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

"*Quote.*"

Abstract

Acknowledgements

Acknowledgements go here.

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 universities campus. 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 outline 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 different 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 systems: 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 parts: Crowdsourced Media Analysis, Image Sentiment Analysis and Video Sentiment Analysis. The dissertation structure is shown in Figure 1.1. In Chapter 2 we start our discussion by an introduction

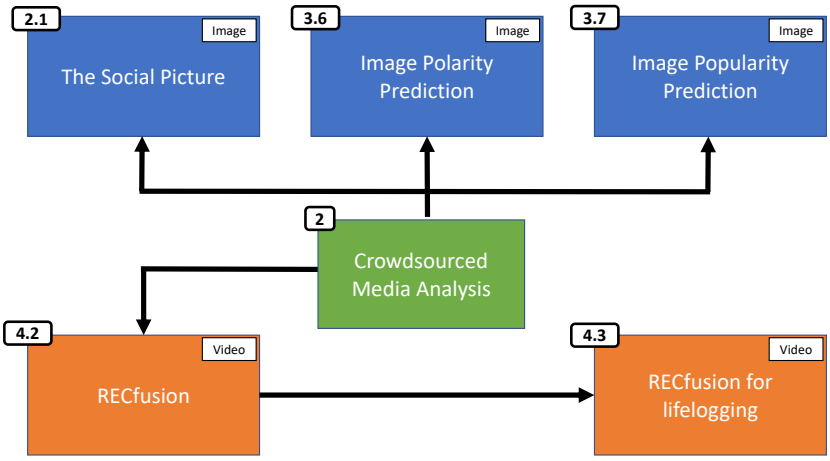

Figure 1.1: Dissertation structure. Numbers depicts the Sections which describe the proposed work. Blue blocks represent algorithms that work on images, whereas orange blocks represent algorithms that work on videos. All the media involved in the works presented in this dissertation are produced by users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

on Crowdsourced Media Analysis, which 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 fo comments, etc. Given an images posted on a social media, the aim of popularity prediction system is to predict a popularity score, which is based on a function of one engagement values (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 List of Publications

In this dissertation we present 5 real use-cases and related papers published in journals and conferences.

- 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:
  - CBMI
  - ??
  - Survey??
- Image Popularity Prediction
  - Popularity??
- 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.

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. Moreover, 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. 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 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 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 live customizable criteria. *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. Nowadays the diffusion of social networks plays a crucial role in collecting information about people opinion and trends. The proliferation of mobile devices and the diffusion of social media have changed the communication paradigm of people that share multimedia data by allowing new interaction models (e.g., social networks). In social events (e.g., concerts), the audience typically produces and share a lot of multimedia data with mobile devices (e.g., images, videos, geolocation, tags, etc.) related to what has captured their interest. The redundancy in these data can be exploited to infer social information about the attitude of the attending people. For example, systems such as RECfusion [5] (detailed in Section 4.2) can be developed to understand if there are groups of people interested to specific scenes. 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 [6], 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 kind 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

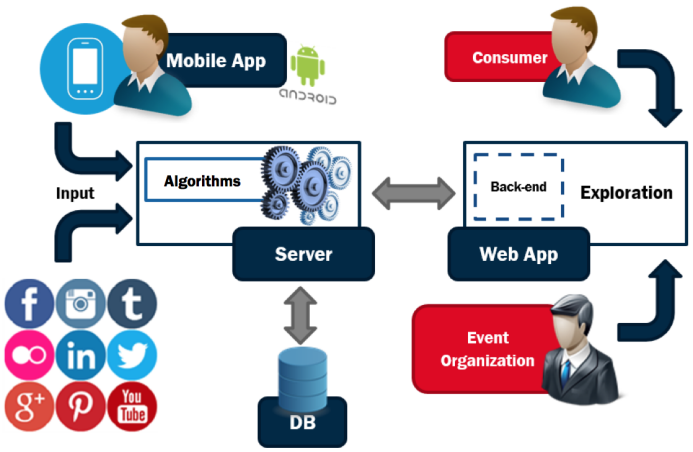

Figure 2.1: The Social Picture's architecture.

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 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 Computer Vision 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

