Reconstructing the World* in Six Days
*(As Captured by the Yahoo 100 Million Image Dataset)

Jared Heinly, Johannes L. Schönberger, Enrique Dunn, Jan-Michael Frahm
The University of North Carolina at Chapel Hill

For decades, modeling the world from images has been a major goal of computer vision. One of the most diverse data sources for modeling is Internet photo collections, and the computer vision community has made tremendous progress in large-scale structure-from-motion (LS-SfM) from Internet datasets over the last decade. However, utilizing this wealth of information for LS-SfM remains a challenging problem due to the ever-increasing amount of image data on the Internet. In a short period of time, research in large-scale modeling has progressed from modeling using several thousand images [5, 6] to modeling from city-scale datasets of several million [3]. Major research challenges that these approaches have focused on are:

- **Data Robustness**: Enable the modeling from unorganized and heterogeneous Internet photo collections.
- **Compute & Storage Scalability**: Achieve efficiency to meet the true scale of Internet photo collections.
- **Registration Comprehensiveness**: Identify as many camera-to-camera associations as possible.
- **Model Completeness**: Build 3D scene models that are as extensive and panoramic as possible.

In practice, these goals have been prioritized differently by existing LS-SfM frameworks [2, 3, 5, 6]. In this work, we propose a novel structure-from-motion framework that advances the state of the art in scalability from city-scale modeling to world-scale modeling (several tens of millions of images) using just a single computer. Moreover, our approach does not compromise model completeness, but achieves results that are on par or beyond the state of the art in efficiency and scalability of LS-SfM systems. In order to achieve a balance between registration comprehensiveness and data compactness, we employ an adaptive, online, iconic image clustering approach based on an augmented bag-of-words representation. The new image cluster representation overcomes several limitations of previous representations, which tended to partition images of the same scene into multiple independent models. In achieving more large-scale scene integrity, our novel cluster representation also avoids needlessly increasing the size of the indexing structure, which previously prohibited the use of datasets of tens of millions of images. We demonstrate this scalability of our framework by performing 3D reconstructions from the 100 million image world-scale Yahoo Flickr Creative Commons dataset [1, 7]. Our method reconstructs models from a world-scale dataset on a single computer in six days leveraging approximately 96 million images (see examples in Figure 1).

Our framework achieves this level of scalability by adopting a streaming-based paradigm for connected component discovery, where images are loaded and processed in a sequential fashion. Moreover, given the constantly increasing size of available photo collections, we posit streaming-based processing as a natural compute paradigm for world-scale structure-from-motion (WS-SfM). Our proposed streaming imposes the constraint on the processing that, in one pass through the data, an image is only loaded once from disk (or other input source) and the image is discarded after a limited period of time (much smaller than the overall computation time). The major challenge posed by stream processing for image overlap detection is to ensure that overlap is detected even when the images are not concurrently loaded. To meet these constraints, we propose to maintain and update in realtime a concise representation of our current knowledge of the images' connectivity. Upon discovering the sets of connected images (referred to as connected components), we then perform incremental SfM to recover the 3D geometry of the scenes contained within the dataset.

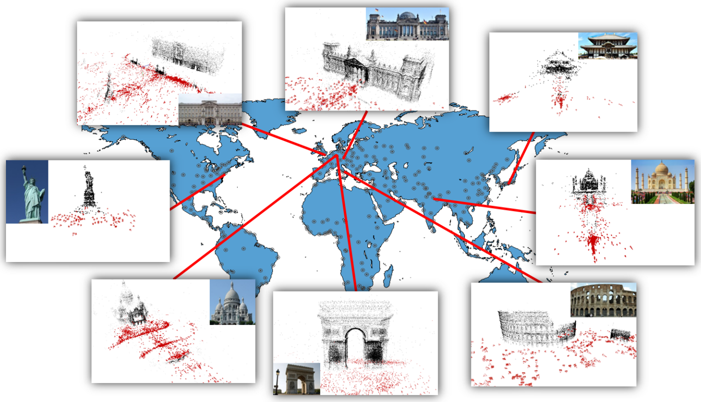

Figure 1: Examples of our world-scale reconstructed models.

Dataset	Number of Images				Time (hours)	
	Input	Registered	Iconics	SfM	Stream	SfM
Roman Forum [4]	74,388	45,341	3,408	20,176	0.35	0.79
Berlin [3]	2,704,486	702,845	42,612	194,870	7.89	6.72
Paris	10,390,391	2,492,310	131,627	858,134	29.16	68.85
London	12,327,690	3,078,303	228,792	566,778	38.29	34.86
Yahoo [1]	96,054,288	1,499,110	74,660	168,099	105.4	20.9

Table 1: Statistics for our tested datasets. Iconics are for clusters of size ≥ 3, and the SfM results report on the 32 largest components (or components with ≥ 50 images for the Yahoo dataset).

To test our approach, we ran our method on datasets of widely varying sizes (see Table 1), using a single computer to process each dataset. For comparability, we leveraged two existing datasets (Roman Forum [4] and Berlin [3]), on which we achieve results that exceed previous works in terms of completeness and efficiency. To demonstrate the true world-scale processing of our approach, we processed 96 million images spanning the globe from the Yahoo Flickr webscope dataset [1, 7]. The processing time was approximately 5.26 days, and our pipeline is the first system to be able to process and reconstruct from such a diverse, world-scale dataset. Example models are shown in Figure 1 and the detailed statistics are provided in Table 1. This clearly demonstrates the scalability of our newly proposed reconstruction system, which enables us to reconstruct the world in six days on a single computer.

[1] Yahoo! webscope. 2014. yahoo! webscope dataset yfcc-100m. http://labs.yahoo.com/Academic_Relations.

[2] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S.M. Seitz, and R. Szeliski. Building Rome in a Day. *Comm. ACM*, 2011.

[3] J.M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a Cloudless Day. *ECCV*, 2010.

[4] Y. Lou, N. Snavely, and J. Gehrke. MatchMiner: Efficient Spanning Structure Mining in Large Image Collections. *ECCV*, 2012.

[5] N. Snavely, S.M. Seitz, and R. Szeliski. Photo Tourism: Exploring Photo Collections in 3D. *SIGGRAPH*, 2006.

[6] N. Snavely, S.M. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 2007.

[7] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The New Data and New Challenges in Multimedia Research. *arXiv:1503.01817 [cs.MM]*, 2015.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.