Learning Deep Representations for Ground-to-Aerial Geolocalization

Tsung-Yi Lin[†], Yin Cui[†], Serge Belongie[†], James Hays[§]
[†]Cornell Tech [§]Brown University

The recent availability of geo-tagged images and rich geospatial data has inspired a number of algorithms for image based geolocalization. Most approaches predict the location of a query image by matching to ground-level images with known locations (e.g., street-view data). However, most of the Earth does not have ground-level reference photos available. Fortunately, more complete coverage is provided by oblique aerial or "bird's eye" imagery. In this work, we localize a ground-level query image by matching it to a reference database of aerial imagery Fig. 1. We use publicly available data to build a dataset of 78K aligned cross-view image pairs. The primary challenge for this task is that traditional computer vision approaches cannot handle the wide baseline and appearance variation of these cross-view pairs. We use our dataset to learn a feature representation in which matching views are near one another and mismatched views are far apart. Our proposed approach, *Where-CNN*, is inspired by deep learning success in face verification and achieves significant improvements over traditional hand-crafted features and existing deep features learned from other large-scale databases. We show the effectiveness of Where-CNN in finding matches between street view and aerial view imagery and demonstrate the ability of our learned features to generalize to novel locations.

For the experiments in this paper, we collect Google street-view and 45° aerial view images from seven cities – San Francisco, San Diego, Chicago, and Charleston in the United States as well as Tokyo, Rome, and Lyon to test how well our method generalizes to novel locations. The data includes both urban and suburban areas for each city.

Before we attempt cross-view matching, a major challenge is simply to establish ground-truth correspondences between street-level and aerial views. This is challenging due to unknown geolocation, scale, and surface orientation in the raw images. For our study, we show that pairs of coarsely aligned cross-view images can be generated with the depth estimates and meta data of street-view images provided by Google [1]. We calibrate the scale, orientation, and geolocation of street-view depth estimates and aerial imagery then reproject street-view images on aerial image plane. For our experiments, we generate cross-view image pairs with area of 15×15 meters (256×256 pixels). Fig.3 shows a snapshot of cross-view image pairs.

There are no "standard" feature representations for the ultra-wide baseline cross-view matching, because it is a relatively unexplored task. In this paper, we investigate three types of feature representations: (1) hand-crafted features; (2) generic deep feature representations; and (3) learned deep feature representations. Inspired by the early "Siamese Network" [2] approach and the more recent DeepFace [3] and Deep Ranking [4] methods, we use a pair-based network structure to learn deep representations from data for distinguishing matched and unmatched cross-view image pairs.

During training, the input to the network is a pair of images $x \in X$ and $y \in Y$, where X and Y are street-view imagery and aerial view imagery in the training set, respectively. The input pair x and y are fed into two deep convolutional neural networks (CNN) A and B , which have same architecture. Fig 2 shows the network architecture for training.

The goal of our convolutional network is to learn a feature representation (non-linear embedding) $f(.)$ that embed raw input images $x, y \in \mathbb{R}^n$ from different views to a lower dimensional space as $f_A(x), f_B(y) \in \mathbb{R}^d$ where images from matched pairs are pulled closer whereas images from unmatched pairs are pushed far way from each other. After learning the parameters of the deep network which produces our feature representation $f(\cdot)$, we can pre-compute $f(y)$ offline for all $y \in Y$, where Y is our aerial imagery. Then, for a given query x, we can rank $f(y)$ by the pair distance $\|f(x) - f(y)\|_2$.

Our experiments compare the effectiveness of the feature representation learned from our database using a deep convolutional network against traditional hand-crafted features and deep features not optimized for our problem. We generate a dataset contains 1M image pairs, with ratio 1 (matched) to 20 (unmatched). Our method achieves 43.6 Average precision (AP) compared to 11.3 for generic deep learned feature (ImageNet-CNN), 7.6 for hand-crafted feature (HoG2x2), and 4.8 for random. In addition, we measure how well a representation learned on some cities generalizes to testing on an unseen city. Please visit our poster for more details in both qualitative and quantitative results in geolocalization.

We have presented the first general technique for the challenging problem of matching street-level and aerial view images and evaluated it for the task of image geolocalizaiton. While standard keypoint matching or bag-of-words approaches barely outperform chance, our learned representations show promise.

While we train and test on cross-view pairs that have been roughly aligned according to aerial and street-view metadata, a limitation of the current approach is the need to estimate scale and dominant depth at test time for ground-level queries with no metadata. This is plausible either through manual intervention or automatic estimation. Another limitation is that the absolute orientation of a ground-level query could be unknown (and difficult to estimate) and would require a sweep over orientations at test time.

Figure 1: Given a query street-view image, this paper aims to find where it was taken by matching it to a city-scale aerial view image database.

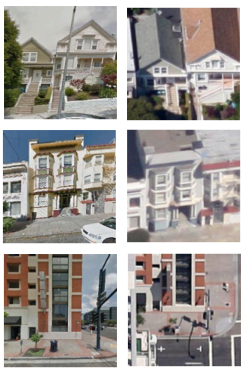

Figure 2: Cross-view pairs.

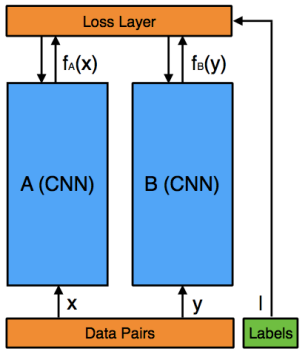

Figure 3: Siamese network.

[1] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. Google street view: Capturing the world at street level. *Computer*, 2010.

[2] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[4] Jiang Wang, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, Ying Wu, et al. Learning fine-grained image similarity with deep ranking. *arXiv preprint arXiv:1404.4661*, 2014.

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