SentiCart: Cartography and Geo-contextualization for Multilingual Visual Sentiment

Brendan Jou*
Electrical Engineering
Columbia University
New York, NY 10027
bjou@ee.columbia.edu

Margaret Yuying Qian*
Computer Science
Columbia University
New York, NY 10027
margaret.qian@columbia.edu

Shih-Fu Chang
Electrical Engineering
Columbia University
New York, NY 10027
sfchang@ee.columbia.edu

ABSTRACT

Where in the world are pictures of cute animals or ancient architecture most shared from? And are they equally sentimentally perceived across different languages? We demonstrate a series of visualization tools, that we collectively call `SentiCart`, for answering such questions and navigating the landscape of how sentiment-biased images are shared around the world in multiple languages. We present visualizations using a large-scale, self-gathered geodata corpus of >1.54M geo-references coming from over 235 countries mined from >15K visual concepts over 12 languages. We also highlight several compelling data-driven findings about multilingual visual sentiment in geo-social interactions.

CCS Concepts

•Information systems → Geographic information systems; Multimedia databases; Data mining; Web interfaces; •Human-centered computing → Visualization; •Applied computing → *Psychology; Sociology;*

Keywords

affective computing; geodata; multilingual; visual affect; sentiment; ontology; visualization; GIS

1. INTRODUCTION

How geographically diverse are our sentiments in social multimedia? And specifically, how diverse (or localized) are our sentiments in the images and concepts we use everyday along linguistic and geographical lines? Following trends of other fields, the advent of high-volume and weakly-supervised data are driving increased interest in *large-scale* sentiment studies in affective computing. These two data properties, sometimes referred to as "volume" and "veracity" respectively, are key elements of any Big Data problem.

*Denotes equal contribution.

ICMR'16, June 6–9, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912022

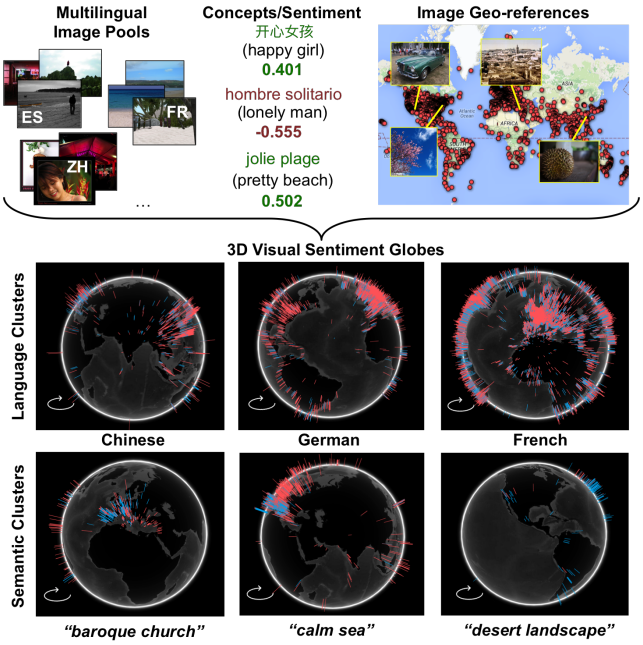

Figure 1: Example globe visualizations of geographical data from sentiment-biased images around the world in `SentiCart` by language and semantics.

In this work, we focus on unpacking another tenant of "Big Affective Computing" that deals with the multi-faceted or multimodal nature of visual affect data (i.e. "variety"). We build on the recent work of [6] which presented a Multilingual Visual Sentiment Ontology (MVSO) consisting of >15K sentiment-biased visual bi-concepts and >7.3M images over 12 different languages. Though it is only a small sampling of web image content, it represents the largest affect-biased unconstrained image dataset from multiple languages to date. The goal of [6] was to expand affective variety by introducing cultural diversity, accomplishing this via language diversity. We believe that *language is only one of many signals to capture multicultural visual sentiment*; and geographical data (geodata) is another signal not previously explored for visual sentiment understanding, to the best of our knowledge. Since any one language may be spoken across a multitude of countries and with varying density, to truly understand visual sentiment around the world, geodata is a necessary complement to the language dimension. Here, to assist the study of visual sentiment variety in geodata, we developed a

visualization system called SentiCart to chart the landscape of multilingual visual sentiment around the world.

