AAAI Technical Report WS-12-03
Social Media Visualization

Cultural Analytics of Large Datasets from Flickr

Daniela Ushizima[1], **Lev Manovich**[2], **Todd Margolis**[3], **Jeremy Douglass**[4]
[1]Computational Research Division,
Lawrence Berkeley National Laboratory, 1 Cyclotron, Berkeley, CA 94117
[2]California Institute for Telecommunication and Information, Visual Arts Dept.,
University of California, San Diego, CA
[3]Center for Research in Computing and the Arts,
University of California, San Diego, CA
[4]Liberal Arts, Ashford University, San Diego, CA

Abstract

Deluge became a metaphor to describe the amount of information to which we are subjected, and very often we feel we are drowning while our access to information is rising. Devising mechanisms for exploring massive image sets according to perceptual attributes is still a challenge, even more when dealing with user-generated social media content. Such images tend to be heterogenous, and using metadata-only can be misleading. This paper describes a set of tools designed to analyze large sets of user-created art related images using image features describing color, texture, composition and orientation. The proposed pipeline permits to discriminate Flickr groups in terms of feature vectors and clustering parameters. The algorithms are general enough to be applied to other domains in which the main question is about the variability of the images.

Introduction: image deluge

While technologies to acquire and record images have allowed us to yield ever more data, algorithms for image reasoning and how to sift through massive image datasets still need much advancement. The development of better techniques to perform these tasks is crucial for data-intensive science, given that some fields are facing hundred to thousand-fold increases in image data volumes from high-throughput instruments, sensor networks, accelerators, and supercomputers, compared to the volumes generated only a decade ago. Humanities are now beginning to face the same problem. In the 1990s and first part of 2000s, most digital work in humanities focused on digitization of cultural heritage and presenting these digitized collections online. However, in the last few years, there is a growing emphasis on using computational techniques to analyze these large collections. Cultural Analytics (Manovich 2009) is a new research area which emerged to address these challenges. It uses digital image processing and high resolution visualization for the analysis of large image and video collections, covering a range from antiquity to contemporary media, including images and video uploaded by users of social media communities.

Social media services and media sharing sites such as Flickr, Twitter and YouTube offer lots of opportunities to study cultural and social behavior and patterns. These data sources attracted attention of many researchers interested in analyzing messages, conversations, favorite rankings, and transactional data available on these social media sites (Anderson et al. 2012). For Flickr, some papers also undertook analysis of the content of images as in (Crandall et al. 2009), who combined the analysis of the content of 35 million Flickr photos, their tags, and upload times to predict their geospatial locations. Siersdorfer and Pedro (Siersdorfer and Pedro 2009) investigated the relations between favorite rankings and image features for a set of Flickr photos to determine what image qualities are responsible for perceived attractiveness of the photos. They discovered that Flickr users prefer pictures with higher color saturation, higher contrast, and that are sharper than less favorite photos.

This paper proposes a novel way of using social media data. By analyzing large collections of images uploaded to Flickr cultural groups, we can test hypotheses about these communities in ways impossible with the traditional humanities approach where researchers examine manually only small numbers of works. We offer a new way to analyze the structure of the overall "landscapes" of cultural fields as represented by these much large samples. For example, we may find that one field is divided into a small number of genre or style categories with sharply delineated boundaries, while another field has such a stylistic diversity that it is meaningless to divide into categories.

Our goal is to use computational methods to analyze large samples of cultural fields, map these samples according to visual style, content and other dimensions, by using metrics in high dimensional spaces, and visualization tools. To the best of our knowledge, there is no system that provides

