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,
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

Abstract

This work addresses the problem of camera elevation estimation from a single photograph in an outdoor environment. We introduce a new benchmark dataset of one-hundred thousand images with annotated camera elevation called Alps100K. We propose and experimentally evaluate two automatic data-driven approaches to camera elevation estimation: one based on convolutional neural networks, the other on local features. To compare the proposed methods to human performance, an experiment with 100 subjects is conducted. The experimental results show that both proposed approaches outperform humans and that the best result is achieved by their combination.

1 Introduction

In outdoor environments one of the more important and informative attributes is the elevation: the height of a geographic location above the sea level. Estimation of elevation has a long history [7]. Nowadays, elevation data are important for a number of applications, including earth sciences, global climate change research, hydrology, and outdoor navigation. Traditionally the assessment of elevation was the domain of geodesy, which offered several means to measure altitude. Among the most popular methods are barometric altimeters, trigonometric or leveling measurements, and Global Positioning System (GPS).

A rich natural heritage is captured and expanded everyday in the form of landscape photos and videos with an ever-growing geographic coverage of outdoor areas, see Figs. 1 and 2. Unfortunately, almost all currently available photos and videos lack elevation information. Moreover, a majority of them do not even contain the GPS coordinates. In this paper, we tackle the problem of estimating the elevation from visual data only. Automatic annotation

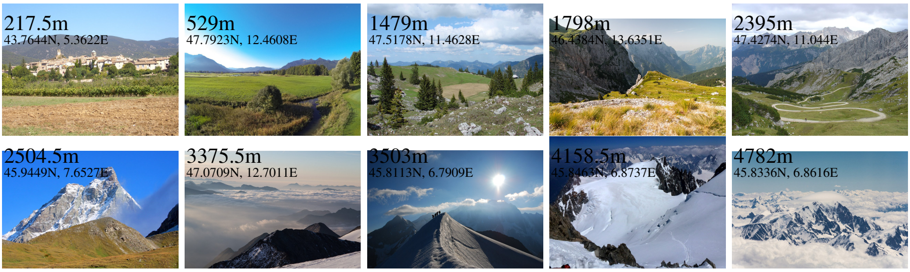

Figure 1: A sample from the new benchmark dataset Alps100K [6]. Image credits - flickr users: Allie_Caulfield, Golf Resort Achental Team, Erik, Guillaume Baviere, Tadas Balčiunas, antoine.pardigon, twiga269, Karim von Orelli.

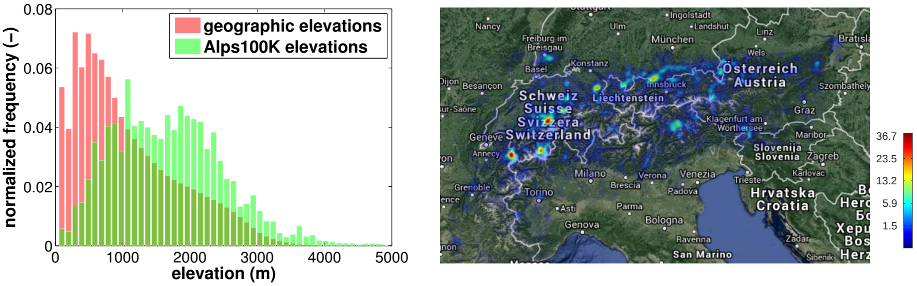

Figure 2: Left: normalized elevation histogram of the Alps mountain range (red) and the distribution of elevations in the Alps100K dataset (green). Right: geographic coverage of the new dataset.

of images with an accurate estimate of the elevation can be exploited in a number of applications ranging from leisure activities and tourism applications to image database exploration, education purposes and geographic games.

Contributions. In order to evaluate elevation estimation methods, we introduce a new dataset containing approximately 100K images of natural environments accompanied with the GPS and elevation information (Section 3). We propose two methods of elevation estimation from image content: one based on convolutional neural networks (CNN) [31], the other exploiting bag-of-words (BOW) based image retrieval [15, 26]. The proposed methods are compared to a human performance in Section 6. To estimate the human performance on this task, an experiment was conducted counting 100 subjects (Section 4). The proposed automatic elevation estimation methods outperform knowledgeable humans, and, moreover, the hybrid combination of BOW with CNNs results in the best predictions.

