Exploration of Unsplash dataset

Tianyuan Zheng
UW-Madison
Madison, USA
tzheng44@wisc.edu

Weijie Chen
UW-Madison
Madison, USA
wchen376@wisc.edu

Yu Li
UW-Madison
Madison, USA
li728@wisc.edu

Zan Zhang
UW-Madison
Madison, USA
zzhang786@wisc.edu

ABSTRACT

People's preference for an image might be reflected by the image information they attached alone the image they shot or downloaded. Understanding the relationships between these tags and images can help to understand how people view the world and how different places of people view the same object differently. Unsplash dataset is utilized in this study to explore this relationship and visualize the prediction model based on the information the user-provided. Generally, people enjoy the color with black or gray style and the keyword with natural scenery all across the world. Canon, Nikon, and Sony were the top three camera brands people chose for photographing. A prediction model is developed based on this information and provides a photo recommendation. Low-dimensional visualization is also provided in this study for a fuzzy explanation of how the different labels of images related to one another. This study could deploy the data visualization to a user-friendly interface on information conveying and provide a generic method for exploring large image datasets.

INTRODUCTION

The Unsplash dataset is a large open-source dataset with over 25,000 images and the corresponding image information, including five files. The photo file contains information about the properties of the photo, the name of the contributor, the unique downloadable link for the image, and overall statistics. The keyword file contains the keyword and conversions of the photo linked. The conversion file is the largest file with how the keywords were searched and the corresponding photo downloaded according to the specific keyword. The color file is also utilized and it presents the name of the closest color as a CSS color keyword. The collection file is not utilized in this study. In this assignment, we utilize this Unsplash research dataset to explore what are the relationships between different labels of photos, establish prediction models, and execute the low-dimension visualization.

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Figure 1. A sample of geography visualization.

OVERALL GOALS AND AIMS

Our overall goal is to explore the relationships among different parameters of the images and based on the modeling of the image parameters relationships, we give a nation prediction based on the filtering model and provide photo recommendation for the corresponding image parameters. Furthermore, this project also involves low-dimensional image data processing for visualizing the image information from different perspectives. This step-by-step process provides the user an interface to explore the Unsplash dataset in a detailed as well as interesting way.

NARRATIVE

The narrative of this project includes three sections: the general information of the dataset, the prediction and photo recommendation process, and the low-dimensional dataset analysis. These three sections follow the order from easy-reading to deep-understanding, presenting the relationships among different parameters of images. Our designed data-driven narrative visualizations combine texts, interactive visualization, and animation to help people visualize people's photo preferences among different regions. The Map shows a general look of the most popular photo among different countries, and the interactive visualization allows the user to click on different counties and get the photo being displayed on the map. The second feature is collecting users' preference among color,

keyword, and camera brand to predict user's preferences fit in the most top five countries and display it by the bar chart. Also, by collecting users' preferences, we display the production of the best-fit picture based on the user's taste preferences.

IMPLEMENTATIONS AND SOLUTIONS

The data sorting, prediction model, and photo recommendation process were pre-calculated in python and four tables were generalized for the implementation in javascript. The highest occurred six keywords were extracted from the conversion file and so do the top three camera brands and top six colors. Based on the loop of these variables, we could generate a series of top five countries that publish these photos fit with the criteria. Taking the number of photos fit in the criteria divided by the total image number published in that region, we could get a rate of the possibility that the people from that region is more likely to publish the photo in the corresponding criteria. In this case, the top five countries got a sum of possibilities and the corresponding weight of the possibilities was generalized for the bar chart. After selecting a specific country, the corresponding photo fit with all criteria could be recommended for the user.

The developed HTML page starts from a general look of the featured photos all around the world. A world map presents the highest downloaded rate of photos uploaded in sixteen countries. Briefly, one of the best ways to have insights into the whole dataset is low-dimensional visualization. Thanks to the preprocessing step done by Unsplash, the coverage of each CSS-keyword color has been evaluated for each photo, by the 3rd party AI. Hence, each image is described as a color coverage vector and each element indicates the coverage of the corresponding color. After that, the PCA algorithm is used to reduce the dimension and finally, the t-SNE algorithm creates 2-d visualization. In each visual, three popular keywords are selected and the t-SNE plot is plotted to demonstrate the relationship between images with those three labels.

Briefly, one of the best ways to have insights into the whole dataset is low-dimensional visualization. Thanks to the preprocessing step done by Unsplash, the coverage of each CSS-keyword color has been evaluated for each photo, by the 3rd party AI. Hence, each image is described as a color coverage vector and each element indicates the coverage of the corresponding color. After that, the PCA algorithm is used to reduce the dimension and finally, the t-SNE algorithm creates 2-d visualization. In each visual, three popular keywords are selected and the t-SNE plot is plotted to demonstrate the relationship between images with those three labels.