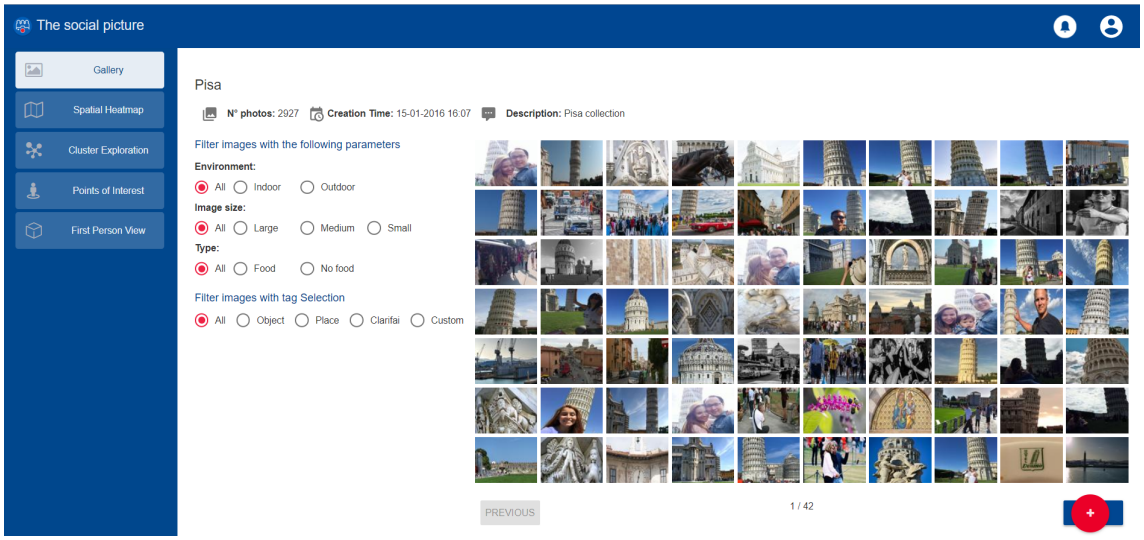

Figure 2.2: Example of exploration interface. It is composed by three main areas: the gallery (right part) shows the collection's pictures according to the selected filters (middle) which allows the users to explore the collection. By selecting an image the system shows all the extracted information 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.

of pictures to infer social behavioural information about the event considering the popularity of the uploaded scenes [5].

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 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

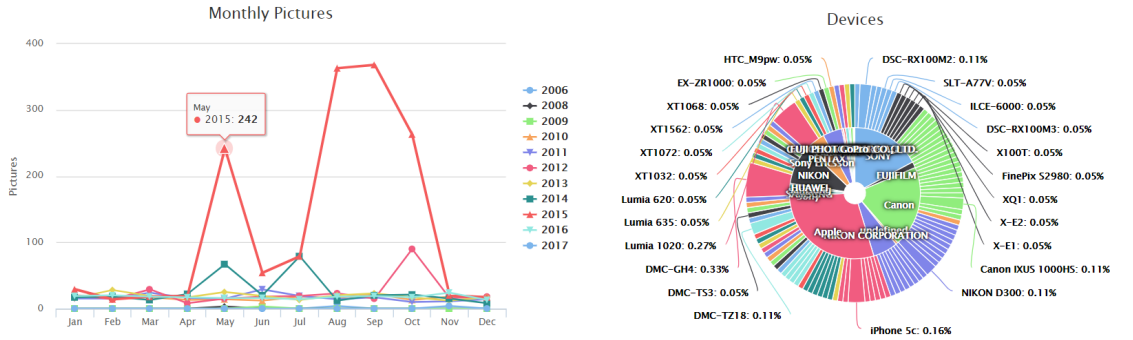

Figure 2.3: Some statistics computed by the framework. The left plot shows the amount of pictures monthly uploaded for a set of selected years. The right plot shows the distribution of the devices used to take the pictures included in the connection.

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). 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.

The visual analysis module of the system feds all the images to two different CNNs [8, 9], 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* [8] and *Places205-AlexNet* [9]. The CNN used in [8] consist of seven internal layers

with a final 1000-way softmax which produces a distribution over the 1000 predefined classes of the ImageNet dataset [10]. We considered the feature activations induced at the last hidden layer, which consists of 4096 dimensional feature (fc-7 feature), as an image representation to be further used with t-SNE algorithm [11] for visualization purposes. We also fed the images to the *Places205-AlexNet* CNN [9]. 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.

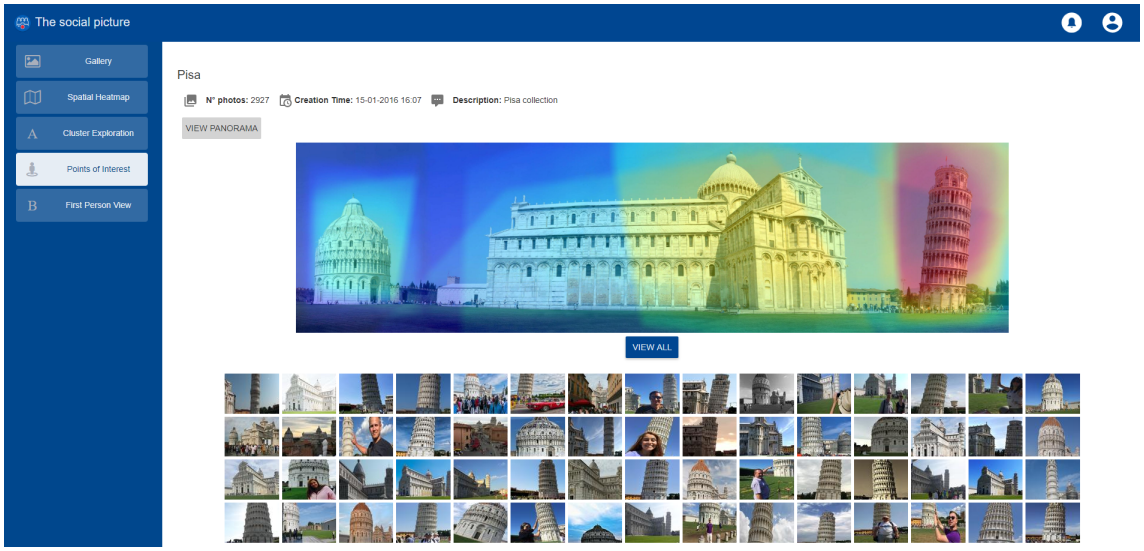

Figure 2.4: heatmap exploration tool. By selecting a point in the heatmap, the system visualizes all the photos that contributed to that point in the bottom part of the interface.

