The Social Picture

ABSTRACT

This paper describes a framework to collect and present huge amount of crowdsourced pictures about social events, cultural sites and other customized events, which invove the creation of a "flow of social pictures". *The Social Picture* aims to create social communities of users that contribute to the creation of image collections related to common events and interests. The collections can be explored through a number of filtering and aggregation tools, such as image content filters, saliency heat maps, embedding and exploration tools. The interfaces available in *The Social Picture* allow users to create customized collections from image subsets by exploiting filters based on visual features, social tags, geolocation, and device gathered information.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Social Media, Crowdsourcing, Multimedia, Image Collections

Keywords

ACM proceedings, LATEX, text tagging

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 if there are groups of people interested to specific contents [2]. In the context of big social data analysis, Machine Learning and Computer Vision can be used to develop new advanced analysis systems to automatically infer knowledge from large scale visual data and other multimedia information gathered by multiple sources.

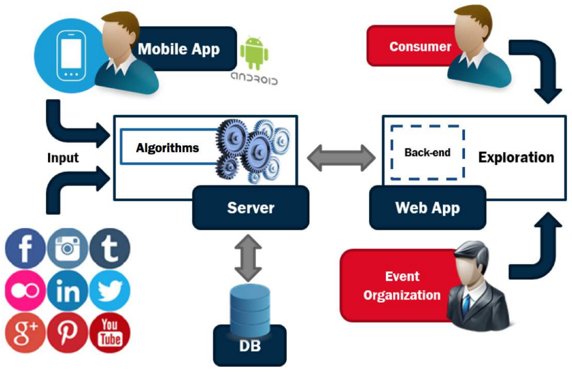
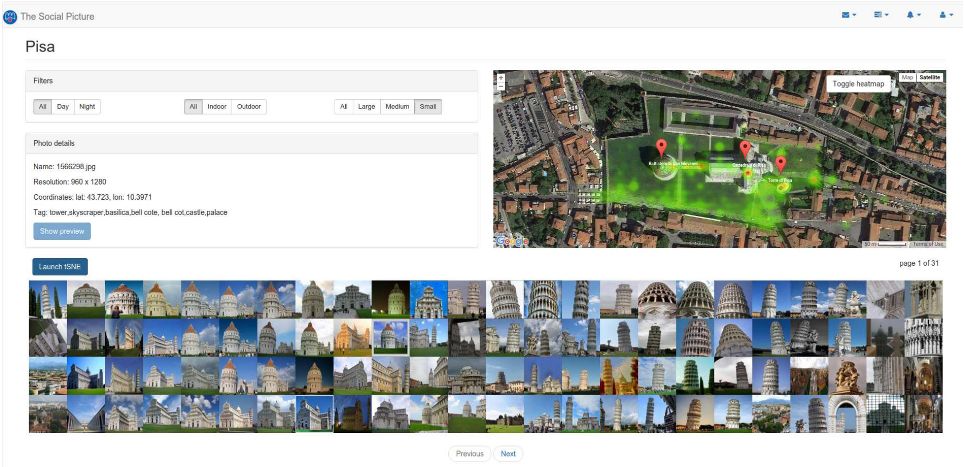


Figure 1: The Social Picture's architecture.

The aim of *The Social Picture* is to collect, analyze and organize these huge flows of visual data, allowing users to exploit specific image collection features. We designed this system to be applied on three main scenarios: social events, cultural sites, private events. The social peculiarities of such collections can be exploited not only by the people who partecipate at an event, in fact each scenario distinguish two kind of users: event organizer and event partecipants. Imagine an art-gallery manager who leases a famous Picasso's 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?

All these information can be inferred by the analysis of the above mentioned multimedia audience activity. The collection of the produced images gives some hints to give an

WOODSTOCK '97 El Paso, Texas USA



Name: 1566298.jpg

Resolution: 960 x 1280

Coordinates: lat: 43.723, lon: 10.3971

Tag: tower,skyscraper,basilica,bell cote, bell cot,castle,palace

Figure 2: Example of exploration interface.

answer to each of the above questions, and this can be exploited by the event organizers for their decision making.

