using Equation 3.8, in order to obtain the scale estimation $\hat{s}_{scale}$. Finally, the estimated shape $\hat{s}_{shape}$ and the estimated scale $\hat{s}_{scale}$ are combined to obtain the predicted sequence $\hat{s}$ (Figure 3.22). The aforementioned procedure is described in the Algorithm 1. For evaluation purposes, the predicted sequence $\hat{s}$ is compared with respect to the Ground Truth s by means of the RMSE measure. We repeated the whole pipeline (i.e., shape clustering, shape prediction, scale estimation) and the evaluation procedure considering n = 10, n = 20 and n = 30.

Baseline Performances

With the aim to perform a better evaluation of the proposed method, we defined three baselines useful to perform a comparison. Each baseline exploits some Ground Truth knowledge about either the scale and the shape of the sequences. Therefore, the baselines error rates can be considered as lower bounds for the proposed approach. The considered baselines are the following:

- **Baseline A:** the inferred sequence is obtained by considering the shape of the Ground Truth (s^*_{shape}) and the Ground Truth scale value (s_{scale}). This method achieves the minimum possible error as both values are taken from the Ground Truth. The measured error is due to the clustering approximation of the sequences.

- **Baseline B:** in this case the shape is predicted ($\hat{s}_{shape}$), whereas the scale value is taken from the Ground Truth (s_{scale}). The measured error is due to the clustering approximation of the sequences and the error of the classification to assign the shape prototype.

- **Baseline C:** this baseline combines the shape of the Ground Truth (s^*_{shape}) and the predicted scale value ($\hat{s}_{scale}$). The measured error is due to the clustering approximation of the sequences and the error in the estimation of the scale.

A summary of the evaluated approach and baselines is reported in Table 3.7. For each Baseline, the known information is reported in brackets (i.e., the data taken from the Ground Truth). Note that the performances of Baselines A and B are not affected by the scale estimation. Table 3.8 shows the RMSE errors of Baseline A and Baseline B, for the three dataset scenarios. These measures are not affected by the prediction of the scale and depends only on the clustering and shape prototyping steps. When both the shape and scale information are known in the baseline (i.e., Baseline A), the most challenging scenario is the prediction of sequences of 30 days. When only the scale is known (i.e., Baseline B), the prediction of 10 days and 20 days sequences are more challenging than the prediction of 30 days sequences. This happens due to the higher variability of the shapes in the first days, whereas the most of the sequences have a flat-like shape in the last 10 days (see Figure 3.24).

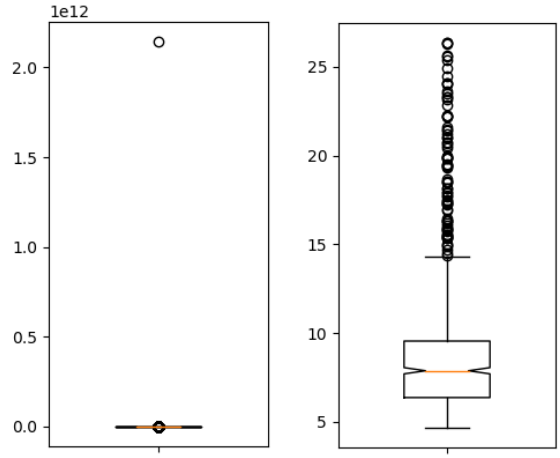


Figure 3.27: Box-and-whisker plot of an example of RMSE values computed on a test set (left). The mean RMSE value is skewed by only one value that is very large compared to others. The right plot shows the same distribution after removing the values lower than the first quartile and higher than the third quartile. In this example, the trimmed mean is 9,00 and the median is 7,87.

Popularity Dynamic Results

As detailed above, the errors concerning the baselines are related to either the shape and the scale estimations, depending on the definition of the baseline. However, the combination of the two predictions in our method is more sensitive to errors in the scale predictions. Indeed, an error in the scale affects all the elements of the sequence and hence the RMSE value. Looking at the distribution of the test errors, we found that the mean RMSE is skewed by a few values that are larger than the others by several orders of magnitude.

The left plot in Figure 3.27 shows the box-and-whisker plot of the computed RMSE values. From this plot, it's clear that the mean of the RMSE errors is not a good method for the evaluation of the performance, as it is skewed by few large values. Figure 3.23 shows the presence of several s_{scale} values which are outliers with respect their distributions. Indeed, the dataset includes few examples with very large scales. For instance, there are only 11 sequences with s_{scale} between 10.000 and 30.000 views, whereas the median values (depicted by the orange lines in Figure 3.23) are all lower than 100. The presence of such few uncommon examples with very large magnitudes caused the differences in the test performances, as observed in the example of Figure 3.27.

For these reasons we considered two more performance measures for the evaluation of either the proposed method and the Baseline C (i.e., the two methods that infer the scale of the sequences). Specifically, we considered the following measures:

- **Truncated RMSE:** the truncated mean (or trimmed mean) of RMSE involves the calculation of the mean of RMSE after discarding an equal amount of either high and low tails of a distribution. In our experiments, when computing the trimmed RMSE measure (tRMSE), we considered the 25% trimmed mean (i.e., the lowest 25% and the highest 25% are discarded), also known as the interquartile mean. The right plot in Figure 3.27 shows the error distribution after removing the higher and lower 25% of values.

- **Median test error:** after performing the prediction of the sequences of the test set, the median of the test RMSE is considered. Indeed, the median is less susceptible to outliers than the mean and can provide more insights for the evaluation.

In the following paragraph, the performances of the proposed method on the prediction of the 30 days sequences are detailed. Then, the prediction of sequences with length 20 and 10 are discussed. Finally all the results are compared and commented.

Performances on 30 days prediction

Tables 3.9 and 3.10 show the results for Baseline C and for the proposed approach in terms of tRMSE and Median RMSE at varying of the input feature used by the scale regressor (i.e., the SVR). The results obtained by feeding the SVR with a single feature are detailed in Table 3.9. In particular, the Spearman's correlation between the input feature and the Ground Truth popularity is reported in the fourth column. As we can observe, the feature with the higher correlation is the mean number of user's photo views (MeanViews). The second ranked feature is the number of the groups the user is enrolled in (GroupsCount). Considering the achieved Spearman's values, one can observe that the features related to the user have higher correlation values with respect to the others. The other columns in Table 3.9 report the error rates on the estimation of the prediction of the whole sequence in terms of trimmed RMSE (tRMSE) and Median RMSE (RMSE MED)

Table 3.9: Results obtained by considering a single feature approach for the predictions of 30 days sequences. For each experimental setting, the fourth column reports the Spearman's correlation of the predicted s_{scale} value, the others columns report the trimmed RMSE (tRMSE 0.25) and the median RMSE (RMSE MED) considering the prediction of the whole temporal sequence. The results achieved by the proposed method are compared with respect to the Baseline C in which the shape prototype is known.

				Proposed Method		**Baseline C (shape)**	
Feature Source	**Feature ID**	**Features**	**Spearman**	**tRMSE 0.25**	**RMSE MED**	**tRMSE 0.25**	**RMSE MED**
User	**0**	**Ispro**	0,25	11,74	8,87	11,50	8,79
User	**1**	**Contacts**	0,52	10,48	8,96	10,22	8,69
User	**2**	**PhotoCount**	-0,06	11,00	9,65	10,70	9,31
User	**3**	**MeanViews**	**0,73**	**9,68**	**8,19**	9,44	7,95
User	**4**	**GroupsCount**	0,53	10,42	8,55	10,18	8,42
User	**5**	**GroupsAvgMembers**	0,44	11,19	9,37	10,91	9,11
User	**6**	**GroupsAvgPictures**	0,46	11,57	8,96	11,31	8,90
Photo	**7**	**Size**	0,05	11,03	9,45	10,72	9,18
Photo	**8**	**Title**	0,03	11,01	9,63	10,72	9,30
Photo	**9**	**Description**	0,06	10,96	9,58	10,66	9,21
Photo	**10**	**NumSets**	0,20	11,57	9,33	11,26	9,14
Photo	**11**	**NumGroups**	0,34	10,95	9,63	10,65	9,32
Photo	**12**	**AvgGroupsMemb**	0,34	10,88	9,54	10,59	9,24
Photo	**13**	**AvgGroupPhotos**	0,34	10,91	9,58	10,61	9,28
Photo	**14**	**Tags**	0,11	11,27	9,36	10,97	9,20
Visual	**hybrid_f1**	**Hybridnet fc7**	0,22	21,61	17,97	20,08	17,44
Visual	**hybrid_f2**	**Hybridnet fc8a**	0,26	13,37	11,76	12,95	11,38
Visual	**senti_f1**	**DeepSentiBank fc7**	0,25	21,16	18,27	20,60	17,82
Visual	**senti_f2**	**DeepSentiBank fc8**	0,30	16,52	14,32	15,99	13,77
Visual	**google_f1**	**GoogleNet pool5/7x7_s1**	0,26	13,85	12,15	13.43	11,74
Visual	**google_f2**	**GoogleNet loss3/classifier**	0,27	13,18	11,61	12,76	11,17

for the proposed method (columns 5 and 6) and the Baseline C (columns 7 and 8). Each row reports the results obtained by varying the features used in the estimation of the scale. The experimental results show that the best features are the MeanViews and GroupsCount, which allow to obtain evaluation performances very close to the Baseline C. Some interesting results have been obtained considering photo's features such as the number of groups the photo has been shared (NumGroups), and the statistics of such groups (AvgGroupsMemb and AvgGroupPhotos).

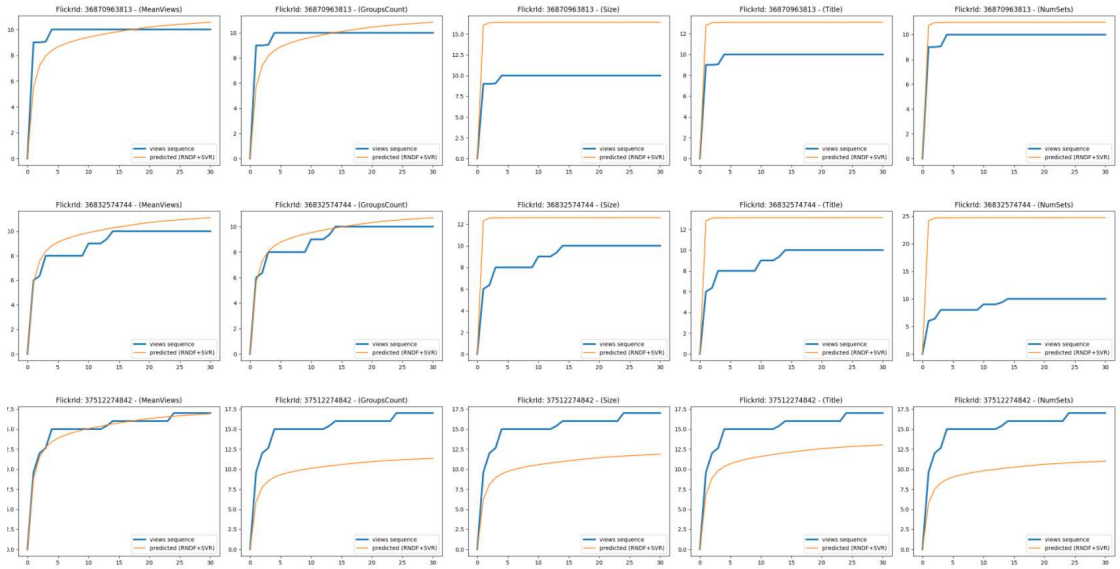
The experiments pointed out that the visual features achieves higher error rates. Indeed, the popularity of a photo in terms of number of views is directly related to the capability of the user and the photo to reach as many users as possible in the social platform. Based on the results obtained in the evaluation of the single features (Table 3.9), we further considered the combination of the most effective

Table 3.10: Evaluation results for the prediction of the 30 days sequences obtained by combining the features.

			Proposed Method		**Baseline C (shape)**	
Feature ID	**Features**	**Spearman**	**tRMSE 0.25**	**RMSE MED**	**tRMSE 0.25**	**RMSE MED**
all	**concat(0-14)**	0,63	9,59	7,43	9,38	7,26
user	**concat(0-6)**	0,66	9,86	**7,03**	9,67	6,92
photo	**concat(7-14)**	0,28	10,80	8,58	10,50	8,30
best_photo	**concat(11, 12, 13)**	0,34	10,63	9,09	10,35	8,84
user2	**concat(3,4)**	0,71	9,60	7,66	9,38	7,54
user3	**concat(1,3,4)**	0,71	9,48	7,51	9,28	7,37
user5	**concat(1, 3, 4, 5, 6)**	0,68	9,77	7,30	9,59	7,24
	concat(user2, best_photo)	**0,72**	9,49	7,55	9,27	7,43
	concat(user3, best_photo)	0,71	**9,37**	7,38	9,16	7,27
	concat(user5,best_photo)	0,69	9,48	7,28	9,28	7,20
	Late Fusion 1 (AVG)	0,59	10,70	9,34	10,43	9,06
	Late Fusion 2 (SVR)	0,46	11,91	8,91	11,66	8,86
	Borda Count	0,63	10,60	9,16	10,33	8,87
	Weighted Fusion 2	0,71	9,88	8,34	9,63	8,18
	Weighted Fusion 3	0,71	9,95	8,49	9,69	8,26
	Weighted Fusion 4	0,68	10,30	8,65	10,05	8,49
	Weighted Fusion 5	0,67	10,44	8,79	10,18	8,63
	Weighted Fusion 6	0,67	10,47	8,88	10,20	8,67
	Weighted Fusion 7	0,67	10,51	8,92	10,24	8,64
	Weighted Fusion 8	0,67	10,52	8,96	10,25	8,69
	Weighted Fusion 9	0,66	10,51	8,99	10,24	8,73
	Weighted Fusion 10	0,65	10,52	9,03	10,25	8,80
	Weighted Fusion 11	0,65	10,53	9,06	10,26	8,80
	Weighted Fusion 12	0,65	10,53	9,07	10,26	8,81
	Weighted Fusion 13	0,65	10,53	9,08	10,27	8,82
	Weighted Fusion 14	0,65	10,54	9,08	10,27	8,82
	Weighted Fusion 15	0,65	10,54	9,08	10,27	8,81

features for the estimation of the sequence scale. To this aim, we evaluated several early and late fusion strategies. The early fusion consists on creating a new input for the SVR, obtained by the concatenation of the selected features. In particular, we evaluated 10 different combinations of features (see the first 10 rows in Table 3.10). Each combination is assigned to an identifier for readability (first column in Table 3.10). The obtained results show that in the experiments which involve user related features, the achieved error rates are lower and similar. Indeed, the higher error rates are obtained by the combinations that do not consider user features (i.e., combinations with IDs "photo" and "best_photo"). The best results in terms of tRMSE are obtained by combining the three best user's features (i.e., MeanViews, GroupsCount and Contacts) and the best photo's features (i.e., NumGroups, AvgGroupsMemb and AvgGroupsPhotos). Whereas the best results in terms of Median

Figure 3.28: Comparison between the Ground Truth sequence (blue) and the predicted sequence (orange) related to test examples obtained by using five different features. Namely the mean number of views for the user (MeanViews), the number of groups the user is enrolled in (GroupsCount), the number of pixels of the original image (Size), the length of the photo title (Title) and the number of albums the photo is shared in (NumSets).



RMSE are obtained by using only the user related features (i.e., the concatenation of the features with indices from 0 to 6, identified by the ID "user" in Table 3.10). In Figure 3.28 the results related to three sequences achieved by five different features are shown. In this Figure, the proposed method is compared with respect to the sequence Ground Truth. Here, is possible to observe how the estimation changes by considering different input features.

We further evaluated the daily error rates obtained by the proposed approach. Instead of computing the mean error between the predicted sequence and the Ground Truth, the daily squared errors are collected, and the daily tRMSE is then computed for all the error rates of the same day. The feature "MeanViews" obtains the lower error over all the period of observation by a certain margin. The features "GroupsCount" and "Contacts" achieves similar results in the early period (i.e., the first week), then their performances become slightly different. Substantially, the daily evaluation confirms the results of in Table 3.9 and Table 3.10.

Table 3.11: Evaluation results for the prediction of the 10 days sequences.

				Proposed Method		Baseline C (shape)	
Feature Source	**Feature ID**	**Features**	**Spearman**	**tRMSE 0.25**	**RMSE MED**	**tRMSE 0.25**	**RMSE MED**
User	**0**	**Ispro**	0,24	9,12	**6,66**	8,18	6,29
User	**1**	**Contacts**	0,52	7,88	7,08	7,33	6,41
User	**2**	**PhotoCount**	-0,08	8,33	7,65	7,72	6,78
User	**3**	**MeanViews**	**0,71**	**7,60**	6,84	7,02	6,08
User	**4**	**GroupsCount**	0,53	7,93	6,89	7,36	6,12
User	**5**	**GroupsAvgMembers**	0,43	8,43	7,34	7,80	6,62
User	**6**	**GroupsAvgPictures**	0,45	8,77	7,24	8,11	6,43
Photo	**7**	**Size**	0,06	8,38	7,53	7,72	6,69
Photo	**8**	**Title**	0,02	8,31	7,64	7,72	6,79
Photo	**9**	**Description**	0,06	8,30	7,63	7,71	6,76
Photo	**10**	**NumSets**	0,21	8,72	6,68	7,90	6,49
Photo	**11**	**NumGroups**	0,33	8,33	7,65	7,72	6,78
Photo	**12**	**AvgGroupsMemb**	0,33	8,33	7,65	7,71	6,78
Photo	**13**	**AvgGroupPhotos**	0,33	8,34	7,65	7,72	6,78
Photo	**14**	**Tags**	0,12	8,54	7,54	7,88	6,66
Visual	**hybrid_f1**	**Hybridnet fc7**	0,22	16,01	13,96	13,75	11,68
Visual	**hybrid_f2**	**Hybridnet fc8a**	0,27	10,84	9,53	9,35	8,11
Visual	**senti_f1**	**DeepSentiBank fc7**	0,25	16,74	14,12	14,24	11,91
Visual	**senti_f2**	**DeepSentiBank fc8**	0,30	13,45	11,57	11,39	9,70
Visual	**google_f1**	**GoogleNet pool5/7x7_s1**	0,27	11,38	10,19	9,83	8,66
Visual	**google_f2**	**GoogleNet loss3/classifier**	0,28	10,71	9,58	9,33	8,14

Performances on 10 and 20 days prediction

The resulting clusters in Figure 3.24 shows that the most of the sequences have a flat shape in the last days. By the other end, we can observe an higher shape variability in the first days. Such observation has been confirmed in the cluster analysis (see Section 3.7.4), which suggest a slightly lower number of clusters for the sequences with length 10 and 20 (45 clusters) with respect to the sequences with length 30 (50 clusters).

For this reason, we further evaluated our system on the sequences with length of 10 and 20 days. Note that both groups of data include the sequences of the 30 days set. Figure 3.29 shows the Spearman's correlations values between the $\hat{s}_{scale}$ estimated by the SVR properly trained with a specific feature and the Ground Truth s_{scale} value, for the 10 days, 20 days and 30 days scenarios. The features are ranked by the mean correlation among the three cases. As we can observe, the ranking of the features is similar for the three scenarios. Most of the user's features (MeanViews, GroupsCount, Contacts, GroupsAvgPictures and GroupsAvgMembers) achieve the

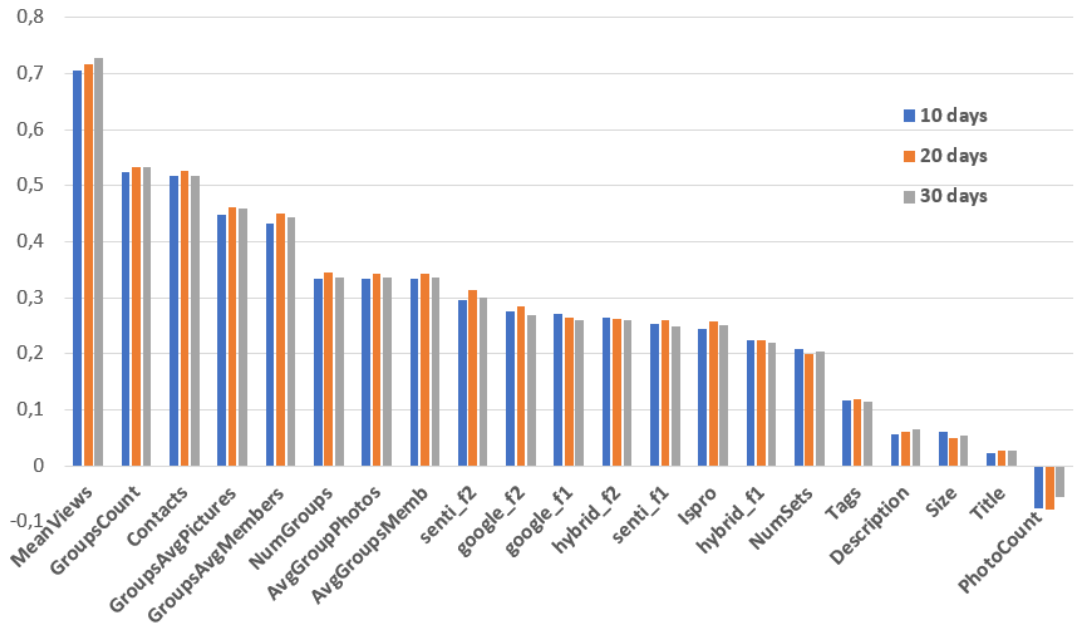
Table 3.12: Evaluation results for the prediction of the 10 days sequences obtained by combining the features.

			Proposed Method		**Baseline C (shape)**	
Feature ID	**Features**	**Spearman**	**tRMSE 0.25**	**RMSE MED**	**tRMSE 0.25**	**RMSE MED**
all	**concat(0-14)**	0,61	7,73	6,18	6,84	5,37
user	**concat(0-6)**	0,64	8,00	6,00	7,11	5,48
photo	**concat(7-14)**	0,28	8,34	6,84	7,58	6,10
best_photo	**concat(11,12,13)**	0,33	8,10	7,30	7,54	6,59
user2	**concat(3,4)**	0,69	7,56	6,44	6,93	5,66
user3	**concat(1,3,4)**	0,69	7,50	6,37	6,84	5,63
user5	**concat(1,3,4,5,6)**	0,66	7,82	5,92	7,07	5,61
	concat(user2,best_photo)	**0,70**	7,46	6,41	6,86	5,60
	concat(user3,best_photo)	0,69	**7,39**	6,37	6,78	5,59
	concat(user5,best_photo)	0,67	7,53	**5,91**	6,86	5,55
	Late Fusion 1 (AVG)	0,57	8,20	7,53	7,53	6,68
	Late Fusion 2 (SVR)	0,47	9,45	6,87	8,37	6,62
	Borda Count (wAVG)	0,62	8,16	7,39	7,48	6,60
	Weighted Fusion 2	0,70	7,74	6,93	7,11	6,15
	Weighted Fusion 3	0,69	7,77	6,95	7,13	6,23
	Weighted Fusion 4	0,66	8,06	7,15	7,38	6,30
	Weighted Fusion 5	0,65	8,11	7,20	7,43	6,35
	Weighted Fusion 6	0,65	8,13	7,22	7,46	6,45
	Weighted Fusion 7	0,66	8,14	7,25	7,48	6,54
	Weighted Fusion 8	0,66	8,15	7,28	7,50	6,59
	Weighted Fusion 9	0,64	8,14	7,31	7,46	6,55
	Weighted Fusion 10	0,63	8,13	7,33	7,45	6,53
	Weighted Fusion 11	0,63	8,13	7,35	7,45	6,54
	Weighted Fusion 12	0,63	8,13	7,35	7,45	6,54
	Weighted Fusion 13	0,63	8,13	7,34	7,46	6,55
	Weighted Fusion 14	0,63	8,13	7,34	7,46	6,55
	Weighted Fusion 15	0,63	8,13	7,35	7,45	6,54

highest correlations values. Among the photo's feature, the ones with highest performances are the features related to statistics of the groups in which the photo is shared in (i.e., NumGroups, AvgGroupPhotos and AvgGroupMemb). The PhotoCount feature (i.e., number of photos of the user) is the only one that achieves a negative value of correlation. The visual features are placed in the middle positions of the ranking.

Table 3.11 and Table 3.13 show the experimental results obtained by training an SVR on single features, on dataset of 10 days and 20 days sequences respectively. The experimental results in terms of Trimmed RMSE (tRMSE) and Median RMSE (RMSE MED) confirm the behaviour previously observed in the prediction of the 30 days sequences, and suggested by the correlation analysis. Indeed, user's features such as MeanViews, Contacts and GroupsCount resulted useful to achieve

Figure 3.29: Spearman's correlation values between the $\hat{s}_{scale}$ estimation performed by a SVR trained with a single feature and the Ground Truth value, for the 10 days, 20 days and 30 days scenarios.



good performances. Among the photo's features, the NumSets achieves good results in terms of RMSE MED, whereas visual features don't obtain good performances.

Table 3.12 and Table 3.14 show the experimental results obtained by combining the input features, on the 10 days and 20 days datasets respectively. Also in these cases, proper combinations of features help to further improve the obtained results. In either the 10 days and the 20 scenarios, the best results have been obtained by combining the best user's features (i.e., Contacts, MeanViews, GroupsCount, GrpipsAvgMembers and GroupsAvgPictures) and the best photo features (i.e., NumGroups, AvgGroupsMemb and AvgGroupPhotos).

3.7.6 Final Remarks

In this work introduced a new challenging task of estimating the popularity dynamics of social images. To benchmark the problem, a new publicly available dataset is proposed. We also describe a method to forecast the sequence of views over a period of 30 days of a photo shared on Flickr. In particular, the proposed approach combines the results obtained by two different algorithms aimed to estimate the

Table 3.13: Evaluation results for the prediction of the 20 days sequences.

