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

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. The aim of Sentiment Analysis is 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

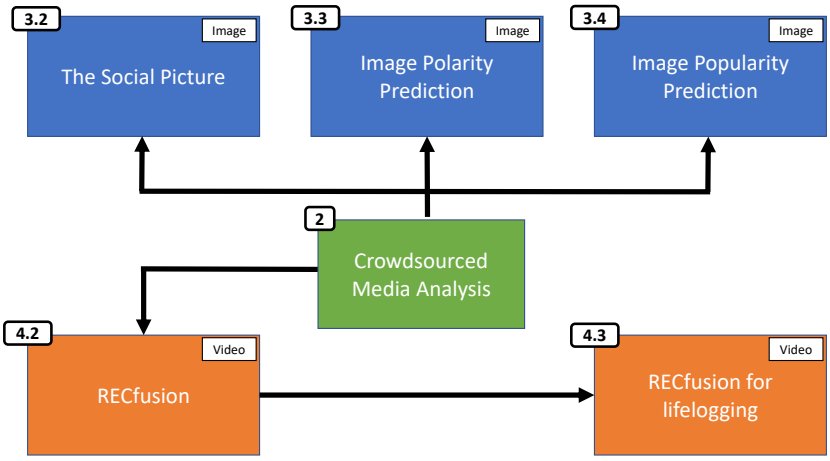

Figure 1.1: Dissertation structure. Numbers depicts the Sections which describe the proposed work. Blue blocks represent algorithms that work on images, whereas orange blocks represent algorithms that work on videos. All the media involved in the works presented in this dissertation are produced by users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

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. This dissertation we present several scientific works that analyses images and videos produced by users in a 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.). In Section

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. We start our discussion by an introduction on Crowdsourced Media Analysis, which brings together all the works presented in this dissertation. Each algorithm is presented in a different Section. In Chapter 3, we present all the methods that analyses images for the tasks of users behaviour analysis, image polarity prediction and image popularity prediction. In Chapter 4, we present our works related to the analysis of video contents. 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* and *RECfusion*. 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.

1.2 List of Publications

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. It allows organizations to develop solutions for both strategic issues and a method to find new ways to reach audiences on a broader scale. Moreover, the growing industry of online communication through smartphones provides a way for people of all backgrounds to give input on a project or research. Social media, blogs, forums, 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 like Facebook where people chat, read news, and share their experiences. 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, specially if they are expert users of the platform who provided useful and fair feedbacks). 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 [1].

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.

PARLARE DI IMMAGINI E VIDEO INTRODUCENDO I LAVORI SUCCESSIVI

Chapter 3

Image Sentiment Analysis

3.1 Introduction

TSP –¿ POPOLARITA IMMAGINI SU BASE EVENTI/LUOGHI –¿ IMMAGINI RAPPRESENTATIVE
POLARITY –¿ SENTIMENT EVOCATO
POPULARITY –¿ POPOLARITA' SU PIATTAFORMA SOCIAL I.E. QUANTE PERSONE RAGGIUNGE
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. 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 [2].

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 [3]. Historically, Sentiment Analysis techniques have been developed for the analysis of text [4], 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. In this Chapter we present three main inference tasks related to the image sentiment analysis that have been investigated in this research work: crowds behaviour, sentiment polarity, sentiment popularity.

First, we present *The Social Picture* [5], 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.

Then we present our work on the classic task of sentiment polarity prediction. Given an image, the proposed method 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. ABSTRACT

The third work presented in this Chapter is related to the task of image popularity prediction. ABSTRACT

3.2 State of the Art

3.3 The Social Picture

3.3.1 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 [6] 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 [7], and other multimedia information gathered by multiple sources.

In this paper 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

Figure 3.1: Heat map visualization.

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 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 in a visualization tool which uses 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.

In Figure 3.1, an example of visual output produced by the proposed framework. It is computed by considering images automatically gathered from social media and related to the cultural heritage site of Teatro Massimo, in Catania. As detailed in the following sections, the interaction with the heat map allows the users to explore all the images contributing in each part of the analyzed cultural heritage site.

3.3.2 Architecture

The architecture of the developed framework is shown in Figure 3.2. 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 of pictures to infer social behavioural information about the event considering the popularity of the uploaded scenes [6].