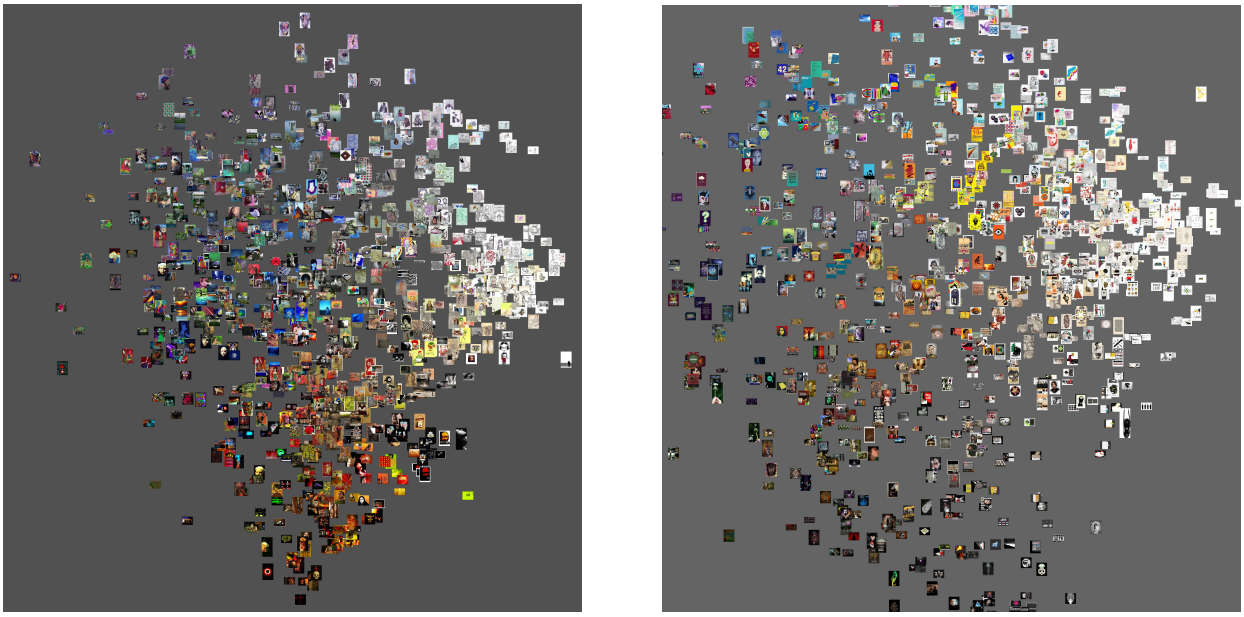
"exploratory data analysis" with large image sets created by users and available on social media networks. We propose a framework for identifying clusters based on visual and/or semantic similarity that uses digital image analysis to describe the content. Different from content-based image retrieval, and assumptions about existing labels (Gu and Ren 2010) such as "styles" and "genres," we want to map the full visual variability of large cultural datasets. Such datasets seldom allow well-defined hypothesis definition with clear labels before analysis, and require alternative approaches (Yamaoka et al. 2011). This is the grand idea of our project, and this paper lays out some preliminary, but fundamental steps towards clustering without "labels". We present a set of tools applied to image datasets from Flickr, proposing a pipeline to organize images according to visual attributes as image illumination, composition, color and texture. Also, we propose the use of image clustering in the context of the cultural analytics for calculating summary statistics of elements in each image group. First, we describe the Flickr image datasets, and present some statistics based on their metadata. Next, we explain the process of extracting the image "fingerprints" to be used later as grouping criteria, from which meaninful subsets should emerge after using datamining algorithms. Finally, we calculate a cluster quality indicator to illustrate how the Flickr datasets deviate from each other, and how dissimilarity measures can support understanding variability in image repositories from social media. Supplemental material for this paper containing additional visualizations which compare the two image sets is available online at (Ushizima et al. 2011).

Methods: sifting through large datasets

Our image datasets come from two Flickr groups: a) Art Now, described as "A group for displaying, fostering awareness and discussing the emerging relevant art and artists of today"; b) Graphic Design, described as "Anything from drawings you did in Paint to photoshoped images. If you made it, put it in the pool". At the time of our download (08/2011), these groups contained approximately the same number of images (170,000). Along with the images, we also downloaded publicly available metadata - image tags, titles, and available info about the authors. Analysis of image tags alone suggests that the two groups have both different and overlapping content. The top 5 tags in Graphic Design group are "design", "art", "illustration", "graphic", and "graphic design". The top 5 tags in Art Now group are "art", "panting", "drawing", "illustration", and "abstract". The overall number of tag assignments for all images in two groups is very similar: 1,777,751 tags in Graphic Design group; 1,700,516 tags in Art Now group. However, if we count the numbers of unique tags in each group, we find a significant difference: 145,124 tags in Graphic Design group vs. 77,008 tags in Art Now group. This suggests that participants in the first group use more specific terms in describing their images. The particular challenge posted by these and similar image sets is their semantic and visual diversity. While many research papers use homogeneous image sets such as user captured photos (Tsay et al. 2009) or a particular element from natural scenes (Gokberk et al. 2003), image sets from web communities often contain hand-drawn images, computer-generated images, the mixture between the two, artistic photographs, photographs of objects, and so on.