DISCUSSION

Generally, it's hard to distinguish different labels via t-SNE plots. There are some potential reasons. One is that the color-coverage representation maybe loses location information, so it's a sub-optimal representation. Another potential reason is that there are not generally shared features for each category. Furthermore, the sample population for each category is near 500, which is too small.

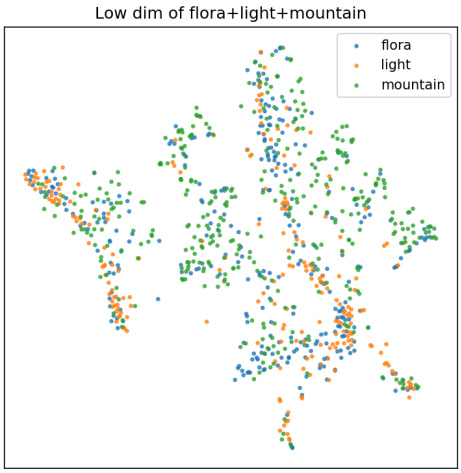

Figure 2. The t-SNE figures about keywords "flora", "lights" and "mountain".

CONCLUSION

After finishing our data visualization, we have seen that designing efficient and effective data visualization applications is a complex process. The process involves representing the data, filtering the data, processing the data, designing a visual representation, and combining all the functionality in an easy to understand and used platform. By dealing with different parameters of the images, modeling the image parameter relationships, and using low-dimensional image data processing, our data visualization allows users to explore the Unsplash dataset in an interesting way.

PAGE SIZE AND COLUMNS

On each page your material should fit within a rectangle of 7 $\times$ 9.15 inches (18 $\times$ 23.2 cm), centered on a US Letter page (8.5 $\times$ 11 inches), beginning 0.85 inches (1.9 cm) from the top of the page, with a 0.3 inches (0.85 cm) space between two 3.35 inches (8.4 cm) columns. Right margins should be justified, not ragged. Please be sure your document and PDF are US letter and not A4.

TYPESET TEXT

The styles contained in this document have been modified from the default styles to reflect ACM formatting conventions. For example, content paragraphs like this one are formatted using the Normal style.

LaTeX sometimes will create overfull lines that extend into columns. To attempt to combat this, the `.cls` file has a command, `\sloppy`, that essentially asks LaTeX to prefer underfull lines with extra whitespace. For more details on this, and info on how to control it more finely, check out http://www.economics.utoronto.ca/osborne/latex/PMAKEUP.HTM.

Title and Authors

Your paper's title, authors and affiliations should run across the full width of the page in a single column 17.8 cm (7 in.) wide. The title should be in Helvetica or Arial 18-point bold. Authors' names should be in Times New Roman or Times Roman 12-point bold, and affiliations in 12-point regular.

See `\author` section of this template for instructions on how to format the authors. For more than three authors, you may have to place some address information in a footnote, or in a named section at the end of your paper. Names may optionally be placed in a single centered row instead of at the top of each column. Leave one 10-point line of white space below the last line of affiliations.

Abstract and Keywords

Every submission should begin with an abstract of about 150 words, followed by a set of Author Keywords and ACM Classification Keywords. The abstract and keywords should be placed in the left column of the first page under the left half of the title. The abstract should be a concise statement of the problem, approach, and conclusions of the work described. It should clearly state the paper's contribution to the field of HCI.

Normal or Body Text

Please use a 10-point Times New Roman or Times Roman font or, if this is unavailable, another proportional font with serifs, as close as possible in appearance to Times Roman 10-point. Other than Helvetica or Arial headings, please use sans-serif or non-proportional fonts only for special purposes, such as source code text.

First Page Copyright Notice

This template include a sample ACM copyright notice at the bottom of page 1, column 1. Upon acceptance, you will be provided with the appropriate copyright statement and unique DOI string for publication. Accepted papers will be distributed in the conference publications. They will also be placed in the ACM Digital Library, where they will remain accessible to thousands of researchers and practitioners worldwide. See http://acm.org/publications/policies/copyright_policy for the ACM's copyright and permissions policy.

Figure 3. Insert a caption below each figure. Do not alter the Caption style. One-line captions should be centered; multi-line should be justified.

		Test Conditions	
Name	*First*	*Second*	*Final*
Marsden	223.0	44	432,321
Nass	22.2	16	234,333
Borriello	22.9	11	93,123
Karat	34.9	2200	103,322

Table 1. Table captions should be placed below the table. We recommend table lines be 1 point, 25% black. Minimize use of table grid lines.

Subsequent Pages

On pages beyond the first, start at the top of the page and continue in double-column format. The two columns on the last page should be of equal length.

