DEEPFOCAL: A METHOD FOR DIRECT FOCAL LENGTH ESTIMATION

Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, Nathan Jacobs

Department of Computer Science, University of Kentucky

{scott, connor, ted, rbalten, jacobs}@cs.uky.edu

ABSTRACT

Estimating the focal length of an image is an important preprocessing step for many applications. Despite this, existing methods for single-view focal length estimation are limited in that they require particular geometric calibration objects, such as orthogonal vanishing points, co-planar circles, or a calibration grid, to occur in the field of view. In this work, we explore the application of a deep convolutional neural network, trained on natural images obtained from Internet photo collections, to directly estimate the focal length using only raw pixel intensities as input features. We present quantitative results that demonstrate the ability of our technique to estimate the focal length with comparisons against several baseline methods, including an automatic method which uses orthogonal vanishing points.

Index Terms— focal length estimation, camera calibration, convolutional neural network

1. INTRODUCTION

Camera calibration, that is estimating the intrinsic and extrinsic parameters relating the 3D world to a 2D image, is a fundamental first step for many vision problems. Often this step is overlooked as there exist many standard techniques for calibrating a camera in a laboratory setting. However, for images captured "in the wild", such as those collected from photo sharing websites [1] and publicly available outdoor webcams [2], it is not possible to calibrate using such traditional methods. This challenge motivates new problems in camera calibration and requires new methods. In this work, we investigate the application of a deep convolutional neural network (CNN) for estimating the focal length of the camera from a single image. Our approach operates directly on raw image intensities, requires no per-image parameter setting, and is very fast.

Our motivation is threefold. First, while the problem of single-image camera calibration is well-studied and existing methods perform exceptionally well, in general they assume that specific calibration objects or configurations of objects are present in the image. However, relatively few images captured "in the wild" satisfy these assumptions. In such cases, existing methods, such as those that require the presence of orthogonal vanishing points [3], either are unable to provide an estimate of the focal length or simply give an incorrect estimate.

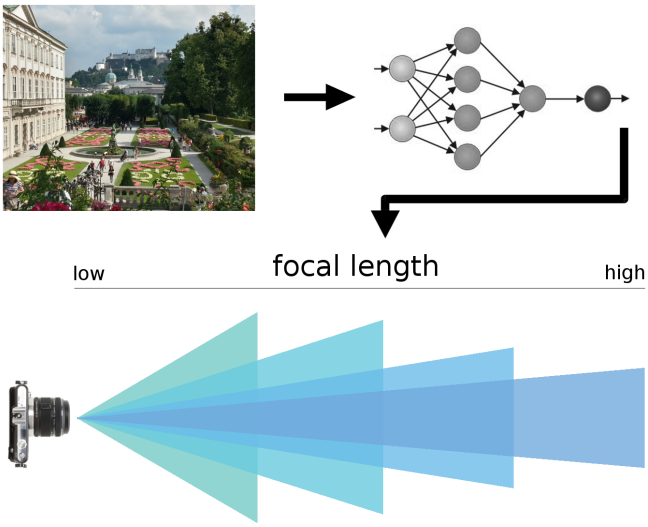


Fig. 1: We estimate the focal length of the camera from a single image using a deep convolutional neural network.

Second, in many problem settings a precise estimate of the focal length is not required. There are many methods that would benefit from a quick technique that provides a rough estimate to serve as a starting point for more computationally expensive search algorithms. For example, Sattler et al. [4] show that a guided sampling scheme is superior to slow minimal solvers when estimating the pose for a camera with unknown focal length.

Finally, recent advances in CNNs have shown dramatic performance improvements for a wide variety of vision tasks, including object classification and detection [5], face recognition and verification [6], and scene parsing [7]. These successes have laid the groundwork for exploring the ability of deep CNN architectures applied to these types of geometric image problems. Figure 1 gives an overview of our approach.

2. RELATED WORK

Interpreting geometric scene information from a single image is a widely studied problem that has significance for many problem domains. Methods have been proposed for camera

calibration [8, 3, 9, 10], inferring 3D layout [11, 12, 13], and performing metrology [14]. These methods impact a variety of fields, including environmental monitoring [15] and forensic analysis [16].

Many methods exist for determining the intrinsic calibration of a camera from a single image. These methods generally rely on detecting reference objects such as a planar calibration grid [17, 18, 19, 20], coplanar circles [8] or concentric circles [21]. Other methods take advantage of the properties of vanishing points [3, 22, 9], which provide a strong characterization of geometric scene structure. As the need to calibrate images captured "in the wild" has grown, many more methods have been introduced which take advantage of natural cues, such as sun position [23] and solar refractive phenomena [24].

