Automatic Photo-from-Panorama for Google Maps

Sema Berkiten
Google, Inc.

Rosália G. Schneider
Google, Inc.

Jared M. Johnson
Google, Inc.

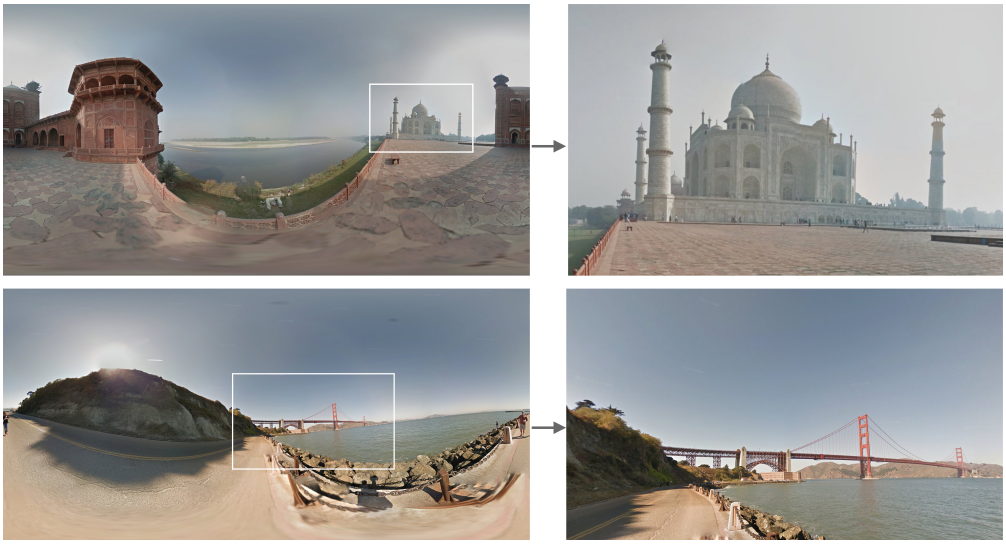

Figure 1: A full 360^o panorama (left) holds a lot of information about a scene, but it is not informative when seen as a picture or thumbnail. Our pipeline selects a meaningful portion to be shown in these situations (right).

ABSTRACT

We introduce a technique for extracting interesting photographs from 360^o panoramas. We build on the success of convolutional neural networks for classification to train a model that scores a given view, using this score to find a best view. Training data for this classification model is generated automatically from landmark detections within Street View panoramas. We validate that our selected views are often preferred over manually chosen ones and have experienced an increase in user interaction when automatically selected views are shown on Google Maps.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**;

KEYWORDS

classification, scene analysis, panorama

ACM Reference Format:
Sema Berkiten, Rosália G. Schneider, and Jared M. Johnson. 2018. Automatic Photo-from-Panorama for Google Maps. In *Proceedings of SIGGRAPH '18 Talks.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3214745.3214802

SIGGRAPH '18 Talks, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5820-0/18/08.
https://doi.org/10.1145/3214745.3214802

1 INTRODUCTION AND RELATED WORK

With the development of automatic panorama generation in modern cameras, special-purpose panorama cameras, and sophisticated panorama stitching software, increasingly many users have begun capturing 360^o views of a scene. Compared to traditional photography, this kind of imagery can provide a more immersive experience, as it allows the viewer to look in any direction. However, in many contexts, 360^o panoramas have to be shown as regular, static pictures, e.g. thumbnail previews in Google Maps. Viewing all of a 360^o panorama at once results in an unfamiliar, distorted image that is frequently not representative of the scene (left on Figure 1).

Our task in this work is to find the best possible photograph inside a panorama, to be used in situations that require a traditional, 2D representation (right on Figure 1). This is equivalent to finding camera parameters (**yaw**, **pitch** and **field-of-view**) that can be used to project the image in 2D.

We define a panorama view as ***good*** if it has two main components: First, we want our view to be representative of the panorama, *i.e.* we want the camera to be pointed at the most important or interesting elements. Second, we want the final composition to be aesthetically pleasing. Our algorithm follows these two aims directly: First, we use our trained Convolutional Neural Network (CNN) to find a representative view. Second, we refine the composition of this view through cropping to produce aesthetically pleasing final output.

Our contributions are development and validation of a pipeline for automatic best view selection in panoramas. While methods exist for interestingness [Dhar et al. 2011; Isola et al. 2014] and cropping of images [Fang and Zhang 2017; Zhang et al. 2014], we apply

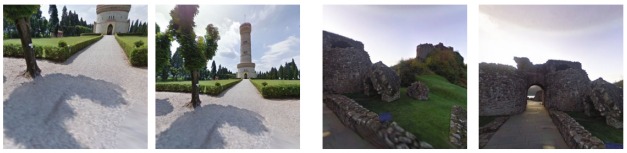

Figure 2: Photographs generated from panoramas with our pipeline (right side of each view pair) vs the manually chosen ones (left side of each view pair).

these techniques in the context of panoramas. Note that panoramas look distorted when projected in 2*D*, such that we cannot apply the same techniques directly. We have also shown the effectiveness of our approach for automatically generating large-scale training data for defining interestingness within panoramas.