Quantifying image properties In order to quantify image properties, we extract 257 features from each image of each repository using FeatureExtractor (Cheamanunkul et al. 2011), a system for batch image processing which uses Python and Matlab's Image Processing Toolbox. Measurements are organized in 7 sets of image attributes, with gray scale, color, composition, lines, and texture information. The features used in each set and the respective numbers of dimensions (in parentheses) are as follows: a) Gray scale channel statistics such as mean and standard deviation (7 features); gray scale histograms for 8 and 32 bins (40 features); b) RGB channels statistics (21); RGB color histograms for 4 and 8 bins (36 features); c) HSV channels statistics (21); HSV color histograms for 4 and 8 bins (36 features); (d) Gabor features for 4 orientations (0,45,90,135 angle) and 4 sizes (original image size, 0.5, 0.25, 0.125) (16 features); (e) Spatial measurements with block difference of inverse probabilities (2x2, 3x3, 4x4) (29 features); (f) Image segments, using similarity-based color segmentation algorithm (full image, 2x2, 3x3, 4x4 blocks) (34 features); (g) Texture features calculated from Grey-Level Co-occurrence Matrix measurements (GLCM) as contrast, correlation, and energy (16 features); Sobel edge energy (1 feature). Since our approach uses distance measure to indicate similarity among samples, and some image descriptors may overwhelm others, we need to transform the data using Z-score to center and rescale the data to be in the range from 0 to 1 (Williams 2011). Early on in the project, we reviewed the variables, cleaning up the data in case of missing feature or non-valid values (some features were not defined for a few images). We analyzed features distributions using graphical techniques, which revealed high correlations among several of the values, as shown in supplemental material (Ushizima et al. 2011). This suggests that the high dimensionality of the feature space may be represented with a reduced number of dimensions, obtained by linear combinations of the original feature variables using principal component analysis (PCA), an unsupervised dimension reduction algorithm that finds a new coordinate system that maximizes the variation in the feature set. Usually, the first few PCA components contain most of the total variation in the data.

Clustering images We investigate the existence of image grouping into compact regions using the multidimensional space defined by the first PCA components. The groups are estimated by using k-means clustering, an heuristic algorithm for data partitioning via an iterative refinement scheme. This method finds a partition of the observations for a given number of clusters by minimizing the total within-group sum of squares over all variables. The number of expected partitions is an input to this algorithm. While the decision on the "optimal" number of clusters (k) is seldom easy, an approach to this problem is to evaluate the within group sum of squares for each partition by detecting a sharp variation in the resulting curve (Everitt and Hothorn 2006). This sharp variation in the within-group sum of squares curve usually occurs around $k = 10$, although we compute



(a) Art Now Flickr Group

(b) Graphic Design Flickr Group

Figure 1: ImagePlot visualizations showing subsets of randomly selected images from each Flickr group (1000 images per group), organized according to dominant PCA dimensions of color features (HSV). X-axis: PCA dimension 1; y-axis: PCA dimension 2. Note the larger dominance of black and white backgrounds and shapes in (b).

cluster partitions for k in [2,50]. In addition, we calculate the dissimilarity between the k groups based on summary statistics as the Calinski-Harabasz (CH) index (Calinski and Harabasz 1974), which relates the sum of squares among clusters (SSA) to the sum of squares within each cluster (SSW) with the equation: $SSA/SSW * (k-1)/(n-k)$, where n is the number of samples and k is the number of clusters. CH quantifies the cluster performance for a particular number of partitions and for the different set of features extracted from Art Now and Graphic Design image datasets.