				Proposed Method		**Baseline C (shape)**	
Feature Source	**Feature ID**	**Features**	**Spearman**	**tRMSE 0.25**	**RMSE MED**	**tRMSE 0.25**	**RMSE MED**
User	**0**	**Ispro**	0,26	11,39	8,76	10,13	7,83
User	**1**	**Contacts**	0,53	9,73	8,67	8,96	7,70
User	**2**	**PhotoCount**	-0,08	10,25	9,68	9,43	8,31
User	**3**	**MeanViews**	**0,72**	**9,11**	**8,10**	8,32	7,16
User	**4**	**GroupsCount**	0,53	9,73	8,30	8,96	7,53
User	**5**	**GroupsAvgMembers**	0,45	10,54	9,03	9,61	8,22
User	**6**	**GroupsAvgPictures**	0,46	10,93	8,74	10,01	7,78
Photo	**7**	**Size**	0,05	10,26	9,37	9,39	8,27
Photo	**8**	**Title**	0,03	10,21	9,44	9,42	8,24
Photo	**9**	**Description**	0,06	10,19	9,50	9,39	8,22
Photo	**10**	**NumSets**	0,20	11,04	8,77	9,89	8,12
Photo	**11**	**NumGroups**	0,34	10,24	9,74	9,39	8,33
Photo	**12**	**AvgGroupsMemb**	0,34	9,91	9,07	9,18	7,91
Photo	**13**	**AvgGroupPhotos**	0,34	9,87	8,97	9,16	7,86
Photo	**14**	**Tags**	0,12	10,54	9,30	9,64	8,31
Visual	**hybrid_f1**	**Hybridnet fc7**	0,22	20,49	18,06	17,43	14,93
Visual	**hybrid_f2**	**Hybridnet fc8a**	0,26	13,86	12,28	11,79	10,27
Visual	**senti_f1**	**DeepSentiBank fc7**	0,26	21,45	18,67	18,03	15,49
Visual	**senti_f2**	**DeepSentiBank fc8**	0,31	16,45	14,28	13,91	11,74
Visual	**google_f1**	**GoogleNet pool5/7x7_s1**	0,26	14,07	12,63	11,98	10,37
Visual	**google_f2**	**GoogleNet loss3/classifier**	0,28	13,22	11,84	11,29	9,71

maximum number of the number of views reached by the photo in the period of observation (scale) and the shape of the sequence. Furthermore, we evaluated our approach for the prediction of 10 days and 20 days sequences, which are characterized by an higher variability with respect to the 30 days sequences.

Future works can be devoted to the extension of the dataset by taking into account other social platforms. Furthermore, additional time-aware features can be considered, such as the day of the week and the hour of the day. Also, different approaches to treat the problem of popularity dynamics prediction as a time series forecasting task can be taken into account.

Table 3.14: Evaluation results for the prediction of the 20 days sequences obtained by combining the features.

			Proposed Method		**Baseline C (shape)**	
Feature ID	**Features**	**Spearman**	**tRMSE 0.25**	**RMSE MED**	**tRMSE 0.25**	**RMSE MED**
all	**concat(0-14)**	0,62	9,16	7,57	8,16	6,37
user	**concat(0-6)**	0,66	9,44	7,02	8,42	6,34
photo	**concat(7-14)**	0,28	10,10	8,39	9,12	7,38
best_photo	**concat(11,12,13)**	0,33	9,74	8,78	9,08	7,87
user2	**concat(3,4)**	**0,71**	8,96	7,63	8,22	6,78
user3	**concat(1,3,4)**	**0,71**	8,88	7,52	8,15	6,67
user5	**concat(1,3,4,5,6)**	0,68	9,23	**7,01**	8,41	6,49
	concat(user2,best_photo)	**0,71**	8,87	7,56	8,15	6,75
	concat(user3,best_photo)	**0,71**	**8,80**	7,48	8,10	6,66
	concat(user5,best_photo)	0,68	8,98	6,99	8,21	6,42
	Late Fusion 1 (AVG)	0,59	10,03	9,14	9,16	8,09
	Late Fusion 2 (SVR)	0,47	11,34	8,11	10,11	7,71
	Borda Count (wAVG)	0,63	9,96	8,94	9,09	7,95
	Weighted Fusion 2	**0,71**	9,27	8,24	8,52	7,34
	Weighted Fusion 3	0.70	9,39	8,35	8,60	7,48
	Weighted Fusion 4	0,67	9,75	8,57	8,92	7,64
	Weighted Fusion 5	0,67	9,88	8,67	9,03	7,71
	Weighted Fusion 6	0,67	9,90	8,70	9,06	7,74
	Weighted Fusion 7	0,67	9,90	8,71	9,06	7,77
	Weighted Fusion 8	0,67	9,91	8,75	9,06	7,75
	Weighted Fusion 9	0,66	9,90	8,79	9,05	7,83
	Weighted Fusion 10	0,65	9,90	8,84	9,05	7,87
	Weighted Fusion 11	0,65	9,91	8,85	9,06	7,88
	Weighted Fusion 12	0,65	9,91	8,86	9,06	7,89
	Weighted Fusion 13	0,65	9,91	8,86	9,06	7,89
	Weighted Fusion 14	0,65	9,92	8,87	9,06	7,89
	Weighted Fusion 15	0,65	9,91	8,87	9,06	7,89

3.8 Conclusions

In this Chapter we have summarized the main issues and techniques related to Visual Sentiment Analysis. The current state of the art has been analysed in detail, highlighting pros and cons of each approach and dataset. Although this task has been studied for years, the field is still in its infancy. Visual Sentiment Analysis is a challenging task due to a number of factors that have been discussed.

The results discussed in this study, such as [39], agree that the semantic content has a great impact on the emotional influence of a picture. Images having similar color histograms and textures could have completely different emotional impacts. As result, a representation of images which express both the appearance of the whole image and the intrinsic semantic of the viewed scene is needed. Early methods in

the literature about Visual Sentiment Analysis tried to fill the so called *affective gap* by designing visual representations. Some approaches build systems trained with human labelled datasets and try to predict the polarity of the images. Other approaches compute the polarity of the text associated to the images (e.g., post message, tags and comments) by exploiting common Sentiment Analysis systems that works on textual contents [38, 157], and try to learn Machine Learning systems able to infer that polarity from the associated visual content. These techniques have achieved interesting improvements in the tasks of image content recognition, automatic annotation and image retrieval [127, 83, 133, 142, 158, 159, 125]. However, is not possible to know if such user provided text is related to the image content or to the sentiment it conveys. Moreover, the text associated to images is often noisy. Therefore, the exploitation of such text for the definition of either polarity ground truth or as an input source for a sentiment classifier have to address with not reliable text sources.

Furthermore, some approaches exploit a combination of feature modalities (often called views) to build feature space embeddings in which the correlation of the multi-modal features associated to the images that have the same polarity is maximized [55]. The results achieved by several discussed works suggest that exploiting multiple modalities is mandatory, since the sentiment evoked by a picture is affected by a combination of factors, beside the visual information. Studies in psychology and art theory suggested some visual features associated to emotions evoked by images. However, the most promising choice is given by representations automatically learned through neural networks, autoencoder and feature embedding methods. These approaches are able to find new feature spaces which capture contributes from the different input factor which the sentiment is affected by. The recent results in representation learning confirm this statement.

To this end, one important contribution is given by the availability of large and robust datasets. Indeed, in this study, we highlighted some issues related to the existing datasets. Modern social media platforms allows the collection of huge amount of pictures with several correlated information. These can be exploited to define either input features and "ground truth". However, as highlighted before, these textual information need to be properly filtered and processed, in order to avoid the association of noisy information to the images.

In Section 3.6 we presented our work on the task of sentiment polarity prediction, by combining textual and visual features. After a study of the common approaches, which exploit user provided text (Subjective Text), we identified several drawbacks brought by the such a text source due its intrinsic nature. Indeed, the subjective text associated to images by users presents very noisy terms. For this reason, the proposed framework exploits a set of features extracted from text automatically inferred from the image visual content (Objective Text). Our study demonstrates that the exploitation of Objective Text associated to images provides better results than the use of the Subjective Text provided by the user. Furthermore, it brings several advantages. Indeed it has a pre-defined structure, providing the same quantity of textual objective information for all the images. Each exploited deep learning architecture used to extract the objective text contributes to the description of the image from a different perspective.

Experiments confirmed that the textual features extracted from the proposed Objective Text outperform the ones based on the Subjective Text provided by users by considering different combinations of features.

In particular, our approach to the challenge of popularity prediction has been presented. In particular, we defined and proposed an even more challenging task, that is the prediction of the popularity score over time, named Popularity Dynamics Prediction. A study about the popularity dynamics of social images shared on Flickr is presented. Starting from a critical study of the current state of the art on the task of image popularity prediction we observed that the change of the popularity score over time is a critical factor. Therefore, we decided to address the problem of the popularity dynamics prediction of an uploaded photo, by designing a framework able to predict the number of views reached by the photo for each day after posting for an extension of 30 days. This work introduces a new challenging task in the field which paves the way to new applications. At the same time a new benchmark dataset is released. The proposed dataset includes $\sim$ 20K Flickr images, with related user and photo meta-data. The daily information are related to three engagement score that have been tracked for 30 days. Namely the number of views, number of comments and number of favorites. In addition to the common used features, the proposed dataset also includes additional statistics about the groups of the user and the groups in which the photo has been shared in the moment of its upload (i.e.,

number of the member of the groups and number of photos in the groups).
This Chapter aimed to give a complete overview of the Visual Sentiment Analysis problem, the relative issues, and the algorithms proposed in the state of the art. Relevant points with practical applications in business fields which would benefit from studies in Sentiment Analysis on visual contents have been also discussed.

Chapter 4

Video Segmentation and Clustering for Scene Popularity Detection

4.1 Introduction

In social events (e.g., concerts), the automatic video understanding goal includes the interpretation of which visual contents are the most relevant (i.e., popular). In this context, the popularity of a visual content depends on how many people are looking at that scene, and therefore it could be obtained through the "visual consensus" among multiple video streams acquired by the different users devices. In live social events such as concerts and sport performances, people capture and collect a lot of multimedia data (and specifically videos) which are related to the event. These data contain a certain amount of redundancy related to interesting scenes which have captured the attention of many individuals (e.g., fireworks during a folkloristic event). We are interested in detecting and summarizing these "popular scenes" in a single video. Section 4.2 we present RECfusion, a system able to automatically create a single video from multiple video sources by taking into account the popularity of the acquired scenes. The frames composing the final popular video are selected from the different video streams by considering those visual scenes which are pointed and recorded by the highest number of users devices. The developed system is hence able to detect and summarize the "popular scenes" captured by users with a mobile camera during social events.

The RECfusion approach is able to produce the "popular video" from the analysis

of a set of video sources about the same event. However, in the experiments we observed that the RECfusion has a drop in performances when the quality of the videos is scarce, for instance when the camera is constantly moving such as in the case of wearable devices (i.e., egocentric videos). Another drawback of RECfusion is related to the computational costs. For these reasons, we decided to define an improved version of the RECfusion approach, able to work in the wearable domain with acceptable time needs. To this aim, we addressed the problem of the clustering of wearable videos taken by the same person among different days. The proposed system is described in Section 4.3. Egocentric videos are becoming popular since the possibility to observe the scene flow from the user's point of view (First Person Vision). Among the different applications of egocentric vision is the daily living monitoring of a user wearing the camera. We propose a system able to automatically organize egocentric videos acquired by the user over different days. Through an unsupervised temporal segmentation, each egocentric video is divided in chapters by considering the visual content. The obtained video segments related to the different days are hence connected according to the scene context in which the user acts. Experiments on a challenging egocentric video dataset demonstrate the effectiveness of the proposed approach that outperforms with a good margin the state of the art in accuracy and computational time.

4.2 RECfusion

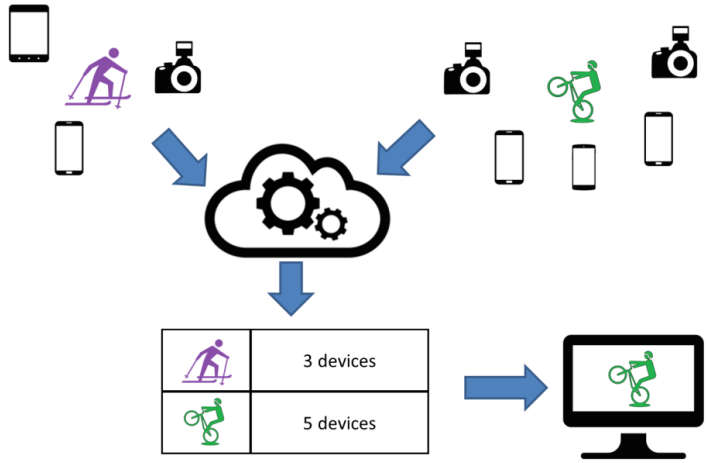


Figure 4.1: RECfusion - Automatic summarization of the popular scenes of a social event in a single video.

4.2.1 Introduction

In social events like a concert, a maratona or a car ride, the audience gathers multimedia information with its mobile devices (e.g., images, video, geolocation, tags, etc.) which are related to what captured the people's interest. Popular scenes (e.g., fireworks in a folkloristic event) are often observed and acquired simultaneously by multiple end-users with different devices. This redundancy in such set of video sequences can be exploited to infer which groups of people are interested to specific visual contents over time, and hence which visual contents are the most popular. In Figure 4.1 the pipeline of the proposed method is sketched.

This work proposes a system that automatically processes multiple video flows from different devices to understand the most popular scenes for a group of end-users. The scenes observed by the different devices are grouped by visual content. Then, the clusters of the different scenes are tracked over time by a comparison of the scene clusters built at different times, giving the possibility to introduce a scenes story log. The scene grouping performed over time allows an automatic video curation process to obtain a single video as output, by mixing the different inputs and taking into account the most popular scenes, i.e. those scenes acquired

by the most of devices over time (see Figure 4.1). The task of establishing the popularity of a scene is challenging because of the variability of the visual content observed by multiple devices: different points of view, pose and scale of the objects, lighting conditions, occlusions, viewing quality, as well as different device models. The Imaging Generation Pipeline (IGP) can vary from device to device and even on an per-image basis [160]. The difference between responses depends essentially on the characteristics of the lenses, filters (e.g., Bayer pattern), sensors, and IGP algorithms [161].

The proposed video curation system works without specific priors on the input, as well as without knowledge of which and how many devices are involved in the curation process. A related work is described in [162], where a 3D reconstruction of the scene and the relative pose of the devices are exploited. However, the reconstruction of the whole scene is computationally expensive and the video curation algorithm can be applied only after such a preprocessing stage. In [163] it is proposed an approach which exploits multiple devices of the same model. Also in this case the algorithm needs a calibration phase to reconstruct the camera poses. The method proposed in [164] tries to create a single popular video of an event from three egocentric videos. All scenes were taken using the same camera model by different participants close each other (few meters). The approach assumes that the number of the different popular scenes and the number of the devices is known a priori. These information are used to recognize an exact number of regions of interest used as prototypes in the grouping phase. The algorithm proposed in [165] combines several videos of the same scene taken from different perspectives. Despite the final video is not based on the popularity (as in the case we are interested in), this framework is able to produce an unique final video related to the considered scenario. The main goal of the approach in [165] is to build an automatic system which tries to imitate a professional video editor by choosing automatically which shooting angle and distance should be used and how long the selected configuration should persists. However, classification of the videos (distance and position) is performed manually.

The aforementioned existing approaches achieve significant results but the mentioned assumptions and simplifications contrast with a real application of such systems. Differently than previous approaches, the method proposed in this work (REC-fusion) combines several videos from unknown different devices based on the popularity of the acquired scenes without any prior knowledge or training stage.

4.2.2 Proposed system

The multiple video streams acquired by the different devices are analysed by using three algorithms. The intraflow analysis segments the different scenes, transitions and the unstable intervals within each video, whereas the interflow analysis performs the grouping of the involved devices over time by taking into account the visual content of the previously defined video segments. The popularity of the obtained clusters over time is used to produce the final video. Moreover, after each clustering step, each cluster is compared with respect to the clusters analyzed in the past, with the aim to track the groups of scenes over time (i.e., cluster tracking). The cluster tracking procedure allows to have an identification of the scenes over time, by applying a time-wise re-labelling of the different scene clusters.

Intraflow Analysis

During intraflow analysis each video is processed comparing its frames in order to segment the video based on the visual content. For each frame, selected by sampling the video, we extract keypoints using the well-known SIFT detection algorithm [166]. The set of SIFT features extracted from a frame are used as a template for the acquired scene. In our experiments we excluded the SIFT extracted near the border of the considered frames to make more robust the feature matching among frames. The intraflow analysis consists on comparing the templates extracted from different frames of a video to split it in blocks (i.e., video segments) coherently with the visual content. During this process the system keeps a reference template regarding the last known scene (i.e., the last stable set of SIFT features extracted from the last detected scene) and compares this template with respect to the features extracted from the current frame under analysis. When a sensible variation of features is observed (i.e., low matching score), the algorithm refreshes the reference template and splits the video producing a new segment (see the example scheme in Figure 4.2).

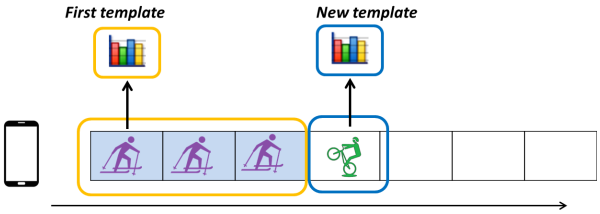


Figure 4.2: Template updating performed by the intraflow analysis.

Given a frame, the number of matchings between its SIFT keypoints and the once of the reference template is considered as a similarity index between the involved scenes. To make the matching more reliable, we excluded the matchings where the keypoints are too far in terms of spatial coordinate by assuming smooth transition between frames (we used a threshold distance of 100 pixels for images with resolution 1280×720 or 1920×1080). In order to detect the sudden changes of the number of matchings we defined a slope function which is computed on a frame at time T as in Equation 4.1.

$$slope(T) = \frac{h}{w} = \frac{l \sin\theta}{l \sin\theta} = \tan\theta \tag{4.1}$$

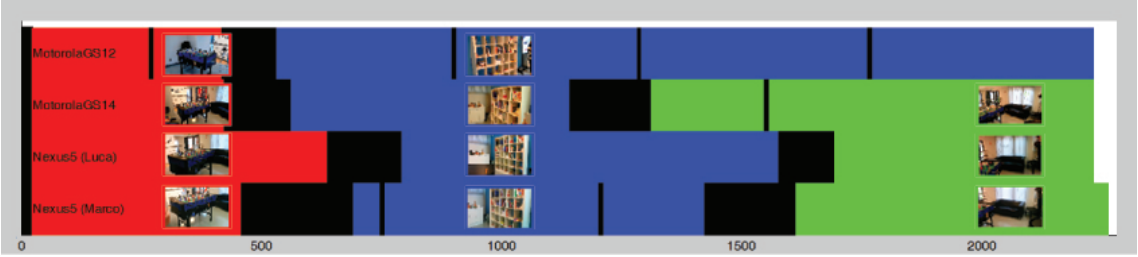
Where l is the length of the segment between the two consecutive matching counts considered to compute the slope. This function represents the variation of the number of matchings h in a range interval w centered in a frame at time T. This value is related to the tangent of an angle θ which is proportional to the gradient of the matching curve. The algorithm asks for a new template (i.e., set of SIFT features) when the slope function has a peak greater than 10 (i.e., θ= 85°). In order to define only reliable templates the algorithm checks if the computed template is stable for at least 2 seconds (i.e., the number of matchings do not change too much). When a new stable template is defined, the algorithm compares it with respect to the past templates in order to understand if it regards a new scene or it is related to a known previously observed scene. This backward checking is done starting from the last found template. Two different templates of the same scene could be rather different due to the elapsed time between them. During this step, to check if two templates describe the same scene we use a geometric verification to exclude the spatial matchings with distance higher than one-third of the height of the image. This approach aims to ideally subdivide the scene in three horizontal

parts and exclude the matchings between the upper side of one image with the lower one of the other image, and vice versa. Two templates are assigned to the same scene if the percentage of the matchings after the geometric verification is greater than 50% of the original matchings. Each reference template is assigned to a scene ID, named *SceneCounter*, and all the video frames which achieve a robust match with a reference template are classified as of that scene (i.e., are assigned to the same *SceneCounter*). All the frames between the instant when a new scene template have to be upgraded by the system and the instant when that template is finally upgraded are classified as a transition intervals.

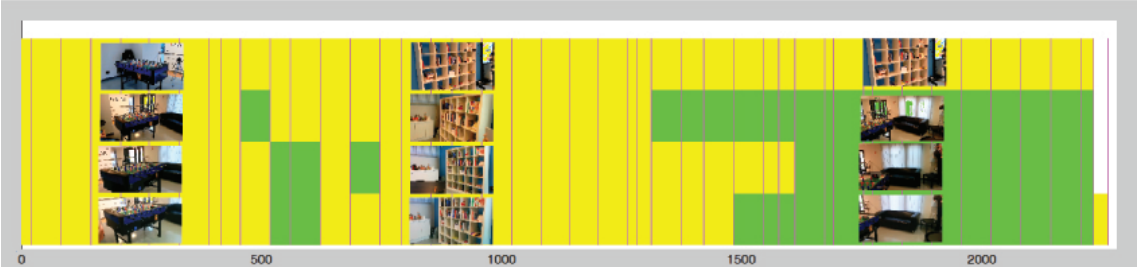
Figure 4.3(a) shows an example of the result of the intraflow analysis applied on four input videos as a coloured chronogram: each scene is identified by a colour (red, blue and green in the figure), whereas the transition intervals and the unstable frames (e.g., shaking frames) are identified by black colour. The intraflow analysis allows an automatic segmentation of each video in several intervals depending on the visual content and locates correctly the transition intervals.

Interflow Analysis

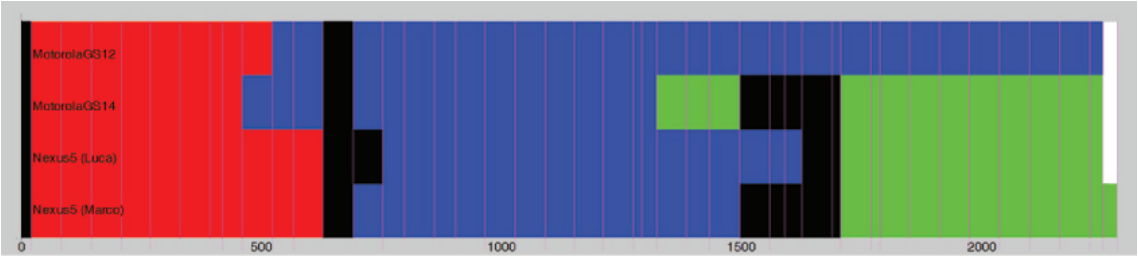
Given two images acquired with different devices it is very challenging to understand if they are related to the same scene using only visual information [167]. In our case, given frames of different videos which have been segmented, as described in Section 4.2.2, we want to understand which of the different devices are simultaneously looking at the same scene over time. The most popular scene over time is then used to produce the final video. In the interflow analysis we defined a frame descriptor based on a weighted colour histogram. In order to address the device invariance issue, we first apply the histogram equalization described in [160]. This method consists on the application of an histogram equalization to each RGB channel and allow us to reduce the variability introduced by the IGPs related to the different devices. After equalization we compute a weighted colour histogram quantizing the color space (8 color for each channel). The weights are obtained by using a gradient map as suggested in [168]. The gradient map highlights the structures of the objects involved in the scene making more robust the descriptor. To compare histograms



(a) Intraflow Analysis



(b) Interflow Analysis



(c) Cluster Tracking

Figure 4.3: RECfusion segmentation and clustering approach applied on Foosball dataset. The chronograms show the results of the three main steps of RECfusion (intraflow analysis, interflow analysis and cluster tracking). Foosball dataset is composed by 4 video streams having a duration of $\sim$ 2300 frames ($\sim$ 90 seconds). Each video stream is represented as a row in the chronogram. Vertical red lines mark the end of time-slots (i.e., a new template is defined). (a) Intraflow analysis: red, blue and green frames are respectively the first, second and third scene of each video stream. Noisy frames are depicted in black. (b) Interflow analysis: yellow and green clusters are respectively the first and second cluster of each time-slot. (c) Cluster tracking: red, blue and green clusters are respectively the first, second and third cluster of the whole video set. Noisy clusters are depicted in black.

we use the same distance employed in [168]:

$$d(h_{D_a}, h_{D_b}) = \sum \frac{\sum (h_{D_a} - h_{D_b})^2}{\sum (h_{D_a})^2} \tag{4.2}$$

Algorithm 2: Devices' clustering

Data: set of devices D
Result: set of clusters C

```
1  Choose an unclustered device d;
2  Insert d in a new cluster C_d;
3  C = C ∪ {C_d};
4  foreach device b in D s.t. b ≠ d do
5      if d(h_d,h_b)<1 then
6          if b is unclustered then
7              Insert b in the cluster C_d;
8          else
9              Find the cluster containing b;
10             Compare b with the elements of the two contending clusters;
11             Insert b in the closer cluster;
12         end
13     end
14     if all devices are clustered then
15         End procedure;
16     else
17         Return to 1;
18     end
19 end
```

where h_{D_a} and h_{D_b} are the weighted histograms related to the two frames of two different devices D_a and D_b.

To cluster the devices accordingly to visual content at every instance of time, we first segment all the videos exploiting the intraflow analysis (Section 4.2.2), and then we use the weighted color histogram representation to compare the obtained video segments (Figure 4.3(b)). At each instant T, the set of frames of the different devices are clustered by taking into account the corresponding weighted histogram representation. Each cluster is assigned to a *ClusterCounter* that is exploited by the cluster tracking procedure.

The different scenes are considered as a complete graph where each node is a device and the arches are labelled with the interflow measure (Equation 4.2) between the scenes taken by the devices. The interflow measures are used to establish similarity during the clustering of the devices. The final video produced as output by the proposed approach is obtained by considering the most popular cluster (i.e.,

the once with the highest number of devices). Specifically, for the output at each instant T we consider the video belonging to the most popular cluster which is closest to the cluster centroid (i.e., average among all histograms). The pseudocode describing the method to cluster the devices at each instant of time considering visual information is reported in Algorithm 2.

Cluster Tracking

The intraflow analysis segments the sequence of frames of a single video stream, and assigns a *SceneCounter* to each segmented scene. However, frames taken by two different video streams labeled with the same *SceneCounter* can represent different scenes, since *SceneCounters* are discriminative only within a single video stream. The interflow analysis takes the segments defined at each time-slot and assigns the same *ClusterCounter* to the scenes (i.e., segments) of the different videos that are grouped in the same cluster by the procedure. The *ClusterCounters* are to be considered only within a single time-slot, since the cluster identifiers assigned at different time-slots are not related one each other. Therefore, we developed a cluster tracking procedure aimed to track the clusters representing the same scene in any video stream and time-slot (Figure 4.3(c)). In [169] a Graphical User Interface (GUI) implementing the cluster tracking typical video player commands (e.g., Start, Pause, Stop, etc.) is described (Figure 4.4).