In contrast to this previous work, where the goal is to detect specific geometric objects in the image, we propose to use a deep convolutional neural network, trained on thousands of images from natural scenes, to directly estimate the focal length of the camera using only raw pixel intensities as input features. This is in line with an emergent research direction exploring the ability of deep learning techniques for estimating geometric image properties, such as estimating a metric depth map from a single image [25].

3. DIRECT FOCAL LENGTH ESTIMATION

Our work takes advantage of one of the most commonly used CNN architectures [26], often referred to as *AlexNet*. Originally designed for multi-class object classification, this architecture has been successfully adapted for tasks such as image style recognition [27] and scene characterization [28]. In this section, we describe our approach for adapting this architecture for estimating the focal length of a camera from a single image.

3.1. Problem Statement

We assume a simplified pinhole camera model. Given the camera intrinsics, $\mathbf{K}$, extrinsic rotation, $\mathbf{R}$, and translation, $\mathbf{t}$, a world point, $P = [X, Y, Z]^\mathsf{T}$, projects to an image location, $\lambda\vec{p} = [\lambda u, \lambda v, \lambda]^\mathsf{T} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]P$. Assuming zero skew, principal point at the center of the image, square pixels, and camera frame aligned to the world frame, this can be reduced to a simple pinhole camera model, $\lambda\vec{p} = \mathbf{K}P = diag([f, f, 1])P$, where f is the focal length. Our goal is to estimate the horizontal field of view, H_θ, which has a one-to-one mapping with focal length:

$$H_\theta = 2tan^{-1}(\frac{w}{2f}) \tag{1}$$

for a given image width, w.

Fig. 2: Example images from the 1DSfM [29] datasets.

3.2. Approach

Given an image, we propose to estimate the horizontal field of view, H_θ using the *AlexNet* architecture. We formulate this as a regression problem with H_θ represented as a continuous label. *AlexNet* was originally designed for single image object classification and is comprised of five convolutional layers and three fully-connected layers. To adapt the network for our regression problem, the only change necessary is to update the final fully-connected layer to have a single output node corresponding to H_θ. We call the resulting network *DeepFocal*.

Instead of training the network from scratch, we use a technique referred to as transfer learning [30]. Transfer learning exploits a previously trained base network by *transferring* the learned features to a target network that will be retrained for a new task. Practically, this means the first n layers of the target network are initialized with the weights from the first n layers of the base network, with the remaining layers weights randomly initialized. The network is then trained toward the new task using the target dataset. Yosinski et al. [30] show quantitatively that transferred features are often better than random weights, even for very different tasks, and that they can improve generalization even after a large amount of fine-tuning. As opposed to *AlexNet* [26], we use the full image as input to our networks. A consequence of this is that it is only possible to transfer features for the five convolutional layers, which are size invariant. We discuss our choice of initialization in Section 4.2.

We consider two approaches for handling images of dif-

ferent orientations. The first simply ignores the aspect ratio of an individual image and resizes all images to be the same size. For the second approach, we train a network specific to each orientation, landscape and portrait. Given a test image, we select the appropriate network using its aspect ratio. We refer to these as *DeepFocal (combined)* and *DeepFocal (best)*, respectively.

3.3. Implementation Details

Our networks are implemented and trained using Caffe [31], an open source deep learning framework. All networks are trained on NVIDIA M2075 GPUs for twenty four hours using stochastic gradient descent with an L_2 loss function and a base learning rate of 10^{-5}. We freeze the transferred features [30] and randomly initialize the final three fully connected layers.

4. EXPERIMENTS

We performed several experiments to evaluate the ability of our methods to estimate the focal length.

4.1. Datasets

Collecting a large number of images with known and varying focal lengths is a difficult task. Therefore, existing datasets have either been limited to a single camera [32], or have required post-hoc manual calibration (which is often infeasible) [33]. To overcome this, we constructed a dataset by combining images and camera models estimated using 1D structure from motion (1DSfM) [29]. Example images from the 1DSfM datasets are shown in Figure 2. We defined our training and testing split such that there are no overlapping fields of view. Table 1 visualizes the split used. From the entire set of 1DSfM images, we selected images that were within 4:3 $\pm$ 0.1 or 3:4 $\pm$ 0.1 aspect ratio, for which the focal length was provided. This resulted in a total of 7076 landscape images with sizes ranging from 480×360 to 1600×1269, and 4246 portrait images of size 334×500 to 1314×1600. Of these, 564 and 186 images, respectively, were used for testing.