2 Related work

We are not aware of any attempt to predict the camera elevation from visual information contained in a landscape photograph of a natural environment. However, the related field of research is the one of *visual geo-localization*. If we were able to localize the position of the camera, the task of elevation estimation would reduce to a simple query into a georeferenced terrain model. Unfortunately existing geo-localization methods [2, 3, 14, 29, 30] are neither robust nor sufficiently accurate for this task. Conversely, the methods presented in this paper may improve geo-localization techniques by reducing the search space to only probable camera elevations.

Height above the ground is among the most important information for navigation of *Unmanned Aerial Vehicles* (UAV). The problem of elevation estimation for UAVs has been attacked using computer vision because it has several advantages: it is a passive system, has low energy consumption, and visual information can be reused for navigation or localization. The proposed solutions are based on artificial ground-located landmarks [12, 25], optical flow [5, 27], stereoscopic vision [11, 16, 20], and machine learning [8]. Most of these methods assume that the camera's viewpoint is oriented to the ground and that the camera parameters are known. In contrast, in this work we aim to assess the *absolute* elevation of the camera. Moreover, the hereby proposed methods work on ordinary photographs of natural outdoor environments. Our scenario is more challenging, since both camera orientation and calibration information is missing.

The lack of work devoted to elevation estimation may be explained by the absence of a suitable dataset. The recently published Places 205 [31] is a dataset for training scene classifiers. It contains 205 scene categories and almost 2.5 million images of which a subset could be selected for our task. Unfortunately, Places 205 contains neither elevation information nor GPS coordinates. IM2GPS dataset [14] does contain GPS coordinates for each image; however, these images are mostly captured in cities with a significant bias towards landmarks like the Eiffel tower or the Sydney opera house. This renders IM2GPS dataset useless for our purpose because we focus on landscape photos of outdoor environments, like the ones shown in Fig. 1.

3 Alps100K: a new dataset

We introduce a new dataset of almost 100K annotated (GPS coordinates, elevation, EXIF if available) outdoor images from mountain environments. The collection covers vast geographic area of the Alps, the highest range in Europe; therefore we name it Alps100K. The images exhibit high variation in elevation as well as in landscape appearance. Furthermore, the collection spans all the seasons of the year. To the best of our knowledge, this is the first dataset of this kind. It contains test sets to evaluate elevation estimation performance (see Section 4 for human performance and Section 6 for results of the proposed automated methods). A large proportion of the dataset serves as a training set for the data-driven approaches.

Dataset acquisition. First we create a list of all hills and mountain peaks located in the seven Alpine countries (Austria, France, Germany, Italy, Liechtenstein, Slovenia, Switzerland) from the OpenStreetMap database [9]. The list of hill names is used to query the Flickr[1] photo hosting service. In order to increase the ratio of outdoor images certain tags, such as wedding, family, indoor, still life, are excluded. Only images containing information about the camera location are kept. Out of 1.2M crawled images, about 400K are unique and inside the Alps region.

To cull irrelevant (non-landscape) images a state-of-the-art scene classifier [31] is applied. Probabilities of 205 scene categories are assigned to each image. We experimentally select 28 categories (mountain, valley, snowfield, etc.) and keep only those images whose

[1] http://www.flickr.com

cumulative probability in those categories exceeds 0.5. This step significantly improves the relevance of the dataset at the expense of reducing the number of images to circa 25%.

Finally the elevation of the camera is inferred from the GPS coordinates via the digital elevation model[2]. This model covers the Alps with 24 meter spaced samples. The collection contains *98136 outdoor images* that span almost all possible elevations observed in the Alps [0, 4782m]. Fig. 2 left compares the elevation distribution of the Alps surface and the elevations in the dataset, Fig. 2 right shows the spatial distribution of the collected images. Geographically the dataset covers virtually all the regions of the Alps with obvious concentrations in tourist spots (e.g., around Zermatt village in Switzerland). The EXIF information is available for 41364 images, which is 42% of the Alps100K dataset. The the presented dataset along with elevation and GPS meta-data is available at the project webpage [6].