1.1 Architecture

Users can add an image to an event's collection by using a specific mobile application or one of the most common social networks. The images are first analysed by a set of Computer Vision algorithms, and then stored in the Database with all their extracted features. These information are exploited by the system to create smart interfaces, available to the users during the exploration of the image collection. The event's collection can be explored through a web application, which exhibits a range of interfaces and tools to better explore the huge amount of data.

The whole architecture is described in Figure 1.

1.1.1 Database

When a new image is added to the system, it is analysed by a set of algorithms with the aim to extract pre-defined features (both visual and textual), as well as semantic information. The image is collected, entered, processed and all the analysis results are stored in the Database.

1.1.2 Mobile App

The mobile application allows users to upload their visual contributes to the event's collection. Other users can also contribute to the creation of the collection by posting their contents through the considered social netwoks, by adding a specific event's hashtag. The system collects all these data, and exploits this crowdsourced multimedia flow of pictures to infer social information about the event, that is to say to perform Sentiment Analysis.

1.1.3 Web Application

The web application shows different interfaces depending on the specific user and event. Collections can be explored by several data visualization enviroments, the available interfaces are defined by the event manager or organizator ("Event Organization" in Figure 1). This system is characterized by a modular architecture: new visualization interfaces, as well as new collection filters can be independently created and then added to the system. Thus, when an event manager creates a new collection, he is allowed to specify several options related to customizing the image gathering, the analysis to be performed, the available visualization modes and tools enabled to the consumers. The event manager is also allowed to set a range of analysis results and statistics, which will be available after the collection analysis. These statistics helps organizators to extract useful inferences from the crowdsourced pictures. For example, what is the most photographed artwork of a museum? What is the least? From which perspective these pictures were taken?

These inferences could be exploited, for example, to perform targeted 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, supporting merchandising strategies. Information about what is the most interesting part of a landmark building can help on taking decisions about renovating the most important parts of the building, where the connotation of importance is achieved by the crowd.

1.1.4 Analysis engine

The several analysis/visualization 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 minimum analysis performed by the system feds all the images to two CNNs, extracting both their classification results and an image representation. One of them, is the so called *AlexNet* [1]. The CNN used in [1] consist of seven internal layers with a final 1000-way softmax which produces a distribution over the 1000 predefined classes of the ImageNet dataset [3].We considered the feature activations induced at the last hidden layer, which consists of 4096 dimensional feature (fc-7 feature), as an image representation. By this CNN architecture, we extract both the image fc7 representation and the classification results, over the 1000 ImageNet's classes. We also fed the images to the *Places205-AlexNet* CNN [5], this CNN has the same architecture of *AlexNet* CNN, but it is trained on 205 scene categories of Places Database *Places205-AlexNet* with 2.5 million images.

The results of the image evaluation by the above CNNs

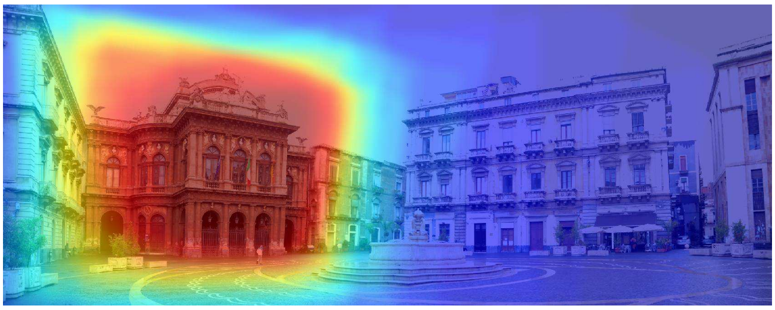


Figure 3: Heat map visualization.

consist of the three classification predictions with the highest scores, given by each CNN, and the fc7 image representation.