2 METHOD

2.1 Data Collection

For training, we need a considerable number of labeled examples for each class. We have approximately 12, 500 **manually** annotated panoramas, which we will use for validating our results, but this is not enough for training our CNN. As such, we had to find data that could be automatically collected to use as ground truth in training.

Views with landmarks. Landmarks are places in the world which are the most photographed: monuments like the Eiffel Tower and the Great Wall of China, but also natural wonders like the Matterhorn. By relying on automated landmark detections within Street View panoramas, we obtained the volume of training data we needed and a reasonable assurance that the positive training data was an interesting view within the panorama. We used a random view within the same panorama to designate a negative example.

Landmark detections allowed us to generate hundreds of thousands of positive and negative examples, without relying on human labelings. We found that these detections generalized well to panoramas without landmarks. Meaning, the most interesting views in panoramas *look* more like landmarks than less interesting views.

2.2 Algorithm

To find the best possible photograph inside a panorama, we propose a two-step algorithm where we first roughly find the most interesting area in a panorama and then refine the result.

Classification. To leverage advances in CNNs, we pose the problem of finding the best view inside a panorama as finding a quality score for any possible view. We train a binary classifier between two classes: ***good view*** and ***bad view***. In this case, the quality score we are looking for will be the probability of a given view belonging to the ***good*** category. We extract a number of views with wider field-of-view than our target photograph from the panorama and feed them to a CNN, which gives back a quality score for each view. Our selected photograph will be the best scoring view. More specifically, we select the set of views to analyze with fixed field-of-view of $120^o \times 80^o$, with 8 views spaced 45^o apart along the horizon. Even though using a wider field-of-view than our target photograph does hurt the accuracy slightly (only around 5^o), it gives us extra freedom to fine-tune the composition which leads to a better end result.

Table 1: User preference (%) on the different views.

	(1)	(2)	No Preference
(1) Manual vs. (2) Ours	24.6%	19%	56.4%
(1) Manual vs. (2) Random	29%	17%	54%
(1) Ours vs. (2) Random	23.9%	17.5%	58.6%

Refinement. In the second step of our algorithm, given the best part of a panorama with a wider view than our final target photograph, we rely on an automated cropping technique described in [Fang and Zhang 2017]. This step improves the view chosen by classification because the classifier cannot score every yaw, pitch, and field-of-view combination efficiently. Also, the automatically generated training data does not include a ground-truth field-of-view. Consequently, after applying auto-cropping, the coarse ultra-wide angle best view from the classifier is refined into an wide-angle, aesthetically-pleasing composition, with a precise yaw and pitch optical axis, and a well-defined field-of-view.

3 RESULTS

Comparison to Human Selected Views. The first evaluation we performed was measuring the angle distance between our best view and a human-chosen one. We compared **pitch** and **yaw** separately, using only the classification method with the target field-of-view and the full pipeline. For the classification method, average angular distance is around 69^o for yaw and 11^o for pitch while for full method it is around 73^o for yaw and 7^o for pitch.

It is expected that cutting bigger panoramas at coarser intervals will make the results worse quantitatively however we do believe it improved the results qualitatively, such as avoiding cutting the tops of some buildings, as shown in Figure 2.

Human Evaluation. As a more rigorous evaluation, we showed users 2 viewcodes and asked them which one they preferred. This helps identify cases where the ground truth was bad, or in where more than one ground truth would be acceptable.

We performed 3 different types of comparisons: (a) Our method vs. Ground truth (b) Our method vs. Random (c) Ground truth vs. Random. Three different users evaluated each result (all panoramas in the test data set were evaluated). Our results are shown in Table 1. The users were instructed to choose **No preference** unless one of the views was clearly better than the other.

These results show that our method performs significantly better than randomly choosing a view, while still not quite as well as humans. Since we cannot have humans select views for the billions of panoramas we have collected, this completely-automatic pipeline has proven invaluable.

REFERENCES

S. Dhar, V. Ordonez, and T.L. Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 1657–1664. https://doi.org/10.1109/CVPR.2011.5995467

Hui Fang and Meng Zhang. 2017. Creatism: A deep-learning photographer capable of creating professional work. *CoRR* abs/1707.03491 (2017). arXiv:1707.03491 http://arxiv.org/abs/1707.03491

P. Isola, Jianxiong Xiao, D. Parikh, A. Torralba, and A. Oliva. 2014. What Makes a Photograph Memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36, 7 (July 2014), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and N. Sebe. 2014. Weakly Supervised Photo Cropping. *Multimedia, IEEE Transactions on* 16, 1 (Jan 2014), 94–107. https://doi.org/10.1109/TMM.2013.2286817