Heamap 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 2.4. 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 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 [12].

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

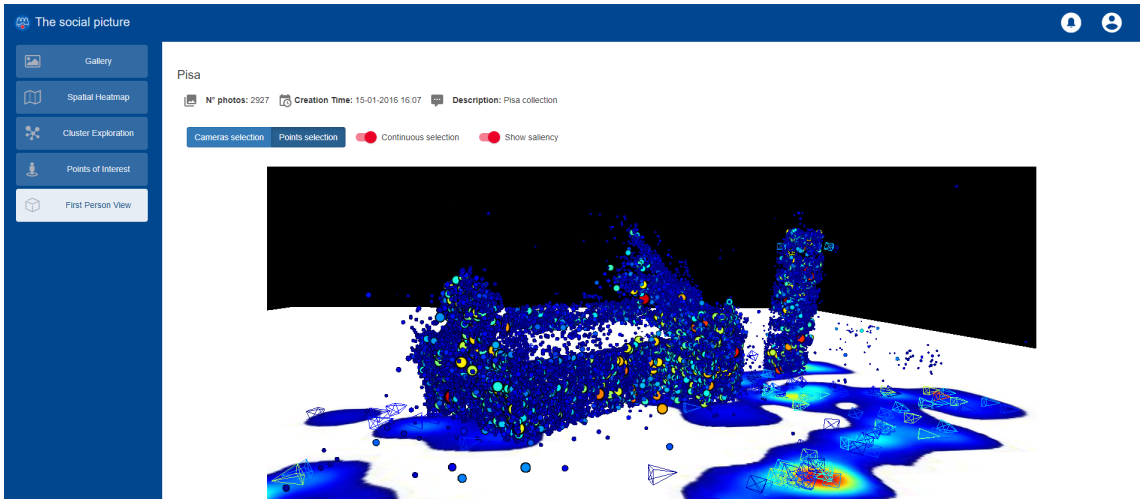

Figure 2.5: 3D sparse reconstruction of a cultural heritage site based on the photos of the collecction. Each point in the 3D space is estimated by considering the projections of the images depicting that part of the scene.

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) [13], we are able to compute a 3D sparse reconstruction of large photos collections. The 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.

Figure 2.6: 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.

Embedding Exploration

We exploit the fc7 feature extracted with the *AlexNet* architecture [8] for each image and use the t-SNE embedding algorithm [11] 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 2.6 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 2.6). 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 [8] has been trained using 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.

Figure 2.7: photos embedding based on the t-SNE coordinates.

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 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

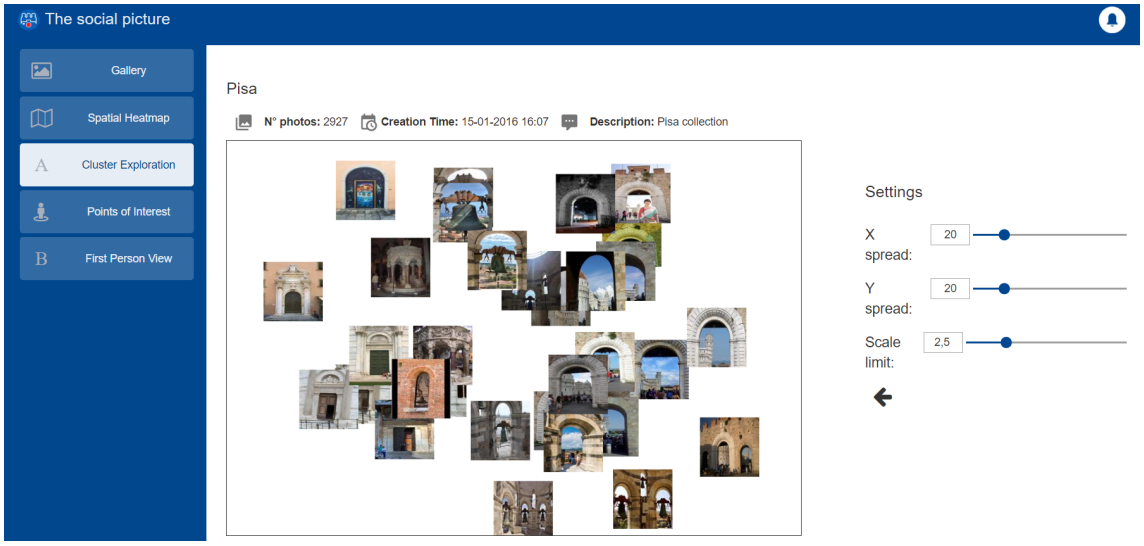

Figure 2.8: visualization interface for t-SNE based image embedding. In this example, a subset of images is shown. All the images are related to photos depicting "archs", however the group if images has been automatically created without taking into account the semantic of the images.