Experiments

Flickr offers different resolutions for its images, and we downloaded 500 pixel versions using Flickr API. During data preparation, we eliminated invalid records, scaled and centered the data sets. After calculating PCA of each set of image attributes separately, the first five principal components accounted for approximately 50% of the total variation; they were input to the image clustering. We use ImagePlot to visualize the resulting groups, an ImageJ tool which renders high resolution scatter plots with images positioned on top of the points (Manovich et al. 2012). Fig.1 presents two visualizations of 1,000 image samples from Art Now and Graphic design groups, rendered with ImagePlot. X axis and Y axis represent the first two most significative PCA components for HSV features. Notice that the images from Art Now are closer to each other, although they also cover a wider color spectrum. This distribution suggests that the colors of Art Now images present smoother color gradients than those from images in Graphic Design. We can also see that many Graphic Design images are defined by a small number of single color areas, with the particular dominance of white and/or black areas. Visualizations which show PCA for other feature sets are available at (Ushizima et al. 2011). Fig.2 summarizes the clustering results for both Flickr groups, and indicates the dissimilarity values among the obtained clusters in terms of CH index calculated for each k. The x-axis indicates the computation of unsupervised k-means clustering for different values of k, and the y-axis shows the CH result. Each color corresponds to a set of image attributes and different symbols correspond to the different image datasets. The optimal number of classes is often taken to be the value that maximizes the CH index. The curves show that HSV attributes promote the best separation of classes, with a higher dissimilarity among groups in Art Now than in Graphic Design, which is confirmed by the dispersion of the two samples of images from those groups, as illustrated in Fig.1. Other sets of attributes showed high values of CH for Art Now as opposed to Graphic Design, suggesting that there is more variability among elements within Art Now group. The PCA, clustering and CH computations were performed using R's *stats* and *multicore* packages, along with customized code. This processing part used two 2.4 GHz Quad-Core Intel Xeon with 32GB 1066 MHz. Computing times, including saving partial results and figure generation were as follows: a) PCA was approximately 5 min and b) clustering for 49 different values of k, including CH calculation, were approximately 3 min total.

Conclusion/Discussions

We proposed a new pipeline to process, analyze, and visualize social media data. This pipeline was used to compare

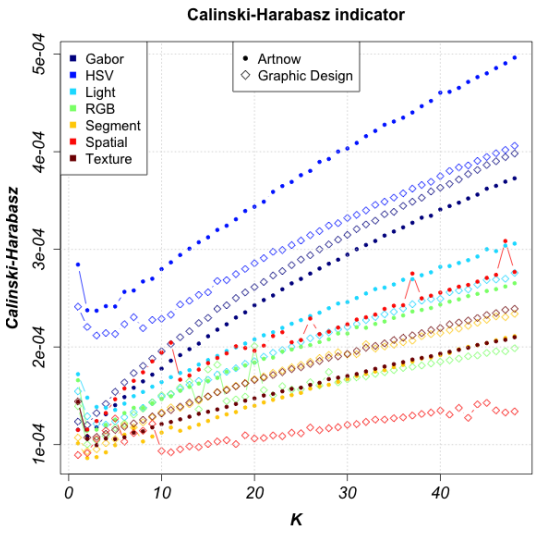


Figure 2: CH index calculated for each feature set for each k of k-means.

340,000 images from Art Now and Graphic Design groups on Flickr. This study also reflects how services such as Flickr influences artists to pool their work together. We speculate that artists often decide to join a particular group by browsing its content and doing the mental computation to understand patterns in this content, the process that we aim at imitating with our algorithms. Using automatic analysis of large image datasets, our methods allow judging quickly how diverse or homogeneous are their contents. The purpose of clustering the images in the context of the cultural analytics was to derive summary statistics as Calinski-Harabasz index to address image dispersion, according to visual features, for different groups on Flickr. Visual differences can be synthesized using sets of feature attributes, and it can be helpful in distinguishing two image sets by comparing features individually or sets of features. Image plots allow us to better understand the reasons behind these differences. The surprising overall result of our analysis is the difference between image distributions of the two groups (see Fig. 1), particularly noticing that Graphic Design group has more images with white/black background, images divided into a small number of single color areas, and the images tend to be geometrically more structured than those from Art Now group, which consists mostly of paintings, characterized by the presence of finer details. In both groups, users submitted images varying on lots of visual and semantic dimensions at the same time (style, composition,lines, textures, presence of text, faces, etc) and combinations of features not explored in this paper may cluster these image sets in different ways.