4 Human performance in elevation estimation task

In this section we measure the ability of humans to estimate camera elevation from an image. The achieved accuracy is subsequently compared to results achieved by the methods proposed in this paper.

During the experiment 100 participants were asked to estimate the *camera* (not the depicted scene) elevation for each of the 50 test images. We utilized a custom web-based interface where the participants assessed the elevation of each test image (for an example see Fig. 1) using a slider (see more details on the experiment in the supplementary material available at the project webpage [6]). The elevation of the test images ranged in [79m, 4463m] (see green crosses in Fig. 3). The images were presented in randomized order, at the resolution of 750×500px. After the experiment was finished participants filled out a questionnaire where additional information about their age, experience with the Alps, highest reached elevation, etc., was gathered. The subjects needed 10 minutes on average to complete the experiment.

4.1 Experimental results

Are humans able to estimate the elevation from an image? We use the analysis of variance (ANOVA) test [4] which determines whether there is a systematic effect of an image (*i.e.*, the elevation) in human elevation predictions, or whether the predictions are random. Formally, we state the null hypothesis H_0 as follows: there is no significant difference between elevation predictions for the test images. This hypothesis is clearly rejected ($F(49,4950) = 165.09, p < 0.001$), meaning that humans *can* indeed estimate the camera elevation from the visual information in images.

How well can humans predict the camera elevation? The predictions for each test image along with the ground truths are plotted in Fig. 3. The overall root-mean-squared error (RMSE) of human elevation predictions is $\mathrm{RMSE}(H) = 879.95\mathrm{m}$. It can be observed that people underestimate high elevations, *i.e.*, elevations above 3000m. In accordance with this, variance of the elevation predictions grows for high altitudes as well.

The effect of human age, sex, experience and other factors collected in post-experiment questionnaire was also analyzed. None of those factors were found statistically significant.

[2] Available from http://www.viewfinderpanoramas.org

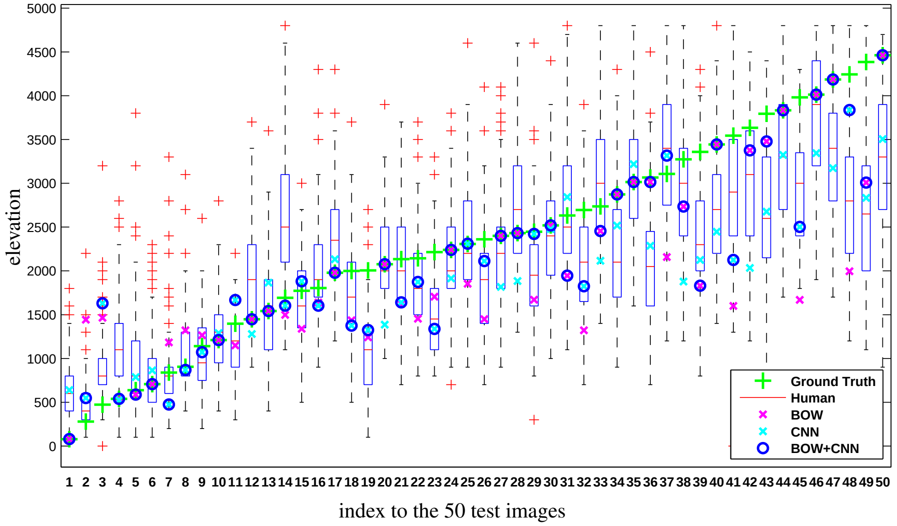

Figure 3: Comparison of the elevation estimation by humans and the proposed methods (CNN, BOW and combination). Blue boxes show the span of the human predictions: the red mark is the median, the edges of the box are the 25th and 75th percentiles respectively, the whiskers extend to extreme human guesses that are not considered outliers, and outliers are plotted individually as '+'.

However, it is worth noting that our participants were either living in the Alps, had exceptional experience with outdoor sports, or both. Accordingly, the reported average prediction errors should be taken as rather conservative ones.

5 Automatic elevation estimation from landscape photo

In this section we propose two approaches to estimate elevation from the visual content. We use popular convolutional neural networks and methods based on local features and combine them to exploit each approach's strengths in a single hybrid method.