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 can be continued by selecting one of the shown pictures and computing the t-SNE embedding for its sub-elements in the hierarchy. Figure 2.8 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 (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 [14] 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 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.,

Figure 2.9: 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 point of views. By providing the positions of the users during the event it gives some hints about the most interesting parts of the site.

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 [15]. 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 [16]. Historically, Sentiment Analysis techniques have been developed for the analysis of text [17], 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. Section 3.2 reviews relevant publications and present a complete view of 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 discussed 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, the evaluation from the viewpoint of progress toward more sophisticated systems and related practical applications, as well as a summary of the insights resulting from this study. 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. *Objective Text* is obtained from the visual content of the images through recent deep learning architectures which are used to classify object, scene and to perform image captioning. *Objective Text* features are then combined

with visual features in an embedding space obtained with Canonical Correlation Analysis. The sentiment polarity is then inferred by a supervised Support Vector Machine. 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.

Then, in Section 3.7, we present a work which 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.

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 [18, 19, 20, 21], 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 [22, 23, 24, 25, 26, 27]. Other research fields close to Visual Sentiment Analysis are those considering the analysis of the image aesthetic [28, 29, 30, 31], interestingness [32], affect [33] and popularity [34, 35, 36, 37].

The first paper on Visual Sentiment Analysis aims to classify images as "positive" or "negative" and dates back on 2010 [38]. 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 [39] 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 [40] 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 the 8 emotional output categories as defined in [41] (i.e., Awe, Anger, Amusement, Contentment, Excitement, Disgust, Sad, and Fear).

In [42] 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 [43], 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. 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 end, 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 [40], obtaining better results.

In 2013, Yuan et al. [44] 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. [45] 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 [46].

Chen et al. [47] 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 [42]). This approach significantly improved the ANP detection with respect to [42]. Similarly to [42], 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 [48] tries to predict what concepts will be evoked

to the image viewers.

In [49] 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 [42], 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 [49] outperform the baseline methods in visual sentiment prediction.

The authors of [50] 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 [42]. 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 [43]. An implementation of the architecture proposed in [50] 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 [42] and [44].

Considering that the emotional response of a person viewing an image may include multiple emotions, the authors of [51] aimed to predict a distribution representation of the emotions rather than a single dominant emotion from (see Figure 3.1). The authors compared three methods to predict such emotion distributions: a Suppor Vector Regressor (based on hand crafted features related to edge, color, texture,

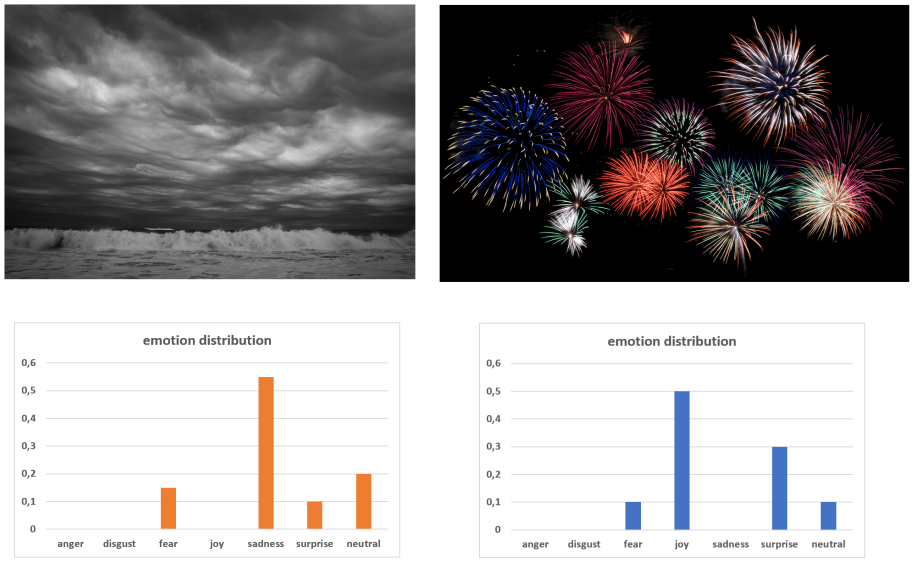

Figure 3.1: Examples of image emotion distributions.

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 [46] 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 [52] 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 [33], mid-level features [42, 44] and Deep Learning architectures [50, 51], the framework in [52] implements an unsupervised approach (USEA - Unsupervised SEntiment Analysis). In [53] 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 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. [54] extended the ANP ontology defined in [42] 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. [55] 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 [42] and [50] on three widely used datasets presented in [42] and [50].

In [56] 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 [57] and [42] 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 [39], a well-known lexical resource used in opinion mining to assign sentiment scores to words.