We propose a cluster tracking procedure based on a voting routine that combines the results of the intraflow and interflow analyses. Once the interflow procedure has assigned a *ClusterCounter* to groups of *SceneCounters*, this set of scenes will characterize the same cluster also in further time-slots. To this aim, the cluster tracking procedure assigns an unique $LoggedCluster_{ID}$ to this set of scenes. Differently from the *ClusterCounters*, the $LoggedCluster_{IDs}$ are intended to be discriminative time-wise. The cluster tracking procedure tracks the clusters of each time-slot assigning them $TrackedCluster_{IDs}$ corresponding to the most similar $LoggedCluster_{ID}$ among the ones previously defined. In order to define the most similar $LoggedCluster_{ID}$, the cluster tracking procedure requires an initialization phase (at first time-slot). In this phase, the assigned $LoggedCluster_{IDs}$ corresponds to the *ClusterCounters* defined by the interflow analysis. Then, starting from the second time-slot, the clusters will be associated to an existent $LoggedCluster_{ID}$ or

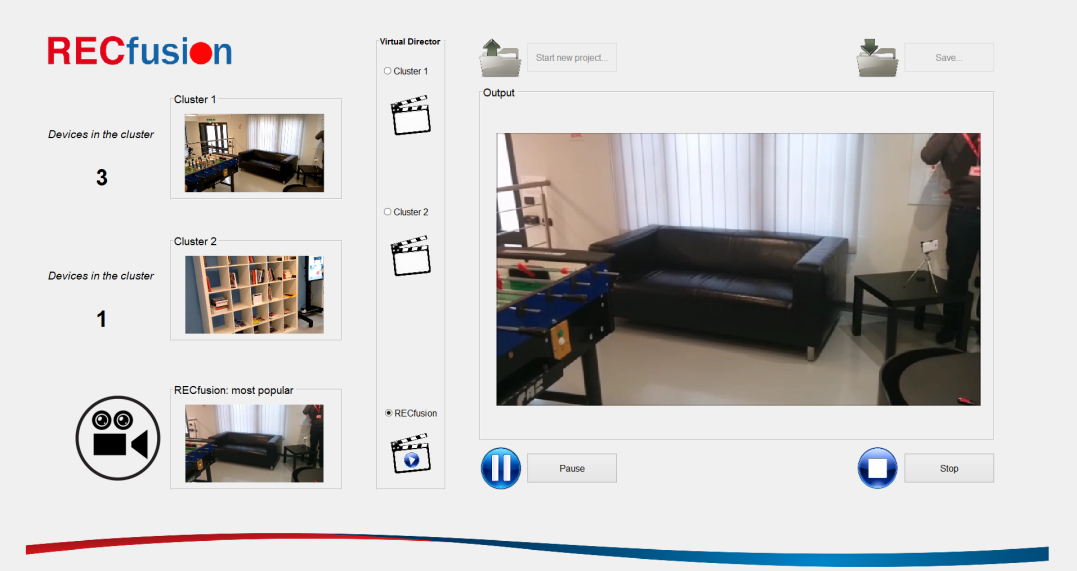


Figure 4.4: RECfusion Graphical User Interface showing the Cluster Tracking framework. On the left, active clusters with respective amount of recording devices and automatically suggested video stream (called *RECfusion: most popular*) are shown. User can browse the "Virtual Director" panel to dynamically select the active video stream. On the right side, active video stream with classic video player commands is shown.

to a new one, depending on a voting routine. The same routine is also used to track the $LoggedCluster_{IDs}$ with proper $TrackedCluster_{IDs}$.

The voting routine can be divided into 2 phases: casting of vote and voting decision. In the former phase, for each time-slot, each scene votes with three different possible values: $TrackedCluster_{ID}$ at the previous time-slot, $LoggedCluster_{ID}$ or unlogged scene (V_N), if the scene is *Noise*, already logged or unlogged, respectively. Once all the votes are cast in a time-slot, the majority is found. Majority of unlogged scenes is not admitted, so in this case we simply ignore these votes in the voting decision, taking into account the second ranked majority. Depending on the reached decision, new $LoggedCluster_{IDs}$ might be instantiated, while $TrackedCluster_{IDs}$ at current time-slot is eventually updated. In the experiments we compare the proposed method with respect to a cluster tracking method based on a threshold T_{CT} [169]. This threshold was used as an hyperparameter to determine whenever to create a new $LoggedCluster_{ID}$ or not. The main issue related to the threshold employed in [169] is that its value should be fine tuned for each video set in order to achieve

the best results in cluster tracking procedure.

4.2.3 Datasets

To demonstrate the effectiveness of the proposed system we have performed experiments on the RECfusion dataset [144] that is available online [1]. The considered dataset includes different scenarios acquired with multiple devices of different models:

- Foosball - Indoor context, some people appear in the scene. In this scenario each device takes a view of a Foosball room switching among three different regions of interest. This scenario is useful to highlights the behaviour of the system when the popularity of a scene leaves a subject advocating a new one.
- Meeting - Indoor context, two people appear in the scene. There were four people sitting around a table. Each person records one of the participants using a mobile device (smartphone or tablet). One more device is placed in order to look constantly to a subject. We tested this scenario varying the number of the devices (2, 4 or 5 devices).
- SAgata - Outdoor context, lots of people appear in the scene (i.e., crowded scene). This set of videos have been taken in a real scenario during a folkloristic event (Saint Agata in Catania). Seven people recorded the event while the main subject (the Agata's statue) was carried along the city's streets.

The aforementioned scenarios include challenging scenes with crowd, large perspective variation, occlusion, periods of tilt and shaking, etc. We have also tested the proposed approach on the benchmark dataset proposed in [164]. This dataset has been acquired with wearable head mounted cameras and is challenging since every video is strongly affected by fast head motion.

For testing purposes each video has been manually segmented and for each instant of time we known the exact number of clusters and the most popular scene.

In the experiments we exploit also a video set from the dataset used in Ballan et al. [170]. This dataset is called *Magician*. It is related to an indoor context, where

[1] http://iplab.dmi.unict.it/recfusionICIAP17.

one person appear in the foreground. There are two main points of view in this video set: one above and one in front of the magician. We have chosen *Magician* video set because it is slighty different from the videos currently in RECfusion dataset. In *Magician* all the video streams are focused on a single target and are acquired as a "casual multi-view video collection" [170]. This means that backgrounds in the video streams are very different from each other and that severe camera motion could often appear. The casually filmed events represent a challenging scenario for detector like SIFT (exploited in our intraflow analysis, see Section 4.2.2). Indeed, we considered the *Magician* video set with the aim to stress and challenge the scene analysis and the cluster tracking performances of the proposed approach.

4.2.4 Experimental settings and results

To evaluate the performances of the proposed method on each scenario we compute two measures obtained from the following scores computed on each clustering over time:

- P_{r}: ground truth popularity score (number of cameras looking at the most popular scene) obtained from manual labelling;
- P_{a}: popularity score computed by our method (number of the elements in the popular cluster);
- P_{g}: number of correct videos in the popular cluster computed by our method.

From the above scores, we compute the weighted mean of the ratios $\mathrm{P_a/P_r}$ and $\mathrm{P_g/P_r}$ over all the segmented blocks of a video, where the weights are given by the length of the blocks (i.e., the number of frames). When $\mathrm{P_a/P_r}$ is close to 1 the popularity score computed by our method is similar to the ground truth popularity. When this number is greater than 1 it means that the most popular cluster obtained with our approach is affected by outliers, whereas when this number is less than 1 it means that the our method missed some element of the ground truth popular cluster. Since $\mathrm{P_a/P_r}$ deal just with the number of video in the popular cluster, it is useful to look also at the ratio $\mathrm{P_g/P_r}$. Indeed, $\mathrm{P_g/P_r}$ assesses the visual content of the videos in the popular cluster (true positive). This score have to be close to 1 to indicate accuracy in the popular cluster computed by our method.

Figure 4.5: Output video example. On the left and right the input videos, the final output in the center.

Table 4.1 shows the obtained results. The first five rows are related to the scenarios of the proposed dataset, row six reports the results obtained with the *Magician* scenario [170], whereas the last three rows are related to the dataset proposed in [164]. The results show the effectiveness of our approach. Difficulties appear when some video regarding the most popular subject are taken with a quite different scale factors. This can be noted comparing the third and the fourth row in Table 4.1. In the meeting scenario with 5 devices of different models there is a huge difference in the scale of the acquired subjects in the scene. In the videos proposed in [164] the camera is constantly moving due to the shake induced by the natural head motion of the wearer. Despite we achieve good performances on wearable egocentric videos, we believe that there is still space for further improvements in such a video category (e.g. by filtering out the head motion). In Section 4.3 the RECfusion approach is evaluated and further improved with the aim to work in the wearable egocentric video domain.

In order to better asses the results obtained by the proposed system, the reader can perform a visual inspection of the videos produced by our approach at the following URL: http://iplab.dmi.unict.it/recfusionICIAP17.

In the proposed cluster tracking procedure we removed the threshold T_{CT} used in [169] as an hyperparameter to decide whenever to create a new logged-cluster or not. In [169] the value of T_{CT} was empirically set to 0.15, after a grid search of the best value that optimizes a set of scores, namely the True Positive Rate,

Table 4.1: Validation Results of Popularity Estimation.

Scenario	**Devices**	**Models**	P_a/P_r	P_g/P_r
Foosball	4	2	1.02	1
Meeting	2	2	1.01	0.99
Meeting	4	4	0.99	0.95
Meeting	5	5	0.89	0.76
SAgata	7	6	1.05	1
Magician [170]	6	6	0.73	0.73
Concert [164]	3	1	1.06	1
Lecture [164]	3	1	1.05	0.86
Seminar [164]	3	1	0.62	0.62

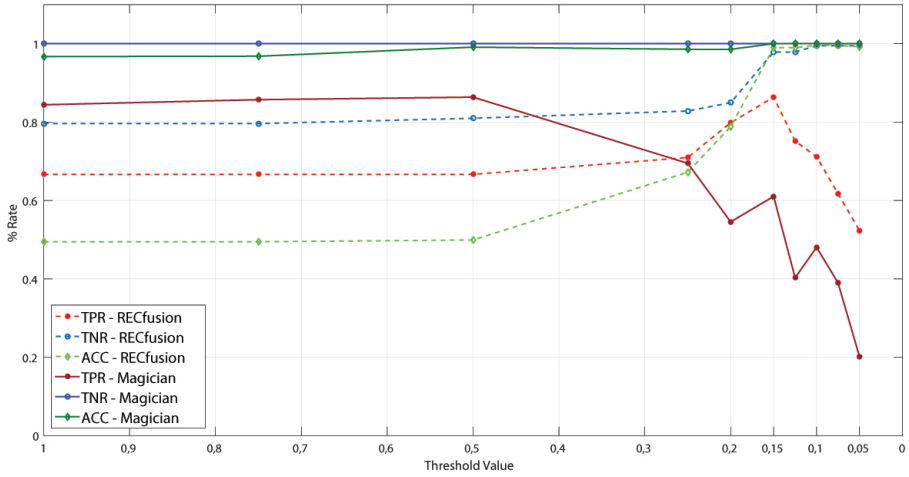


Figure 4.6: A comparison of *TPR (True Positive Rate, or Recall)*, *TNR (True Negative Rate, or Specificity)* and *ACC (Accuracy)* between RECfusion dataset [6] and *Magician* video set cluster tracking validations using the threshold-based procedure from [169]. The video set *Magician* requires a fine tuned threshold to increase *TPR*, *TNR* and *ACC* values.

True Negative Rate and Accuracy of clustering tracking procedure on RECfusion dataset [6]. In Figure 4.6 a comparison between the average values of *TPR (True Positive Rate, or Recall)*, *TNR (True Negative Rate, or Specificity)* and *ACC (Accuracy)* of RECfusion dataset and *Magician* video set with a range of T_{CT} values is shown. As we can observe, the value of T_{CT} equals to 0.15 is not the best value to be used by cluster tracking procedure, while $T_{CT} = 0.5$ should be used instead. For this reason we proposed the new threshold independent cluster tracking procedure described above. The *TPR*, *TNR* and *ACC* values for each video set described

Table 4.2: Validation results between cluster tracking procedure threshold-based [169] and vote-based (proposed).

		TPR (RECALL)		TNR (SPECIFICITY)		ACC (ACCURACY)	
DS	Scene	[169]	PROPOSED	[169]	PROPOSED	[169]	PROPOSED
Foosball	1	0,91	0,92	0,70	1,00	0,69	1,00
	2	0,69	0,97	0,98	0,91	0,99	0,97
	3	0,41	0,74	1,00	1,00	0,50	1,00
	MEAN	0,67	0,87	0,89	0,97	0,73	0,99
Meeting	1	0,99	1,00	1,00	1,00	1,00	1,00
	2	0,80	1,00	0,95	0,93	0,83	0,67
	3	0,43	0,50	1,00	1,00	0,70	1,00
	MEAN	0,74	0,83	0,98	0,98	0,84	0,89
S.Agata	1	0,71	1,00	1,00	1,00	1,00	1,00
	2	0,87	0,97	0,49	0,14	0,80	0,68
	3	0,48	0,00	1,00	1,00	0,60	0,00
	MEAN	0,69	0,66	0,83	0,71	0,80	0,56
Magician	1	0,73	1,00	1,00	1,00	1,00	1,00
	2	0,45	0,56	1,00	1,00	0,98	0,91
	MEAN	0,59	0,78	1,00	1,00	0,99	0,96

in Section 4.2.3 are computed and compared with respect to the results obtained in [169]. The comparative validation results are shown in Table 4.2.

These results show that the proposed vote-based cluster tracking procedure reaches *TPR* values much higher than the threshold-based procedure, while results on TNR and ACC are comparable between the two procedures. Only in the *Meeting* video set the proposed vote-based procedure is slightly outperformed: this is a limitation of the procedure. Indeed, cluster tracking procedure relies on intraflow analysis. Thus, if the latter defines N scenes, then the former is able to distinguish at most N scenes. Hence, differently by threshold-based procedure used in [169], that can generate a bunch of small sparse clusters if T_{CT} is not fine tuned, in this case only a limited number of clusters is tracked. In the *Meeting* video set two people are recorded and there are only two distinguished clusters focusing on each one of them. Sometimes interflow analysis generates a cluster containing both of the two people. This is treated by the cluster tracking vote-based procedure as *Noise*, since intraflow analysis has never labeled a scene in which the people are recorded together.

A final remark is about *Magician* video set. We added it to our dataset in order to evaluate scene analysis and cluster tracking performances in a video collection with a single scene, where all the users are focused on the same target and videos are affected by severe camera motion and scale differences. Cluster tracking results with threshold-based procedure from [169] are really bad, indeed we got the worst

average performance on this video set (Table 4.2). By the other hand, the proposed vote-based procedure reached good values of *TPR*, further assessing the soundness of this new cluster tracking approach. The output videos showing the result of cluster tracking vote-based procedure could be found at the following http://iplab.dmi.unict.it/recfusionICIAP17.

4.2.5 Final Remarks

In this Section we described an automatic video curation method driven by the popularity of the scenes acquired by multiple devices. Although some errors could occur during the clustering of the devices, the system rarely chooses video frames outliers as the output for the proposed dataset. When the algorithm works with a few number of input video these errors could affect the popularity of the clusters. However, if the number of the devices increases, then the effect of the clustering errors is reduced due consensus on visual contents.

The developed system is also able to track the scene clusters over time, by finding relationships between the different clustering outputs obtained over time. In particular, we proposed a novel and alternative vote-based cluster tracking procedure and compared it with the one, threshold-based, described in [169]. From this comparison we found that vote-based procedure reaches very good results totally automatic and independently by a hyperparameter fine tuning phase. In the Assistive Technology research field, the proposed framework could be useful to highlight bad behaviour in the life style (e.g., the user watches too much television, the user spends lots of time driving a car), to track behaviours in the day-time, to log the visited places, and so on. In the security field it can be exploited to automatically cluster several video streams filtering them semantically: this could make faster the search of something or someone that appears in the scene and has been recorded in the video collection.

4.3 RECfusion for lifelogging

4.3.1 Introduction and Motivations

In the last years there has been a rapid emerging of wearable devices, including body sensors, smart clothing and wearable cameras. These technologies can have

a significant impact on our lives if the acquired data are considered to assist the users in tasks related to the monitoring of the quality of life [171, 172, 173, 174]. In particular, egocentric cameras enabled the design and development of useful systems that can be organized into three general categories with respect to the assistive tasks:

(i) User Aware - systems able to understand what the user is doing and what/how he interact with, by recognizing actions and behaviours from first person perspective.

(ii) Environment/Objects Aware - systems able to understand what objects sourround the user, where they are with respect the user's perspective and what the environment looks like.

(iii) Target Aware - systems able to understand what others are doing, and how they interact with the user that is wearing the device.

The egocentric monitoring of a person's daily activities can help to stimulate the memory of users that suffer from memory disorders [175]. Several works on recognition and indexing of daily living activities of patients with dementia have been recently proposed [176, 177, 178, 179]. The exploitation of aids for people with memory problems is proved to be one of the most effective ways to aid rehabilitation [180]. Furthermore, the recording and organization of daily habits performed by a patient can help a doctor to have a better opinion with respect to the specific patient's behaviour and hence his health needs. To this aim, a set of egocentric videos recorded among different days with a camera weared by a patient can be analysed by experts to monitor the user's daily living activities for assistive purposes. The live recording for life logging applications poses challenges on how to perform automatic index and summarization of big personal multimedia data [181].

Beside assistive technologies, the segmentation and semantic organization of egocentric videos is useful in many application scenarios where wearable cameras have recently become popular, including lifelogging [181], law enforcement [182] and social cameras [6]. Other applications of egocentric video analysis are related to action and activity recognition from egocentric videos [183, 184], and recognition of interaction-level activities from videos in first-person view [185] (e.g., recognition of human-robot interactions). For all these applications, the segmentation of daily egocentric videos into meaningful chapters and the semantic organization of such video

segments related to different days is an important first step which adds structure to egocentric videos, allowing tools for the indexing, browsing and summarization of egocentric video sets.

In the last years, several papers have addressed different problems related to vision tasks from first person perspective. The work in [186] proposes a temporal segmentation method of egocentric videos with respect to 12 different activities organized hierarchically upon cues based on wearer's motion (e.g., static, sitting, standing, walking, etc.). A benchmark study considering different wearable devices for context recognition with a rejection mechanism is presented. The system discussed in [187] aims at segmenting unstructured egocentric videos to highlight the presence of given personal contexts of interest. In [174] the authors perform a benchmark on the main representations and wearable devices used for the task of context recognition from egocentric videos. A method that takes a long input video and returns a set of video subshots depicting the essential moments is detailed in [188]. The method proposed in [189] learns the sequences of actions involved in a set of habitual daily activities to predict the next actions and generate notifications if there are missing actions in the sequence. The framework presented in Section 4.2 (i.e., RECfusion [6]), is able to automatically process multiple video flows from different mobile devices to understand the most popular scenes for a group of end-users. The output is a video which represents the most popular scenes organized over time.

In this work, we build on the RECfusion method [6] improving it in the domain of wearable cameras, in the context of daily living monitoring from egocentric videos. Then we apply the achieved insights and the developed methods to the original domain (i.e., mobile camera) and task (i.e., popularity estimation). In RECfusion multiple videos are analysed by using two algorithms: the former is used to segment the different scenes over time (intraflow analysis). The latter is employed to perform the grouping of the videos related to the involved devices over time. As reported in the experimental results of [6], the intraflow analysis of RECfusion suffers when applied to egocentric videos because they are highly unstable due to the user's movements. The framework proposed in this work allows to have better performances for egocentric videos organization, on both segmentation accuracy and computational costs. The proposed method takes a set of egocentric videos regarding the daily living of a user among different days, and performs an unsupervised

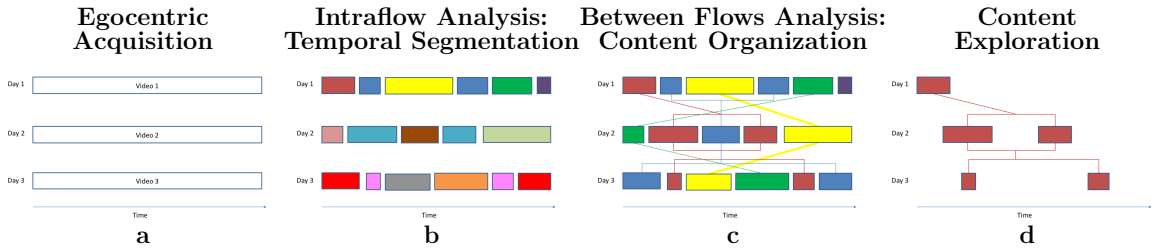


Figure 4.7: Overall scheme of the proposed framework.

segmentation of them. The obtained video segments among different days are then organized by contents. The video segments of the different days sharing the same contents are then visualized by exploiting an interactive web-based user interface. In our framework we use a unique representation for the frames which is based on CNN features [7] (for both intraflow and between flows analysis) instead of the two different representations based on SIFT and color histograms as proposed in RECfusion. Experiments show that the proposed framework outperforms RECfusion for daily living egocentric video organization. Moreover, the approach obtains better accuracy than RECfusion for the popularity estimation task for which RECfusion has been designed.

The rest of the Section is organized as follows. Section 4.3.2 presents the designed framework. Section 4.3.3 describes the considered wearable dataset. The discussion on the obtained results is reported in Section 4.3.4. In Section 4.3.5 the developed system is compared with respect to RECfusion [6] on mobile videos. Finally Section 4.3.6 concludes the Section and gives insights for further works.

4.3.2 Proposed Framework

The proposed framework performs two main steps on the videos acquired by a wearable camera: temporal segmentation and segment organization. Figure 4.7 shows the scheme of the overall pipeline related to the our system.

Starting from a set of egocentric videos recorded among multiple days (Figure 4.7 (a)), the first step performs an intraflow analysis of each video to segment it with respect to the different scenes observed by the user (Figure 4.7 (b)). Each video is then segmented by employing temporal and visual correlations between frames of the same video (Figure 4.7 (b): the colour of each block identifies a scene observed within the same video). Then, the segments obtained over different days referred to

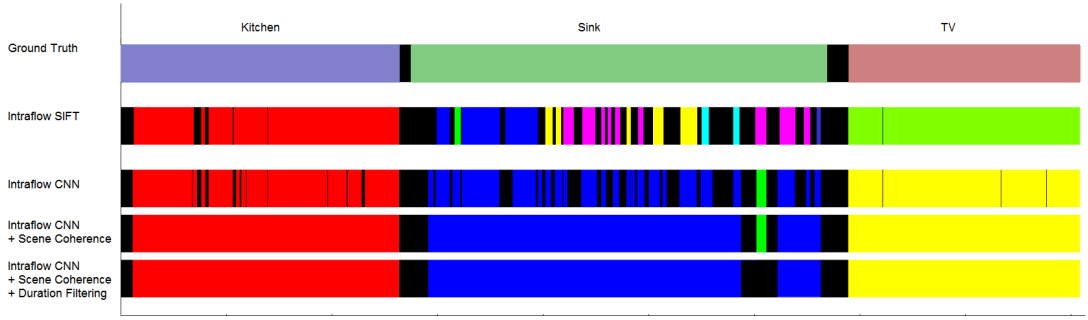


Figure 4.8: Output of the intraflow analysis using SIFT and CNN features when applied to the video *Home Day 1*. The first row is the Ground Truth of the video. The second row shows the result of the intraflow analysis with the SIFT-based method discussed in [6]. The third row shows the result of the intraflow analysis of our method, whereas the last two rows show the results of our method with the application of the proposed scene coherence and duration filtering criteria.

the same contents are grouped by means of a between flows analysis which implements an unsupervised clustering procedure aimed to semantically associate video segments obtained from different videos (Figure 4.7 (c): the colour of each block now identifies video segments associated to the same visual content, observed in different days). The system hence produces sets of video clips related to each location where the user performs daily activities (e.g., the set of the clips over days where the user washes dishes in the kitchen, the set related to the activity of the user of playing piano, and so on). The clips are organized taking into account both, visual and temporal correlations. Finally, the framework provides a web based interface to enable a browsing of the organized videos (Figure 4.7 (d)). In the following subsections the details on the different steps involved into the pipeline are given.

Intraflow Analysis

The intraflow analysis performs the unsupervised temporal segmentation of each input video, as well as associates a scene ID to each video segment. Segments with the same content have to share the same scene ID. This problem has been addressed in [6] for videos acquired with mobile devices. To better explain the problem, in the following we focus our analysis on the issues related to the intraflow algorithm detailed in [6] when applied on first-person videos. This is useful to introduce the main

problems of a classic feature based matching approach for temporal segmentation in wearable domain. Then we present our solution for the intraflow analysis. Furthermore, in Appendix C the pseudocode describing the proposed intraflow analysis approach is reported.

Issues Related to SIFT Based Templates: The intraflow analysis used in [6] compares two scenes considering the number of matchings between a reference template and the frame under consideration (current frame). The scene template is a set of SIFT descriptors that must accomplish specific properties of "reliability". When the algorithm detects a sudden decrease in the number of matchings, it refreshes the reference template extracting a new set of SIFTs and splits the video. In order to detect such changes, the system computes the value of the slope in the matching function (i.e., the variation of the number of matchings in a range interval). When the slope is positive and over a threshold (which correspond to a sudden decrease of the number of matchings between the SIFT descriptors) the algorithm finds a new template (i.e., a new block). When a new template is defined, it is compared with the past templates in order to understand if it regards a new scene or it is related to a known one (backward search phase). Although this method works very well with videos acquired with mobile cameras, it presents some issues when applied on videos acquired with wearable devices. In such egocentric videos, the camera is constantly moving due to the shake induced by the natural head motion of the wearer. This causes a continuous refresh of the reference template that is not always matched with similar scenes during the backward search. Hence, the method produces an oversegmentation of the egocentric videos. Furthermore, it requires to perform several SIFT descriptor extraction and matching operations (including geometric verifications) to exclude false positive matchings. This have a negative impact on the realtime performances. The first row of Figure 4.8 shows the Ground Truth segmentation of the video acquired with a wearable camera in an home environment[2]. An example of a temporal segmentation obtained with the SIFT based interflow analysis proposed in [6] on a egocentric video is reported in the second row of Figure 4.8. The algorithm works well when the scene is quite stable (e.g., when the user is watching a TV program), but it performs several errors when the scene

[2]The video related to the example in Figure 4.8 is available for visual inspection at the URL http://iplab.dmi.unict.it/dailylivingactivities/homeday.html.

is highly unstable due head movements. In fact, in the middle of the video related to Figure 4.8, the user is whashing dishes at the sink, and he is continuously moving his head. In this example the intraflow approach based on SIFT features detects a total of 8 different scenes instead of 3. The algorithm cannot find the matchings between the current frame and the reference template due to two main reasons:

1. when the video is unstable, even though the scene content doesn't change, the matchings between local features are not reliable and stable along time;

2. In a closed space such as an indoor environment, the different objects of the scene can be very close to the viewer. Hence a small movement of the user's head is enough to cause an high number of mismatches between local features.