4.2. Identifying the Best Initialization

Our first experiment examined the impact of transfer learning starting from different sets of initial network weights. We evaluated three different publicly available pretrained networks that use the *AlexNet* architecture. The first, CaffeNet, was trained for the task of object classification using 1.2 million images from the ImageNet ILSVRC-2012 challenge [34]. The second, Places, was trained on 2.5 million images for scene categorization [28]. The third, Hybrid, was trained using a combination of both object and scene categories using labeled images from both the Places Database [28] and the ILSVRC-2012 challenge [34]. These networks are available as Caffe [31] model files.

Table 1: Non-overlapping train/test splits.

Dataset Location	Training	Testing
Alamo, San Antonio, USA	X	
Ellis Island, New York City, USA	X	
Madrid Metropolis, Madrid, Spain	X	
Montreal Notre Dame, Montreal, Canada	X	
Notre Dame, Paris, France	X	
NYC Library, New York City, USA	X	
Piazza del Popolo, Rome, Italy		X
Picadilly, London, England	X	
Roman Forum, Rome, Italy	X	
Tower of London, London, England	X	
Trafalgar, London, England	X	
Union Square, San Francisco, USA		X
Vienna Cathedral, Vienna, Austria	X	
Yorkminster, York, England	X	

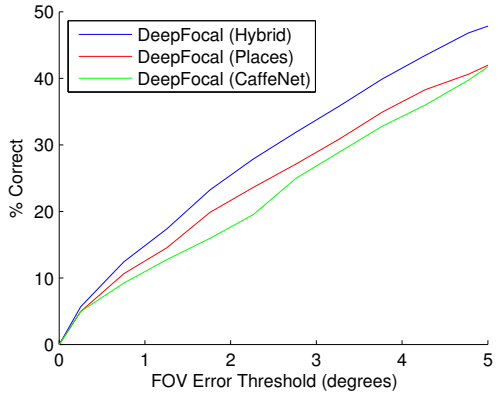
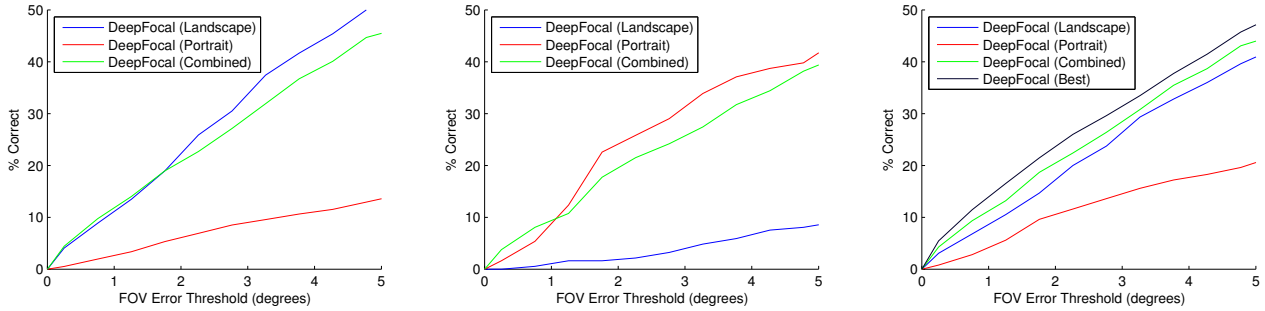


Fig. 3: The choice of initialization when transferring features has an impact on performance.

For each of these base networks, we followed the approach outlined in Section 3.2 to train a new network using the landscape training and testing split defined in Section 4.1. The results of this experiment are visualized in Figure 3. The performance of each network is comparable, but the features transferred from Hybrid perform slightly better. This implies that a combination of object-centric and scene-centric features are useful for estimating field of view. We use Hybrid as our initialization for the remainder of our experiments.

4.3. Examining the Important of Aspect Ratio

We hypothesized that for this geometric task, aspect ratio would have an impact on performance. This forms the basis of our second experiment. We trained three separate networks: the first using only the landscape training images, the second using the portrait training images, and the third a combination of both landscape and portrait images. We then evaluated the performance of each of these networks on a different set of test images using the best performing iteration



(a) Landscape Test Set (b) Portrait Test Set (c) Combined Test Set

Fig. 4: Evaluating the impact of aspect ratio on network performance. (a-b) When restricting the test set to a specific aspect ratio, the network trained solely on images of that aspect ratio performs best. (c) The "best" strategy when evaluating on a combined set of images of multiple aspect ratios is to select the network matching the aspect ratio of each test sample.