5.1 Convolutional neural networks (CNN)

The elevation estimation task can be treated as a standard regression problem where the goal is to directly predict the real-valued elevation given the pixel data of a single photo. Convolutional neural networks have proven to be the state-of-the-art in various image-based machine learning tasks including object and scene classification [31], object detection [13], semantic segmentation [18], and facial recognition [28]. We build upon the previous successes and apply large convolutional networks to the Alps100K dataset. Considering the relatively small size of the dataset, we initialize the networks from a network previously trained on the Places205 dataset [31], which includes a rich collection of outdoor scenes among its 2.5 million images. Additionally, we extend the network inputs by EXIF data, which carries information indicative of possible weather conditions and camera settings.

Convolutional network architecture. The basic network architecture follows previous

Table 1: Architecture of the convolutional network. Layers up to fc6 were initialized from Places-CNN network [31]. Resolution of the input is 227 × 227 pixels.

Layer	conv1	pool1	conv2	pool2	conv3	conv4	conv5	pool5	fc6	fc7	fc8
units	96	96	256	256	384	384	256	256	4096	2048	1
kernel	11×11	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
features	55×55	27×27	27×27	13×13	13×13	13×13	13×13	6×6	-	-	-

successful work on image classification. Specifically, it is the same as the one used in the Caffe reference network[3] and the Places-CNN network [31], which are in turn very similar to the network used by Krizhevsky et al. [17] to win the ImageNet challenge in 2012.

The main building blocks of the network (see Tab. 1) are convolutions followed by Rectified Linear Units (ReLU). 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 [17]. The output of the convolutional part of the network is fed into a fully connected layer (fc6) with 4096 units. Weights of this part of the network are initialized from network Places-CNN[4] [31].

The final two layers of the network are fully connected and contain 2048 and 1 neurons, respectively. Weights of these layers were initialized from a normal distribution with standard deviation 0.005 and 0.02, respectively. The final activation function is linear and the optimization objective is Mean Squared Error (MSE). The network was trained by mini-batch Stochastic Gradient Descent with momentum.

5.2 Local features

An alternative approach is based on the k-NN classifier [10]. Two efficient methods of obtaining nearest neighbours are considered: sparse high-dimensional bag-of-words (BOW) based image retrieval [26], and image retrieval with compact image representation [15, 24]. In both cases, the number of neighbours k used to estimate the elevation is a function of the confidence in the retrieved nearest neighbour.

The BOW approach has been shown to perform well for specific object and place recognition, especially when combined with a spatial verification step [23], while the short-vector image representations [15, 24] obtained by a joint dimensionality reduction from multiple vocabularies show certain level of generalization power.

High-dimensional image representation (BOW). The majority of image retrieval methods based on BOW representation follow the same procedure as introduced in [26]. First, local features [21] such as multi-scale Hessian-Affine [22] are detected and described by an invariant d-dimensional descriptor such as SIFT [19] or RootSIFT [1] for all images in the dataset. Then the descriptor space is vector-quantized into a visual vocabulary: a K-means algorithm is performed on an independent training dataset to create K clusters representing the visual words. In our paper, K=1M visual words was used. Finally, a histogram of occurrences of visual words is generated for each image followed by the inverse document frequency weighting (*idf*), and sparse BOW vectors are obtained with dimensionality D=K.

To estimate the elevation of a photograph the photograph is used to query an elevation-annotated image database (the training part of Alps100K). Efficient retrieval via the inverted

[3] Available from the Caffe package http://caffe.berkeleyvision.org/model_zoo.html

[4] Available from http://places.csail.mit.edu

file structure [26] is followed by a spatial verification step [23] to re-rank the results. If the top ranked image is likely to be from the same location as the query image, *i.e.*, the top ranked image is spatially verified with high confidence (more than t_{sp} features pass the verification test), we use its elevation as the estimate. Otherwise, we use median of elevations of all retrieved k-NN images.

Short-vector image representation (mVocab). Large scale image retrieval with short vectors has recently became popular as a method for reducing high computational or memory costs. In [15, 24] concatenated vocabularies of different origins followed by a joint dimensionality reduction are used as short image descriptors.