Contextualizing social multimedia with explicit location or implicit geographical metadata has been a topic of research interest for over a decade [1][5][7]. Among the first for digital photography, [10] presented several frameworks for location tag acquisition, geo-referencing and image media browsing interfaces. In computer vision, [12] used geotagged photos to assist visual landmark detection over about 20M images. In [4] and [8], Twitter posts are used to study geographic sentiment and music preference differences, respectively, while in [9], Twitter and Flickr posts are analyzed spatiotemporally to map out social visual interests. These geographical information systems (GIS) each lack a multilingual, visual grounding, or sentiment biasing component we seek here.

One close relative to our system is [11] where a large-scale collection of >1.5M facial expression videos sourcing from over 94 countries was presented. Subjects were shown one of about 8K online videos as stimuli and their reactions were captured via a web camera. Although there is a wide variety in originating locations of subjects, this study focuses more on region-based differences rather than cultural differences. In addition, [11] also relies on explicit sentiment in facial expression videos as a form of sentiment feedback while we use weakly supervised semantic cues for sentiment.

In [6], an ontology of 15,630 visual concepts coming from 12 different languages was presented along with accompanying social photos and metadata called Multilingual Visual Sentiment Ontology (MVSO). Each visual concept in the ontology is semantically structured and sentimentally biased using a construct called adjective-noun pairs (ANPs). The noun component provides a visual grounding to the concept and the adjective defines a sentimental attribute, e.g. *fluffy dog* or *abandoned railroad*. MVSO aimed at gathering culturally diverse visual concepts for studying sentiment and emotion in social multimedia; however, the ontology only strikes at one aspect of cultural diversity: language. We augment the work in [6] by also incorporating geographical data signals toward multicultural affect research.

The contributions of our work include: (1) an intuitive, interactive and fluid browsing visualization system called SentiCart for uncovering geographical insights in visual sentiment data, (2) an implementation using large-scale visual sentiment geographical data with over 1.54M geo-references covering 237 countries, and (3) the public release of this geographical data over a multilingual ontology.

2. GEODATA ACQUISITION

In order to collect geo-localization data for social photos, we use a combination of two multi-source methods – one with high reliability, but low coverage and another with lower reliability, but higher coverage. We root our geodata collection and analysis on top of MVSO [6] because of publicly available auxiliary data streams that can be mined from the same social media platform, i.e. Flickr. The multilingual nature of the data also lends itself toward being geographically spread, and the large-scale nature of over 7.3M images maximizes our opportunity to also study language and culture together on a significant geographic scale.

2.1 GPS Coordinate Data

Global positioning system (GPS) geo-localization provides highly reliable latitude-longitude coordinates (usually within several meters worst-case depending on satellite-receiver precision) and may be encoded in image headers of some digital photos. In November 2015, we queried the Flickr API[1] over the entire MVSO corpus [6] of 7,368,364 images and acquired GPS coordinate data for 1,410,892 images. The remaining images were either not GPS-tagged or privacy permissions were not granted for public querying. The top three languages with GPS-tagged images in order are: Persian (32.24%), French (25.35%) and Italian (24.40%). Although the largest language by image count is English (4,049,507), it only had a 17.48% coverage in GPS-tagged images.

Language	*#ANPs*	*#Georefs*	*Language*	*#ANPs*	*#Georefs*
Arabic	22	99	Italian	3,184	206,315
Chinese	395	11,553	Persian	10	92
Dutch	315	16,292	Polish	67	3,873
English	4,407	707,846	Russian	95	2,014
French	2,241	163,193	Spanish	3,241	259,138
German	717	38,544	Turkish	137	1,933

Table 1: Number of GPS-based geo-references (georefs) collected from MVSO [6] images by language. Since not every visual concept (ANP) had images with GPS data, we also show the number of remaining ANPs with at least one geo-referenced image.