The works in [58] and in [59] perform emotional image classification of images considering multiple emotional labels. As previously proposed in 2015 by Peng [51], 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 [58] 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 [59] 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 [59] 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 [43]. The experimental results show that the approach proposed in [59] outperforms eleven baseline Label Distribution Learning (LDL) methods, including BCPNN and ACPNN proposed in [58].

In [60] the authors extended their previous work [53] 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 experimental results and observations. Then, the authors propose an improved CNN architecture based on the empirical insights.

The authors of [61] 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 [62] and HybridNet [9]). The proposed system has been compared with the CNN and PCNN presented in [50], DeepSentiBank [47] and MVSO [54] obtaining better results on the built dataset.

The system proposed in [63] 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 [42] as mid-level representations. In particular, the sentiment value of an extracted ANP is defined by summing the sentiment scores defined in SentiWordNet [39] and SentiStrength [64] 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 [50]. 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 [50] exhibit better performances than the late fusion method.

Lastly, is worth mentioning a survey recently presented by Soleymani et al. [65] which provides a review of the latest works in the research community about three major fields:

- multimodal spoken reviews and vlogs (audio-visual context);
- sentiment analysis in the interactions between humans and machines (face to face interactions);
- sentiment analysis of images and tags shared in social media (photo sharing platforms).

The survey presents a brief overview of the methods adopted to address the three different tasks.

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. 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.

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) [42, 38, 50]. However, there are also systems that adopt more then 3 levels, such as the 5-level sentiment scheme used by Xu et al. [49] or the 35 "impression words" used by Hayashi et al. [66]. 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 [40]. Thus, these systems can be evaluated also for the task of polarity estimation. Generally, there are two main approaches for emotion modelling:

- ***Dimensional approach:*** this model represents emotions as points in a 2 or 3 dimensional space. Indeed, as discussed in several studies [22, 24, 25, 26], emotions have three basic underlying dimensions: valence, arousal and control (or dominance). However, 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 valence axis. Indeed, for example, Hanjalic et al. [67] considered the VA (Valence-Arousal) space to model the affective video content.

- ***Category approach:*** this model defines a set of descriptive words that are assigned to regions of the VAC (Valence-Arousal-Control) space. Thus it can be considered a quantized version of the dimensional approach.

Considering the category 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 in psychology can be beneficial for this problem.

What are the basic emotions? There are several works that aim to ask this question. As we observed in Section **??**, the most adopted model is the Plutchnik's Wheel of Emotions [43]. This model defines 8 basic emotions with 3 valences each (see Figure 3.2). Thus it defines a total of 24 emotions:

- "ecstasy" → "joy" → "serenity"

- "admiration" → "trust" → "acceptance"

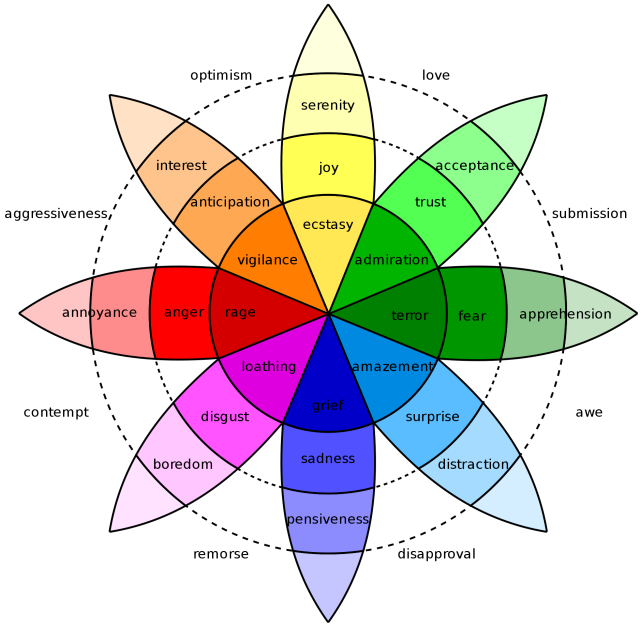

Figure 3.2: Plutchik's Wheel of Emotions.

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

According to Ekman's theory [46] 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. [41]. In this work the authors perform an intensive study on the *International Affective Picture System* (IAPS) in order to extract a categorical structure of such dataset [68]. 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. [69], where emotion concepts are organized

in a hierarchical structure.

As we 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 [43] is a well established psychological model of emotions. This model 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) [68]. This 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. This dataset has been used in [40] in combination with other two datasets built by the authors, which are publicly available[1].

In [70] the authors considered a subset of IAPS extended with subject annotations to obtain a training set categorized in distinct emotions according to the

[1] www.imageemotion.org

emotional model described in [41] (see Section 3.3.1). However, the number of images of this dataset is very low.

In [40], 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 [41].

The authors of the dataset presented in [38] considered the top 1.000 positive and negative words in SentiWordNet [39] 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) [71] 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. The dataset is available for research purposes [3].

In 2013 Borth et al. [42] 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 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). This 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 [50] and in [44] for instance, due to the large scale of this dataset.

