Constructing an Anonymous Dataset From the Personal Digital Photo Libraries of Mac App Store Users

Jesse Prabawa Gozali[1,2*] Min-Yen Kan[1,2] Hari Sundaram[3]
[1]Department of Computer Science, National University of Singapore, Singapore
[2]NUS Interactive and Digital Media Institute, Singapore
[3]Arts Media & Engineering, Arizona State University, USA
{jprabawa, kanmy}@comp.nus.edu.sg hari.sundaram@asu.edu

ABSTRACT

Personal digital photo libraries embody a large amount of information useful for research into photo organization, photo layout, and development of novel photo browser features. Even when anonymity can be ensured, amassing a sizable dataset from these libraries is still difficult due to the visibility and cost that would be required from such a study.

We explore using the Mac App Store to reach more users to collect data from such personal digital photo libraries. More specifically, we compare and discuss how it differs from common data collection methods, *e.g.* Amazon Mechanical Turk, in terms of time, cost, quantity, and design of the data collection application.

We have collected a large, openly available photo feature dataset using this manner. We illustrate the types of data that can be collected. In 60 days, we collected data from 20,778 photo sets (473,772 photos). Our study with the Mac App Store suggests that popular application distribution channels is a viable means to acquire massive data collections for researchers.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Measurement, Human Factors

Keywords

Personal digital library, Photography, Data collection, Ground truth, Crowd-sourcing

*This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

1. INTRODUCTION

Researchers in personal photo digital libraries (DLs) require access to such DLs to conduct their studies. For example, works on photo summarization [10], photo stream alignment [11], automatic albuming [9], and event photo stream segmentation [3] require various features and ground truth annotations from DLs. Acquiring such personal data, however, tends to be a challenging process – especially when sizable data is desired. Common methods to obtain photos, such as from volunteers or from study participants, do not scale well to thousands of photo sets due to the remuneration costs and the limited reach that study advertisement has in gathering interested participants.

To collect ground truth annotations on such collected photos, even more human effort is required. For example, in automatic albuming, the ground truth is the true grouping of photos into separate events. In some works, the authors themselves produced the ground truth [9] or external annotators were employed [8], which may be problematic due to unfamiliarity, bias, or ignorance of events that transpired in the photos. The semantics associated with personal photos render these tasks difficult to annotate by parties not privy with the context of the photos. In our own experience with event photo stream segmentation [3], we find that the average agreement of external annotators to be unsatisfactory.

For such reasons, studies often require that the photo owners themselves produce the ground truth [7]. Data collection thus involves both 1) accessing the DLs, and 2) acquiring the efforts of the photo owners themselves to produce the ground truth annotations. These two issues exacerbate the difficulty in scaling up the data collection process. We encountered this exact problem in building upon our previous work on event photo stream segmentation.

We propose using the popular application distribution channel, the Mac App Store (hereafter, MAS), to alleviate issues with cost and reaching potential study participants. We use our own research needs as a case study to explore using the MAS as a platform to acquire the needed data. To the best of our knowledge, this is the first study to explore collecting anonymous data from personal digital photo libraries at a large scale, *i.e.* our data collection application was downloaded by over 2,500 users in 60 days.

The contributions of our study is two-fold. First, we report and discuss our experiences with the design of the data collection application, timeline, visibility, and cost in using the MAS in Section 2. Secondly, we present the large collected data to the research community in Section 3, providing an in-depth analysis of a few pertinent features.

2. DATA COLLECTION

The goal of our study is to explore the MAS for data collection in personal digital photo libraries. Primarily, we were motivated by its large user base: on Jan 7th, 2011, after only 24 hours of being available, the MAS had received over one million downloads[1]. We hypothesize that with its large user base in multiple countries, using the MAS will increase the visibility of our study and thus yield more collected data.

To facilitate the study, we used our own data collection needs. In our previous work[3], we developed an event photo stream segmentation method that relied on a dataset of photo features to compute smoothing weights and parameters. We also collected the ground truth segmentation from these sets to evaluate against baselines: between each pair of consecutive photos, we need to know if a segment boundary exists. To obtain better smoothing weights and parameters, we need to collect features from a larger dataset. Additional ground truth segmentations would also expand our evaluation and strengthen the validity of the results.