CNN Based Image Representation: To deal with the issues described in the previous section, we exploit an holistic feature to describe the whole image rather than an approach based on local features. In particular, in the intraflow analysis we represent frames by using features extracted with a Convolutional Neural Network (CNN) [190]. Specifically, we consider the CNN proposed in [7] (*AlexNet*). In our experiments, we exploit the representation obtained considering the output of the last hidden layer of *AlexNet*, which consists of a 4096 dimensional feature vector (*fc7* features). We decided to use *AlexNet* representation since it has been successfully used as a general image representation for classification purpose in the last few years [191] [192]. Moreover, the features extracted by *AlexNet* have been successfully used for transfer learning [193] [194] [195]. Finally, *AlexNet* architecture is a small network compared to others (e.g., VGG [196]). Thus, it allows to perform the feature extraction very quickly. The proposed solution computes the similarity between scenes by comparing a pair of *fc7* features with the *cosine similarity* measure. The cosine similarity of two vectors measures the cosine of the angle between them. This measure is independent of the magnitude of the vectors, and is well suited to compare high dimensional sparse vectors, such as the *fc7* features. The cosine similarity of the two *fc7* feature vectors v_1 and v_2 is computed as following:

$$CosSimilarity(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \, \|v_2\|} \tag{4.3}$$

Proposed Intraflow Analysis: During the intraflow analysis, the proposed algorithm computes the cosine similarity between the current reference template and the *fc7* features extracted from the frames following the reference. When the algorithm detects a sudden decrease in the cosine similarity, it refreshes the reference template selecting a new *fc7* feature corresponding to a stable frame. As in [6], to detect such changes the system computes the value of the slope (i.e., the variation of the cosine similarity in a range interval). When the slope has a positive peak (which correspond to a sudden decrease of the cosine similarity) the algorithm finds a new template and produces a new video segment. There are two cases in which the intraflow analysis compares two templates:

1. a template is compared with the features extracted from the forward frames, when the algorithm have to check its eligibility to be a reference template for the involved scene;

2. A template is compared with the past templates during the backward checking phase to establish the scene ID.

In the first case, the elapsed time between the two compared frames depends on the sampling rate of the frames (in our experiments, we sampled at every 20 frames for videos recorded at 30 fps). Differently, the frames compared during the backward checking could be rather different due to the elapsed time between them. For this reason, when we compare a new template with a past template, we assign the templates to the same scene ID by using a weaker condition with respect to the one used in the forward verification. In the forward process, the algorithm assigns the same scene ID to the corresponding frames if the cosine similarity between their descriptors is higher than a threshold T_f (equal to 0,60 in our experiments). When the algorithm compares two templates in the backward process, it assign the same scene to the corresponding frames if the similarity is higher than a threshold T_b (equal to 0.53 in our experiments).

Besides the image representation, our intraflow algorithm introduces two additional rules:

1. In [6] each video frame is assigned to a scene ID or it is classified as Noise and hence rejected. Our approach is able to distinguish the rejected frame among the ones caused by the movement of the head of the user (Head Noise) and the

frames related to the transition between two scenes in the video (Transition). When a new template is defined after a group of consecutive rejected frames, the frames belonging to the rejected group are considered as "Transition" if the duration of the block is longer than 3 seconds (i.e., head movements are faster than 3 seconds). Otherwise they are classified as "Head Noise". In case of noise, the algorithm doesn't assign a new scene ID to the frame that follows the "Head Noise" video segment because the noise is related to head movements, but the user is in the same context. When the user changes its location he changes his position in the environment. Thus, the transition between different scenes involves a longer time interval.

2. The second rule is related to the backward verification. In [6] it is performed starting from the last found template and proceeding backward. The search stops when the process finds the first past template that have a good matching score with the new template. Such approach is quite appropriate for the method in [6] because it compares sets of local features and relies on the number of achieved matchings. The approach proposed in this work compares pairs of feature vectors instead of sets of local descriptors, and selects the past template that yields the best similarity to the new one. In particular, the method compares the new template with all the previous ones, considering all the past templates that yields a cosine similarity greater than T_b. From this set of positive cases, the algorithm selects the one that achieves the maximum similarity, even if it is not the most recent in the time.

Considering the example in Figure 4.8, the segmentation results achieved with the proposed intraflow approach (third row) are much better than the ones obtained using SIFT features [6] (second row).

After the discussed intraflow analysis a segmentation refinement is performed as detailed in the following subsection.

Intraflow Segmentation Refinement: Starting from the result of the proposed intraflow analysis (see the third row of Figure 4.8), we can easily distinguish between "Transition" and "Noise" blocks among all the rejected frames. A block of rejected frames is a "Noise" block if both the previous and the next detected scenes are related to the same visual content, otherwise it is marked as a "Transition block". We

refer to this criteria as *Scene Coherence*. The result of this step on the example considered in Figure 4.8 is shown in the fourth row. When comparing the segmentation with the Ground Truth (first row), the improvement with respect to [6] (second row) is evident. Moreover, many errors of the proposed intraflow analysis (third row) are removed. The second step of the segmentation refinement consists in considering the blocks related to user activity in a location with a duration longer than 30 seconds (*Duration Filtering*), unless they are related to the same scene of the previous block (i.e., after a period of noise, the scene is the same as before but it has a limited duration). We applied this criteria because, in the context of Activities of Daily Living (ADL), we are interested to detect the activities of a person in a location that have a significant duration in order to be able to observe the behavior. This refinement step follows the *Scene Coherence* one. The final result on the considered example is shown in the last row of Figure 4.8. Despite some frames are incorrectly rejected (during scene changes) the proposed pipeline is much more robust than [6] (compare first, second and fifth rows of Figure 4.8). This outcome is quantitatively demonstrated in the experimental section of this work on a dataset composed by 13 egocentric videos.

Between Video Flows Analysis

When each egocentric video is segmented, the successive challenge is to determine which video segments among the different days are related to the same content. Given a segment block b_{v_A} extracted from the egocentric video v_A, we compare it with respect to all the segment blocks extracted from the other egocentric videos v_{B_j}. To represent each segment, we consider again the CNN $fc7$ features extracted from one of the frames of the video segment. This frame is selected considering the template which achieved the longer stability time during the intraflow analysis. For each block b_{v_A}, our approach assigns the ID of b_{v_A} (obtained during intraflow analysis) to all the blocks $b_{v_{B_j}}$ extracted from the other egocentric videos v_{B_j} such that

$$b_{v_{B_j}} = \underset{\bar{b}_{v_{B_j}} \in v_{B_j}}{\arg\max} \{CosSimilarity(b_{v_A}, \bar{b}_{v_{B_j}}) \mid \sigma_{(b_{v_A}, v_{B_j})} \geq T_\sigma\} \ \forall v_{B_j} \tag{4.4}$$

where $\sigma_{(b_{v_A}, v_{B_j})}$ is the standard deviation of the cosine similarity values obtained by considering the segment block b_{v_A} and all the blocks of v_{B_j}, and T_σ is the decision threshold. This procedure is performed for all the segment blocks b_{v_A} of the video

v_A. When all the blocks of v_A have been considered, the algorithm takes into account a video of another day and the matching process between video segments of the different days is repeated until all the video segments in the pool are processed. In this way a scene ID is assigned to all the blocks of all the considered videos. The pairs of blocks with the same ID are associated to the same scene, even if they belong to different videos, and all the segments are connected in a graph with multiple connected components (as in Figure 4.7 (c)). When there is a high variability in the cosine similarity values (i.e., the value of $\sigma_{(b_{v_A}, v_{B_j})}$ is high), the system assigns the scene ID to the segment block that achieved the maximum similarity. When a block matches with two blocks related to two different scene IDs, the system assigns the scene ID related to the block which achieved the highest similarity value. When a block isn't matched, it means that all the similarity values of the miss-matched blocks are similar. This causes low values of σ and helps the system to understand that the searched scene ID is not present (i.e., scenes with only one instance among all the considered videos). To better explain the proposed approach, the pseudocode related to the above described between video flows analysis is reported in Appendix C.

4.3.3 Dataset

To demonstrate the effectiveness of the proposed approach, we have considered a set of representative egocentric videos to perform the experiments. The egocentric videos are related to different days and have been acquired using a Looxcie LX2 wearable camera with a resolution of 640x480 pixels. The duration of each video is about 10 minutes. The videos are related to the following scenarios:

- **Home Day:** a set of four egocentric videos taken in a home environment. In this scenario, the user performs typical home activities such as cooking, washing dishes, watching TV, and playing piano. This set of videos has been borrowed from the *10contexts* dataset proposed in [187] which is available at the following URL: http://iplab.dmi.unict.it/PersonalLocations/segmentation/.

- **Office Day:** a set of six egocentric videos taken in a home and in different office environments. This set of videos concerns several activities performed during a tipical day. Also this set of videos belongs to the *10contexts* dataset [187].

Table 4.3: Comparison of the performances, on the considered egocentric video dataset, achieved by the intraflow approach in [6] and the proposed one. Each test is evaluated considering the accuracy of the temporal segmentation (Q), the computation time (Time) and the number of the scenes detected by the algorithm (Scenes). The accuracy is measured as the percentage of correctly classified frames with respect to the Ground Truth. The measured time includes the feature extraction process (i.e., image preprocessing and *Alexnet* forward pass.).

			Intraflow proposed in [6]			**Proposed Intraflow Approach**		**Proposed Interflow Approach with Segmentation Refinement**		
Video	**Scenario**	**Scenes**	**Q**	**Scenes**	**Time**	**Q**	**Scenes**	**Q**	**Scenes**	**Time**
1	HomeDay1	3	62,5%	8	20'45"	77,5%	4	**92,5%**	**3**	**1'23"**
2	HomeDay2	3	71,6%	3	20'18"	80,3%	4	**94,5%**	**3**	**1'46"**
3	HomeDay3	3	64,3%	5	19'03"	79,7%	5	**94,3%**	**3**	**1'21"**
4	HomeDay4	3	84.4%	3	8'36"	**91.8%**	**3**	85,4%	2	**36"**
5	WorkingDay1	4	95,7%	5	16'16"	98,4%	5	**99,5%**	**4**	**1'22"**
6	WorkingDay2	4	82,5%	5	15'15"	98,9%	5	**100%**	**4**	**1'08"**
7	WorkingDay3	5	98,7%	6	19'02"	99,2%	6	**99,4%**	**5**	**1'29"**
8	OfficeDay1	3	23,0%	5	24'8"	55,3%	19	**66,9%**	**2**	**2'39"**
9	OfficeDay2	2	59,7%	2	10'25"	90,0%	3	**98,7%**	**2**	**1'26"**
10	OfficeDay3	3	57,2%	4	13'28"	83,6%	10	**96,3%**	**3**	**1'49"**
11	OfficeDay4	3	52,0%	4	11'37'	79,5%	5	**84,1%**	**4**	**1'41"**
12	OfficeDay5	3	70,7%	3	8'35"	86,7%	4	**95,9%**	**3**	**1'21"**
13	OfficeDay6	3	78,8%	**3**	9'33"	61,5%	5	**94,5%**	4	**1'34"**
	Average		69,3%		15'9"	83,3%		**91,8%**		**1'31"**

- **Working Day:** a set of three videos taken in a laboratory environment. The activities performed by the user in this scenario regards reading a book, working in a laboratory, sitting in front of a computer, etc.

Each video has been manually segmented to define the blocks of frames to be detected in the intraflow analysis. Moreover, the segments have been labeled with the scene ID to build the Ground Truth for the between video analysis. The Ground Truth is used to evaluate the performances of the proposed framework. The used egocentric videos, as well as the Ground Truth, are available at the following URL: http://iplab.dmi.unict.it/dailylivingactivities/.

4.3.4 Experimental Results

In this section we report the temporal segmentation and the between flows video analysis results obtained on the considered dataset.

Temporal Segmentation Results

Table 4.3 shows the performances of the proposed temporal segmentation method (see Section 4.3.2). We compared our solution with respect to the one adopted by RECfusion [6]. For each method we computed the quality of the segmentation as the percentage of the correctly classified frames (Q), the number of detected scenes and the computational time [3]. The proposed approach obtains strong improvements (up to over 30%) in segmentation quality with respect to RECfusion (e.g., results in rows eight and nine of Table 4.3). Furthermore, the application of the segmentation refinements provides improvements up to 43% in segmentation quality (results at row eight in Table 4.3). In the fourth row of Table 4.3 (related to the analysis of the video *Home Day 4*), we can observe that the application of the segmentation refinements causes a decrease in performances. This video is very short compared to the other videos of the dataset. It has a duration of just 4'28" and consists of a sequence of three different scenes (piano, sink and TV). The scene blocks are correctly detected by the proposed intraflow approach, which finds exactly 3 different scenes and achieves a 91,8% of accuracy without refinement. However, the middle scene (i.e., sink) has a duration of just 22 seconds according to the Ground Truth used in [187], thus the refinement process rejects this block due to the application of the *Duration Filtering* criteria. Considering the mean performances (last row in Table 4.3) our system achieves an improvement of over 14% without segmentation refinements, with over 22% of margin after the segmentation refinement. The proposed method also reduces the computational time of more than 21 minutes in some cases (eighth row in Table 4.3). It has an average computational time saving of about 13 minutes with respect to the compared approach [6]. The results of Table 4.3 show that the application of the *Scene Coherence* and the *Duration Filtering* criteria used in the segmentation refinement step (Section 4.3.2) allows to detect the correct number of scenes.

In sum, considering the qualitative and quantitative results reported respectively in Figure 4.8 and Table 4.3, the proposed system is demonstrated to be robust for the temporal segmentation of egocentric videos, and it provides high performances with a lower computational time with respect to [6].

[3]The experiments have been performed with Matlab 2015a, running on a 64-bit Windows 10 OS, on a machine equipped with an Intel i7-3537U CPU and 8GB RAM.

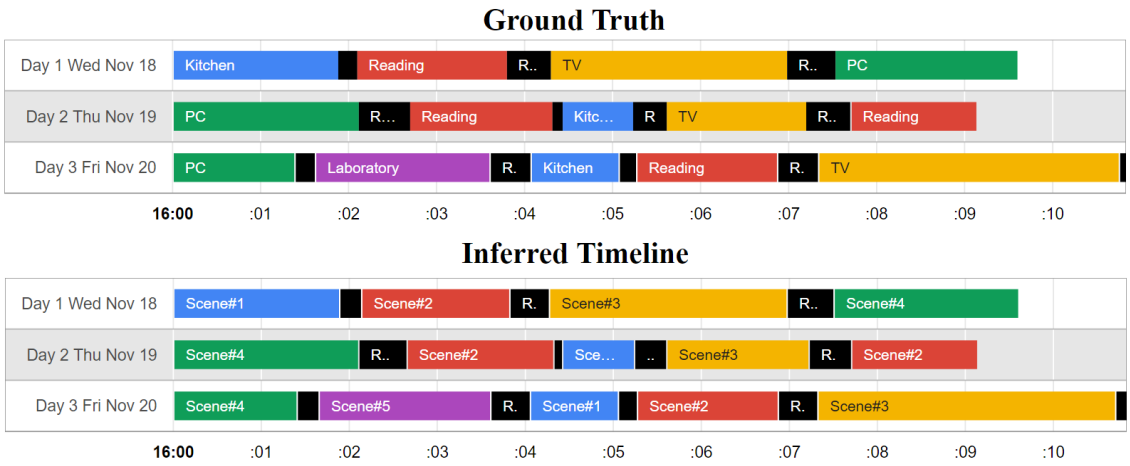


Figure 4.9: The two timelines show the Ground Truth segmentation of egocentric videos related to the *Working Day* scenario and their organization (top). The inferred segmentation and organization obtained by the proposed method is reported in the bottom.

Between Flows Video Analysis Results

In our experiments, all the segments related to the *Home Day* and *Working Day* scenarios have been correctly connected among the different days without errors. Figure 4.9 shows two timelines related to the videos of the scenario *Working Day*, whereas Figure 4.10 shows the timelines related to the scenario *Home Day*. The first timeline in Figure 4.9 and 4.10 shows the Ground Truth labeling. In the timeline, the black blocks indicate the transition intervals (to be rejected). The second timeline shows the result obtained by our framework. In this case, the black blocks indicate the frames automatically rejected by the algorithm. In order to better asses the results obtained by the proposed system, the reader can perform a visual inspection of all the results produced by our approach at the following URL: http://iplab.dmi.unict.it/dailylivingactivities/. Through the web interface the different segments can be explored.

Differently than the *Home Day* and *Working Day* scenarios, some matching error occurred in the between flow analysis of the *Office Day* scenario (see Figure 4.11). Since we are grouping video blocks by contents, to better evaluate the performance of the proposed between flow analysis we considered three quality scores usually used in clustering theory. The simplest clustering evaluation measure is the *Purity* of the clustering: each cluster of video segments obtained after the between flow

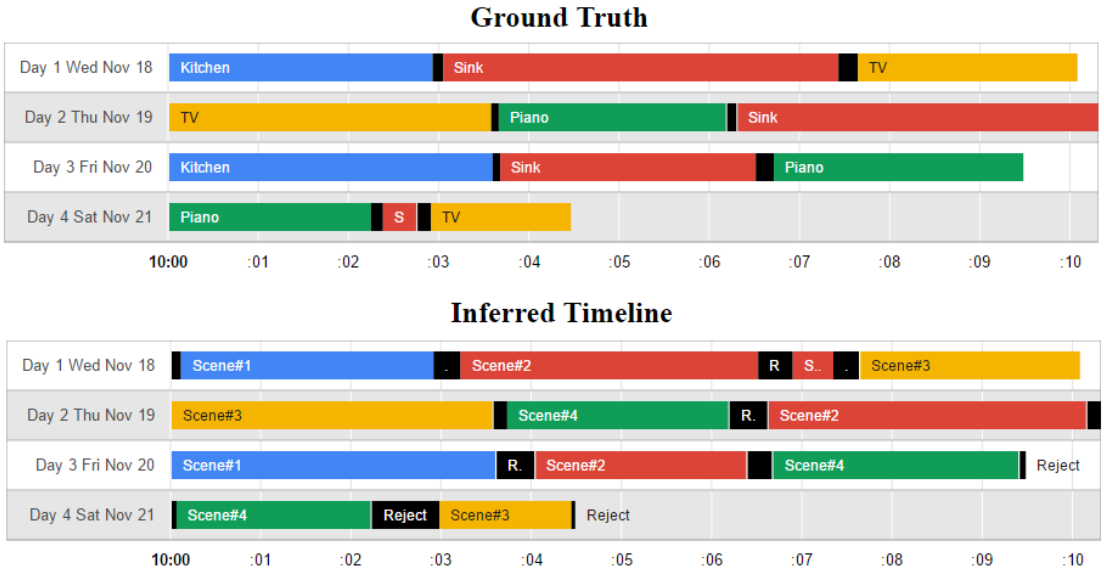


Figure 4.10: The two timelines show the Ground Truth segmentation of egocentric videos related to the *Home Day* scenario and their organization (top). The inferred segmentation and organization obtained by the proposed method is reported in the bottom.

analysis is assigned to the most frequent class in the cluster. The *Purity* measure is hence the mean of the number of correctly assigned video blocks within the clusters.

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |C_i \cap T_j| \tag{4.5}$$

where N is the number of video blocks in the cluster, k is the number of clusters, C_i is the i-th cluster and T_j is the set of elements of the class j that are present in the cluster C_i. The clustering task (i.e., the grouping performed by the between flow analysis in our case) can be viewed as a series of decisions, one for each of the $N(N-1)/2$ pairs of the elements to be clustered [197].

The algorithm obtains a true positive decision (TP) if it assigns two videos of the same class to the same cluster, whereas a true negative decision (TN) if it assigns two videos of different class to different clusters. Similarly, a false positive decision (FP) assigns two different videos to the same cluster and a false negative decision (FN) assigns two similar videos to different clusters. With the above formalization we can compute the confusion matrix associated to the pairing task.

From the confusion matrix we compute the *Rand Index* (RI), which measures

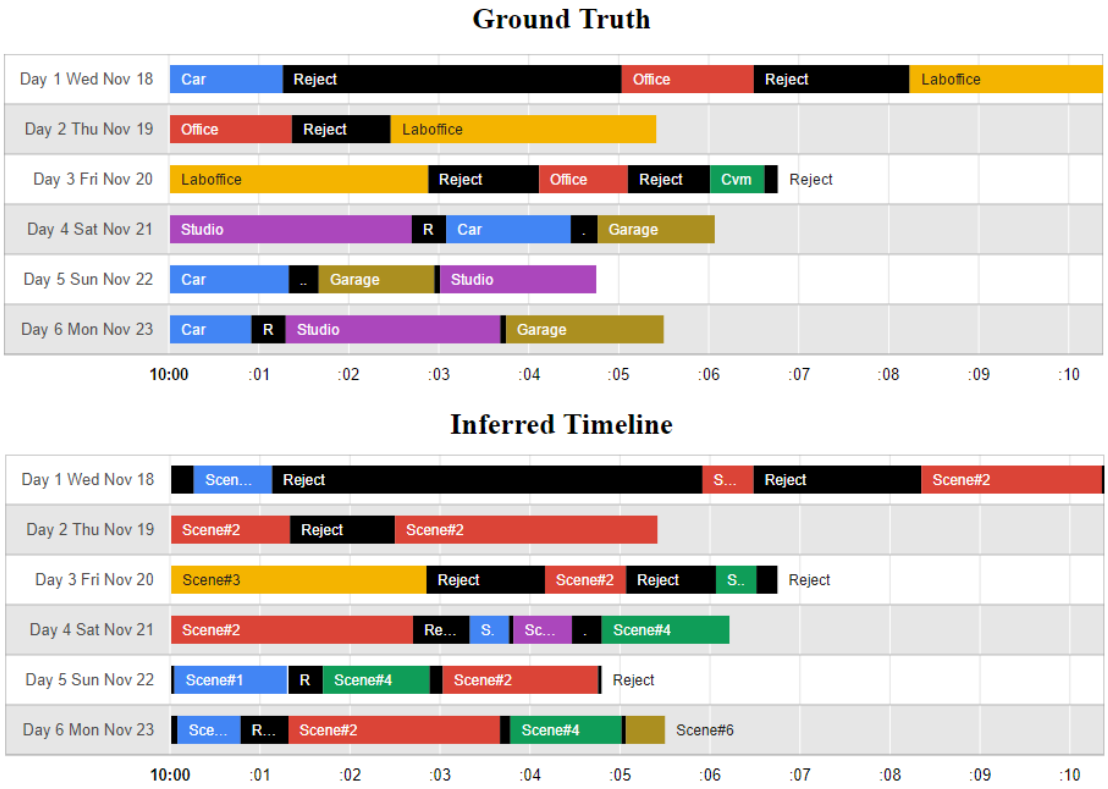


Figure 4.11: The two timelines show the Ground Truth segmentation of egocentric videos related to the *Office Day* scenario and their organization (top). The inferred segmentation and organization obtained by the proposed method is reported in the bottom. In this case, the blocks related to the scenes 'office', 'studio' and 'laboffice' are clustered in the same cluster (identified by the color red), with the exeption of only one 'laboffice' block.

the percentage of the correct decisions (i.e., the pairing accuracy) of the between flow analysis:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.6}$$

In order to take into account both precision and recall, we also considered the F_β measure defined as following:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 Precision + Recall} \tag{4.7}$$

Specifically, we considered the F_1 measure that weights equally precision and recall, and the F_2 measure, which weights recall higher than precision and, therefore, penalizes false negatives more than false positives. Table 4.4 shows the results of the proposed between flow analysis approach considering the aforementioned measures.

Table 4.4: Between video clustering results.

		Original GT				New GT			
Dataset	**# of Videos**	**Purity**	**RI**	**F_1**	**F_2**	**Purity**	**RI**	**F_1**	**F_2**
WorkingDay	3	1	1	1	1	--	--	--	--
HomeDay	4	1	1	1	1	--	--	--	--
OfficeDay	6	0,68	0,81	0,49	0,57	0,95	0,89	0,80	0,75
Office+Home	10	0,58	0,83	0,41	0,54	0,74	0,87	0,61	0,67

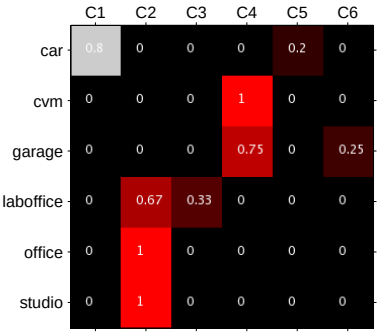


Figure 4.12: Co-occurrence matrix obtained considering the *Office Day* scenario.

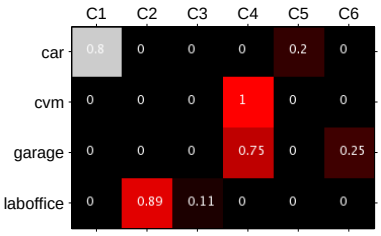


Figure 4.13: Co-occurrence matrix obtained considering the *Office Day* scenario with the "New GT".

For the *Home Day* and *Working Day* scenarios our approach achieves the maximum scores for all the considered evaluation measures (column "Original GT" of Table 4.4). Regarding the *Office Day* scenario, we obtain lower scores of Purity and Rand Index. The obtained values of F_1 and F_2 measures indicate that the proposed between flow approach achieves higher recall values than precision. This can be further verified by observing the Figure 4.12. This figure shows the co-occurrence matrix obtained considering the *Office Day* scenario. The columns of the matrix represent the computed graph components (clusters) obtained with the between flow analysis, whereas each row represents a scene class according to the Ground Truth. The values of this matrix express the percentage of video blocks belonging to a specific class that are assigned to each graph component (according to the Ground Truth used in [187]). This figure shows that even if different blocks are included in the same graph component (FP), the majority of the blocks belonging to the same class are assigned to the same graph component (TP). The second column of the co-occurrence matrix in Figure 4.12 shows that the "laboffice", "office" and "studio" blocks have been assigned to the graph component C2. This error is due to strong

Figure 4.14: Some first person view images from the *Office Day* scenario depicting frames belonging to the classes "laboffice", "office" and "studio".

ambiguity in the visual content of these three classes. Figure 4.14 shows some example of the frames belonging to these classes. We can observe that these scenes are all related to the same activity (i.e., working at a PC) and the visual content is very similar. We repeated the tests considering an alternative Ground Truth that considers the three above mentioned scene classes to be a unique class (Laboffice). The results of this test are reported in the column labeled as "New GT" of Table 4.4. In this case the *Office Day* scenario result achieves significant improvements for all the considered evaluation measures. Figure 4.13 shows the co-occurrence matrix obtained considered the new Ground Truth.