The Emotion6 dataset, presented and used in [51], has been built considering the Elkman's 6 basic emotion categories [46]. 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 [50] 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,

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

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

Table 3.1: 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 [68]	716 photos	Pleasure, arousal and dominance	✗	✗	✗
2005	Mikels et al. [41]	369 photos	Awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
2010	Affective Image Classification Dataset [40]	228 paintings 807 photos	Awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
2010	Flickr-sentiment [38]	586.000 Flickr photos	Positive, negative.	✓	✓	✓
2011	GAPED [71]	730 pictures	Positive, negative, neutral.	✗	✓	✗
2013	VSO [42]	0,5 M Flickr Photos 603 Twitter Images	- Adjective-Noun Pairs - Positive or negative	✓	✓	✓
2015	Emotion6 [51]	1.980 Flickr photos	- Valence-Arousal score - 7 emotions distribution	✓	✗	✗
2015	You et al. [50]	1.269 Twitter images	Positive, negative.	✓	✓	✗
2016	CrossSentiment [56]	90.139 Flickr photos 65.439 Instagram images	Positive, negative, neutral.	✓	✓	✗
2017	T4SA [61]	1,5 M Twitter images	Positive, negative, neutral.	✓	✓	✗

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 [42].

In [56] the authors crawled two large sets of social pictures from Instagram and Flickr (CrossSentiment). The list of labelled Instagram and Flickr image URLs is available on the Web [4].

Vadicamo et al. [61] 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

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

obtain impressive results as long as these systems are trained with very large scale datasets (e.g., VSO [42]). 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 scale datasets. This furthered the building of very large datasets such as T4SA in the last few years. Table 3.1 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 for 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 of semantics:

- ***Low-level features*** **-** 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 [40].

- ***Mid-level features*** **-** This group of features bring more semantic, thus they are more interpretable and have stronger associations with emotions [72]. One example is given by the scene-based 102-dimensional feature defined in [44]. 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. [42].

- ***High-level features*** **-** 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 [56].

In 2010 Machajdik and Hanbury [40] 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 [23, 27], or exploited in image retrieval [73] and image classification [28, 20] tasks. They selected 17 visual features, categorized in 4 groups:

- **color:** mean saturation and brightness, 3-dimensional emotion representation by Valdez et al. [27], hue statistics, colorfulness measure according to [28], number of pixels of each of the 11 basic colors [74], Itten contrast [23], color histogram designed by Wang Wei-ning et al. [20];
- **texture:** wavelet textures for each HSB channel, features by Tamura et al. [75], 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 [40]), depth of field (DOF) [28], 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 in Visual Sentiment Analysis 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 [47, 49, 50]. By employing such representation methods image features are learned from the data. This avoid the designing of a proper feature for the task of Visual Sentiment Analysis, 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 [56]. 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 [76], 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 [77];
- define a set of image aspects, in terms of low level features, each one associated to a sentiment polarity based on previous studies [40, 70]. 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 [76], an opinion (or sentiment) is a quintuple

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

where $\boldsymbol{e_i}$ is the name of an entity, $\boldsymbol{a_{ij}}$ is an aspect related to $\boldsymbol{e_i}$, $\boldsymbol{s_{ijkh}}$ is the sentiment score with respect to the aspect $\boldsymbol{a_{ij}}$, $\boldsymbol{h_k}$ is the opinion holder, and $\boldsymbol{t_l}$ is the time when the opinion is expressed by the opinion holder $\boldsymbol{h_k}$. The sentiment score $\boldsymbol{s_{ijkh}}$ 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 $\boldsymbol{e_i}$ and the aspect $\boldsymbol{a_{ij}}$ represent the opinion target.

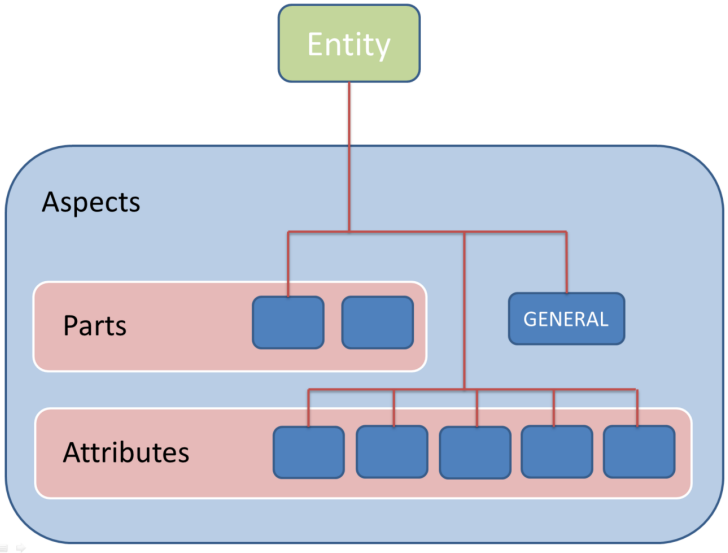

Figure 3.3: Relationship between an entity and its aspects.

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 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" [78].