Design. With any data collection method, a means for the collection needs to be designed and created. Even when the data to collect is small in scale, researchers still need to create a way to collect the data (*e.g.* from the volunteers) and a way for annotators to provide ground truth (*e.g.* for parameter tuning, supervised learning, or evaluation). When large-scale data collection is necessary, other scaling issues arise. For example, with crowd-sourcing platforms like Amazon Mechanical Turk (MTurk), recent works [6] have noted that verification questions or qualification task is necessary to ascertain if the annotators are suited for the task. Results also often have to be filtered to remove fake data from cheating crowd-sourcing users [1].

For the MAS, its review guidelines outline very specific *functionality requirements* for any application it distributes. One of the requirements states that applications *"that are not very useful"* may be rejected. As such, in the design of our application, we needed to relegate the data collection to a secondary function. While this seems counter-intuitive, we argue that generally, the data is collected to ultimately serve some practical purpose for the users; this purpose is a natural fit as the primary function of the application.

In our case, we need the data to improve our event photo stream segmentation algorithm. In our previous study on chapter-based photo organization [4], we developed Chaptrs, a photo browser that utilizes the algorithm to automatically group users' event photos into chapters. Thus for the MAS, we improved upon Chaptrs from the feedback and results of our previous study. At the same time, we can use Chaptrs for data collection, *i.e.* as a secondary function.

When Chaptrs is launched for the first time, a window appears and explains how the automatic segmentation works and then appeals to the user to participate in the study to help improve the algorithm (Figure 1). Participation is voluntary and opt-in, but we entice users by stating that a future improved algorithm would be provided exclusively to participants. We also explained that the data is anonymized and the 60-day study was approved by our Institutional Review Board, as described in detail in a provided hyperlink.

[1] `http://apple.com/pr/library/2011/01/07macappstore.html`

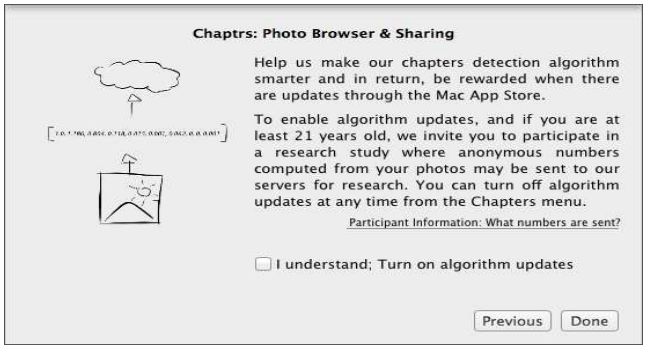

Figure 1: Window (sheet) inviting users to participate in a study to help improve our algorithm

Cost. Currently, there is no mechanism in the Mac Software Development Kit (SDK) to allow MAS developers to send money to their users, and thus we opted not to remunerate participants monetarily. This reduces the overall cost of the study as it no longer grows with the number of participants. At the same time, the participants are more likely to be users who are genuinely interested in helping to improve the algorithm so they can benefit from the future algorithm, unlike many crowd-sourced users who may cheat.

We made Chaptrs a free application to maximize number of downloads. All the cost in the study is then attributed to the Mac Developer Program annual fee of 99 USD. Past works with MTurk [6] have reported paying about 0.02 USD per annotationm on top of the 60.50 USD Amazon fee, while others paid 0.10 USD per translation (Urdu into English) [1]. For data collections that involve no human judgement or annotation, *e.g.* collecting Short Message Service (SMS) messages, recent work [2] reported paying at most 0.01 USD per message.

In our case, we collected features from 20,778 photo sets, comprising of 473,772 photos, of which 60 sets have ground truth segmentations, comprising of 8,107 photos. This translates to 0.0002 USD per photo, or if we attribute all the cost to the collected annotations, 0.012 USD per annotation[2].