2.2 Metadata-inferred Location Data

In most social media platforms, including our setting on Flickr, users can and do often provide image title and descriptions to add additional context for their media posts. Locations are commonly found in these user metadata because they provided a concrete grounding for *where* an event or memory took place. For geo-localization though, this data can often be very uninformative, e.g. user input text like "our new house," as well as arbitrary, e.g. "Little Rock" could refer to a literal small rock in the image content or a town in Arkansas, USA. These two streams of user text from title and description offer greater coverage of images than depending on the presence of GPS data, but introduces noise is less reliable in localizing. As result, when GPS data is available, we always use it to geo-localize a given image regardless of user-provided metadata. Otherwise, we prefer automatically extracted locations in the image title over those in descriptions.

Using image metadata streams, we extracted location fields by performing named entity recognition (NER) [2] on user-provided text. We translated all text into English[2] and used only English NER models. We note that translation allowed us to get a higher recall of tagged locations compared to native-language NER models due to the frequency at which users posted descriptions with mixed languages, e.g. an image title "Small Alley in 香港". We extracted metadata-inferred locations for all languages with under 100K GPS-tagged images to bolster their geo-reference count.

We applied NER-tagged locations as queries to Google Maps' Geocoding API[3] to retrieve latitude-longitude coordinates as well as to filter incorrect NER detections or ambiguous locations. Since geocoding queries can be region-biased, we performed multiple searches over all relevant country

[1]https://www.flickr.com/services/api
[2]https://translate.google.com
[3]https://developers.google.com/maps

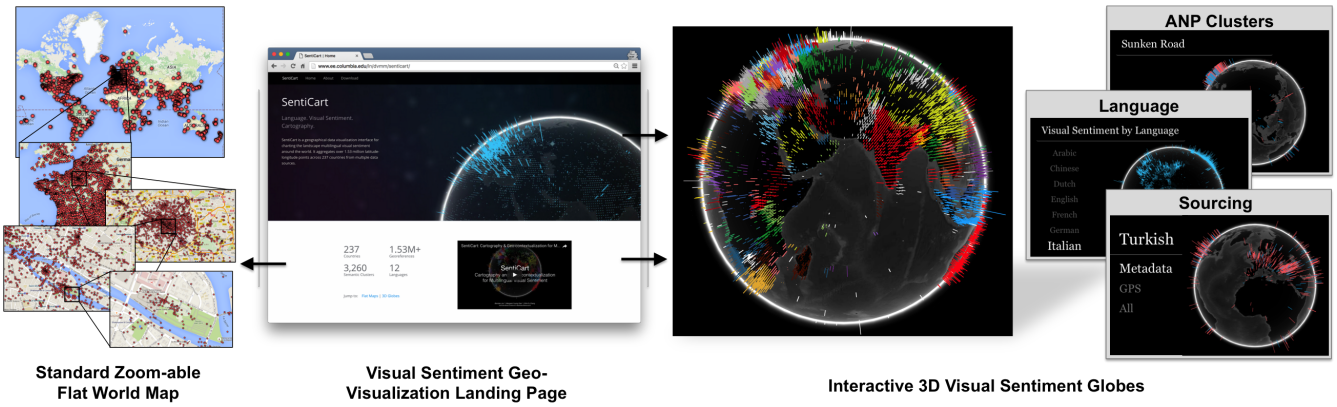

Figure 2: Example visualizations in `SentiCart` of multilingual visual sentiment around the world show the main landing page along with a classic flat world map view, here on the French sub-corpus, and 3D globe visualizations of visual sentiment. The colorful globe shows country-colored sentiment for the English sub-corpus and adjacent cards show additional display modes in `SentiCart` along source, semantics and language.

code top-level domains (ccTLDs) per language according to official ICANN listings, e.g. `.pl` for Polish was included, `.ir` for Persian, `.ru` for Russian, etc. This increased the likelihood that matches were found in regions that speak the target language and also gave queries as many chances to succeed overall as there were ccTLDs per language, e.g. "Main Street"; for example, we queried across over 10 different ccTLDs for Arabic given the diversity of regions it is spoken in. When multiple named locations were detected by NER, we queried for geocodes in descending string length order taking the first query that succeeded. And when multiple geocoding results matched, e.g. the query "Gulf Coast" can match to "Gulf Coast Airport, Tivoli, TX", "Gulf Coast, Missouri City, TX" and "Gulf Coast, Naples, FL", we selected one at random only if all matches came from the same country, and otherwise, omitted the query result altogether for its ambiguity. We also ruled out geo-references to continents, e.g. "South America," and large bodies of water, e.g. "Mediterranean Sea," for being too broad for our setting. If all queries failed to geocode an image, we determined that the image lacked sufficient information needed to localize.