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

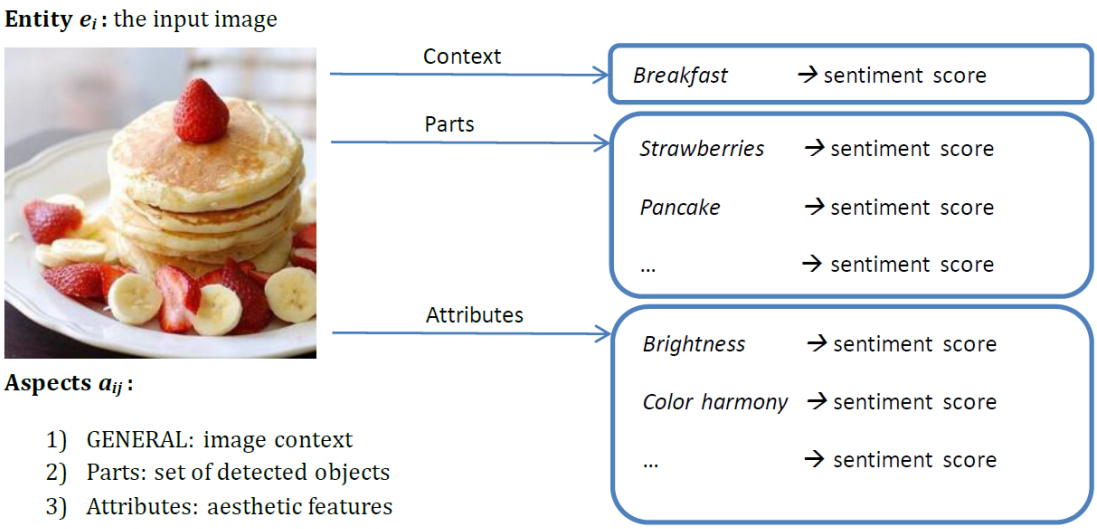

Figure 3.4: Example of the different scores that can be extracted from an image. The entity e_i is the input image, the context (i.e., breakfast), the parts (i.e., objects in the scene) and the attributes (i.e., aesthetic features) represent the different aspects. Each aspect a_{ij} is associated to a sentiment score s_{ijkh}. The sentiment scores can be expressed in terms of polarity (e.g., positive or negative) or bu considering different levels of strength (e.g., a score from 1 to 5).

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 3.3). 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 [79, 80]. 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 achieved for each aspect. Finally, the scores are combined to obtain the sentiment classification (e.g., data can be used as input features of a regression model). As an example, the Figure 3.4 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 [81, 82] and scene recognition can 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 different 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 sentiment analysis of 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. 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. For a deeper analysis, a comprehensive treatise of image tag assignment is presented in [83].

As discussed in [84], the semantic of an image can be expressed by means of an

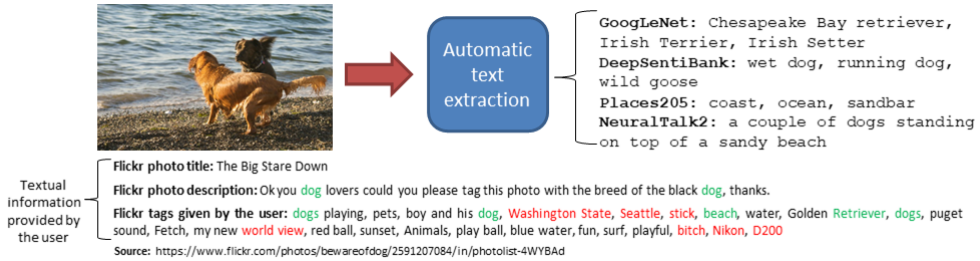

Figure 3.5: Given an image, the text describing the visual content can be extracted by exploiting four different deep learning architectures. The considered architectures are used to extract text related to objects, scene and image description. The figure shows also the text associated to the image by 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.

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 3.5 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 3.5 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. For each considered scene understanding method (i.e., GoogLeNet [85], DeepSentiBank [47] and Places205 [9]). 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. Several papers faced the issues related to the subjective text associated to images, such as tag refinement and completion [83, 86, 87, 88], 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 [89]. 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 [48] 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 [90] 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 [91], 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 discussed 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 over time. For example, the sentiment evoked by an image depicting the World Trade Center (Figure 3.6) 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

Figure 3.6: Picture of the World Trade Center, taken in 1990.

over time focused on specific cognitive and psychology applications. As an example, the work in [92] 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 [93] 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 [40].

3.5 Challenges

So far we discussed on the current state of the art in Visual Sentient 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.).

The popularity of an image is a difficult quantity to define, hence to measure or infer. However human beings are able to predict what visual contents other people will like in specific contexts (e.g., marketing campaigns, professional photography). This suggest 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 quality of the text associated to images is often pre-processed in order to avoid noisy text. On the other hand, the users who want 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, so the use of popular tags highs 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. [94] 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.

The work in [36] 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. The experimental results show that the 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.

In 2014 Khosla et al. [35] proposed a log-normalized popularity score that has been then commonly used in the community. Let $\boldsymbol{c_i}$ 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 $\boldsymbol{i^{th}}$ item is defined as follows:

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

where $\boldsymbol{T_i}$ is the number of days since the uploading of the image on the Social Platform. Equation 3.2 normalizes the number of interactions reached by an image by dividing the engagement measure by the time. However, the measures $\boldsymbol{c_i}$ 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. There are very few works which takes into account the evolution of the image popularity over time. For example, the study presented in [95] 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.