The last row of Table 4.4 reports the results obtained by applying the proposed approach on the set of 10 videos obtained by considering only the *Home Day* and the *Office Day* sets of videos. This test have been done to analyse the robustness of the proposed approach considering more contexts and also scenes that appear only in some videos. Furthermore, to better assess the effectiveness of the proposed approach, we performed a number of incremental tests by varying the number of input videos from 2 to 10 videos of the considered set. This correspond to consider from 2 to 10 days of monitoring. The obtained scores reported in Figure 4.15 shows that the evaluation measures are quite stable. We also evaluated the performance of the proposed approach by varying the threshold T_σ value (see Figure 4.16). Also in this case the results demonstrate that the method is robust. We also tested the between analysis approach by using the color histogram and the distance function used in RECfusion [6]. When the system uses such approach there is a high variance in the obtained distance values even if there isn't any segmentation block to be matched among videos. This causes the matching between uncorrelated blocks and

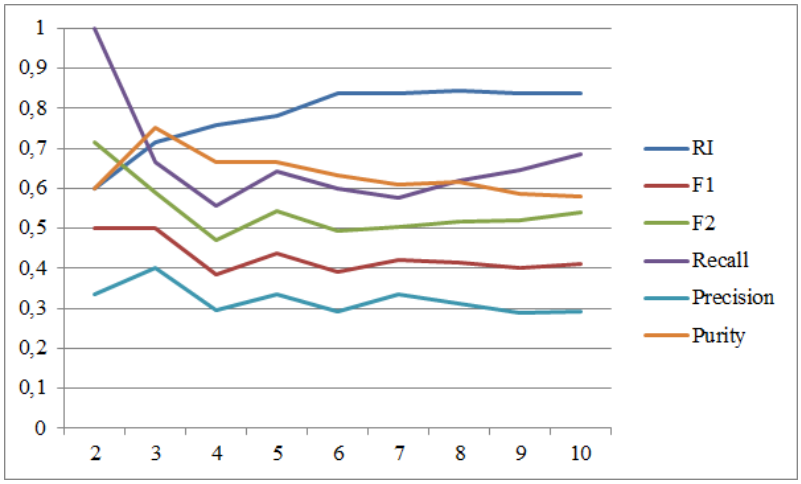


Figure 4.15: Evaluation measures at varying of the number of videos.

hence errors.

4.3.5 Popularity estimation

Considering the improvements achieved by the proposed method in the context of egocentric videos, we have tested the approach to solve the popularity estimation problem defined in [6]. The popularity of the observed scene in a instant of time depends on how many devices are looking at that scene. In the context of organizing the egocentric videos the popularity can be useful to estimate the most popular scene over days.

To properly test our approach for popularity estimation we have considered the dataset introduced in [6]. This allows a fair comparison with the state of the art approaches for this problem because the results can be directly compared with the ones in [6]. Differently than [6], after processing the videos with the intraflow analysis and segmentation refinement proposed in this approach, we have used the proposed CNN features and the clustering approach of [6]. To evaluate the performances of the compared methods we used three measures. For each clustering step we compute:

- P_{r}: ground truth popularity score (number of cameras looking at the most popular scene) obtained from manual labelling;
- P_{a}: popularity score computed by the algorithm (number of the elements in the popular cluster);

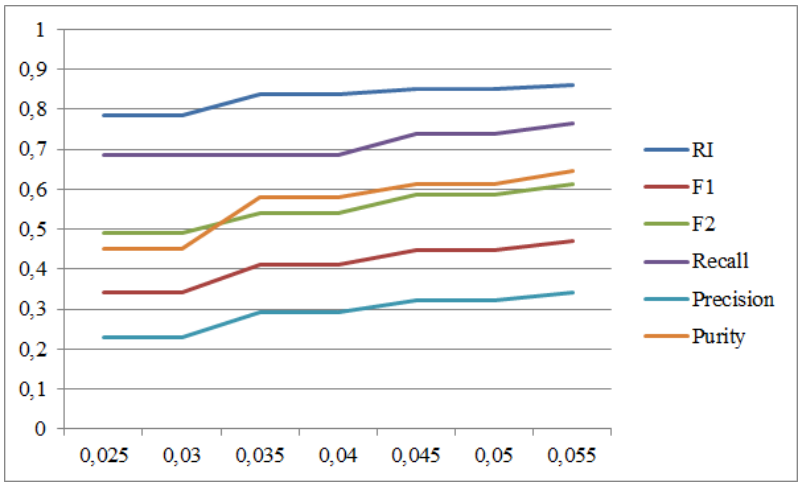


Figure 4.16: Evaluation measures obtained by varying the threshold T_σ.

- P_g: number of correct videos in the popular cluster (inliers).

From the above scores, the weighted mean of the ratios $P_\mathrm{a}/P_\mathrm{r}$ and $P_\mathrm{g}/P_\mathrm{r}$ over all the clustering steps are computed [6]. The ratio $P_\mathrm{a}/P_\mathrm{r}$ provides a score for the popularty estimation, whereas the ratio $P_\mathrm{g}/P_\mathrm{r}$ assesses the visual content of the videos in the popular cluster. These two scores focus only on the popularity estimation and the number of inliers in the most popular cluster. Since the aim is to infer the popularity of the scenes, it is useful to look also at the number of outliers in the most popular cluster. In fact, the results reported in [6] show that when the algorithm works with a low number of input videos, the most popular cluster is sometimes affected by outliers. These errors could affect the popularity estimation of the clusters and, therefore, the final output. Thus, we introduced a third evaluation measure that takes into account the number of outliers in the most popular cluster. Let P_o be the number of wrong videos in the popular cluster (outliers). From this score, we compute the weighted mean of the ratio $P_\mathrm{o}/P_\mathrm{r}$ over all the clustering steps, where the weights are given by the length of the segmented blocks (i.e., the weights are computed as suggested in [6]). This value can be considered as a percentage of the presence of outliers in the most popular cluster inferred by the system. The aim is to have this value as lower as possible.

Table 4.5 shows the results of the popularity estimation obtained by the compared approaches. The first column shows the results obtained by the approach

Table 4.5: Experimental results on the popularity estimation problem.

		[6]			**Proposed method**		
Dataset	**dev**	Pa/Pr	Pg/Pr	Po/Pr	Pa/Pr	Pg/Pr	Po/Pr
Foosball	4	1,02	1	0,023	**1**	**1**	**0**
Meeting	2	1,01	0,99	0,020	**0,99**	**0,99**	**0,018**
Meeting	4	**0,99**	**0,95**	0,033	0,93	0,93	**0**
Meeting	5	**0,89**	**0,76**	0,131	0,70	0,70	**0,006**
SAgata	7	1,05	**1**	0,050	**0,99**	0,99	**0**

proposed in [6]. Although this method achieves good performance in terms of popularity estimation and inliers rate, the measure P_o/P_r highlights that the method suffers from the presence of outliers in the most popular cluster. This means that the popularity ratio (P_a/P_r) is affected by the presence of some outliers in the most popular cluster. Indeed, in most cases the P_a/P_r value is higher than 1 and P_o/P_r is greater than 0. The second column of Table 4.5 is related to the results obtained with the proposed approach. We obtained values of popularity estimation and inliers rate comparable with respect to [6]. In this case, the values of popularity score are all lower than 1, which means that sometimes the clustering approach lost some inliers. However it worth to notice that for all the experiments performed with the proposed approach, the outlier ratio P_o/P_r is very close to zero and lower with respect to the values obtained by [6]. This means that the most popular cluster obtained by the proposed approach is not affected by outliers, and this assure us that the output video belongs to the most popular scene. By performing a visual assessment of the results, we observed that using the proposed CNN features in the clustering phase involves a fine-grained clustering, which better isolates the input videos during the transitions between two scenes or during a noise time interval. This behaviour is not observed in the outputs obtained by [6]. Using the color histogram indeed, the system defines a limited number of clusters that are, therefore, affected by outliers. Some examples about this behaviour are shown in Figure 4.17 and Figure 4.18. In these figures, each column shows the frames taken from the considered devices at the same instant. The border of each frame identifies the cluster. The first column of each figure shows the result of the proposed approach, and the second column shows the result of the method proposed in [6]. Specifically Figure 4.17 shows an example of clustering performed by the compared approaches during the analysis of the scenario *Foosball*. In this example, the first and the third devices are viewing the

Figure 4.17: Examples of clustering performed by our system and the approach in [6]. Each column shows the input frames taken from different devices. The color of the border of each frame identifies the cluster. The proposed method has correctly identified the three clusters.

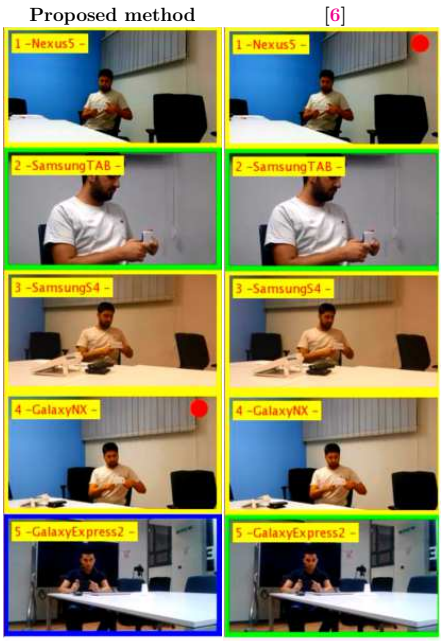


Figure 4.18: Examples of clustering performed by our system and the approach in [6]. Each column shows the input frames taken from different devices. The color of the border of each frame identifies the cluster. The proposed method produces better results in terms of outlier detection.

same scene (the library), the second device is viewing the sofa, whereas the fourth device is performing a transition between the two scenes. In such case, the proposed method creates a different cluster for the device that is performing the transition, whereas the method proposed in [6] includes the scene in the same cluster of the second device. Figure 4.18 shows an example of the popularity clustering performed during the analysis of the scenario *Meeting*. In this case both the approaches fail to insert the second frame in the popular cluster due to the huge difference in scale, but the method in [6] creates a cluster that includes the second and the fifth device (despite they are viewing two different scenes) whereas the proposed method can distinguish the second image from the fifth one.

Besides the improvement in terms of performance, proposed approach provides also an outstanding reduction of the computation time. In fact, the time needed

to extract a color histogram as suggested in [6] is about 1.5 seconds, whereas the time needed to extract the CNN feature used in our approach is about 0.08 seconds. Regarding the popularity intraflow analysis on the mobile video dataset, the proposed approach achieves similar segmentation results compared to the SIFT based approach. However, as in the wearable domain test, the proposed segmentation method strongly outperforms [6] when the system have to deal with noise (e.g., hand tremors).

4.3.6 Final Remarks

In this Section we presented a complete framework to segment and organize a set of egocentric videos of daily living scenes. The experimental results show that the proposed system outperforms the state of the art in terms of segmentation accuracy and computational costs, both on wearable and mobile video datasets. Furthermore, our approach obtains an improvement on outliers rejection on the popularity estimation problem [6].

4.4 Conclusions

In this Chapter we presented RECfusion, a framework designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices. Given a set of video streams as input, the framework can group these video streams by means of similarity and popularity, then it automatically suggests a video stream to be used as output, acting like a "virtual director".

First we focused on the task of the automatic detection of the most popular scene in a social event, where attending people are simultaneously recording a set of unknown scenes with different mobile devices (i.e., smartphones, digital cameras). Several issues related to the differences of the devices, the unknown number and type of scenes as well as the high presence of noise in videos recorded in real case scenarios have been addressed.

Then, considering the low performance of RECfusion with low quality videos, we tackled the problem of applying RECfusion in the wearable domain by improving the approach with the objective to challenge the task of daily living monitoring by egocentric video analysis. The defined approach is able to segment and organize a

set of wearable videos with high segmentation and clustering accuracy as well as low computational costs. The achieved improvements allow RECfusion to outperform previous methods in the task of scene popularity detection, in terms of both accuracy and computational time.

Chapter 5

Final Discussion, Remarks and Future Works

The main contributions of this thesis are related to the investigation of methods for Sentiment Analysis on visual contents (i.e., Visual Sentiment Analysis). Within the broad scope of Visual Sentiment Analysis, we have investigated specific tasks, namely Image Polarity Prediction and Image Popularity Prediction. Furthermore, in the context of videos, we investigated the problem of content popularity estimation (i.e., visual consensus) applied in both mobile and wearable video domains.

The input data for all the addressed tasks comes from large groups of people with common interests, that share their experiences and opinions by publishing visual contents through specific social platforms aimed to spread contents to the largest possible, but targeted, number of people. Chapter 2 presented the crowdsourcing paradigm with a focus on media contents. It further presents *The Social Picture* (TSP), a framework for the inference of the preferences of people visiting a cultural heritage site, or attending an event, by the analysis of crowdsourced collection of photos. By exploiting the crowdsourcing paradigm, *The Social Picture* empowers each gathered image, which contribute to the inferences performed by the system. Indeed, cues about what is perceived as "salient" by the audience may rise from a collection of photos. Moreover, users' pictures can be exploited to build a 3D reconstruction of the scene, which takes into account only the content that is important for users.

Our investigation pointed out the following:

- the large amount of public available data supports the definition of systems able to exploit crowdsourced data coming from real users to get insights and

infer the "sentiment" of people;

- the visual exploration and analysis of very large collections of photos require the development of systems able to automatically infer the semantic from images, properly correlate the information of visual contents, as well as the design of visualization tools that allow the visual exploration of huge amount of images taking into account the computational costs. That is, these systems need to be able to scale with the number of images in the collection;

- the analysis of a large group of photos, collected from people that visit the same heritage site or attend the same event, allows several inferences about users' preferences and behaviours. Such outcomes can be exploited to better organize the site fruition (e.g., by changing the visitors paths), perform focused investments and drive marketing decisions, among others.

Chapter 3 presented the field of Visual Sentiment Analysis applied on still images (referred as Image Sentiment Analysis). First it presents a complete view of the field, by detailing the theoretical models, the state of the art methods, approaches, and available datasets, as well as a number of new challenges related to the Visual Sentiment Analysis field. In particular, the most significant publications are described and compared, considering a chronological order, allowing the reader to follow the evolution of the field. Then, the Chapter presents two works related to Sentiment Polarity Prediction (Section 3.6) and Image Popularity Prediction (Section 3.7).

The approach presented in Section 3.6 is based on the observation that most of the existing methods, that combine visual and textual information, lean on the text provided by users, which is often noisy. The presented approach proposes an alternative source of text, directly extracted from images by means of deep neural networks inferences. Experimental results demonstrated that the the exploitation of the proposed *Objective Text* outperforms the state of the art for sentiment polarity estimation.

Our investigation pointed out that:

- visual and textual information (i.e., views) can be combined to define a new feature space (i.e., embedding), in which the correlation of coupled input views is maximized. The representations defined in such a feature space are able to

embed the factors by which the sentiment evoked by a picture is affected, whose contributes come from the different modalities;

- the text associated to images by users is unreliable, due to its subjective nature, as it is often properly defined with the aim to boost the engagement of the photo in the social platform. Thus, it needs to be properly filtered and processed;
- defining features by exploiting text automatically inferred from the image visual content brings several advantages. Our approach exploits four deep architectures trained on different inference tasks. Hence, this approach provides a fixed structured and reliable source of text associated to images taken from different perspectives, as demonstrated by the conducted experiments;
- the exploitation of deep visual features further boost the performances, however the best results have been obtained by combining deep visual features and the proposed *Objective Text* features.

The work presented in Section 3.7 addresses a challenging task related to the prediction of the popularity evolution of a photo over a period of 30 days. Section 3.7 first presents the state of the art in image popularity prediction. Then, it adds the temporal axis to the problem presenting an even more challenging task. The first dataset related to this task is presented, as well as an approach to the challenge of photo popularity dynamics prediction.

Our study pointed out that:

- the popularity score of a shared image can't be defined as a single value that ignores the temporal evolution of the photo engagement in the platform. Indeed, such a scoring strategy penalizes old contents with respect to new ones;
- the photo lifecycle is characterized by several periods, and not all the images have the same trend. Our study detected a number of shapes for the photo sequences, which can be represented by proper sequence templates;
- the social features associated to a photo (i.e., user and post features) are useful to predict the popularity dynamic of a photo, even at time zero. In particular, the features of the user who shared the photo have been proved to have strong correlations with image popularity.

In Chapter 4 we presented RECfusion, a system able to analyse a set of videos, split each video based on the visual content and correlate the segments extracted from different videos by creating clusters of scenes. The presented system works in a fully-unsupervised way. Indeed, the analysis is performed without any information on the input devices, the number, duration nor type of scenes. In the experiments, we challenged RECfusion to work on both mobile and wearable domains, which have several differences. The developed system is able to estimate the most popular scene in a group of videos, over time, with high accuracy on the selection of the popular scenes, low outlier rates and computational costs.

The main contributions and insights coming from the works presented in Chapter 4 are the following:

- the comparison between images taken with very different devices can be difficult, however deep visual features extracted by modern CNNs (Convolutional Neural Networks) are useful in generalizing the image content by providing an holistic representation related to the semantic of the image;
- the main difficulty in scene popularity estimation is related to the unconstrained number of devices and scenes. Indeed it requires the definition of unsupervised clustering methods, which are often computational heavy and require a final supervision (i.e., at the end one have to choose the number of clusters). In the presented study, a fully unsupervised clustering of the scenes is presented, based on the relative similarity among sub-groups of scenes previously segmented and associated to the other segments within the same video. Such a two-step unsupervised clustering demonstrated high performances in grouping scenes from different videos without any prior on the video contents.

5.1 Future Directions

In this thesis, we have investigated several aspects related to Visual Sentiment Analysis applied on crowdsourced images and videos, with a special focus on image polarity and popularity. However, many challenges still need to be faced. In the context of Image Sentiment Analysis, systems with broader ambitions could be developed to address the new challenges (e.g., relative attributes, popularity prediction, commonsense, etc.) or to focus on new emerging tasks (e.g., image popularity prediction,

sentiment over time, sentiment by exploiting ideograms, etc.). Furthermore, the task of popularity dynamics prediction presented in this thesis finds very practical applications and worth to be further investigated. To this aim, the built dataset is publicly available. Future works can be devoted to the extension of the dataset by taking into account other social platforms (e.g., Facebook, Twitter, Instagram, etc.). Furthermore, additional time-aware features can be considered, such as the day of the week and the hour of the day. Also, different approaches to treat the problem of popularity dynamics prediction as a time series forecasting task can be taken into account, by exploiting techniques that explicitly model the temporal relationship between data (e.g., Hidden Markov Models, Long-Short Term Memory Networks, etc.).

In the context of video analysis, future works can consider to extend RECfusion to perform recognition of scenes or contexts from egocentric images [8, 167, 172, 198] and to recognize the activities performed by the user [183]. Furthermore, recent methods for action anticipation could be also taken into account [199].

Appendices

Appendix A

Preliminary Study on Social Influencers Engagement

This section summarizes the activity done at the Imperial College London, as Visiting Researcher of the Business School, under the supervision of Professor Catarina Sismeiro. The described activity consists on the statistical analysis of social post published on Facebook by specific brands (i.e., brands' official pages) and users who can access to large audiences, also known as "social influencers".

The aim of this preliminary study is to assess the possibility to detect the features which most influence the engagement reached by the social posts, to be able predict or maximize the influencers' social advertising campaigns in terms of reached audiences (i.e., number of users who saw the post) and engagement (i.e., number of likes, comments, and shares). The insights coming from this preliminary investigation have been useful and inspiring for the design of the approach described in Section 3.7.

A.1 Influencer Marketing in Social Media

Nowadays, social networks have a fundamental role for companies and service providers that aim to reach an high number of potential customers with specific characteristic (e.g., gender, age, work, hobbies, etc.) at low costs. In this context, the last years saw the rise of the so-called "social influencers". A social influencer is a user of one or more social platforms with a very large number of followers/fans (called "fanbase") who follows the influencer activity on the social platforms and interact with his posts. Every content posted by a social influencer reaches a large number

of people in very short time, and generates a sequence of reactions of the member of his fanbase. In the early years of social networks, the social profiles with large fanbases were related to professionals and experts (e.g., journalists). However, in the last years, this role has been taken mostly by people from show-business (i.e., models, singers, actors, etc.). In this context, detecting social users able to influence large numbers of people is very important for companies that want to spread their brands and products. Indeed, social influencers are engaged by companies to publish social posts related to specific products and at predefined times.

A.2 Activities

The following activities are related to real data downloaded from the Facebook profiles of a group of social influencers committed to publish specific contents aimed to advertise companies' products.

A.2.1 Problem Analysis

A common way to measure the effect of a social post un users takes into account three main indexes which describe the level of engagement of the users who read the post: number of likes, number of comments and number of shares. However, one can't just consider just the raw values of these scores. Indeed, factors such as the number of members in the fanbase, the time of posting and how old is the post should be taken into account.

Several works in the state of the art normalize the three above indexes with the number of followers [200] or with the number of days passed since the post has been published [33]. This normalization is needed, for instance, to compare the values coming from influencers/pages with very different fanbases or to compare tho posts with different ages. For the aims of this analysis, of course, is convenient to suggest to perform an advertising campaign by committing influencers with the highest number of followers as possible.

Normalizing an engagement value with the age of the post will penalizes contents published in the past with respect to more recent contents. This is evident when the difference between the ages of two compared posts is high. A social post is characterized by a lifecycle. Typically, the most of the engagement is obtained in

Name	ID	@pagenameID
Greta Scarano	134045083337849	
The Style Pusher by Lavinia Biancalani	186732534837251	thestylepusher
Syria	39461368738	syriamusic
Catherine Poulain (Catherine Poulain Official)	229885990518845	CatherinePoulainPage
Andrea Delogu	235334306596863	andreadelogu
Federica Fontana	143548395657271	runfederun
Barilla	11503325699	BarillaIT
Estee Lauder UK	107111802663853	EsteeLauderUK
PUPA Milano Italy	121945384890	PupaMilanoItaly
MAC Cosmetics	16126780553	MACcosmetics
Clinique	147484181947465	cliniqueuk

Table A.1: List of the Facebook profiles related to influencers/brands subject of our analysis.

the first period (early period), then the engagement measures become more stable. This effect is also due to the mechanism of the platform to show the posts on the users' feed pages, that are often not known. As a consequence, old posts will stop to generate engagement at a certain time. By the other end, an analysis performed on a very new post could underestimate the engagement reached by that post in the best point of its lifecycle. All the above factors affect the lifecycle of a post, making difficult the related analysis and any prediction approach.

Based on the above, the first goal of this analysis is the detection of a period in which the posts reach the 90-95% of their maximum engagement, in terms of likes, comments and shares. Each comment is accompanied with the date/time information. Therefore, we can easily observe the temporal evolution of the number of comments generated by the post. However, the number of likes and shares are cumulative values. Thus, we performed a daily crawling of such indexes related to a set of social posts for a period of 40 days. Table A.1 lists the 11 monitored profiles.

A.2.2 Comments Temporal Analysis

First we analyzed the comments related to the posts of the 11 Facebook profiles listed in Table A.1. As previously mentioned, each post comment is accompanied

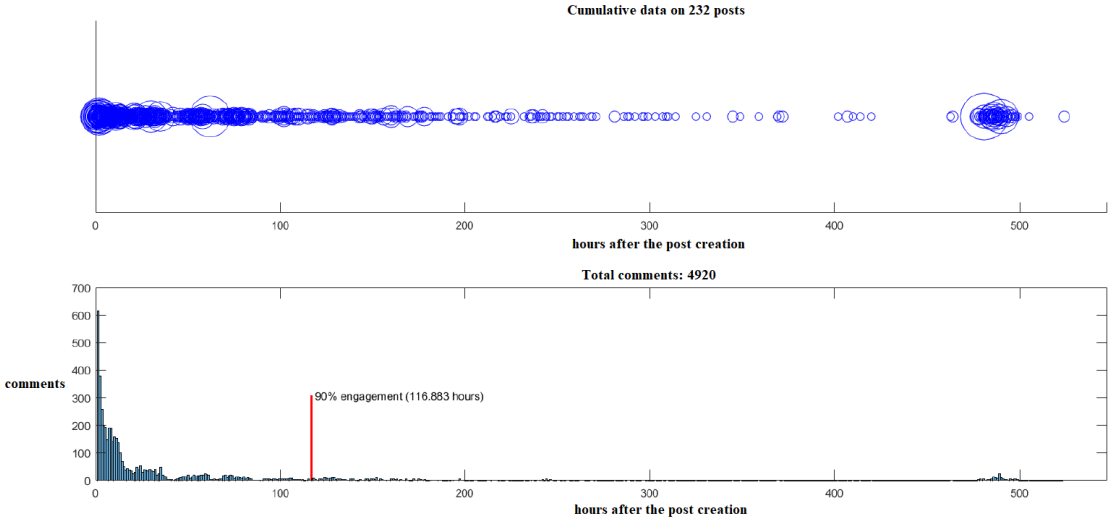


Figure A.1: Temporal sequence of the age of the comments (i.e., hours between the post and the comment creation) related to 232 Facebook posts published by the 11 profiles listed in Table A.1. In the top plot, each comment is depicted with a circle which radius is proportional to the number of sub-comments (i.e., comments to that comment). The bottom plot shows the histogram of the number of comments added to a post with respect to his age. The red line depicts the average number of hours after the post creation needed to reach the 90% of the maximum engagement in terms of number of comments.

with the date/time information. However, Facebook allows the possibility to edit or delete comments [1].

Figure A.1 shows two plots related to the temporal evolution of the 4920 comments considered by our analysis. In both plots the x axis is the number of hours between the comment upload and the post creation, measured in hours. In the upper plot, each comment is depicted as a circle, which radius is proportional to the number of sub-comments (i.e., the engagement generated by that comment). Sub-comments are not considered in the comment count related to the posts. In the bottom plot, the y axis si the number of comments. It shows an histogram, with one bin for each hour. The histogram shows that after 117 hours (i.e., about 5 days) in average, the considered posts reach the 90% of the total engagement (red line in Figure A.1).

[1] During our analysis we found comments dated earlier than the comment they refer to. We noticed this issue to the Facebook platform for developers which classified it as a bug. Such comments have been removed from the analysis here described.