In our experiments we follow [24] and combine eight different visual vocabularies of 8K visual words each. The vocabularies are constructed over two different measurement regions and four power-law normalizations of SIFT descriptors. The region sizes are $r \times s$ and $1.5 \times r \times s$ (s is the scale of the detected feature, and $r = 3\sqrt{3}$ is the standard relative scale change between detected region and the measurement region, as in [22]). The power-law normalization of SIFT descriptors ranges $\beta = 0.4, 0.5, 0.6, 1$, where $\beta = 1$ is the original SIFT descriptors, while $\beta = 0.5$ corresponds to RootSIFT [1].

Camera elevation of landscape photographs is estimated by finding k-NN images from the training dataset and taking the weighted average of their elevations. Weights $\mathbf{w}$ are calculated using image dissimilarities $\mathbf{d}$, $\mathbf{w} = \max[0, 1 - \mathbf{d}/(w_t d_1)]$, where d_1 is the dissimilarity of the top ranked image and w_t is a constant, $w_t \geq 1$. The number of retrieved images is fixed, but the number used from that set depends on their similarity to the top ranked image; in addition, we can control the number used from the retrieved set by the appropriate choice of parameter w_t. Specifically, only images that have dissimilarity up to $w_t d_1$ are used to compute weighted average of elevations.

Parameter choice. We use the training dataset to learn the necessary parameters. For all experiments presented in this paper $t_{sp} = 8$, $w_t = 1.4$ and $k = 100$. Slight changes do not significantly influence the presented results.

5.3 Combining the elevation estimators

The BOW representation with spatial verification has been shown to perform well in a specific object or location recognition. Since people take and share photographs from similar locations, a natural approach is to recognize the specific location. On the other hand, BOW does not generalize meaning that it does not perform well on unseen scenes. Therefore, we propose a hybrid method that first tries to estimate the elevation by recognizing the location, and if that fails, *i.e.*, no spatially verified image is retrieved, then by a secondary estimator: either mVocab or CNN.

6 Results

In this section we experimentally evaluate the proposed methods. A *test set* of 13148 images (13% of the datased) 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 root-mean-square error (RMSE) of elevation predictions with regards to the known ground truth elevations. The outcome is summarized in Tab. 2, where *Baseline* denotes a simple elevation predictor that reports the mean elevation of the training dataset for all queries. For complete-

Table 2: Results (overall root-mean-square error in meters)

Method	test dataset (13148 images)	user experiment set (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**

ness we show the best achievable baseline RMSE values as well, *i.e.*, using the mean of the *testing* elevations (in italics).

All of the proposed methods perform better than the baseline. The best prediction accuracy among the individual methods is achieved by neural networks (CNN); however, this is significantly improved by combining BOW with CNN. For a better understanding of these results, we plot the fraction of correct predictions as a function of the elevation error, see Fig. 4, left. For more than one third of the test images, the BOW approach spatially verifies the query image, which results in extremely low error rate. The specific place recognition performed by the BOW approach seems to work well for this particular task because the geographic distribution of the Alps100K dataset, which corresponds to reality, exhibits many strong peaks at popular places (see Fig. 2, right). CNN, on the other hand, seems to generalize better on previously unseen locations and it surpasses BOW in sparsely covered regions. The combination of BOW and CNN achieves the best RMSE scores, and is more robust as well.

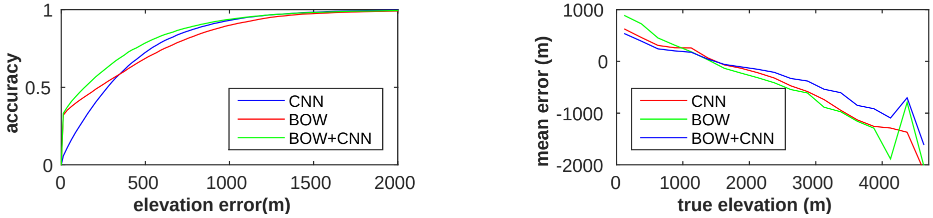

Figure 4: Left: cumulative elevation prediction accuracy. Right: dependence of prediction bias on image elevation.