Language	*#Locs*(T)	*#Georefs*(T)	*#Locs*(D)	*#Georefs*(D)
Arabic	318	102	1,317	137
Chinese	27,631	9,110	179,407	28,579
Dutch	8,537	3,344	73,315	13,848
German	23,867	13,947	207,630	52,159
Persian	124	49	545	140
Polish	4,048	854	23,222	2,721
Russian	2,731	1,168	20,586	3,539
Turkish	4,235	734	28,890	2,781

Table 2: Number of named locations (locs) recognized and geo-references (georefs) coded from user-provided title (T) and description (D) metadata in MVSO [6] images by language. Only *new* geo-references counts are given, i.e. #Georefs columns are mutually exclusive with each other and also mutually exclusive to GPS-tagged images in Table 1.

On the subset of eight languages we extracted metadata-inferred geo-references for, a total of 133,212 new geodata points were extracted compared to 74,400 GPS coordinates (i.e. in relative, 79.05% greater image coverage). The combination of GPS and metadata-inferred geodata accounted for 20.96% of the total 7,368,364 images in the MVSO [6] image dataset.

The geographical data we collected and web visualizations can be accessed at http://www.ee.columbia.edu/ln/dvmm/senticart and a video showcasing the functionality of `SentiCart` can be accessed at https://youtu.be/cI-2llSErSo.

3. VISUAL AFFECT GEOVISUALIZATION

In order to better understand visual sentiment around the world in `SentiCart`, we developed two interactive visualizations. One visualization is a flat, point-wise, low-interaction view of geodata points. The lower interactivity in this visualization allowed for batch processing and thus precise localization to geodata points at our large input scale. The other visualization is a three-dimensional, fluid, high-interaction globe of geo-references. This visualization allowed us to quickly and easily compare across data modalities and also gives us an additional dimension along which to interact with the geodata. The two modes of visualization provide a trade-off in exploring geographical distributions of visual sentiment. In Figure 2, we show examples from our visualizations. For the flat world map view, our interface matches many other canonical map interfaces but scales to hundreds of thousands of geodata points by linking with Google Fusion Tables [3]. The interface allows for a zoomable, albeit flat view of the data and can render elements like political borders without burdening usability.

For the 3D globe view, we enable fast and fluid browser-based rendering using WebGL and a base globe library[4]. To maintain low-latency interaction in this rendering, we reduced the resolution of the geo-references by performing geohashing, where latitude-longitude coordinates are hashed into geodesic spatial bins and where hash lengths, or precisions, correspond directly to geographical distances. In our visualization and scale of 1,544,104 geo-references, we found that quantizing coordinates to within 19.55 km (or 12.14 mi), which corresponds to 10 hash bits for both latitude and longitude, provided the best latency-to-resolution visualization trade-off on most modern browsers and machines. The

[4]https://www.chromeexperiments.com/globe

added dimension of the 3D globe visualization compared to the flat world map, also allowed us to visualize the sentiment magnitudes at given geo-localization points. Since we perform geohashing, to aggregate sentiment in a given spatial bin, we take the weighted average of the image sentiment values from adjective-noun pairs.

In the globe view, we provide multiple slice views of the ontology and geodata we collected. We can visualize sentiment by language around the world and easily compare regional differences. On the surface of the globe, bar heights correspond to sentiment strengths scaled $[0, h \in \mathbb{R}]$ where colors represent positive or negative (usually red/"hot" and blue/"cool"). In addition, since we had multiple sources for our geodata, we can visualize each source's geographical origins to compare localization consistency. We also enable visualization along ANP clusters using the ontology structure in MVSO [6] where ANPs were gathered into semantically coherent groups, e.g. we can visualize geodata and sentiment of images in the multilingual cluster "old town square."