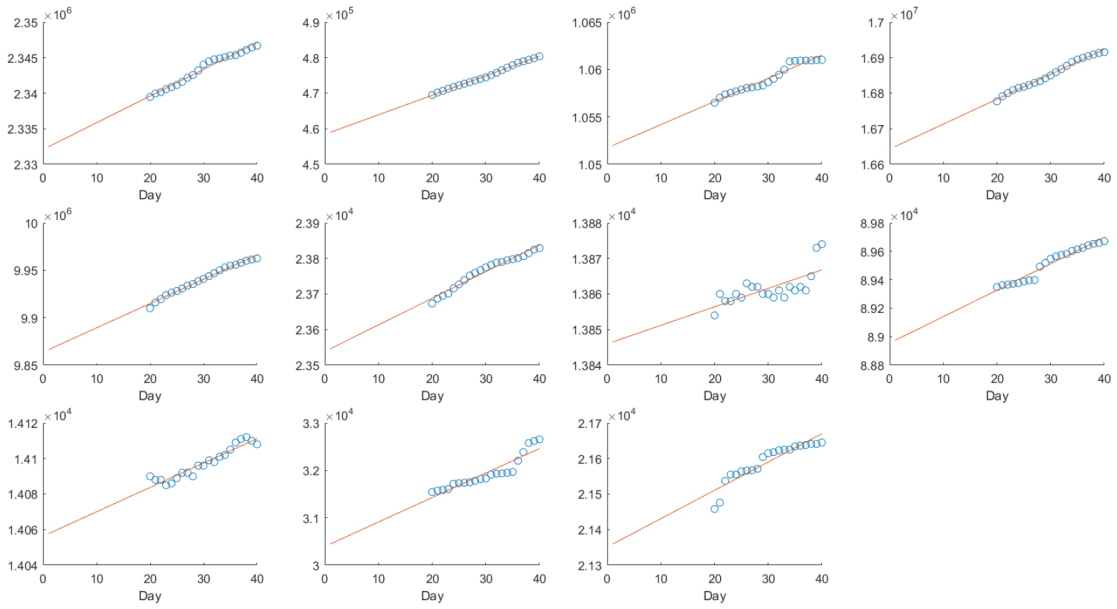


Figure A.2: Number of followers over time (blue circles) and linear fitting of the data (red line). The fitted function is useful to highlight that the points follow a linear law.

A.2.3 Fanbase Analysis

Since the number of followers is often taken into account as a feature or to normalize the popularity score of a social media item [200, 33], in the context of our analysis we analyzed the temporal evolution of the fanbases of the considered profiles. The number of followers (denoted as "friends" in Facebook) changes over time and the value provided by the platform is related to the number of followers at the time of the query. For this reason we downloaded such value for each profile every day for 20 consecutive days. Then, for each profile, a linear interpolation on the 20 related values has been performed. Figure A.2 shows, for each profile, the number of followers downloaded from the platform, depicted by blue circles and arranged starting from day 20 to day 40. The red lines depicts the fitted linear functions. These plots shows how the fanbase evolution follows a linear law, for the majority of the considered profiles. This could help in normalizing popularity scores at past times in which we don't know how many users were in the fanbase of the analyzed profile, or to have an estimate of how many people will be part of the fanbase, hence exploit such estimate to normalize scores in the future.

A.2.4 Web Tool for Downloading Post Information

In the context of the activities described in this Appendix, a web tool for the download of specific post information from Facebook has been developed. This web-based tool aims to download and properly archive the salient information related to a specific post, by providing a textual summary, as well as perform a structured download of the comments. The developed tool takes a Facebook's *post_id*, which identifies a post in the platform and provides the following outputs:

- link to the post;
- picture (if present);
- summary of the salient information related to the post, namely the post_id, textual message, type of post, date/time, number of shares comments and counts of each reaction (i.e., likes, angry, love, haha, wow e sad);
- CSV (Comma Separated Values) file containing all the post comments and sub-comments up to the third level of sub-comments.

Figure A.3 shows an example of output of the tool, with the details of the post with provided *post_id* = 107111802663853_1425282950846725. In particular, the summary shows the engagement statistics, the picture and the URL useful to visually inspect the post. By clicking on the "download comments" button is possible to download the CSV file containing the comments details. Figure A.4 shows the CSV file related to the post with *post_id* = 107111802663853_1425282950846725. For each comment, the column "reply to" specifies the *comment_id* of the parent comment. When this value is empty it means that the row refers to a comment in the first level (i.e., a comment to the post). Note that there is a wide usage of emoji in the comments, since they are very common in social media communication as we detailed in Section 3.5.5.

A.2.5 Post Temporal Analysis

Since the Facebook API provides only the current cumulative values of the number of likes and shares of a post, we downloaded these values related to a number of posts published by the 11 considered profiles. For each user we consider the

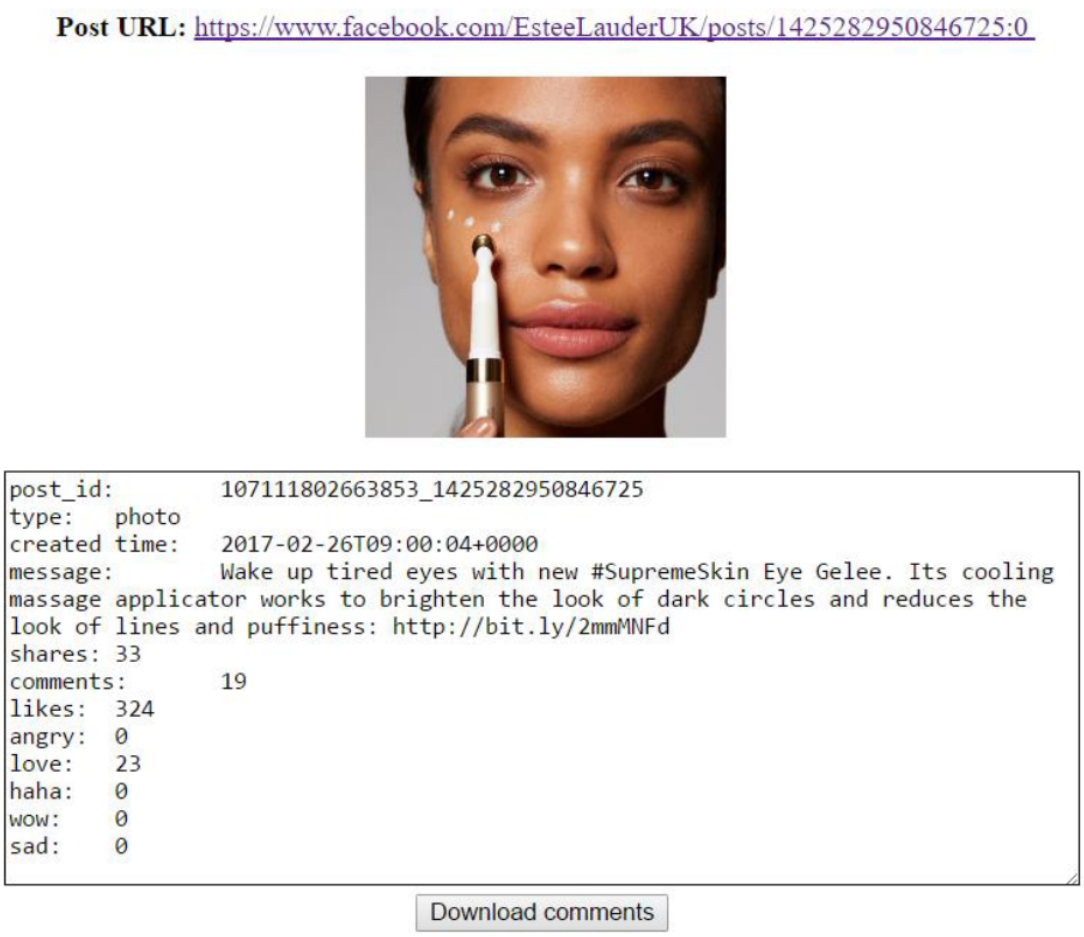


Figure A.3: Summary of the post statistics and content provided by the developed tool.

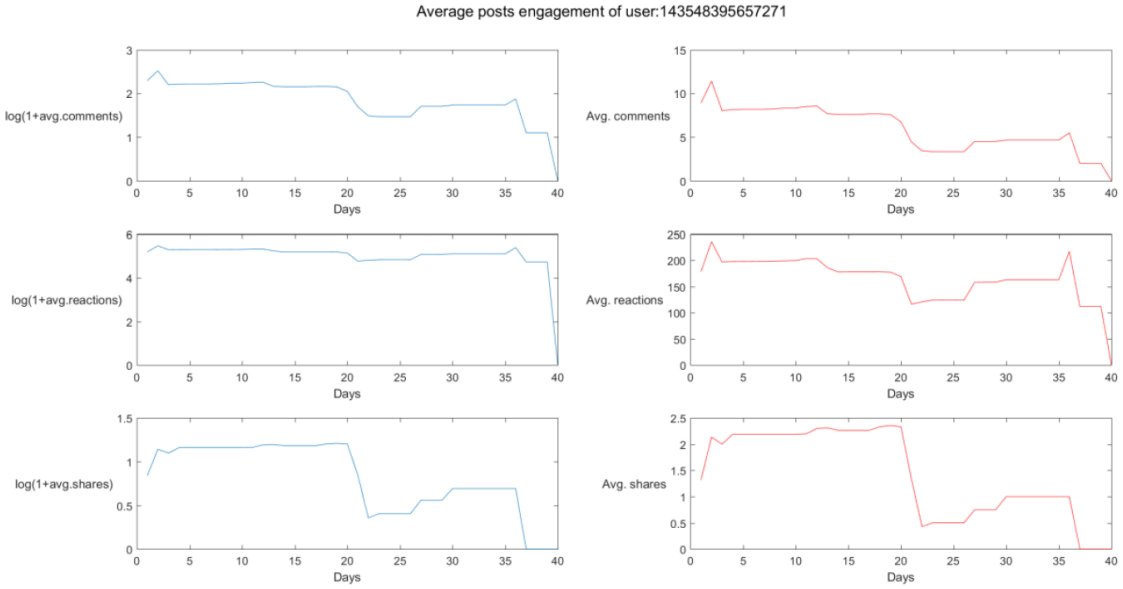
temporal sequence of each score (i.e., comments, likes and shares), by averaging the values of the posts related to the same profile. The following figures show the sequences grouped by users. Specifically, for each figure, the right plots are related to the average score sequence, whereas the left plots show the logarithmic transformation of such score. As we can observe, the score dynamics of two profiles can be very different. However, in general the mean values of comments likes and shares is achieved within the first few days after the post uploading. Then, the scores have a stationary period and sometimes increase or decrease very quickly. This last behavior could be caused by the mechanisms by which Facebook shows the posts on the users' feed pages. Some old posts may be re-posted automatically by the platform or sponsored (i.e., the author payed to propose the post to specific

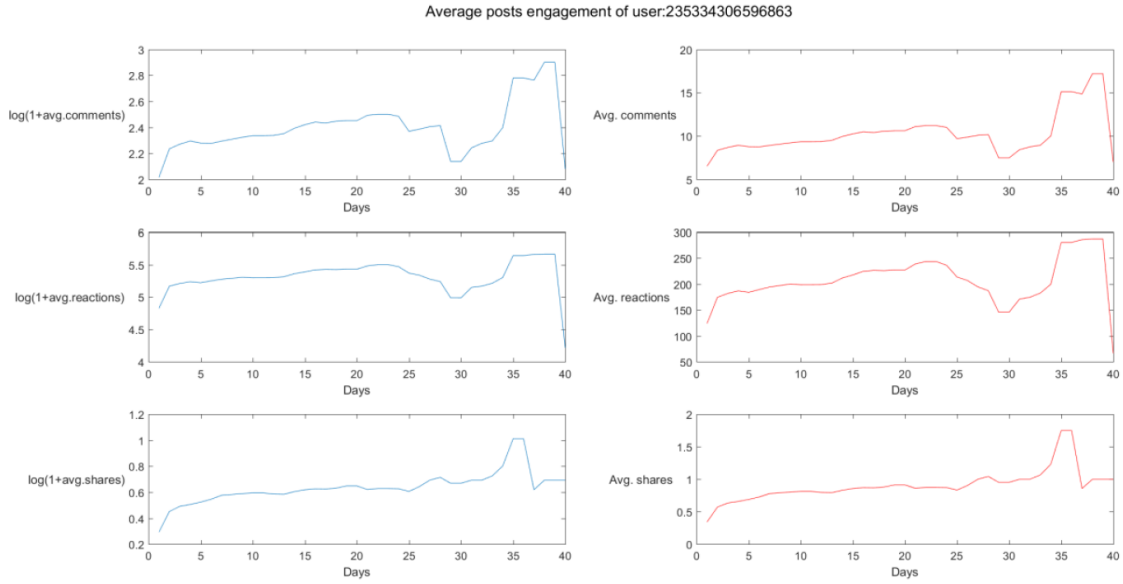
comment_id	reply_to	created_time	comment
1425282950846725_1437338672974486		2017-02-26T13:58:08+0000	I need this in my life! Does it actually work though for the £40 price tag? :/
1425282950846725_1437150199660000		2017-02-26T10:08:44+0000	Have just brought this product this week so look forward to seeing what results it gives
1425282950846725_1437599286281758	1425282950846725_1437150199660000	2017-02-26T18:25:13+0000	Come back & let us know how it was please
1425282950846725_1437108906330796		2017-02-26T09:01:58+0000	It says out of stock?
1425282950846725_1437656592942694		2017-02-26T19:25:38+0000	Does this work?
1425282950846725_1437119569663063		2017-02-26T09:11:17+0000	Ooo I love this and patted on over my makeup also x
1425282950846725_1437713582936995		2017-02-26T20:24:41+0000	We have this in shop Sharon Lawrie. Just came in x
1425282950846725_1437715252936828	1425282950846725_1437713582936995	2017-02-26T20:27:51+0000	'Aww great I''ll find out my days off this week and try and get over x thanks x'
1425282950846725_1437691519605868		2017-02-26T20:06:11+0000	Steven Rowland we both need this x
1425282950846725_1437717276269959	1425282950846725_1437691519605868	2017-02-26T20:31:01+0000	Oh thanks!
1425282950846725_1437717836269903	1425282950846725_1437691519605868	2017-02-26T20:31:56+0000	
1425282950846725_1437426452965708		2017-02-26T15:39:36+0000	LouLou Fontaine one of these my gosh!!!
1425282950846725_1437144736327213		2017-02-26T09:58:42+0000	'Shame it''s out of stock'
1425282950846725_1437334442974909		2017-02-26T13:53:05+0000	Just tried ordering but says out if stock
1425282950846725_1437628962945457		2017-02-26T18:55:45+0000	I need
1425282950846725_1437118392996514		2017-02-26T09:08:32+0000	Chit Chit Win
1425282950846725_1437150242993329	1425282950846725_1437118392996514	2017-02-26T10:08:51+0000	?
1425282950846725_1437807166260970		2017-02-26T22:10:18+0000	Vera Delaney
1425282950846725_1437807686260918	1425282950846725_1437807166260970	2017-02-26T22:11:01+0000	Looks great

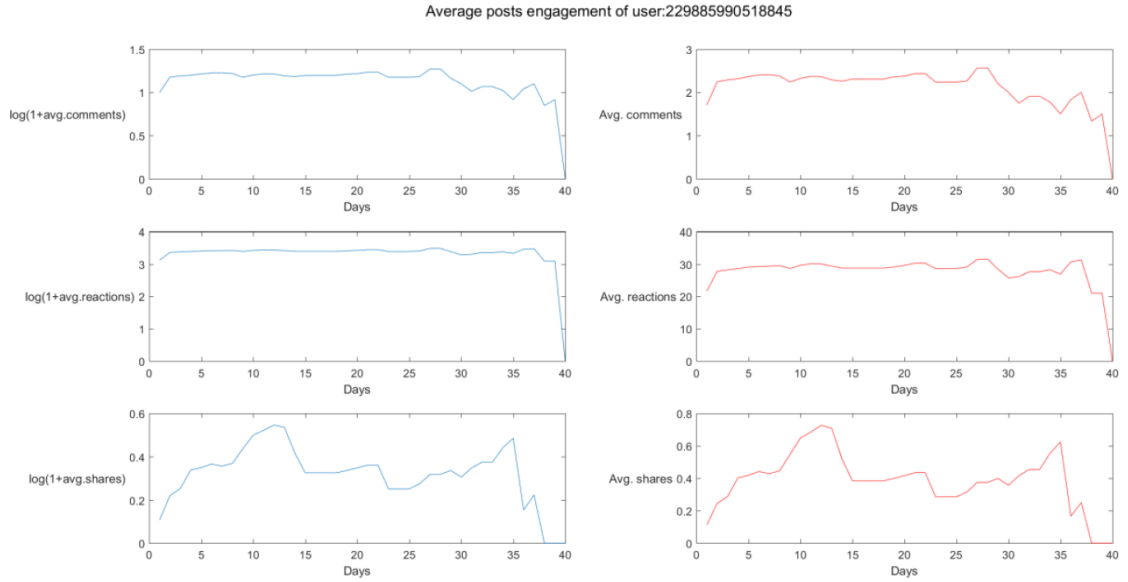
Figure A.4: Structured details of the comments and sub-comments downloaded by the developed tool.

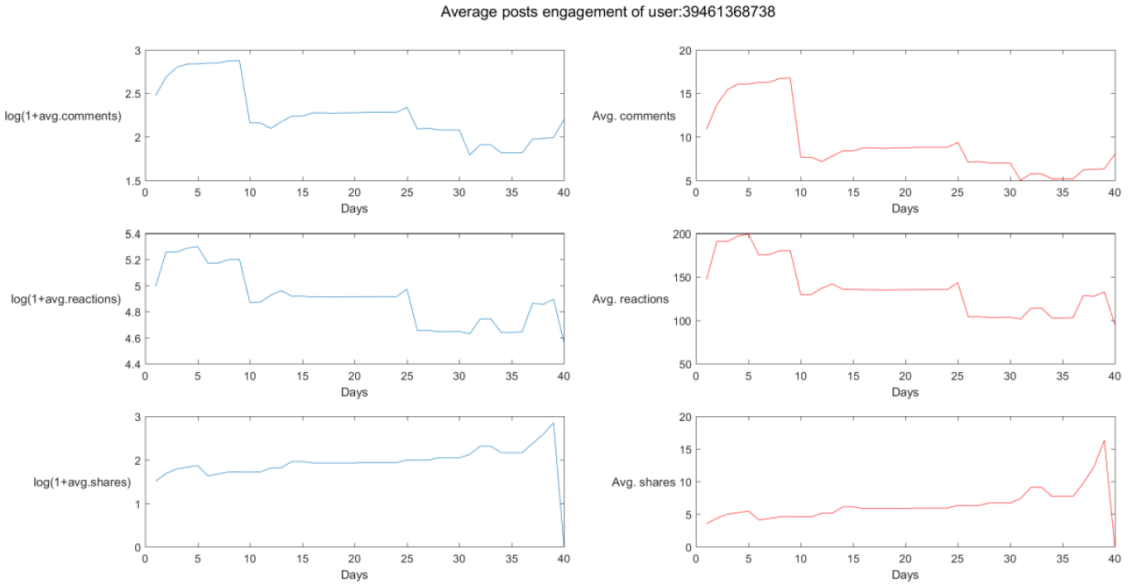
categories of users). Therefore, is not possible to explicitly distinguish between organic reach and paid reach from the engagement score we can retrieve from a post. Another factor is related to the discussions generated in the comments. Indeed, as the number of comments and replies to comments increase, the platform suggest the post to people who participated to the discussion. This of course is aimed to encourage the engagement of the interested users. For all the above reasons, results very difficult to understand and analyze the evolution of the engagement scores related to Facebook posts.

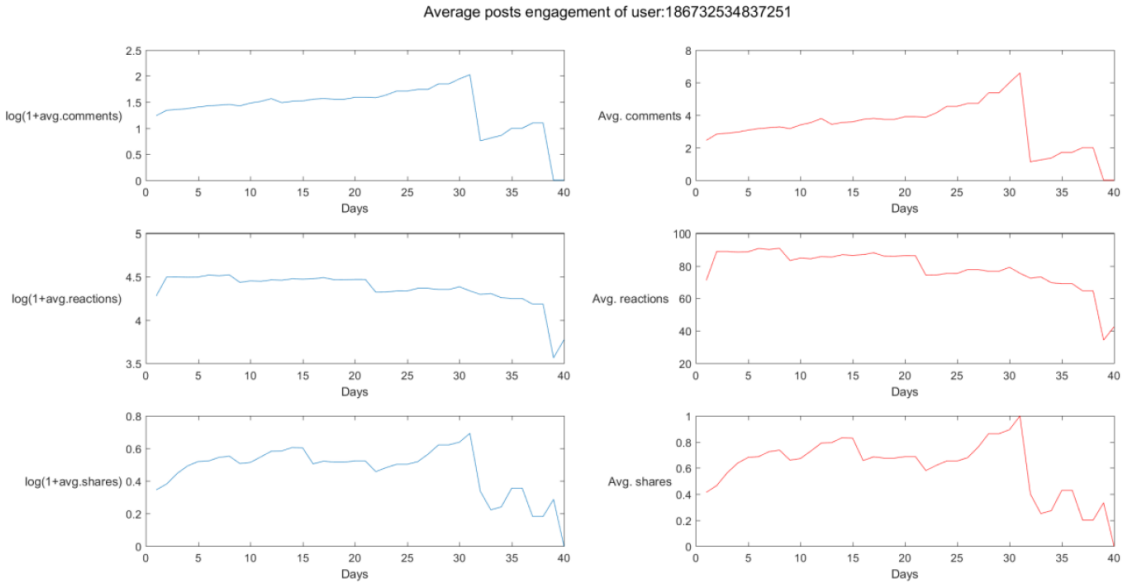
In the following figures, each plot shows a specific score (i.e., number of likes, comments or shares) trend over time for a specific user. The x axis represents the number of days after the post upload. The first row shows the average number of comments of the user's posts (right) and its log-transformed version (left). The second and the third rows show the same statistics related to the number of reactions (i.e., likes, angry, love, haha, wow e sad) and shares respectively.

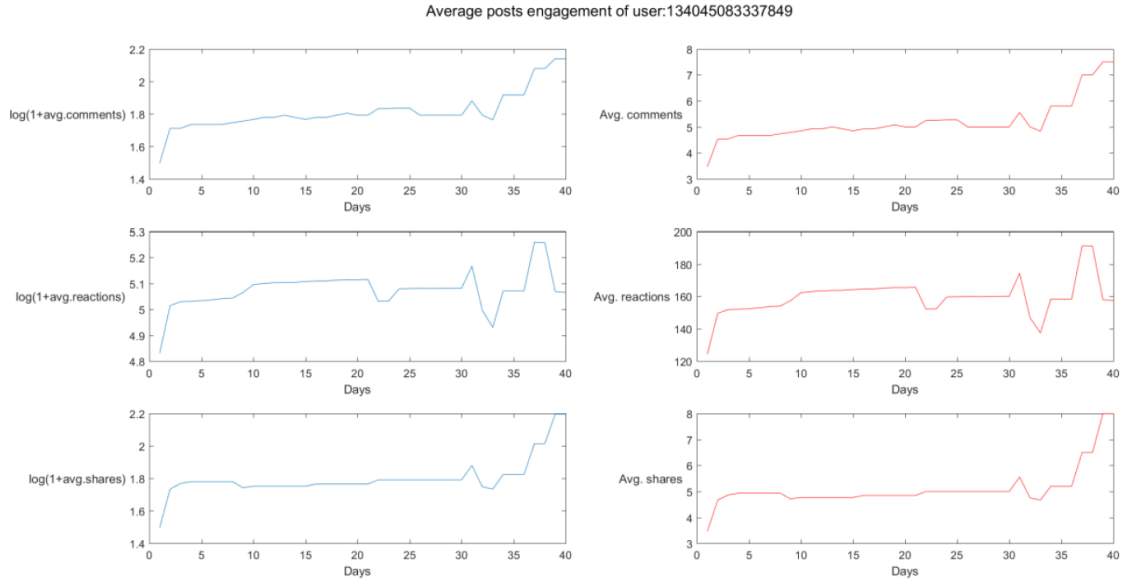

Average posts engagement of user:143548395657271
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days

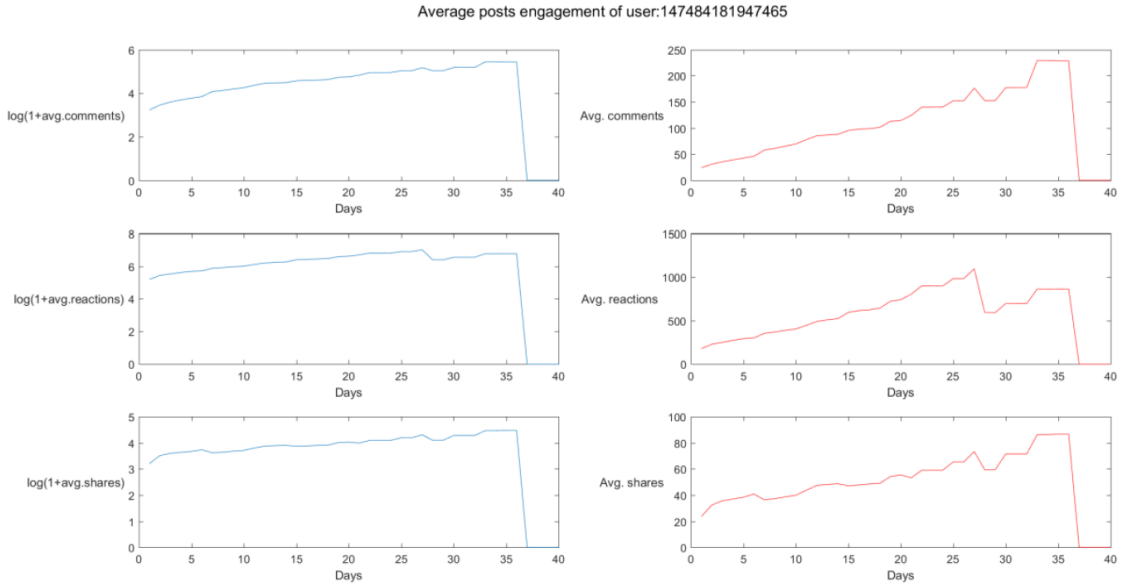

Average posts engagement of user:235334306596863
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days

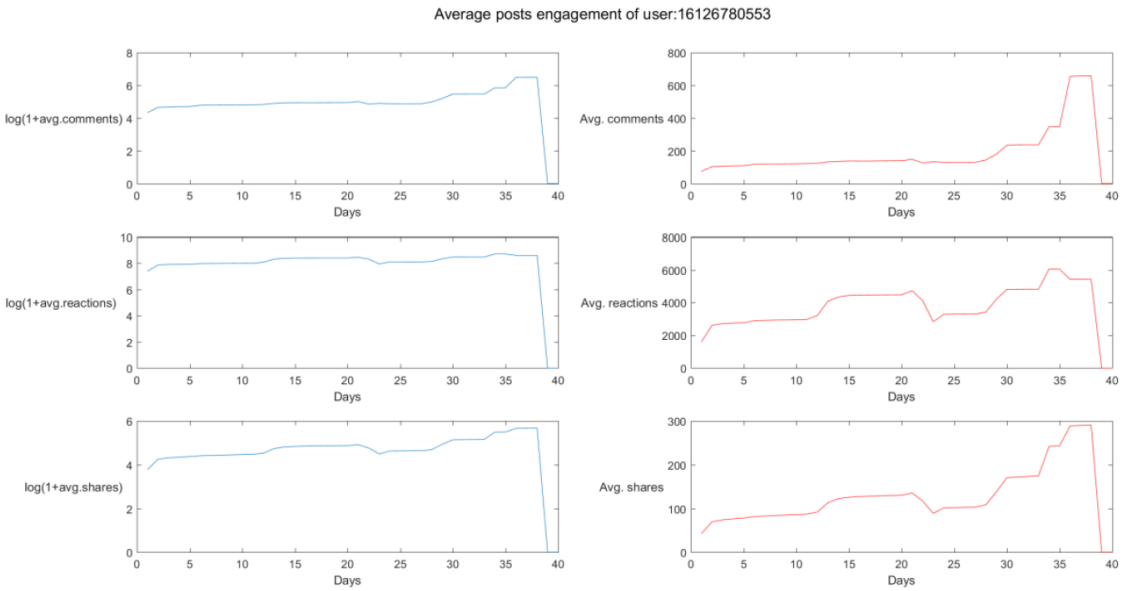

Average posts engagement of user:229885990518845
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days