1.2 User experience

An event manager creates a new event selecting among three kinds of event: social event, cultural site or private event. The available event categorization can be extended to include other customized categories. We imagined these three categories to better focus the aims of the specific analysis, and the inferred information the organizator wants to extract. The data gathering can be performed within a specific time window, also 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 and inferences the system have to perform on the collection.

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 two main exploration systems. One is well suited for outdoor events such as cultural sites, and whenever a large environment image can be considered as a reference for more detailed ones, the other can be exploited to better explore any huge image collection.

1.2.1 Heat map exploration

In a cultural site, people are used to take pictures of the same views regarding the most famous and appreciated attractions and artworks. It would be interesting to infer the attitude of people with respect to the different parts of the sites, in the sense that we want to know which parts of the site captures people's interests. In order to highlight this kind of information, the system creates an heat map computed on a panoramic image of the entire site.

An heat map is a visualization used to depict the intensity of data 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 have taken that area.

By clicking on a point of the heat map, the user can visualize the images that contribuited to 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 users a method to understand the behaviour of the attending people, also it can be considered as a powerful image retrieval tool, over the event's image collection.

1.2.2 t-SNE representation

We exploit the fc7 feature from each image of the dataset and the used the t-SNE embedding algorithm [4] to compute a 2D embedding that respects the pairwise distances between features, forcing to fit a grid layout (Figure 4). The t-SNE (t-Distribuited Stochastic Neighbour Embedding) is a technique for dimensionality reduction that is particularly well suited for the visualization of high dimensional datasets. Note that in the images with the same subject are arranged nearby, moreover the system arranges nearby also images which are not the same but have a similar appearance (i.e., all selfie images). It is also important to highlight that the used CNN 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 an event specific collection.

Another exploration tool, allow users to see the result of t-SNE embedding without using the grid layout. If 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.

For example, the user can apply a translation or a zooming to all the viewed images, by 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. A number of tools enable users to fine-tune the spread of images along X and Y axis, and the zooming effect. These parameters can be used to transform the distribution of the embedded images, and adapt the effect of the mouse commands. For example, if the user fix an upper bound for the zoom effect, when this limit is reached the scrool wheel of the mouse has the effect of separating the images according to the amout of the rotation.

Furthermore, the user can choose a subset of images and compute the t-SNE embedding of them directry on the browser. This tool has been developed by using the *tsnejs* Javascript library by Andrej Karpathy.

2. CONCLUSIONS

The Social Picture helps to distill the essence of a social event, throught the multimedia audience activity.

3. ACKNOWLEDGMENTS

??????????????

4. REFERENCES

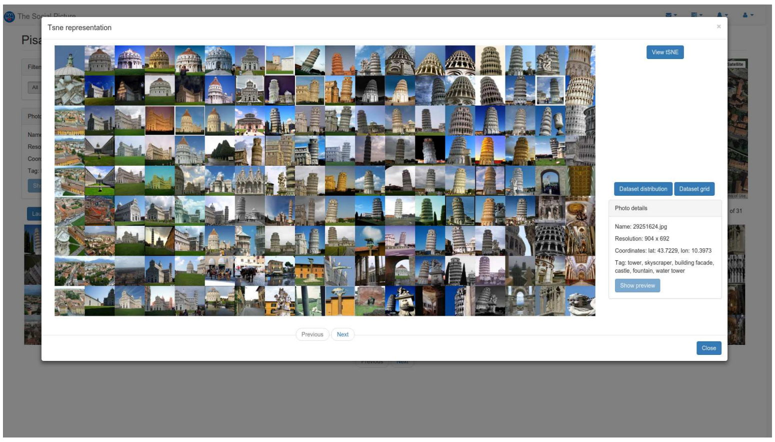
[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] 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*. ACM MM 2015, 2015.

[3] 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. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[4] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[5] 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*, pages 487–495, 2014.



Name: 29251624.jpg

Resolution: 904 x 692

Coordinates: lat: 43.7229, lon: 10.3973

Tag: tower, skyscraper, building facade, castle, fountain, water tower

Figure 4: t-SNE visualization, the images are forced to fit a grid layout.