In [35] the authors analysed the importance of several image and social cues 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, average groups' members, etc.), image context features (e.g., title length, description length, number of tags), as well as image features (e.g., GIST, LBP, BoW color patches, CNN activations features, etc.). Is interesting to notice that, differently than several works on sentiment polarity prediction in which the text concerning the images (i.e., title, description and tags included in the post) is semantically analysed in order to achieve sentiment related insights on the image content, in this work only the length of the text associated to the images is considered.

Cappallo et al. [96] address the popularity prediction problem as a ranking task by exploiting a latent-SVM objective function defined such that the ranking of the popularity scores between pairs of images is maintained. They considered the number of views and comments for Flickr images and the number of re-tweets and favorites for Twitter.

The problem of image popularity prediction is also addressed in [34], whose experiments suggest that some sentiment ANPs defined in VSO [42] have a correlation with popularity.

In [97] the authors considered the number of likes achieved within the first hour after the image posting (early popularity) to predict the popularity after a day, a week or a month. This study has been performed on Instagram images and the dataset is publicly available [5]. The images are categorized as popular or not popular considering a popularity threshold obtained with the Pareto principle (80 % - 20 %). Three features representing information that is retrieved within the first hour of image upload (i.e., early information) are evaluated: social context (based on the user's number of followers), image semantics (based on image caption and NLP), and early popularity in the first hour. The binary classification is performed by using a Gaussian Naive Bayes Model. The experimental results show that the early popularity feature significantly outperforms the other evaluated features. Furthermore, the authors compared the proposed semantic feature with the features

[5] http://www1bpt.bridgeport.edu/∼jelee/sna/pred.html.

proposed by [36] for the task of popularity binary classification considering the MIR-1M Flickr dataset [98], obtaining better accuracy rates.

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.

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 [99] 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 [6]. This dataset consists of over 850K posts and 80K users, including photos from VSO [42] as well as photos collected from personal users' albums [100, 101, 102]. 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:

- **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;

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

- **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. Li et al. [103] extracted multiple time-scale features from a set of timestamps related to the photo post. As instance, the timestamp "postdate" is used to define several features with different time scales: "season of year", "month of year", etc. The framework presented in [104] exploits an ensemble learning method to combine the outputs of an SVR and a CART (Classification And Regression Tree) models, previously trained to estimate the popularity score. The models have been trained by exploiting features extracted from user's information, image meta-data and visual aesthetic features extracted from the image. In particular, the authors take into account the post duration (i.e., the number of days the image was posted), the upload time, day and month.

3.5.2 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 3.7). 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 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.

Figure 3.7: Example of images ranking based on the emotional category "joy".

3.5.3 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].

3.5.4 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 description of these ideograms (see the eleventh column in Figure 3.8). 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

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 3.8: 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.

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 3.8). 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 a precious resource for many useful applications [7].

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

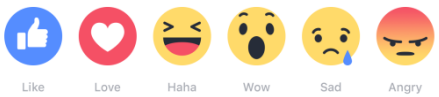

Figure 3.9: Facebook emoji reactions: Like, Love, Haha, Wow, Sad, and Angry.

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 [51].

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 3.9).

3.6 Image Polarity Prediction

3.7 Image Popularity Prediction

In Section 3.5.1 we introduced the task of Image Popularity Prediction.

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 in this paper.

The results discussed in this study, such as [40], 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 [39, 118], 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 [119, 84, 120, 121, 122, 123, 124]. 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 [56]. 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.

Systems with broader ambitions could be developed to address the new challenges (e.g., relative attributes, popularity prediction, common-sense, etc.) or to focus on new emerging tasks (e.g., image popularity prediction, sentiment over time, sentiment by exploiting ideograms, etc.). For instance, ideograms helps people to reduce the gap between real and virtual communications. Thanks to the diffusion of social media platforms, the use of emojis has been growing for years, and they are now integrated to the way people communicate in the digital world. They are commonly used to express user reactions with respect to messages, pictures, or news. Thus, the analysis of such new communication media could help to improve the current state of the art performances.

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 Sentiment Analysis

4.1 Introduction

4.2 RECfusion

4.3 RECfusion for lifelogging

4.4 Conclusions

Chapter 5

Final Discussion, Remarks and Future Works

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] 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.

[6] 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.

[7] 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*. IS&T. 2016. URL: http://recfusionproject.altervista.org/clustertracking.htm.

[8] 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.

[9] 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.

[10] 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.

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

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

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

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

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

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

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

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

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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).

[48] 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.

[49] 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).

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic. "A survey of multimodal sentiment analysis". In: *Image and Vision Computing* 65 (2017), pp. 3–14.

[66] 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.

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

[68] 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).

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

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

[79] 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.

[80] 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.

[81] 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.

[82] 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/.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

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

[94] 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.

[95] 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.

[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] 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.

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

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[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] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.