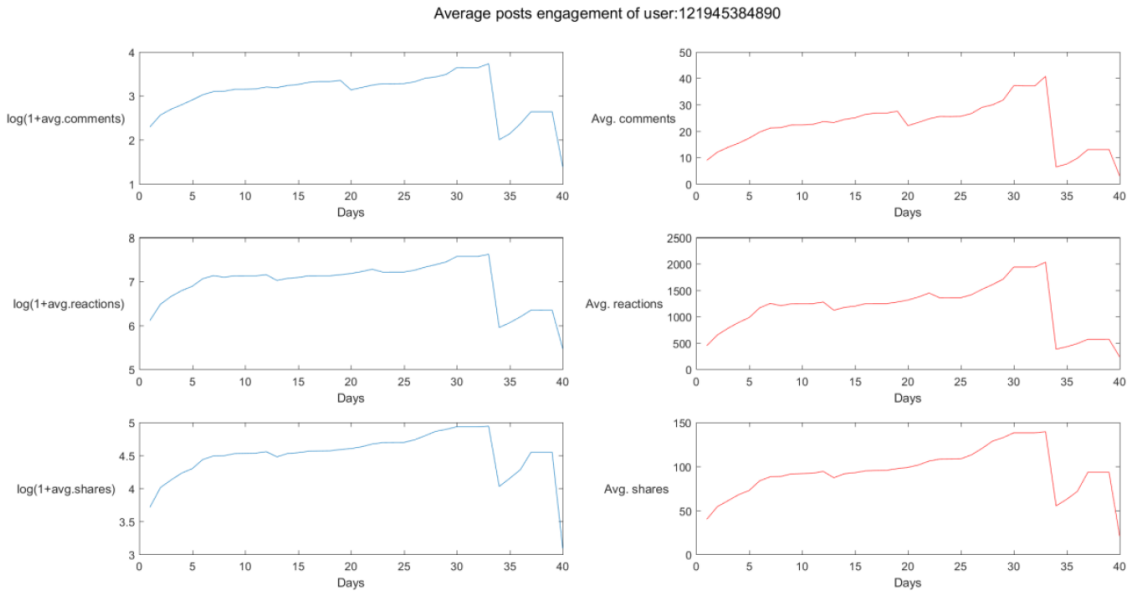

Average posts engagement of user:39461368738
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days

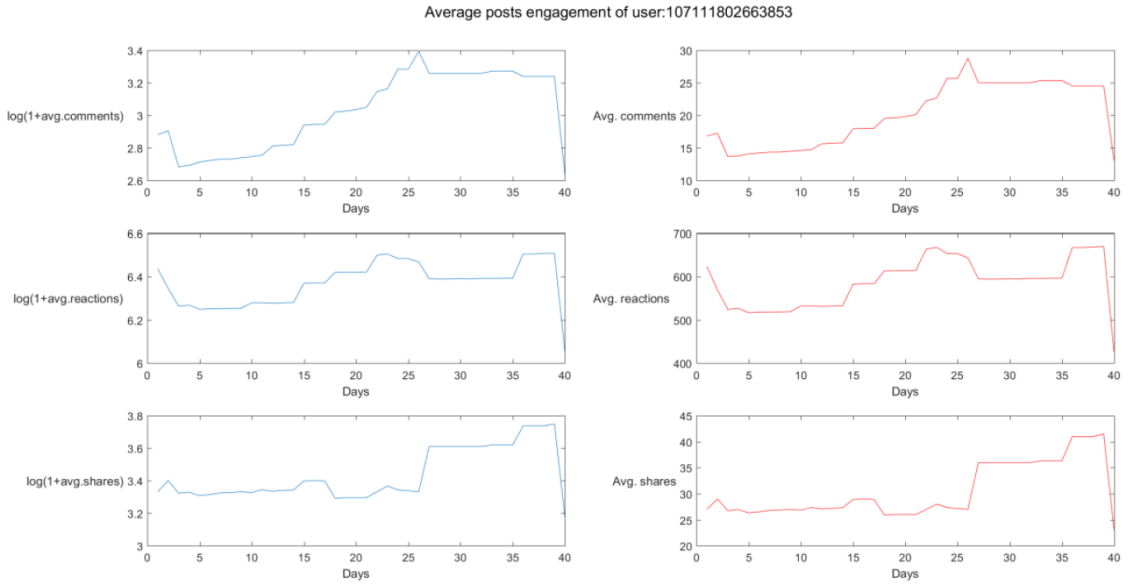

Average posts engagement of user:186732534837251
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days

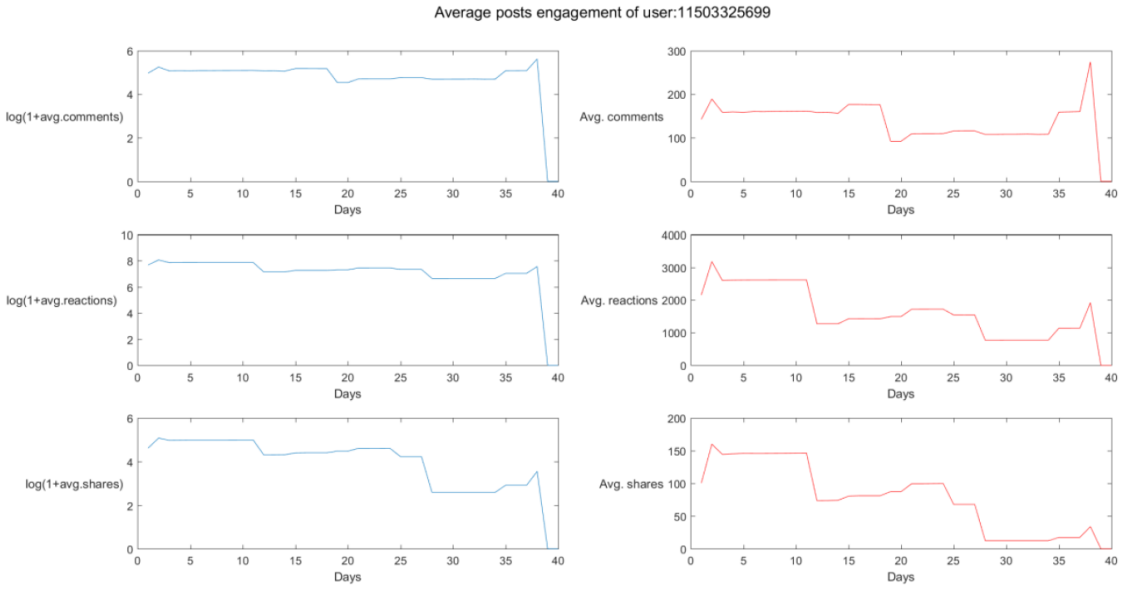

Average posts engagement of user:134045083337849
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days


Average posts engagement of user:147484181947465
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days


Average posts engagement of user:16126780553
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days


Average posts engagement of user:121945384890
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days


Average posts engagement of user:107111802663853
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days


Average posts engagement of user:11503325699
log(1+avg.comments)
Avg. comments
log(1+avg.reactions)
Avg. reactions
log(1+avg.shares)
Avg. shares
Days

Appendix B

The Social Picture Collection Examples

This section presents some additional visual examples related to image collections defined by means of the *The Social Picture* framework. In particular, two example collections taken in the main squares of Catania and Milan are presented.

B.1 Catania - Piazza Duomo

The collection *Catania - Piazza Duomo* has been created by crawling 2355 public photos from Flickr, taken by users in the main square of Catania (Piazza Duomo) Between October 2008 and February 2018. A statue of an elephant with an obelisk on his back is placed in the center of the square, which is surrounded by the facades of four baroque buildings (see Figure B.1). One of them is the city Cathedral. The collected photos are related to several subjects beside the above mentioned buildings. Indeed, as an example, Figure B.2 shows that there are several photos related to the category "food". The photos have been assigned to this category based on the classification given by the CNN trained to distinguish food vs. no-food pictures.

The heatmap shows that the most important building is the Cathedral (Figure B.3). However, by the gathered image is possible to reconstruct the 3D structure of most of the four facades and the elephant statue.

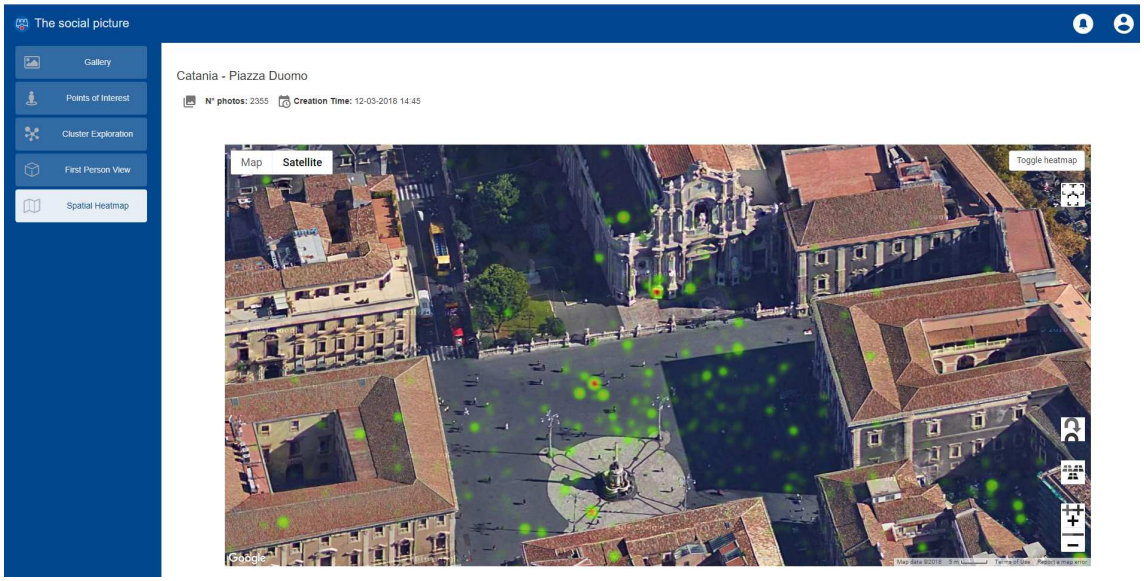


Figure B.1: GPS points related to the locations from which users taken the photos of the collection "*Catania - Piazza Duomo*".

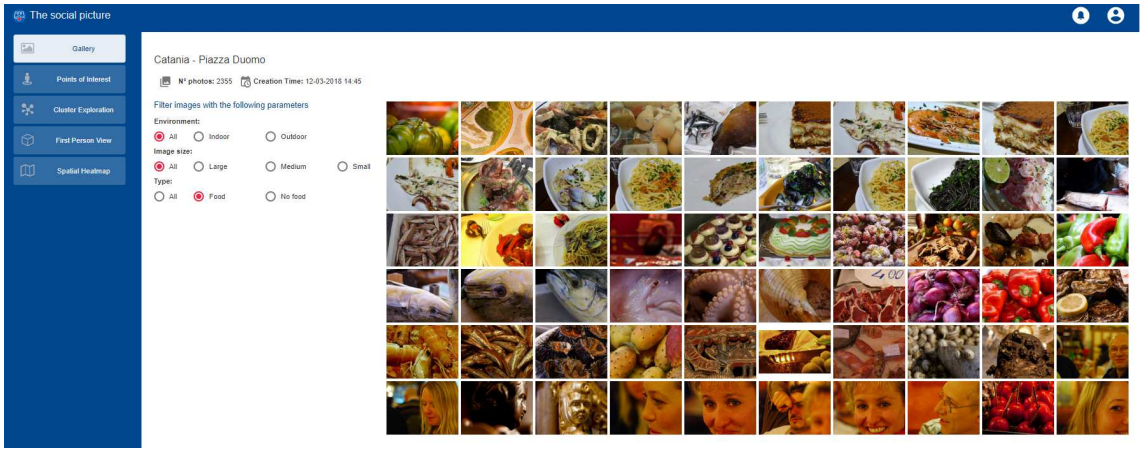


Figure B.2: Subset of images of the 2355 image collection "*Catania - Piazza Duomo*" filtered by selecting the "food" category.

Figure B.3: Heatmap obtained by projecting the photos of the collection "*Duomo di Catania*" on the reference image.

Figure B.4: Example of a user photo, crawled by *The Social Picture*, superimposed on the reference image.

B.2 Milan

The collection *Milan* includes 1836 public photos taken in the main square of the city of Milan and uploaded by users between September 2015 and December 2016.

Despite the low number of pictures, from the heatmap visualization (Figure B.6) is easy to observe that the most interesting building is the Cathedral, even though the presence of several attractive subjects in the same place (see Figure B.5). Indeed, beside the Cathedral, there is a bronze equestrian statue placed opposite to the Cathedral, an arch and the facade of a gallery which is very famous in Milan. Comparing the heatmap of the collection *Milan* with respect to the heatmap of the collection *Pisa* (see Figure 2.5), we can observe that it the latter case there are three main buildings that attract the attention of users (i.e., the Cathedral, the tower and the baptistery).

In the case of Milan, the interest of users toward the Cathedral is definitely higher than the other possible attractive subjects.

Figure B.5: Reference image of a point of interest defined in the collection "*Milan*".

Figure B.6: Heatmap obtained by projecting the photos of the collection "*Milan*" on the reference image.

Appendix C

RECfusion Pseudocode

The Algorithm 3 describes the intraflow analysis process (see Section 4.3.2) exploiting the involved variables and utility functions as reported in the following:

- A template is represented by a `Template Object`. It keeps information about the time of the related image frame, the $fc7$ feature, and the scene ID assigned during the procedure.
- A new `Template Object` is created by the function $newTemplateAt(t)$. This function initializes a `Template Object` with the information related to the time t.
- The flag $SEEK$ is $True$ if the procedure is performing the research of a new stable template. In this case, the `Template Object C` represents the current template candidate, which is assigned to the current template instance T if its stability has been verified according to the conditions defined in [6]. In this case, the new template instance T is added to the Templates Set TS. If the flag $SEEK$ is $False$, the procedure compares the current reference template T with the forward frames until a peak in the slope sequence is detected.
- The function $sceneAssignment()$ performs the scene ID assignment after a new template is defined. In particular, it assigns $Rejected$ to all the frames in the interval between the instant when the new template has been requested to the time when it has been defined. Then, the function finds the scene ID of the scene depicted by the new defined template, eventually performing the backward procedure (as described in Section 4.3.2).

- After the segmentation of the input video, the refinements described in Section 4.3.2 are applied by the function *applySegmentationRefinements()*.

Algorithm 3: Intraflow Analysis Pseudocode

```
Data: Input Video
Result: Segmentation Video
C ← newTemplateAt(0);
SEEK ← True;
t ← 0;
while t ≤ videoLength − step do
    t ← t + step;
    fc7_t ← extractFeatureAt(t);
    if SEEK then
        sim ← cosSimilarity(C.fc7, fc7_t);
        if sim < T_f then
            C ← newTemplateAt(t);
        else
            if C is a stable template then
                SEEK ← False;
                TS ← TS ⋃ {C};
                T ← C;
                T.scene_id ← sceneAssignment();
            end
        end
    else
        sim ← cosSimilarity(T.fc7, fc7_t);
        slope ← slopeAt(t);
        if the slope function has a peak then
            SEEK ← True;
            C.fc7 ← fc7_t;
            T ← C;
        end
        assignIntervalSceneId(T.scene_id, t − step, t − 1);
    end
end
applySegmentationRefinements();
```

The Algorithm 4 describes the between video flows analysis process (see Section 4.3.2). This procedure takes as input a set S of videos, previously processed with the intraflow analysis described in Algorithm 3 . To describe this procedure we defined a data structure called *linkStrength*. This data structure is an array indexed with the blocks extracted from the analysed videos. Considering as example a block segment b_v, the value of $linkStrength[b_v]$ is equal to zero if the block b_v has not yet been assigned to a scene ID. Otherwise, it is equal to the cosine similarity value between b_v and the matched block which caused the assignment. Indeed, all the link strengths are initialized to zero. Then, if the scene ID assigned to b_v changes, the value of $linkStrength[b_v]$ is updated with the new cosine similarity value, as well as the scene ID value assigned to b_v.

Algorithm 4: Between Flows Video Analysis pseudocode.

```
Data: Set S of Segmented Videos
Result: Clusters of the Segments
initialize all link strengths to zero;
for each video v_A in S do
    for each block b_{v_A} in v_A do
        if linkStrength[b_{v_A}] != 0 then
            scene_Id ← b_{v_A}.scene_Id;
        else
            assign a new scene_Id to b_{v_A};
        end
        for each video v_{B_j} in S \ {v_A} do
            b_{v_{B_j}} = arg max_{b̄_{v_{B_j}} ∈ v_{B_j}} {CosSimilarity(b_{v_A}, b̄_{v_{B_j}}) | σ_{(b_{v_A}, v_{B_j})} ≥ T_σ}
            if CosSimilarity(b_{v_A}, b_{v_{B_j}}) ¿ linkStrength[b_{v_{B_j}}] then
                assign scene_Id to b_{v_{B_j}};
                linkStrength[b_{v_{B_j}}] ← CosSimilarity(b_{v_A}, b_{v_{B_j}});
            end
        end
    end
end
```

Appendix D

Other Publications

In the following, it is reported a list of works published during my Ph.D. but not directly related to this thesis.

International Journals:

- Rundo, F., Ortis, A., Battiato, S., & Conoci, S., (2018). Advanced Bio-Inspired System for Noninvasive Cuff-less Blood Pressure Estimation from Physiological Signal Analysis. Computation, 6(3), 46.

- Rundo, Francesco; Conoci, Sabrina; Banna, Giuseppe L; Ortis, Alessandro; Stanco, Filippo; Battiato, Sebastiano: Evaluation of Levenberg-Marquardt Neural Networks and Stacked Autoencoders Clustering for Skin Lesion Analysis, Screening and Follow-up, IET Computer Vision, 2018, DOI: 10.1049/iet-cvi.2018.5195 IET Digital Library, http://digital-library.theiet.org/content/journals/10.1049/iet-cvi.2018.5195.

- Rundo, F., Conoci, S., Ortis, A., & Battiato, S. (2018). An Advanced Bio-Inspired PhotoPlethysmoGraphy (PPG) and ECG Pattern Recognition System for Medical Assessment. Sensors, 18(2), 405.

International Conferences:

- Battiato, S., Cantelli, L., D'Urso, F., Farinella, G. M., Guarnera, L., Guastella, D., Melita, C. D., Muscato, G., Ortis, A., & Santoro, C. (2017, September). A System for Autonomous Landing of a UAV on a Moving Vehicle. In International Conference on Image Analysis and Processing (pp. 129-139). Springer, Cham.

- CANTELLI, L., GUASTELLA, D., MELITA, C., MUSCATO, G., BATTIATO, S., D'URSO, F., FARINELLA, G.M., ORTIS, A., & SANTORO, C. (2017, August). AUTONOMOUS LANDING OF A UAV ON A MOVING VEHICLE FOR THE MBZIRC. In Human-centric Robotics-Proceedings Of The 20th International Conference Clawar 2017 (p. 197). World Scientific.

International Conferences:

- Armano G., Battiato S., Bennato D., Boratto L., Carta S. M., Di Noia T., Di Sciascio E., Ortis A., & Recupero D. R., (2018). NewsVallum: Semantics-Aware Text and Image Processing for Fake News Detection system. Proceedings of the 26th Italian Symposium on Advanced Database Systems (Discussion Paper), Castellaneta Marina (Taranto), Italy, June 24-27, 2018.

Bibliography

[1] H. Chesbrough, W. Vanhaverbeke, and J. West. *Open innovation: Researching a new paradigm*. Oxford University Press on Demand, 2006.

[2] S. Sesia, M. Baker, and I. Toufik. *LTE-the UMTS long term evolution: from theory to practice*. John Wiley & Sons, 2011.

[3] J. M. Dolmaya. "The ethics of crowdsourcing". In: *Linguistica Antverpiensia, New Series–Themes in Translation Studies* 10 (2011).

[4] S. Battiato, G. M. Farinella, F. L. Milotta, A. Ortis, L. Addesso, A. Casella, V. D'Amico, and G. Torrisi. "The Social Picture". In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM. 2016, pp. 397–400.

[5] T. Weyand and B. Leibe. "Visual landmark recognition from Internet photo collections: A large-scale evaluation". In: *Computer Vision and Image Understanding* 135 (2015), pp. 1–15.

[6] A. Ortis, M. Farinella Giovanni, V. D'Amico, L. Addesso, G. Torrisi, and S. Battiato. "RECfusion: Automatic Video Curation Driven by Visual Content Popularity". In: *ACM Multimedia*. 2015.

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

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. "Learning deep features for scene recognition using places database". In: *Advances in neural information processing systems*. 2014, pp. 487–495.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal*

of Computer Vision 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[10] L. Van der Maaten and G. Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.2579-2605 (2008), p. 85.

[11] A. Mikulík, F. Radenović, O. Chum, and J. Matas. "Asian Conference on Computer Vision". In: 2015. Chap. Efficient Image Detail Mining, pp. 118–132.

[12] C. Wu. "Towards linear-time incremental structure from motion". In: *3D Vision-3DV 2013, 2013 International Conference on*. IEEE. 2013, pp. 127–134.

[13] J. Johnson, A. Karpathy, and L. Fei-Fei. "DenseCap: Fully Convolutional Localization Networks for Dense Captioning". In: *arXiv preprint arXiv:1511.07571* (2015).

[14] B. Liu and L. Zhang. "A survey of opinion mining and sentiment analysis". In: *Mining text data*. Springer, 2012, pp. 415–463.

[15] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". In: (2010).

[16] B. Pang and L. Lee. "Opinion mining and sentiment analysis". In: *Foundations and trends in information retrieval* 2.1-2 (2008), pp. 1–135.

[17] C. Colombo, A. Del Bimbo, and P. Pala. "Semantics in visual information retrieval". In: *IEEE Multimedia* 6.3 (1999), pp. 38–53.

[18] S. Schmidt and W. G. Stock. "Collective indexing of emotions in images. A study in emotional information retrieval". In: *Journal of the American Society for Information Science and Technology* 60.5 (2009), pp. 863–876.

[19] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. "Image retrieval by emotional semantics: A study of emotional space and feature extraction". In: *IEEE International Conference on Systems, Man and Cybernetics*. Vol. 4. IEEE. 2006, pp. 3534–3539.

[20] S. Zhao, H. Yao, Y. Yang, and Y. Zhang. "Affective image retrieval via multi-graph learning". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 1025–1028.

[21] M. M. Bradley. "Emotional memory: A dimensional analysis". In: *Emotions: Essays on emotion theory* (1994), pp. 97–134.

[22] J. Itten. *The Art of Color: The Subjective Experience and Objective Rationale of Color*. John Wiley & Sons Inc, 1973. ISBN: 0442240376.

[23] P. J. Lang. "The network model of emotion: Motivational connections". In: *Perspectives on anger and emotion: Advances in social cognition* 6 (1993), pp. 109–133.

[24] C. E. Osgood. "The nature and measurement of meaning." In: *Psychological bulletin* 49.3 (1952), p. 197.

[25] J. A. Russell and A. Mehrabian. "Evidence for a three-factor theory of emotions". In: *Journal of research in Personality* 11.3 (1977), pp. 273–294.

[26] P. Valdez and A. Mehrabian. "Effects of color on emotions." In: *Journal of experimental psychology: General* 123.4 (1994), p. 394.

[27] R. Datta, D. Joshi, J. Li, and J. Z. Wang. "Studying aesthetics in photographic images using a computational approach". In: *European Conference on Computer Vision*. Springer. 2006, pp. 288–301.

[28] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. "Aesthetics and emotions in images". In: *IEEE Signal Processing Magazine* 28.5 (2011), pp. 94–115.

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

[30] F. Raví and S. Battiato. "A Novel Computational Tool for Aesthetic Scoring of Digital Photography". In: *Proceedings of 6th European Conference on Colour in Graphics, Imaging, and Vision*. Amsterdam: SPIE-IS&T, 2012, pp. 1–5. published.

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

[32] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. "Can we understand van gogh's mood?: learning to infer affects from images in social networks". In: *Proceedings of the 20th ACM international conference on Multimedia*. ACM. 2012, pp. 857–860.

[33] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang. "Image popularity prediction in social media using sentiment and context features". In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 907–910.

[34] A. Khosla, A. Das Sarma, and R. Hamid. "What makes an image popular?" In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 867–876.

[35] P. J. McParlane, Y. Moshfeghi, and J. M. Jose. "Nobody comes here anymore, it's too crowded; Predicting Image Popularity on Flickr". In: *Proceedings of International Conference on Multimedia Retrieval*. ACM. 2014, p. 385.

[36] L. C. Totti, F. A. Costa, S. Avila, E. Valle, W. Meira Jr, and V. Almeida. "The impact of visual attributes on online image diffusion". In: *Proceedings of the 2014 ACM conference on Web science*. ACM. 2014, pp. 42–51.

[37] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. "Analyzing and predicting sentiment of images on the social web". In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010, pp. 715–718.

[38] A. Esuli and F. Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining". In: *Proceedings of LREC*. Vol. 6. Citeseer. 2006, pp. 417–422.

[39] J. Machajdik and A. Hanbury. "Affective image classification using features inspired by psychology and art theory". In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010, pp. 83–92.

[40] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. "Emotional category data on images from the International Affective Picture System". In: *Behavior research methods* 37.4 (2005), pp. 626–630.

[41] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. "Large-scale visual sentiment ontology and detectors using adjective noun pairs". In: *Proceedings of the 21st ACM international conference on Multimedia*. ACM. 2013, pp. 223–232.

[42] R. Plutchik. "A general psychoevolutionary theory of emotion". In: *Theories of emotion* 1 (1980), pp. 3–31.

[43] J. Yuan, S. Mcdonough, Q. You, and J. Luo. "Sentribute: image sentiment analysis from a mid-level perspective". In: *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM. 2013, p. 10.