The collections can be explored from 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 3.3). The web application shows different interfaces depending of the specific user and the event in which he has joined after an invitation from the event manager (the person who created the event). To join an event's collection, the user must upload at least one picture related to that event. Collections can be explored by several data visualization enviroments, which are selected by the event manager. Anyone registered to *The Social Picture* can become event manager and start a social collection: this follows the "prosumer" paradigm, where the users are both producers and consumers of the service. The developed framework is characterized by a modular architecture: new visualization interfaces, as well as new semantic filters can be independently created and futher added to the system. Thus, when an event manager creates a new collection, he is allowed to specify several options to customize the image gathering, the social analysis to be performed and the visualization tools for the users of that collection. The event manager is also allowed to set a range of statistics, which will be available after the analysis of the collected images. These statistics helps organizators to extract useful social information from the crowdsourced pictures [8]. For example, what is the most popular artwork of a museum? What is the least considered? From

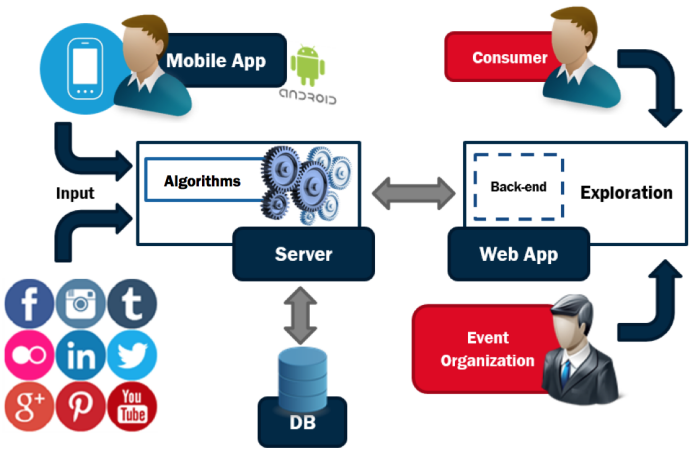

Figure 3.2: The Social Picture's architecture.

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 photo captured) 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, JPEG compression, 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 [9, 10], 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* [9] and *Places205-AlexNet* [10]. The CNN used in [9] consist of seven internal layers with a final 1000-way softmax which produces a distribution over the 1000 predefined classes of the ImageNet dataset [11].We considered the feature activations induced at the last hidden layer, which consists of 4096 dimensional feature (fc-7 feature), as an

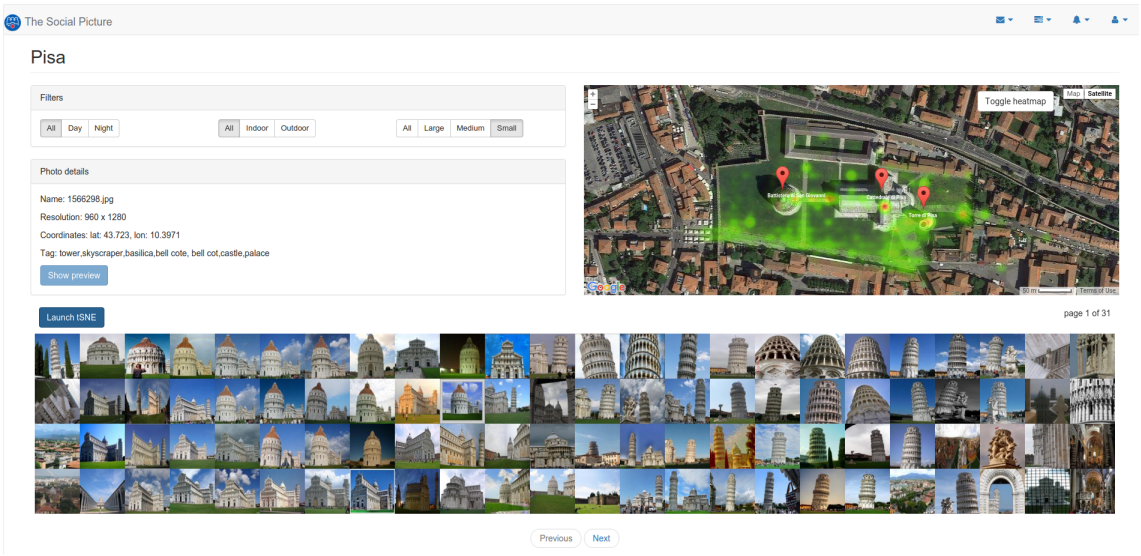

Figure 3.3: Example of exploration interface. It is composed by three main areas: the map area (upper-right) shows the positions where the images have been taken from. This give the positions of the users during the event and some hints about the most interesting parts of the site. The interaction with this map allows the users to know more details by selecting the images from their positions. The gallery (bottom) shows the collection's pictures organized by using t-SNE algorithm according to the selected filters and allows the users to explore the collection. The detail area (upper-left) shows the details of a picture selected from the gallery, and presents the available filters. From this interface, the user can launch the t-SNE algorithm on a customized set of images, apply one or more filters, and know more about the different statistics and inferences performed by the system.

image representation to be further used with t-SNE algorithm [12] for visualization purposes. We also fed the images to the *Places205-AlexNet* CNN [10]. 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.