The results for a subset of 50 images selected from the test dataset (used in user experiment described in Section 4) are shown in the right column of Tab. 2, which compares the performance of automatic elevation estimation to the performance of humans. Generally, all of the proposed methods achieve better scores than humans. The best RMSE is again obtained by the hybrid combination of BOW and CNN, which is on average significantly better than humans. In Fig. 3 we plot the predictions for each of the 50 images separately. Interestingly, BOW+CNN method exhibits a similar bias as humans; i.e., it tends to underestimate the highest elevations (images #49, 48, 45, 41, 39, 38) and overestimate lowest elevations (#2, 3). This tendency actually holds true for the whole test dataset, as illustrated in Fig. 4 right. We attribute this behavior to the distribution of elevations in Alps100K dataset, which is less populated in both elevation extremes, as shown in Fig 2, left.

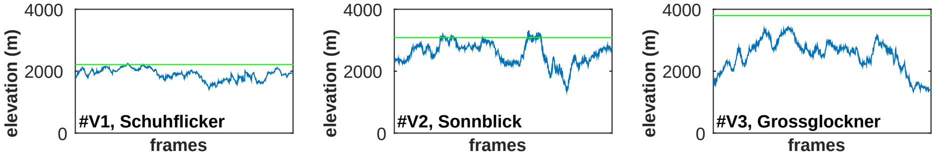

Figure 5: BOW+CNN elevation predictions (blue) for three video sequences (green) captured using a camera with constant elevation and varying yaw and pitch.

CNN+EXIF information. Visual appearance of photographs is strongly influenced by the season, the time of the day, and camera field of view. We encode time of the day, time of the year, exposure coefficient, and camera field of view as a sparse binary vector and input it to the first fully connected layer of the CNN (fc6). Each of the values is quantized to 16 discrete levels and encoded as 1-of-N. The exposure coefficient (EC) combines relative aperture N, exposure time t, and sensor sensitivity ISO into a single value that represents “sensitivity of the photograph” to the light in the scene. Assuming that photos are properly exposed, high EC values imply low-light conditions and low EC values imply bright conditions. EC is calculated as $EC = \log_2 N^2 - \log_2 t \cdot \frac{ISO}{100}$. Camera *field of view* (FOV) indicates possible composition of photographs. As most cameras do not store their field of view explicitly in EXIF data, it needs to be computed from sensor size S and focal length f as $FOV = \arctan(0.5S/f)$.

The combination of CNN with EXIF information was evaluated on a smaller subset of images with EXIF information available (36050 training and 5314 test images respectively). On this subset, CNN+EXIF achieves RMSE=510m, compared to pure CNN with RMSE=550m on the same subset. We conclude that the EXIF data bring a small improvement in elevation predictions.

Video sequences. We evaluated the BOW+CNN method using three video sequences (#V1-3 available at the project webpage [6]) in Fig. 5. The videos were acquired by a hand-held point-and-shoot camera from spots of constant elevation (green lines), while changing camera yaw and pitch. These videos represent challenging scenarios, in particular, due to the varying pitch and high camera elevations (#V1=2215m, #V2=3088m, #V2=3789m). Moreover, no frame has been spatially verified by BOW in any case, and thus the prediction accuracy depends solely on CNN. Nevertheless, the proposed method achieves decent elevation predictions for #V1 and #V2. The third video (#V3) illustrates the limitations of the current solution: the prediction accuracy for extremely high elevations is low, suffering from a small number of appropriate images in the training set.

7 Conclusions

Multiple aspects of the camera elevation estimation task were addressed in the paper. A new benchmark dataset of elevation-annotated images Alps100K was collected. Two approaches were proposed to automatically estimate the camera elevation from a single landscape photograph. In an extensive user experiment, human performance on this task was measured. Experimental evaluation showed that the proposed methods outperform human abilities in camera elevation estimation. The best performing method first attempts to recognize specific location via BOW-based image retrieval, and in case of failure, uses CNN to estimate the elevation. In the future, we plan to extend the dataset to different geographic areas and other climate zones.

Acknowledgements

This work was supported by SoMoPro II (financial contribution from the EU 7 FP People Programme Marie Curie Actions, REA 291782, and from the South Moravian Region), SGS15/155/OHK3/2T/13, and MSMT LL1303 ERC-CZ grants. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

References

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, pages 2911–2918, 2012.