References and Citations

Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets [1, 2, 7]. Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e., you provide the address for obtaining the report within your citation) and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced (e.g., "[Borriello, personal communication]").

References should be in ACM citation format: http://acm.org/publications/submissions/latex_style. This includes citations to internet resources [1, 3, 4, 9] according to ACM format, although it is often appropriate to include URLs directly in the text, as above.

SECTIONS

The heading of a section should be in Helvetica or Arial 9-point bold, all in capitals. Sections should *not* be numbered.

Subsections

Headings of subsections should be in Helvetica or Arial 9-point bold with initial letters capitalized. For sub-sections and sub-subsections, a word like *the* or *of* is not capitalized unless it is the first word of the heading.

Sub-subsections

Headings for sub-subsections should be in Helvetica or Arial 9-point italic with initial letters capitalized. Standard `\section`, `\subsection`, and `\subsubsection` commands will work fine in this template.

FIGURES/CAPTIONS

Place figures and tables at the top or bottom of the appropriate column or columns, on the same page as the relevant text (see Figure 3). A figure or table may extend across both columns to a maximum width of 17.78 cm (7 in.).

Captions should be Times New Roman or Times Roman 9-point bold. They should be numbered (e.g., "Table 1" or "Figure 3"), centered (if one line) otherwise justified, and placed beneath the figure or table. Please note that the words "Figure" and "Table" should be spelled out (e.g., "Figure" rather than "Fig.") wherever they occur. Figures, like Figure 4, may span columns and all figures should also include alt text for improved accessibility. Papers and notes may use color figures, which are included in the page limit; the figures must be usable when printed in black-and-white in the proceedings.

The paper may be accompanied by a short video figure (we recommend staying within five minutes in length). However, the paper should stand on its own without the video figure, as the video may not be available to everyone who reads the paper.

Inserting Images

When possible, include a vector formatted graphic (i.e. PDF or EPS). When including bitmaps, use an image editing tool to resize the image at the appropriate printing resolution (usually 300 dpi).

QUOTATIONS

Quotations may be italicized when *"placed inline"*.

> Longer quotes, when placed in their own paragraph, need not be italicized or in quotation marks when indented.

LANGUAGE, STYLE, AND CONTENT

The written and spoken language of SIGCHI is English. Spelling and punctuation may use any dialect of English (e.g., British, Canadian, US, etc.) provided this is done consistently. Hyphenation is optional. To ensure suitability for an international audience, please pay attention to the following:

- Write in a straightforward style.
- Try to avoid long or complex sentence structures.
- Use common and basic vocabulary (e.g., use the word "unusual" rather than the word "arcane".
- Briefly define or explain all technical terms that may be unfamiliar to readers.
- Explain all acronyms the first time they are used in your text—e.g., "Digital Signal Processing (DSP)".
- Explain local references (e.g., not everyone knows all city names in a particular country).
- Explain "insider" comments. Ensure that your whole audience understands any reference whose meaning you do not describe (e.g., do not assume that everyone has used a Macintosh or a particular application).
- Explain colloquial language and puns. Understanding phrases like "red herring" may require a local knowledge of English. Humor and irony are difficult to translate.
- Use unambiguous forms for culturally localized concepts, such as times, dates, currencies, and numbers (e.g., "1–5–97" or "5/1/97" may mean 5 January or 1 May, and "seven o'clock" may mean 7:00 am or 19:00). For currencies, indicate equivalences: "Participants were paid ₩ 25,000, or roughly US $22."
- Be careful with the use of gender-specific pronouns (he, she) and other gendered words (chairman, manpower, man-months). Use inclusive language that is gender-neutral (e.g., she or he, they, s/he, chair, staff, staff-hours, person-years). See the *Guidelines for Bias-Free Writing* for further advice and examples regarding gender and other personal attributes [10]. Be particularly aware of considerations around writing about people with disabilities.
- If possible, use the full (extended) alphabetic character set for names of persons, institutions, and places (e.g., Grønbæk, Lafreniére, Sánchez, Nguyễn, Universität, Weißenbach, Züllighoven, Århus, etc.). These characters are already included in most versions and variants of Times, Helvetica, and Arial fonts.

ACCESSIBILITY

The Executive Council of SIGCHI has committed to making SIGCHI conferences more inclusive for researchers, practitioners, and educators with disabilities. As a part of this goal, the all authors are asked to work on improving the accessibility of their submissions. Specifically, we encourage authors to carry out the following five steps:

1. Add alternative text to all figures
2. Mark table headings
3. Add tags to the PDF
4. Verify the default language
5. Set the tab order to "Use Document Structure"

For more information and links to instructions and resources, please see: http://chi2016.acm.org/accessibility. The `\hyperref` package allows you to create well tagged PDF files, please see the preamble of this template for an example.