When we consider the first 19 days of the study — the time taken by [6] to collect 2,500 annotations from MTurk — we collected 5,787 photo sets, comprising of 227,969 photos, of which 23 sets have ground truth segmentations, comprising of 4,559 photos. This translates to a similar cost of 0.02 USD per annotation, but without any other additional fees.

This illustrates another difference between our study and existing data collection methods. Because the cost of our study does not scale with the amount of data collected, the cost per collected data (*e.g.* photo or annotation) decreases with the duration of the study and with the number of concurrent studies.

Visibility. We define visibility as the exposure obtained by Chaptrs to MAS users. This includes both MAS users who downloaded Chaptrs and those who did not. While visibility is difficult to ascertain, we can produce a lower bound by determining the number of MAS users who downloaded Chaptrs. The daily number of downloads for the study can be seen in Figure 2, where the best-matching

[2] *i.e.* whether there are segment boundaries in the pairs of consecutive photos in a photo set

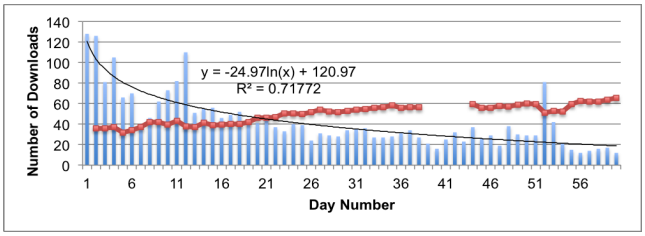

Figure 2: Daily number of downloads (columns) with trendline and average rankings (line) for Chaptrs in the 60 days of study

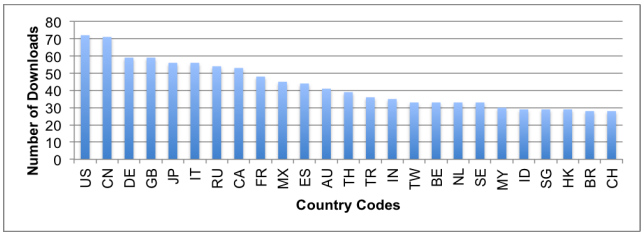

Figure 3: Top 25 Countries with Highest Number of Downloads

trendline is logarithmic: $y = -24.97ln(x) + 120.97$ with a coefficient of determination, $R^2 = 0.71$, where y and x correspond to the number of downloads and the day number respectively.

We note that there are two anomalous spikes in the number of downloads on Days 12 and 52. Both spikes are attributed to the unusually high number of downloads in the Japan MAS on those days (47 and 38). These high number of downloads are caused by a snowballing effect from CHAPTRS taking the number 2 and 4 positions in the top photography category in the Japan MAS. The line graph in Figure 2 plots the average photography category ranking for Chaptrs among various MAS stores. We can observe that the ranking decays linearly with time. Figure 3 shows a time series plot of the top 25 countries with the highest number of downloads. This ranking shows relative market sizes that would be useful for planning pilot studies.

As CHAPTRS is a free application, one tendency is for users to download and delete the application after only a brief experience. This is undesirable especially if the data collection is meant to contribute to a longitudinal study. To estimate the percentage of deletions, we submitted an update to the MAS. As the MAS only notifies updates to users with the application still installed, this gives us a good estimate. The update was released on Day 50 (see Figure 4). Comparing the number of downloads in the first 49 days (2,261) and the number of updates in the last 11 days (2,226), we can estimate that there is only at most a 1.5% deletion rate.

Timeline. It took 19 days to collect 23 photo sets with ground truth annotations, comprising of 4,559 photos. In the same amount of time, [6] collected 2,500 music mood annotations using MTurk. The work on SMS collection [2], which was considerably simpler as it involved no annotations from contributors, reported less success with 43 submissions (over 200 SMS per submission on average) over 40+ days.