[44] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang. "How Do Your Friends on Social Media Disclose Your Emotions?" In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI'14. Québec City, Québec, Canada: AAAI Press, 2014, pp. 306–312. URL: http://dl.acm.org/citation.cfm?id=2893873.2893922.

[45] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras. "Universals and cultural differences in the judgments of facial expressions of emotion." In: *Journal of personality and social psychology* 53.4 (1987), p. 712.

[46] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks". In: *arXiv preprint arXiv:1410.8586* (2014).

[47] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang. "Predicting viewer affective comments based on image content in social media". In: *Proceedings of International Conference on Multimedia Retrieval*. ACM. 2014, p. 233.

[48] C. Xu, S. Cetintas, K.-C. Lee, and L.-J. Li. "Visual sentiment prediction with deep convolutional neural networks". In: *arXiv preprint arXiv:1411.5731* (2014).

[49] Q. You, J. Luo, H. Jin, and J. Yang. "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI'15. Austin, Texas: AAAI Press, 2015, pp. 381–388. ISBN: 0-262-51129-0. URL: http://dl.acm.org/citation.cfm?id=2887007.2887061.

[50] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. "A mixed bag of emotions: Model, predict, and transfer emotion distributions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 860–868.

[51] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li. "Unsupervised Sentiment Analysis for Social Media Images". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, pp. 2378–2379. ISBN: 978-1-57735-738-4. URL: http://dl.acm.org/citation.cfm?id=2832415.2832579.

[52] V. Campos, A. Salvador, X. Giró-i Nieto, and B. Jou. "Diving Deep into Sentiment: Understanding Fine-tuned CNNs for Visual Sentiment Prediction". In: *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*. ASM '15. Brisbane, Australia: ACM, 2015, pp. 57–62. ISBN: 978-1-4503-3750-2. DOI: 10.1145/2813524.2813530. URL: http://doi.acm.org/10.1145/2813524.2813530.

[53] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. "Visual affect around the world: A large-scale multilingual visual sentiment ontology". In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 159–168.

[54] M. Sun, J. Yang, K. Wang, and H. Shen. "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction". In: *IEEE International Conference on Multimedia and Expo*. IEEE. 2016, pp. 1–6.

[55] M. Katsurai and S. Satoh. "Image sentiment analysis using latent correlations among visual, textual, and sentiment views". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2016, pp. 2837–2841.

[56] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. "Designing category-level attributes for discriminative visual recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 771–778.

[57] J. Yang, M. Sun, and X. Sun. "Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network." In: *AAAI*. 2017, pp. 224–230.

[58] M. S. Jufeng Yang Dongyu She. "Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 3266–3272. DOI: 10.24963/ijcai.2017/456.

[59] V. Campos, B. Jou, and X. G. i Nieto. "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction". In: *Image and Vision Computing* 65 (2017). Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing, pp. 15 –22. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2017.01.011. URL: http://www.sciencedirect.com/science/article/pii/S0262885617300355.

[60] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi. "Cross-Media Learning for Image Sentiment Analysis in the Wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 308–317.

[61] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[62] Z. Li, Y. Fan, W. Liu, and F. Wang. "Image sentiment prediction based on textual descriptions with adjective noun pairs". In: *Multimedia Tools and Applications* 77.1 (2018), pp. 1115–1132.

[63] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. "Sentiment strength detection in short informal text". In: *Journal of the Association for Information Science and Technology* 61.12 (2010), pp. 2544–2558.

[64] T. Hayashi and M. Hagiwara. "Image query by impression words-the IQI system". In: *IEEE Transactions on Consumer Electronics* 44.2 (1998), pp. 347–352.

[65] A. Hanjalic and L.-Q. Xu. "Affective video content representation and modeling". In: *IEEE transactions on multimedia* 7.1 (2005), pp. 143–154.

[66] R. Dietz and A. Lang. "Affective agents: Effects of agent affect on arousal, attention, liking and learning". In: *In proceedings of the Cognitive technology conference*. 1999.

[67] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. "International affective picture system (IAPS): Technical manual and affective ratings". In: *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida* (1999).

[68] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor. "Emotion knowledge: further exploration of a prototype approach." In: *Journal of personality and social psychology* 52.6 (1987), p. 1061.

[69] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek. "Emotional valence categorization using holistic image features". In: *15th IEEE International Conference on Image Processing*. IEEE. 2008, pp. 101–104.

[70] E. S. Dan-Glauser and K. R. Scherer. "The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance". In: *Behavior research methods* 43.2 (2011), pp. 468–477.

[71] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. "Exploring principles-of-art features for image emotion recognition". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 47–56.

[72] J. Stottinger, J. Banova, T. Ponitz, N. Sebe, and A. Hanbury. "Translating journalists' requirements into features for image search". In: *15th International Conference on Virtual Systems and Multimedia*. IEEE. 2009, pp. 149–153.

[73] J. Van de Weijer, C. Schmid, and J. Verbeek. "Learning color names from real-world images". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.

[74] H. Tamura, S. Mori, and T. Yamawaki. "Textural features corresponding to visual perception". In: *IEEE Transactions on Systems, Man, and Cybernetics* 8.6 (1978), pp. 460–473.

[75] B. Liu. "Sentiment analysis and opinion mining". In: *Synthesis lectures on human language technologies* 5.1 (2012), pp. 1–167.

[76] B. Alexe, T. Deselaers, and V. Ferrari. "Measuring the objectness of image windows". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2189–2202.

[77] A. Hanjalic. "Extracting moods from pictures and sounds: Towards truly personalized TV". In: *IEEE Signal Processing Magazine* 23.2 (2006), pp. 90–100.

[78] A. Karpathy and L. Fei-Fei. "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[79] A. Karpathy and L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137.

[80] A. Ortis, G. M. Farinella, V. D'Amico, L. Addesso, G. Torrisi, and S. Battiato. "Organizing egocentric videos of daily living activities". In: *Pattern Recognition* 72.Supplement C (2017), pp. 207 –218. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.07.010. URL: http://iplab.dmi.unict.it/dailylivingactivities.

[81] A. Furnari, S. Battiato, and G. M. Farinella. "Personal-Location-Based Temporal Segmentation of Egocentric Video for Lifelogging Applications". In: *Journal of Visual Communication and Image Representation* 52 (2018), pp. 1–12. ISSN: 1047-3203. URL: http://iplab.dmi.unict.it/PersonalLocationSegmentation/.

[82] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo. "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval". In: *ACM Computing Surveys (CSUR)* 49.1 (2016), p. 14.

[83] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. "A multi-view embedding space for modeling internet images, tags, and their semantics". In: *International journal of computer vision* 106.2 (2014), pp. 210–233.

[84] J. Sang, C. Xu, and J. Liu. "User-aware image tag refinement via ternary semantic analysis". In: *IEEE Transactions on Multimedia* 14.3 (2012), pp. 883–895.

[85] H. Xu, J. Wang, X.-S. Hua, and S. Li. "Tag refinement by regularized LDA". In: *Proceedings of the 17th ACM international conference on Multimedia*. ACM. 2009, pp. 573–576.

[86] L. Wu, R. Jin, and A. K. Jain. "Tag completion for image retrieval". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.3 (2013), pp. 716–727.

[87] W. Wang and Q. He. "A survey on emotional semantic image retrieval". In: *15th IEEE International Conference on Image Processing*. 2008, pp. 117–120. DOI: 10.1109/ICIP.2008.4711705.

[88] D. Lockner, N. Bonnardel, C. Bouchard, and V. Rieuf. "Emotion and interface design". In: *Proceedings of the 2014 Ergonomie et Informatique Avancée Conference-Design, Ergonomie et IHM: quelle articulation pour la co-conception de l'interaction*. ACM. 2014, pp. 33–40.

[89] S. Kazim. *An Introduction to Emotive UI*. Accessed: 2018-04-17. Apr. 2016. URL: https://www.hugeinc.com/articles/an-introduction-to-emotive-ui..

[90] A. G. Reece and C. M. Danforth. "Instagram photos reveal predictive markers of depression". In: *EPJ Data Science* 6.1 (2017), p. 15. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-017-0110-z.

[91] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. "Memorability of image regions". In: *Advances in Neural Information Processing Systems*. 2012, pp. 305–313.

[92] R. Hanna, A. Rohm, and V. L. Crittenden. "We're all connected: The power of the social media ecosystem". In: *Business horizons* 54.3 (2011), pp. 265–273.

[93] A. Goeldi. *Website network and advertisement analysis using analytic measurement of online social media content*. US Patent 7,974,983. 2011.

[94] M. Trusov, R. E. Bucklin, and K. Pauwels. "Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site". In: *Journal of marketing* 73.5 (2009), pp. 90–102.

[95] T. Yamasaki, S. Sano, and K. Aizawa. "Social popularity score: Predicting numbers of views, comments, and favorites of social photos using only annotations". In: *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*. ACM. 2014, pp. 3–8.

[96] S. Cappallo, T. Mensink, and C. G. Snoek. "Latent factors of visual popularity prediction". In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM. 2015, pp. 195–202.

[97] J. Lin and M. Efron. *Overview of the trec2013 microblog track*. Tech. rep. 2013.

[98] M. J. Huiskes, B. Thomee, and M. S. Lew. "New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative". In: *Proceedings of the international conference on Multimedia information retrieval*. ACM. 2010, pp. 527–536.

[99] B. Wu, W.-H. Cheng, Y. Zhang, and T. Mei. "Time Matters: Multi-scale Temporalization of Social Media Popularity". In: *Proceedings of the 2016 ACM on Multimedia Conference (ACM MM)*. Amsterdam, The Netherlands, 2016.

[100] B. Wu, W.-H. Cheng, Y. Zhang, H. Qiushi, L. Jintao, and T. Mei. "Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks". In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Melbourne, Australia, 2017.

[101] B. Wu, T. Mei, W.-H. Cheng, and Y. Zhang. "Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*. Phoenix, Arizona, 2016.

[102] M. Valafar, R. Rejaie, and W. Willinger. "Beyond friendship graphs: a study of user interactions in Flickr". In: *Proceedings of the 2nd ACM workshop on Online social networks*. ACM. 2009, pp. 25–30.

[103] X. Alameda-Pineda, A. Pilzer, D. Xu, N. Sebe, and E. Ricci. "Viraliency: Pooling Local Virality". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 484–492. DOI: 10.1109/CVPR.2017.59.

[104] A. Deza and D. Parikh. "Understanding image virality". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1818–1826. DOI: 10.1109/CVPR.2015.7298791.

[105] D. Parikh and K. Grauman. "Relative attributes". In: *IEEE International Conference on Computer Vision*. IEEE. 2011, pp. 503–510.

[106] H. Altwaijry and S. Belongie. "Relative ranking of facial attractiveness". In: *IEEE Workshop on Applications of Computer Vision (WACV)*. 2013, pp. 117–124.

[107] Q. Fan, P. Gabbur, and S. Pankanti. "Relative attributes for large-scale abandoned object detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2736–2743.

[108] A. Yu and K. Grauman. "Just noticeable differences in visual attributes". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2416–2424.

[109] B. Agarwal, N. Mittal, P. Bansal, and S. Garg. "Sentiment analysis using common-sense and context information". In: *Computational intelligence and neuroscience* (2015).

[110] H. Liu and P. Singh. "ConceptNet — A Practical Commonsense Reasoning Tool-Kit". In: *BT Technology Journal* 22.4 (2004), pp. 211–226. ISSN: 1573-1995. DOI: 10.1023/B:BTTJ.0000047600.45421.6d. URL: https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d.

[111] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. "Show, attend and tell: Neural image caption generation with visual attention". In: *International Conference on Machine Learning*. 2015, pp. 2048–2057.

[112] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. "Image captioning with semantic attention". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4651–4659.

[113] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. "Exploiting emoticons in sentiment analysis". In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM. 2013, pp. 703–710.

[114] T Dimson. *Emojineering part 1: Machine learning for emoji trends*. Accessed: 2018-04-17. 2015.

[115] S. Cappallo, T. Mensink, and C. G. Snoek. "Image2emoji: Zero-shot emoji prediction for visual media". In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 1311–1314.

[116] S. Cappallo, S. Svetlichnaya, P. Garrigues, T. Mensink, and C. G. M. Snoek. "The New Modality: Emoji Challenges in Prediction, Anticipation, and Retrieval". In: *IEEE Transactions on Multimedia* (2018). Pending minor revision.

[117] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič. "Sentiment of Emojis". In: *PLOS ONE* 10.12 (Dec. 2015), pp. 1–22. DOI: 10.1371/journal.pone.0144296. URL: https://doi.org/10.1371/journal.pone.0144296.

[118] S. Rudinac, M. Larson, and A. Hanjalic. "Learning crowdsourced user preferences for visual summarization of image collections". In: *IEEE Transactions on Multimedia* 15.6 (2013), pp. 1231–1243.

[119] S. Qian, T. Zhang, C. Xu, and J. Shao. "Multi-modal event topic model for social event analysis". In: *IEEE Transactions on Multimedia* 18.2 (2016), pp. 233–246.

[120] X. Lei, X. Qian, and G. Zhao. "Rating prediction based on social sentiment from textual reviews". In: *IEEE Transactions on Multimedia* 18.9 (2016), pp. 1910–1921.

[121] Q. You, L. Cao, Y. Cong, X. Zhang, and J. Luo. "A multifaceted approach to social multimedia-based prediction of elections". In: *IEEE Transactions on Multimedia* 17.12 (2015), pp. 2271–2280.

[122] G. A. Miller. "WordNet: a lexical database for English". In: *Communications of the ACM*. Vol. 38. 11. ACM, 1995, pp. 39–41.

[123] Z. Yuan, J. Sang, and C. Xu. "Tag-aware image classification via nested deep belief nets". In: *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE. 2013, pp. 1–6.

[124] S. J. Hwang and K. Grauman. "Accounting for the Relative Importance of Objects in Image Retrieval." In: *Proceedings of British Machine Vision Conference, Vol. 1*. 2. 2010.

[125] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. "A new approach to cross-modal multimedia retrieval". In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM. 2010, pp. 251–260.

[126] S. J. Hwang and K. Grauman. "Learning the relative importance of objects from tagged images for retrieval and cross-modal search". In: *International Journal of Computer Vision* 100.2 (2012), pp. 134–153.

[127] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. "Transductive multi-view embedding for zero-shot recognition and annotation". In: *European Conference on Computer Vision*. Springer. 2014, pp. 584–599.

[128] L. Pang, S. Zhu, and C.-W. Ngo. "Deep multimodal learning for affective analysis and retrieval". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 2008–2020.

[129] P. Cui, S. Liu, and W. Zhu. "General Knowledge Embedded Image Representation Learning". In: *IEEE Transactions on Multimedia* (2017).

[130] Z. Yuan, J. Sang, C. Xu, and Y. Liu. "A unified framework of latent feature learning in social media". In: *IEEE Transactions on Multimedia* 16.6 (2014), pp. 1624–1635.

[131] X. Yang, T. Zhang, and C. Xu. "Cross-domain feature learning in multimedia". In: *IEEE Transactions on Multimedia* 17.1 (2015), pp. 64–78.

[132] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[133] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. "Improving image-sentence embeddings using large weakly annotated photo collections". In: *European Conference on Computer Vision*. Springer. 2014, pp. 529–545.

[134] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods". In: *Neural computation* 16.12 (2004), pp. 2639–2664.

[135] T. Mike, B. Kevan, P. Georgios, C. Di, and K. Arvid. "Sentiment in short strength detection informal text". In: *Journal of the Association for Information Science and Technology* 61.12 (2010), pp. 2544–2558.

[136] J. Johnson, L. Ballan, and L. Fei-Fei. "Love thy neighbors: Image annotation by exploiting image metadata". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4624–4632.

[137] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo. "A multimodal feature learning approach for sentiment analysis of social network multimedia". In: *Multimedia Tools and Applications* 75.5 (2016), pp. 2507–2525.

[138] Z. S. Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

[139] G. Wang, D. Hoiem, and D. Forsyth. "Building text features for object image classification". In: *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1367–1374.

[140] A. Rahimi, B. Recht, et al. "Random Features for Large-Scale Kernel Machines." In: *In proceedings of the Neural Information Processing Systems*. Vol. 3. 4. 2007, p. 5.

[141] F. Perronnin, J. Sénchez, and Y. L. Xerox. "Large-scale image categorization with explicit data embedding". In: *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2010, pp. 2297–2304.

[142] M. Guillaumin, J. Verbeek, and C. Schmid. "Multimodal semi-supervised learning for image classification". In: *IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society. 2010, pp. 902–909.

[143] C. Hung and H.-K. Lin. "Using objective words in SentiWordNet to improve sentiment classification for word of mouth". In: *IEEE Intelligent Systems* 28.2 (2013), pp. 47–54.

[144] A. Ortis, G. M. Farinella, V. D'amico, L. Addesso, G. Torrisi, and S. Battiato. "RECfusion: Automatic video curation driven by visual content popularity". In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 1179–1182.

[145] S. Aloufi, S. Zhu, and A. El Saddik. "On the Prediction of Flickr Image Popularity by Analyzing Heterogeneous Social Sensory Data". In: *Sensors* 17.3 (2017), p. 631.

[146] K. Almgren, J. Lee, et al. "Predicting the future popularity of images on social networks". In: *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*. ACM. 2016, p. 15.

[147] B. Wu, W.-H. Cheng, Y. Zhang, and T. Mei. "Time matters: Multi-scale temporalization of social media popularity". In: *Proceedings of the 2016 ACM on Multimedia Conference*. ACM. 2016, pp. 1336–1344.

[148] B. Wu, T. Mei, W.-H. Cheng, Y. Zhang, et al. "Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition." In: *AAAI*. 2016, pp. 272–278.

[149] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei. "Sequential prediction of social media popularity with deep temporal context networks". In: *arXiv preprint arXiv:1712.04443* (2017).

[150] L. Li, R. Situ, J. Gao, Z. Yang, and W. Liu. "A Hybrid Model Combining Convolutional Neural Network with XGBoost for Predicting Social Media Popularity". In: *Proceedings of the 2017 ACM on Multimedia Conference*. MM '17. Mountain View, California, USA: ACM, 2017, pp. 1912–1917. ISBN: 978-1-4503-4906-2. DOI: 10.1145/3123266.3127902. URL: http://doi.acm.org/10.1145/3123266.3127902.

[151] S. C. Hidayati, Y.-L. Chen, C.-L. Yang, and K.-L. Hua. "Popularity Meter: An Influence-and Aesthetics-aware Social Media Popularity Predictor". In: *Proceedings of the 2017 ACM on Multimedia Conference*. ACM. 2017, pp. 1918–1923.

[152] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. "A peek into the future: predicting the evolution of popularity in user generated content". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM. 2013, pp. 607–616.

[153] K. Lerman and T. Hogg. "Using a model of social dynamics to predict popularity of news". In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 621–630.

[154] H. Pinto, J. M. Almeida, and M. A. Gonçalves. "Using early view patterns to predict the popularity of youtube videos". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM. 2013, pp. 365–374.

[155] G. Szabo and B. A. Huberman. "Predicting the popularity of online content". In: *Communications of the ACM* 53.8 (2010), pp. 80–88.

[156] R. Bandari, S. Asur, and B. A. Huberman. "The pulse of news in social media: Forecasting popularity." In: *ICWSM* 12 (2012), pp. 26–33.

[157] T. Wilson, J. Wiebe, and P. Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis". In: *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics. 2005, pp. 347–354.

[158] M. Katsurai, T. Ogawa, and M. Haseyama. "A cross-modal approach for extracting semantic relationships between concepts using tagged images". In: *IEEE Transactions on Multimedia* 16.4 (2014), pp. 1059–1074.

[159] Z. Li, J. Liu, J. Tang, and H. Lu. "Robust structured subspace learning for data representation". In: *IEEE transactions on pattern analysis and machine intelligence* 37.10 (2015), pp. 2085–2098.

[160] G. Finlayson, S. Hordley, G. Schaefer, and G. Y. Tian. "Illuminant and device invariant colour using histogram equalisation". In: *Pattern recognition* 38.2 (2005), pp. 179–190.

[161] G. Finlayson and G. Schaefer. "Colour indexing across devices and viewing conditions". In: *International Workshop on Content-Based Multimedia Indexing*. 2001.

[162] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. "Automatic editing of footage from multiple social cameras". In: *ACM Transactions on Graphics* 33.4 (2014), p. 81.

[163] H. S. Park, E. Jain, and Y. Sheikh. "3d social saliency from head-mounted cameras". In: *Advances in Neural Information Processing Systems*. 2012, pp. 431–439.

[164] Y. Hoshen, G. Ben-Artzi, and S. Peleg. "Wisdom of the Crowd in Egocentric Video Curation". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, pp. 587–593.

[165] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. "MoViMash: online mobile video mashup". In: *ACM International Conference on Multimedia*. 2012, pp. 139–148.

[166] D. G. Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[167] G. Farinella, D Ravì, V Tomaselli, M Guarnera, and S Battiato. "Representing scenes for real-time context classification on mobile devices". In: *Pattern Recognition* 48.4 (2015), pp. 1086–1100.

[168] J. Domke and Y. Aloimonos. "Deformation and Viewpoint Invariant Color Histograms." In: *British Machine Vision Conference*. 2006, pp. 509–518.

[169] F. L. M. Milotta, S. Battiato, F. Stanco, V. D'Amico, G. Torrisi, and L. Addesso. "RECfusion: Automatic Scene Clustering and Tracking in Video from Multiple Sources". In: *EI – Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications*. 2016.

[170] L. Ballan, G. Brostow, J. Puwein, and M. Pollefeys. "Unstructured Video-Based Rendering: Interactive Exploration of Casually Captured Videos." In: *ACM Transactions on Graphics*. 2010, pp. 1–11.

[171] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas. "You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video." In: *British Machine Vision Conference*. 2014.

[172] A. Furnari, G. M. Farinella, and S. Battiato. "Recognizing Personal Contexts from Egocentric Images". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops - Assistive Computer Vision and Robotics*. 2015, pp. 393–401.

[173] T. Kanade. "Quality of life technology". In: *Proceedings of the IEEE* 100.8 (2012), pp. 2394–2396.

[174] A. Furnari, G. M. Farinella, and S. Battiato. "Recognizing personal locations from egocentric videos". In: *IEEE Transactions on Human-Machine Systems* 47.1 (2017), pp. 6–18.

[175] M. L. Lee and A. K. Dey. "Lifelogging memory appliance for people with episodic memory impairment". In: *The 10th ACM International Conference on Ubiquitous Computing*. 2008, pp. 44–53.

[176] V. Buso, L. Hopper, J. Benois-Pineau, P.-M. Plans, and R. Megret. "Recognition of Activities of Daily Living in natural "at home" scenario for assessment of Alzheimer's disease patients". In: *IEEE International Conference on Multimedia & Expo Workshops*. IEEE. 2015, pp. 1–6.

[177] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, and J.-F. Dartigues. "Hierarchical Hidden Markov Model in detecting activities of daily living in wearable videos for studies of dementia". In: *Multimedia Tools and Applications* 69.3 (2014), pp. 743–771.

[178] Y. Gaëstel, S. Karaman, R. Megret, O.-F. Cherifa, T. Francoise, B.-P. Jenny, and J.-F. Dartigues. "Autonomy at home and early diagnosis in Alzheimer's Disease: Utility of video indexing applied to clinical issues, the IMMED project". In: *Alzheimer's Association International Conference on Alzheimer's Disease 2011*, S245.

[179] J. Pinquier, S. Karaman, L. Letoupin, P. Guyot, R. Mégret, J. Benois-Pineau, Y. Gaëstel, and J.-F. Dartigues. "Strategies for multiple feature fusion with Hierarchical HMM: application to activity recognition from wearable audio-visual sensors". In: *21st International Conference on Pattern Recognition*. IEEE. 2012, pp. 3192–3195.

[180] N. Kapur, E. L. Glisky, and B. A. Wilson. "External memory aids and computers in memory rehabilitation". In: *The essential handbook of memory disorders for clinicians* (2004), pp. 301–321.

[181] C. Gurrin, A. F. Smeaton, and A. R. Doherty. "Lifelogging: Personal big data". In: *Foundations and trends in information retrieval* 8.1 (2014), pp. 1–125.

[182] M. D. White. "Police officer body-worn cameras: Assessing the evidence". In: *Washington, DC: Office of Community Oriented Policing Services*. 2014.

[183] A. Fathi, A. Farhadi, and J. M. Rehg. "Understanding egocentric activities". In: *IEEE International Conference on Computer Vision*. 2011, pp. 407–414.

[184] H. Pirsiavash and D. Ramanan. "Detecting activities of daily living in first-person camera views". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2847–2854.

[185] M. S. Ryoo and L. Matthies. "First-person activity recognition: What are they doing to me?" In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 2730–2737.

[186] Y. Poleg, C. Arora, and S. Peleg. "Temporal Segmentation of Egocentric Videos". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[187] A. Furnari, G. M. Farinella, and S. Battiato. "Temporal Segmentation of Egocentric Videos to Highlight Personal Locations of Interest". In: *International Workshop on Egocentric Perception, Interacion, and Computing - European Conference on Computer Vision Workshop*. 2016.

[188] Z. Lu and K. Grauman. "Story-Driven Summarization for Egocentric Video". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[189] B. Soran, A. Farhadi, and L. Shapiro. "Generating Notifications for Missing Actions: Don't Forget to Turn the Lights Off!" In: *International Conference on Computer Vision*. 2015.

[190] L. Deng. "A tutorial survey of architectures, algorithms, and applications for deep learning". In: *APSIPA Transactions on Signal and Information Processing* 3 (2014), e2.

[191] P. Agrawal, R. Girshick, and J. Malik. "Analyzing the performance of multilayer neural networks for object recognition". In: *European Conference on Computer Vision*. 2014, pp. 329–344.

[192] S. Bell, P. Upchurch, N. Snavely, and K. Bala. "Material Recognition in the Wild With the Materials in Context Database". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[193] J. Hosang, M. Omran, R. Benenson, and B. Schiele. "Taking a Deeper Look at Pedestrians". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[194] J. Long, E. Shelhamer, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *The IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[195] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How transferable are features in deep neural networks?" In: *Advances in Neural Information Processing Systems*. 2014, pp. 3320–3328.

[196] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: vol. arXiv:1409.1556. 2014.

[197] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*. 1. Cambridge University Press Cambridge, 2008.

[198] G. M. Farinella and S. Battiato. "Scene classification in compressed and constrained domain". In: *IET Computer Vision* 5.5 (2011), pp. 320–334.

[199] A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella. "Next-active-object prediction from egocentric videos". In: *Journal of Visual Communication and Image Representation* 49 (2017), pp. 401–411.

[200] I. P. Cvijikj and F. Michahelles. "Online engagement factors on Facebook brand pages". In: *Social Network Analysis and Mining* 3.4 (2013), pp. 843–861.