Given our geodata visualizations, we highlight several preliminary insights from the geo-social data we collected using Multilingual Visual Sentiment Ontology [6] imagery. One striking, but in retrospect, intuitive phenomena we observed about geodata in [6] is that despite its focus on multilingual breadth, most language's geodata were not as geographically constrained as we expected. In fact, while there were indeed geo-references that originated from countries where the primary language was the same as the corresponding language source, there were also a sizable number of geo-references that originated from regions outside. We hypothesize that because MVSO collected its imagery from a social multimedia platform, the phenomena is due to photo tourism – that is, photos are geographically tagged in countries that users are just visiting and when posting on social media, they unsurprisingly still use their native tongue to describe the image. As an example, many images from the Chinese ontology branch actually geo-localized to locations in Japan with ANP tags such as 古建築 (*ancient architecture*)[5]. Yet, it is unclear how pervasive this tourism phenomena is in [6] and whether it is stronger in some languages than others, e.g. geo-references for the Spanish ANP for *tropical landscape* (*paisaje tropical*) mostly localized to places in Mexico, Brazil, Spain and Colombia as one might expect.

In general, we found that geo-localizations for many multilingual ANP clusters and their sentiments coincided with intuition. For example, clusters referring to bodies of water like seas and lagoons did mostly geo-localize along coastlines and were composed of positive sentiment ANPs. Some weather-related clusters like *rainy day* and *cloudy landscape* tended to be geographically diverse and overwhelmingly negative in sentiment across languages. In some cases, where geo-references *did not* occur was also telling, e.g. the ANP cluster *stray cat* had fewer points in Asia than Western countries, likely owing to more domesticated cats in the West.

4. CONCLUSIONS

We presented a fluid and interactive visualization system called `SentiCart` for visual sentiment geodata, implemented with 1.54M latitude-longitude points over more than 235 countries originating from visual concepts across 12 languages. We aggregated geodata from multiple sources including native GPS tagging as well as automatically extracted geo-references from image metadata using natural language processing and geocoding. Two modes of visualization were demonstrated in `SentiCart`: a flat world map view and a 3D globe view. We also presented evidence that the multilingual ontology presented in [6] is affected by photo tourism, and suggest that this is consistent with social multimedia and language usage phenomena. Overall, `SentiCart` has provided concrete visual confirmation that multicultural visual sentiment is a deeply geographic-dependent entity as much as it is a semantically-dependent one.

In the future, we seek to expand our geo-reference coverage by training our own or applying higher fidelity native language NER models and incorporate implicit signals from user profiles to better localize social photos. We also plan to make our sentiment scores more fine-grained and accurate. In addition, we will investigate using geodata for regularizing the training of models for culture and language prediction.

[5]Despite these two cultures sharing character sets, we verified from image metadata text that users did use Chinese.

5. ACKNOWLEDGMENTS

We would like to thank Miriam Redi, Mercan Topkara, Nikolaos Pappas and Tao Chen from the MVSO team for their support and insightful discussions. We especially thank Nikolaos Pappas for providing English metadata translations for our implicit location extraction.

6. REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW*, 2010.

[2] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[3] H. Gonzalez, A. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. Google Fusion Tables: Data management, integration, and collaboration in the cloud. In *ACM SOCC*, 2010.

[4] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L.-E. Haug, and M.-C. Hsu. Visual sentiment analysis on Twitter data streams. In *IEEE VAST*, 2011.

[5] R. Ji, Y. Gao, W. Liu, X. Xie, Q. Tian, and X. Li. When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics. *ACM TIST*, 6(1), 2015.

[6] 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 *ACM MM*, 2015.

[7] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision - A survey. *Multimedia Tools and Applications*, 51(1), 2011.

[8] J. L. Moore, T. Joachims, and D. Turnbull. Taste space versus the world: An embedding analysis of listening habits and geography. In *ISMIR*, 2014.

[9] V. K. Singh, M. Gao, and R. Jain. Social pixels: Genesis and evaluation. In *ACM MM*, 2010.

[10] K. Toyama, R. Logan, A. Roseway, and P. Anandan. Geographic location tags on digital images. In *ACM MM*, 2003.

[11] T. Vandal, D. McDuff, and R. E. Kaliouby. Event detection: Ultra large-scale clustering of facial expressions. In *FG*, 2015.

[12] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *CVPR*, 2009.