We note that there is some temporal overhead with using

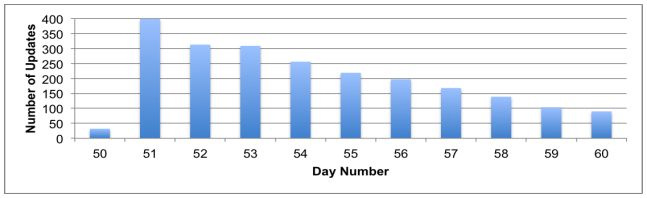

Figure 4: Number of updates from Day 50 to 60

MAS as a distribution channel. This is because applications need to undergo a review process before it becomes available for download. The review time fluctuates over time and usually takes 1-2 weeks[3]. Additional time is required for resubmission if the application is rejected.

3. DATA SET

While there are publicly available datasets (*e.g.* Corel), there are none that are event photos from personal photo libraries. We have previously noted that researchers have so far made use of their own collections to conduct studies. This poses a hurdle for new researchers. In practice, producing a public dataset of personal photos is challenging due to the private nature of the photos and their semantics.

We believe that a compromise is possible. The data we collected is a "blind" dataset of personal photos because the photos themselves are not in the dataset. Instead, only anonymized photo features and annotations are contained[4].

The dataset currently contains features that we use for our own work on event photo stream segmentation: time gap, focal length, aperture diameter, logLight, and an 8-bin color histogram, but can be easily extended to collect others. Here we provide some brief analysis of the data set, details of which are packaged with the dataset[5].

Using k-means, we clustered the color distributions and searched for an optimal value for k, $k < 9$, which was found to be 6. Figure 5 shows the color distributions of the cluster centroids. We observe that there is a large percentage of black in all clusters due to the binning of dark colors to the nearest color, black. We also observe that Cluster 2 represents the blue/cyan photos while the red/yellow photos are represented by Cluster 3. These two clusters thus show the color distribution of the "blue/cyan" photos and "red/yellow" photos in the dataset. The other three clusters seem to represent different ratios of white to black while the ratios of the remaining 6 colors remain fairly constant.

We also analyzed for bursts of photo taking activity [5], *i.e.* a sequence of photos (> 2) taken in succession with a certain average time gap. In our analysis, we looked for 15 kinds of bursts, each with a different average time gap[6]. Figure 6 shows the number of bursts found and the average number of photos for each kind of burst. We observe that the most frequent burst has an average time gap of 9s. Also, the burst with the lowest average time gap in our analysis has

[3]Trend is reported at reviewtimes.shinydevelopment.com

[4]An externally-hosted web application was used to collect the data. As the incoming traffic was rerouted through its servers, the origin IP address is unknown.

[5]`http://wing.comp.nus.edu.sg/~jeprab/chaptrs_dataset/`

[6]While photos taken > 1 min apart can hardly be considered a burst, we analyze such "bursts" for completeness

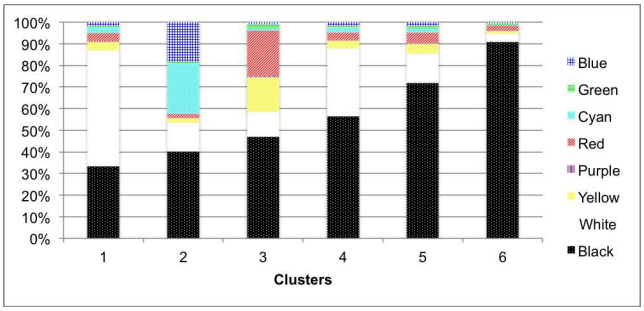

Figure 5: Color distributions of the six cluster centroids in the dataset

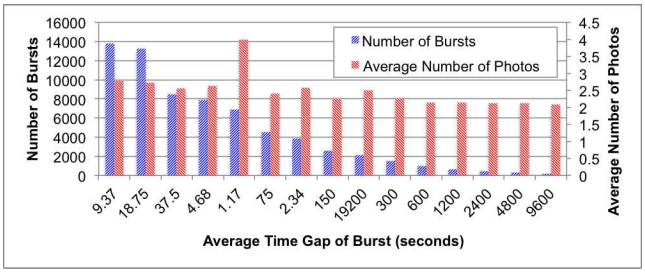

Figure 6: Dataset statistics of photo taking bursts

the highest average number of photos. This suggests that when people take photos in quick succession (∼1s), they do so with 4 photos on average.

