Camera Elevation Estimation from a Single Mountain Landscape Photograph

Martin Čadík[1]
cadik@fit.vutbr.cz

Jan Vašíček[1]
xvasic21@stud.fit.vutbr.cz

Michal Hradiš[1]
ihradis@fit.vutbr.cz

Filip Radenović[2]
filip.radenovic@cmp.felk.cvut.cz

Ondřej Chum[2]
chum@cmp.felk.cvut.cz

[1] CPhoto@FIT, http://cphoto.fit.vutbr.cz/elevation/,
Faculty of Information Technology,
Brno University of Technology,
Brno, Czech Republic

[2] Center for Machine Perception,
Department of Cybernetics,
Faculty of Electrical Engineering,
Czech Technical University in Prague,
Prague, Czech Republic

In outdoor environments one of the most important and informative attributes is the elevation: the height of a geographic location above the sea level. However, almost all currently available photos and videos lack elevation information. Moreover, a majority of them do not even contain the GPS coordinates. This work addresses the problem of camera elevation estimation from visual information contained in a landscape photograph.

We introduce a new **Alps100K dataset** of annotated (GPS coordinates, elevation, EXIF if available) outdoor images from mountain environments. We create a list of all hills and mountain peaks located in the seven Alpine countries from the OpenStreetMap database. The list of hill names is used to query the Flickr photo hosting service. The final collection contains *98136 outdoor images* that span almost all possible elevations observed in the Alps [0, 4782m] and covers vast geographic area of the Alps. The images span all the seasons of the year and exhibit high variation in landscape appearance, see Fig. 1.

Figure 1: A sample from the new benchmark dataset Alps100K. Image credits - flickr users: Allie_Caulfield, Erik, Guillaume Baviere, antoine.pardigon, Karim von Orelli.

To measure the ability of humans to estimate camera elevation from an image, an **experiment with 100 subjects** is conducted. The participants are asked to estimate the *camera* elevation for each of the 50 test images using a web-based interface. The overall root-mean-squared error of human elevation predictions is RMSE(H) = 879.95m. The predictions for each test image along with the ground truths are plotted in Fig. 2.

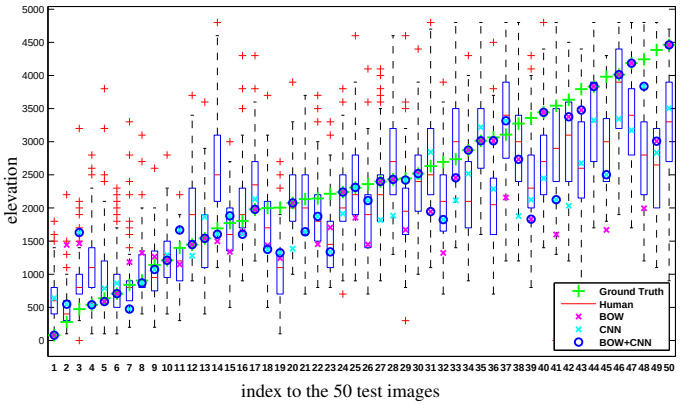

Figure 2: Comparison of the elevation estimation by humans and the proposed methods. Blue boxes show the span of the human predictions: the red mark is the median, the edges of the box are the 25th and 75th percentiles respectively, the whiskers extend to extreme human guesses that are not considered outliers, and outliers are plotted individually as '+'.

First proposed automatic approach to elevation estimation from image content is based on **convolutional neural networks (CNN)**. We initialize CNNs from a network previously trained on the Places205 dataset [2]. The basic network architecture is the same as the one used in the Caffe reference network and the Places-CNN network [2]. The main building blocks of the network are convolutions followed by Rectified Linear Units. First, second, and fifth convolutional layers are followed by max-pooling, each reducing resolution by a factor of two. The activations of the first and second convolutional layers are locally normalized. The output of the convolutional part of the network is fed into a fully connected layer with 4096 units. The network is trained by mini-batch Stochastic Gradient Descent with momentum.

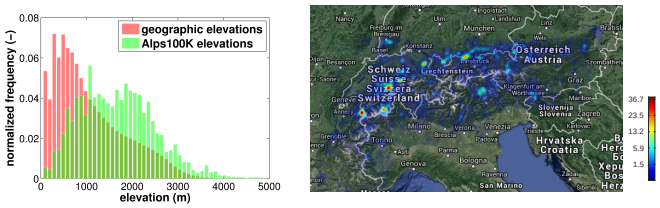

Figure 3: Left: normalized elevation histogram of the Alps mountain range (red) and the distribution of elevations in the Alps100K dataset (green). Right: geographic coverage of the new dataset.

An alternative bag-of-words approach is based on the k-NN classifier, two efficient methods of obtaining k-NN are considered: **sparse high-dimensional bag-of-words (BOW)** based image retrieval, and image retrieval with **compact image representations (mVocab)** [1]. The **BOW** approach using inverted file to efficiently retrieve images has been shown to perform well for specific object and place recognition, especially when combined with a spatial verification step, while the **mVocab** approach using a joint dimensionality reduction from multiple vocabularies shows certain level of generalization power. Therefore, we also propose a **hybrid method** that first tries to estimate the elevation by recognizing the location (using BOW), and if that fails, *i.e.*, no spatially verified image is retrieved, then by a secondary estimator: either mVocab or CNN.

A *test set* of 13148 images is randomly selected from Alps100K, the rest of the dataset (*i.e.*, 84988 images) is used for *training*. The selected measure of performance is an overall RMSE of elevation predictions with regards to the known ground truth elevations, see Tab. 1. The results for a subset of 50 images selected from the test dataset (used in user experiment), which compares the performance of automatic elevation estimation to the performance of humans, are shown in Fig. 2 and Tab. 1. Generally, all of the proposed methods achieve better scores than humans.

Table 1: Results (overall root-mean-square error in meters)

Method	test dataset (13148 images)	user experiment (50 images)
Baseline	801.49; *786.42*	1383.64; *1154.43*
Human	-	**879.95**
CNN	537.11	709.10
BOW	601.63	757.76
mVocab	610.36	811.00
BOW+mVocab	564.14	646.89
BOW+CNN	**500.44**	**531.05**

[1] F. Radenović, H. Jégou, and O. Chum. Multiple measurements and joint dimensionality reduction for large scale image search with short vectors. In *Proc. ICMR*. ACM, 2015.

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.