Although we expected some of these results due to our familiarity with the fields of contemporary art and graphic design, we did not know beforehand which images people upload to a certain Flickr group. The analysis of Graphic Design image set shows that a significant part of its content is more similar to the kinds of images being uploaded to Art Now group, rather than being typical examples of design one finds in professional design journals. Our pipeline gave us opportunity to learn new information. It also made possible identification of image subsets that are stylistically similar, particularly with regards to the use of color palettes as illustrated in Fig. 1. Visualization of image plots exploring image similarities using other feature sets were less useful in identifying stylistically similar images because of their visual complexity, e.g. orientation features, since most images in our groups contain patterns with more than one direction. Further improvements might include analysis and comparison of how the style of pictures (of certain communities) changes over time, how trends could be observed from analyzing the visualizations of Flickr groups, and combination of image descriptors with tag assignments. Our pipeline is general enough to be applied to other domains, and we believe that it could allow analysis of cell images with variability described in terms of textural features as a way of expressing chromatin condensation. We also plan to continue working with large cultural image sets, extending our methods to analyze evolution of cultural "landscapes" over time, with the aim to understand in details how cultural fields evolve in response to new technological, economic and social forces.

Acknowledgments

The research presented in this paper was funded by Interdisciplinary Collaboratory Grant "Visualizing Cultural Patterns" (UCSD Chancellor office, 2008-2010), Humanities High Performance Computing Award (NEH/DOE, 2009), Digital Startup level II grant (NEH, 2010), and CSRO grant Interactive Supervisualization of Large Image Collections for Humanities Research (Calit2, 2010). This research used resources of NERSC and Visualization group, which are supported by the Office of Science of the U.S. DOE under DE-AC02-05CH11231. It was also supported in part by the Applied Mathematical Science subprogram of the Office of Energy Research, U.S. DOE, under DE-AC03-76SF00098.

References

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2012. Effects of user similarity in social media. *Proc. 5th ACM Symp Web Search and Data Mining*.

Calinski, T., and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3(197):127,.

Cheamanunkul, S.; Douglas, J.; Ushizima, D.; and Manovich, L. 2011. Featureextractor. http://code.google.com/p/softwarestudies/wiki/FeatureExtractor.

Crandall, D.; Backstrom, L.; Huttenlocher, D.; and Kleinberg, J. 2009. Mapping the world's photos. *Proc. 18th Int WWW Conf.*

Everitt, B., and Hothorn, T. 2006. *A Handbook of Statistical Analyses Using R*. CRC. ISBN 1-584-88539-4.

Gokberk, B.; Irfanoglu, M. O.; Akarun, L.; and Alpaydin, E. 2003. Optimal gabor kernel location selection for face recognition. In *ICIP*, 677–680.

Gu, C., and Ren, X. 2010. Discriminative mixture-of-templates for viewpoint classication. *ECCV 2010* 6315:408–421.

Manovich, L.; Douglass, J.; Zepel, T.; and Zeng, X. 2012. Imageplot. http://lab.softwarestudies.com/p/imageplot.html.

Manovich, L. 2009. Cultural analytics: Visualing cultural patterns in the era of more media. *Domus* (923).

Siersdorfer, S., and Pedro, J. S. 2009. Ranking and classifying attractiveness of photos in folksonomies. *Proc. 18th Int WWW*.

Tsay, K.-E.; Wu, Y.-L.; Hor, M.-K.; and Tang, C.-Y. 2009. Personal Photo Organizer Based on Automated Annotation Framework. In *V Int Conf on Intelligent Information Hiding and Multimedia Sig Proc*, 507–510.

Ushizima, D.; Manovich, L.; Margolis, T.; and Douglas, J. 2011. Caf - cultural analytics in flickr. http://lab.softwarestudies.com/p/flickr-groups-analytics.html.

Williams, G. J. 2011. *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer.

Yamaoka, S.; Manovich, L.; Douglass, J.; and Kuester, F. 2011. Cultural analytics in large-scale visualization environment. *IEEE Computer* 44(12):39–48.