Figure 7 shows a histogram of LogLight values. Fitting a two-mixture Gaussian to the histogram ($\mu = \{-4.91, -1.47\}$, $\sigma = \{0.74, 2.35\}$, $\lambda = \{0.26, 0.74\}$) yields results that suggest that the LogLight values correspond to two normal distributions, that plausibly represent day (left mixture) and night (right mixture). LogLight values are small (large) for high (low) levels of ambient light, respectively.

4. CONCLUSION

There is a lack of publicly available datasets for personal photos. We believe that the challenge lies is in the issue of privacy and in the difficulty in collecting sizable data. In this paper, we have demonstrated how such a dataset can be constructed by collecting anonymous photo features and ground truth annotations using an application distributed through the Mac App Store.

Aside from the review time overhead and conceptual overhead of designing the data collection application, we have demonstrated that the MAS with its large user base allows ChaptRS to achieve high number of downloads, collects data at a faster rate and with lower cost than the data collection experiences from some recent works.

Ultimately, there is a self-filtering process because only genuinely interested users would volunteer to participate in the studies. This is in contrast with other data collection means, *e.g.* crowd-sourcing platforms where some users may only be interested in the monetary remunerations.

We note that in the works that we have reviewed, the types of data and annotations collected are very different and thus we should not discount the possibility of confounding variables affecting our comparisons. Nonetheless, our experiences with ChaptRS shows that the MAS provides a

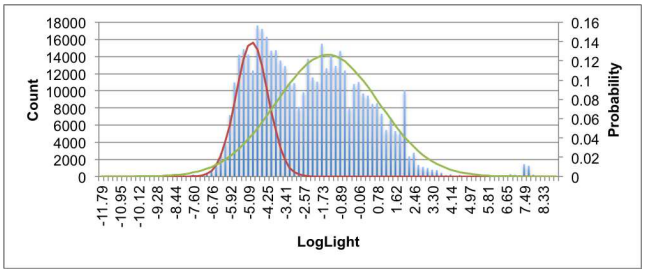

Figure 7: Histogram of LogLight values and the estimated Gaussian mixtures. The probabilities of the mixtures have been multiplied by their mixture ratios (0.26, 0.74) to aid with the visualization.

fruitful and viable alternative for data collection especially in reaching out to personal digital photo libraries. In future work, we can adjust ChaptRS to collect a different set of anonymous features from the photos to expand on the dataset and its analysis.

5. REFERENCES

[1] M. Bloodgood and C. Callison-Burch. Using Mechanical Turk to build machine translation evaluation sets. In *Proc. NAACL 2010 Workshop on AMT*, 2010.

[2] T. Chen and M.-Y. Kan. Creating a live, public short message service corpus: The NUS SMS Corpus. *Language Resources and Evaluation*, pages 1–37, Aug. 2012.

[3] J. P. Gozali, M.-Y. Kan, and H. Sundaram. Hidden Markov Model for event photo stream segmentation. In *Proc. ICME 2012 Workshop on HFC3D*, 2012.

[4] J. P. Gozali, M.-Y. Kan, and H. Sundaram. How do people organize their photos in each event and how does it affect storytelling, searching and interpretation tasks? In *Proc. JCDL*, pages 315–324, 2012.

[5] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. KDD*, pages 91–101, 2002.

[6] J. H. Lee and X. Hu. Generating ground truth for music mood classification using Mechanical Turk. In *Proc. JCDL*, pages 129–138, 2012.

[7] A. C. Loui and A. E. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Trans. Multimedia*, 5(3):390–402, September 2003.

[8] A. Pigeau and M. Gelgon. Spatial-temporal organization of one's personal image collection with model-based ICL clustering. In *Proc. CBMI*, 2003.

[9] J. C. Platt. AutoAlbum: Clustering digital photographs using probabilistic model merging. In *Proc. CAIV*, pages 96–100, 2000.

[10] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *Proc. ICMR*, 2012.

[11] J. Yang, J. Luo, J. Yu, and T. Huang. Photo stream alignment and summarization for collaborative photo collection and sharing. *IEEE Trans. Multimedia*, 14(6):1642–1651, Dec. 2012.