3.3.3 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, which allow us to create several taxonomies in the image collection (panoramic pictures versus close-ups, natural versus artificial, indoor or outdoor, the presence or absence of crowds, selfies versus not selfies, and others). 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. There is a tool for the exploration of outdoor collections such as cultural sites. One more tool 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 video of the demo is available at the following link: http://iplab.dmi.unict.it/TSP

Heat map 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 heat map 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 heat map generated from data in *The Social Picture* is shown in Figure 3.1. 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 heat map by aligning images in *The Social Picture* with respect to panoramic images of the site of interest [13].

The heat map is a visualization used to depict the intensity of images at spatial points. The heat map 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 heat map is given by the number of collected pictures that contain that visual area. By clicking on a point of the heat map, the user can visualize the subject of images that contribuited to generate the map intensity at that point. This set of pictures can be further refined by selecting one of the images and asking the system to search similar pictures, or use the image subset as a starting point for furhter analysis. In other words, the heat map visualization gives the possibility to understand the behaviour of the people. 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) [14], 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 based on Threejs, 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

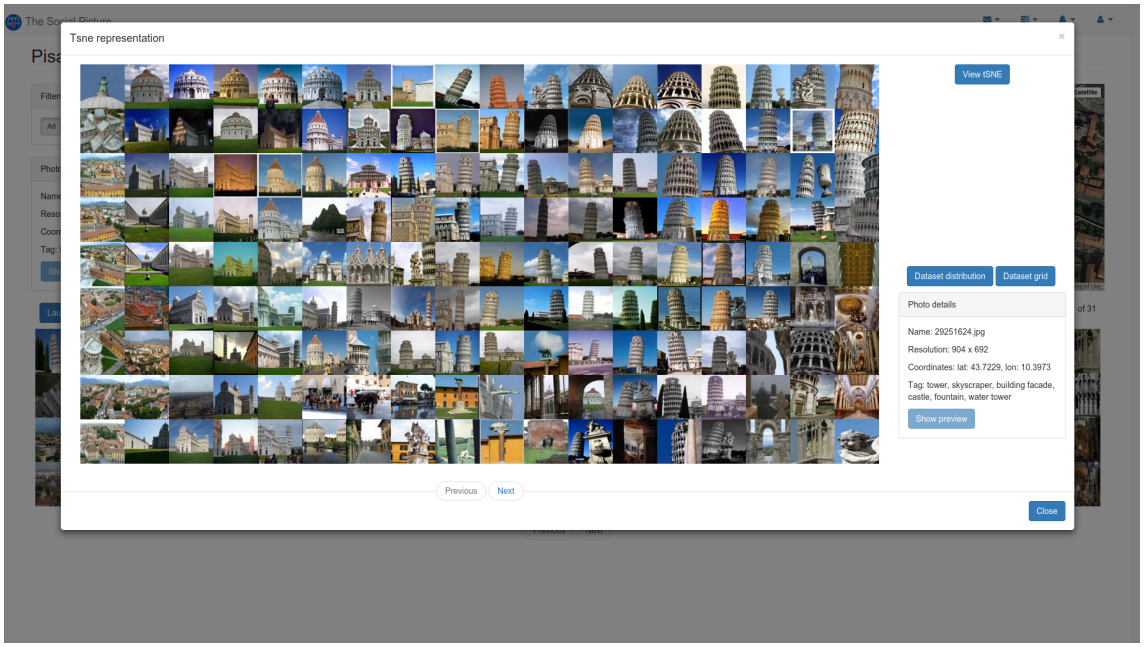

Figure 3.4: t-SNE visualization, the images are forced to fit a grid layout. Images of an event are automatically organized by visual content. Images close in the 2D space of the visualization tool are also close in terms of visual content.

photos. Moreover, the models in the 3D web viewer can also be browsed through Leap Motion system, an intuitive and fast interactive system.

t-SNE exploration

We exploit the fc7 feature extracted with the *AlexNet* architecture [9] for each image and use the t-SNE embedding algorithm [12] to compute a 2D embedding that respects the pairwise distances between visual features, forcing to fit a grid layout (Figure 3.4). 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. Note that images with the same subject are automatically arranged nearby (see Figure 3.4). 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 [9] has been trained using a different dataset concerning 1000 classes of objects, but the fc7 features resulted expressive and representative enough to be applied succesfully to a generic event collection.

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 desidered direction and by using the scroll wheel respectively. 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 (see Figure 3.4).

Hierarchical t-SNE The first implementation of the t-SNE exploration tool in TSP was unable to scale with the number of the collections' images. The new tool presented in this demo implements 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 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.

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).

One more feature included in The Social Picture is the automatic image captioning as described in [15]. With the aim to help the user to include a description to an uploaded image, *The Social Picture* automatically generates and suggests a description to the user that can then refine it. The descriptions of images can be used for text based query performed by the user.

3.4 Image Polarity Prediction

3.5 Image Popularity Prediction

3.6 Conclusions

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

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

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

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

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