[2] G. Baatz, O. Sauer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *Proc. ECCV*, pages 517–530, 2012.

[3] L. Baboud, M. Čadík, E. Eisemann, and H.-P. Seidel. Automatic Photo-to-terrain Alignment for the Annotation of Mountain Pictures. In *Proc. CVPR*, pages 41–48, 2011. ISBN 978-1-4577-0394-2.

[4] V. Bewick, L. Cheek, and J. Ball. Statistics review 9: one-way analysis of variance. *Critical care (London, England)*, 8(2):130–136, 2004.

[5] A. Beyeler, C. Mattiussi, J.-C. Zufferey, and D. Floreano. Vision-based altitude and pitch estimation for ultra-light indoor microflyers. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 2836–2841, May 2006.

[6] M. Čadík, J. Vašíček, M. Hradiš, F. Radenović, and O. Chum. Supplementary material for camera elevation estimation from a single mountain landscape photograph, September 2015. URL http://cphoto.fit.vutbr.cz/elevation/.

[7] F. Cajori. History of determinations of the heights of mountains. *Isis*, 12(3):482–514, 1929.

[8] A. Cherian, J. Andersh, V. Morellas, N. Papanikolopoulos, and B. Mettler. Autonomous altitude estimation of a UAV using a single onboard camera. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3900–3905, Oct 2009.

[9] K. Curran, J. Crumlish, and G. Fisher. OpenStreetMap. *International Journal of Interactive Communication Systems and Technologies*, 2(1):69–78, 2012. ISSN 2155-4218.

[10] O. Duda, R., E. Hart, P., and G. Stork, D. *Pattern classification*. John Wiley & Sons, 2012.

[11] D. Eynard, P. Vasseur, C. Demonceaux, and V. Fremont. UAV altitude estimation by mixed stereoscopic vision. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 646–651, Oct 2010.

[12] J. Garcia-Pardo, P., S. Sukhatme, G., and F. Montgomery, J. Towards vision-based safe landing for an autonomous helicopter. *Robotics and Autonomous Systems*, 38(1): 19–29, 2001.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[14] J. Hays and A. Efros, A. IM2GPS: estimating geographic information from a single image. In *Proc. CVPR*, 2008.

[15] H. Jégou and O. Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *Proc. ECCV*, Firenze, Italy, October 2012.

[16] I.-K. Jung and S. Lacroix. High resolution terrain mapping using low attitude aerial stereo imagery. In *Proc. ICCV*, pages 946–951 vol.2, Oct 2003.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[18] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *CoRR*, abs/1411.4, 2014.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Proc. ICCV*, 60(2):91–110, 2004.

[20] M. Meingast, C. Geyer, and S. Sastry. Vision based terrain recovery for landing unmanned aerial vehicles. In *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, volume 2, pages 1670–1675 Vol.2, Dec 2004.

[21] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004.

[22] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *Proc. CVPR*, 2009.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[24] 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.

[25] S. Saripalli, F. Montgomery, J., and G. Sukhatme. Vision-based autonomous landing of an unmanned aerial vehicle. In *Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on*, volume 3, pages 2799–2804, 2002.

[26] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003.

[27] V. Srinivasan, M., W. Zhang, S., S. Chahl, J., and A. Garratt, M. Landing strategies in honeybees, and applications to uavs. In A. Jarvis, R. and A. Zelinsky, editors, *Robotics Research*, volume 6 of *Springer Tracts in Advanced Robotics*, pages 373–384. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-00550-6.

[28] Y. Taigman, M. Yang, A. Ranzato, M., and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proc. CVPR*, pages 1701–1708. IEEE, June 2014. ISBN 978-1-4799-5118-5.

[29] E. Tzeng, A. Zhai, M. Clements, R. Townshend, and A. Zakhor. User-driven geolocation of untagged desert imagery using digital elevation models. In *Proc. CVPRW*, pages 237–244, June 2013.

[30] R. Zamir, A. and M. Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. *IEEE PAMI*, PP(99):1–1, 2014. ISSN 0162-8828.

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