PAGE NUMBERING, HEADERS AND FOOTERS

Your final submission should not contain footer or header information at the top or bottom of each page. Specifically, your final submission should not include page numbers. Initial submissions may include page numbers, but these must be removed for camera-ready. Page numbers will be added to the PDF when the proceedings are assembled.

Figure 4. In this image, the map maximizes use of space. You can make figures as wide as you need, up to a maximum of the full width of both columns. Note that LaTeX tends to render large figures on a dedicated page. Image: ayman on Flickr.

PRODUCING AND TESTING PDF FILES

We recommend that you produce a PDF version of your submission well before the final deadline. Your PDF file must be ACM DL Compliant. The requirements for an ACM Compliant PDF are available at: http://www.scomminc.com/pp/acmsig/ACM-DL-pdfs-requirements.htm.

Test your PDF file by viewing or printing it with the same software we will use when we receive it, Adobe Acrobat Reader Version 10. This is widely available at no cost. Note that most reviewers will use a North American/European version of Acrobat reader, so please check your PDF accordingly.

CONCLUSION

It is important that you write for the SIGCHI audience. Please read previous years' proceedings to understand the writing style and conventions that successful authors have used. It is particularly important that you state clearly what you have done, not merely what you plan to do, and explain how your work is different from previously published work, i.e., the unique contribution that your work makes to the field. Please consider what the reader will learn from your submission, and how they will find your work useful. If you write with these questions in mind, your work is more likely to be successful, both in being accepted into the conference, and in influencing the work of our field.

ACKNOWLEDGMENTS

Sample text: We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Authors 1, 2, and 3 gratefully acknowledge the grant from NSF (#1234–2012–ABC). *This whole paragraph is just an example.*

REFERENCES FORMAT

Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced (e.g., [Golovchinsky, personal communication]). References must be the same font size as other body text. References should be in alphabetical order by last name of first author. Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets. For papers from conference proceedings, include the title of the paper and the name of the conference. Do not include the location of the conference or the exact date; do include the page numbers if available.

References should be in ACM citation format: http://www.acm.org/publications/submissions/latex_style. This includes citations to Internet resources [4, 3, 9] according to ACM format, although it is often appropriate to include URLs directly in the text, as above. Example reference formatting for individual journal articles [2], articles in conference proceedings [7], books [10], theses [11], book chapters [12], an entire journal issue [6], websites [1, 3], tweets [4], patents [5], games [8], and online videos [9] is given here. See the examples of citations at the end of this document and in the accompanying `BibTeX` document. This formatting is a edited version of the format automatically generated by the ACM Digital Library (http://dl.acm.org) as "ACM Ref." DOI and/or URL links are optional but encouraged as are full first names. Note that the Hyperlink style used throughout this document uses blue links; however, URLs in the references section may optionally appear in black.

REFERENCES

[1] ACM. 1998. How to Classify Works Using ACM's Computing Classification System. (1998). `http://www.acm.org/class/how_to_use.html`.

[2] R. E. Anderson. 1992. Social Impacts of Computing: Codes of Professional Ethics. *Social Science Computer Review December* 10, 4 (1992), 453–469. DOI: http://dx.doi.org/10.1177/089443939201000402

[3] Anna Cavender, Shari Trewin, and Vicki Hanson. 2014. Accessible Writing Guide. (2014). http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/.

[4] @_CHINOSAUR. 2014. "VENUE IS TOO COLD" #BINGO #CHI2014. Tweet. (1 May 2014). Retrieved Febuary 2, 2015 from https://twitter.com/_CHINOSAUR/status/461864317415989248.

[5] Morton L. Heilig. 1962. Sensorama Simulator. U.S. Patent 3,050,870. (28 August 1962). Filed Februrary 22, 1962.

[6] Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal and Ubiquitous Computing* 18, 4 (2014), 765–766. DOI: http://dx.doi.org/10.1007/s00779-014-0773-4

[7] Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where Do Web Sites Come from?: Capturing and Interacting with Design History. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 1–8. DOI:http://dx.doi.org/10.1145/503376.503378

[8] Nintendo R&D1 and Intelligent Systems. 1994. *Super Metroid*. Game [SNES]. (18 April 1994). Nintendo, Kyoto, Japan. Played August 2011.

[9] Psy. 2012. Gangnam Style. Video. (15 July 2012). Retrieved August 22, 2014 from https://www.youtube.com/watch?v=9bZkp7q19f0.

[10] Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing*. ERIC, Bloomington, IN, USA.

[11] Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.

[12] Langdon Winner. 1999. *The Social Shaping of Technology* (2nd ed.). Open University Press, UK, Chapter Do artifacts have politics?, 28–40.