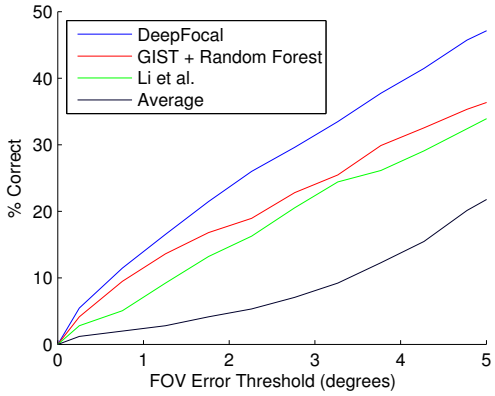


Fig. 5: The results of our method versus several baseline approaches. Our method outperforms the method proposed by Li et al. [35], which uses orthogonal vanishing points.

from each. Figure 4 visualizes the results of this experiment.

The results support our initial hypothesis. In Figure 4(a), the network trained on landscape images outperforms the other two networks when tested solely on landscape images. This trend continues in Figure 4(b), where the network trained using portrait images excels when tested on portrait images. Finally, Figure 4(c) shows that when evaluating on a combined test set containing portrait and landscape images, the "best" strategy is to choose the appropriate network based on the aspect ratio of the test sample.

4.4. Comparison to Baseline Methods

We compared the results of our approach versus several baseline methods, using the combined test set containing landscape and portrait images from Union Square in San Francisco, and Piazza del Popolo in Rome.

The first baseline followed a simple strategy: given a query image, use the average field of view of the training set as the prediction. With this approach, we computed the average for each orientation separately. The second baseline explored the usefulness of off-the-shelf global image descriptors as features for estimating the field of view. For all images we extracted GIST descriptors [36], often used for scene recognition and related tasks, and trained an ensemble of regression trees (10 trees). Similar to our findings in Section 4.3, training a separate ensemble for each image orientation individually outperforms training jointly on images of multiple orientations, and we evaluated versus this strategy. Finally, the third baseline applied the method of Li et al. [35], which simultaneously estimates three orthogonal vanishing points and the focal length from a single image. We used the code provided by the authors and converted focal length to horizontal field of view as in (1).

As demonstrated in Figure 5, our approach outperforms all baseline methods by a wide margin. For example, we correctly predict approximately 30% of images within 3 degrees error as opposed to 20% by Li et al. [35].

5. CONCLUSION

Many high-level vision methods require an estimate of the focal length to function. In spite of this, existing methods for single-view calibration are limited in that they require very specific geometric calibration objects to appear in the image or they fail. We proposed a fast method which overcomes these limitations by directly estimating the focal length from raw pixels using a deep convolutional neural network. Our method outperforms several baselines, including an automated technique based on orthogonal vanishing points, in spite of the images being captured in urban environments. In the future we will explore expanding our methods to a more complex camera model.

Acknowledgments This research was supported by DARPA CSSG D11AP00255.

6. REFERENCES

[1] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg, "Mapping the world's photos," in *WWW*, 2009.

[2] Nathan Jacobs, Nathaniel Roman, and Robert Pless, "Consistent temporal variations in many outdoor scenes," in *CVPR*, 2007.

[3] Bruno Caprile and Vincent Torre, "Using vanishing points for camera calibration," *International Journal of Computer Vision*, vol. 4, no. 2, pp. 127–139, 1990.

[4] Torsten Sattler, Chris Sweeney, and Marc Pollefeys, "On sampling focal length values to solve the absolute pose problem," in *ECCV*, 2014.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

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

[7] Pedro Pinheiro and Ronan Collobert, "Recurrent convolutional neural networks for scene labeling," in *ICML*, 2014.

[8] Qian Chen, Haiyuan Wu, and Toshikazu Wada, "Camera calibration with two arbitrary coplanar circles," in *ECCV*, 2004.

[9] Jonathan Deutscher, Michael Isard, and John MacCormick, "Automatic camera calibration from a single manhattan image," in *ECCV*, 2002.

[10] Stephen Lin, Jinwei Gu, Shuntaro Yamazaki, and Heung-Yeung Shum, "Radiometric calibration from a single image," in *CVPR*, 2004.

[11] Maxwell B Clowes, "On seeing things," *Artificial intelligence*, vol. 2, no. 1, pp. 79–116, 1971.

[12] Derek Hoiem, Alexei A Efros, and Martial Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.

[13] David F Fouhey, Abhinav Gupta, and Martial Hebert, "Data-driven 3d primitives for single image understanding," in *ICCV*, 2013.

[14] Antonio Criminisi, Ian Reid, and Andrew Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[15] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kylia Miskell, Bobby H Braswell, Andrew D Richardson, and Robert Pless, "The global network of outdoor webcams: properties and applications," in *ACM GIS*, 2009.

[16] Abby Stylianou, Austin Abrams, and Robert Pless, "Finding jane doe: A forensic application of 2d image calibration," in *ICDP*, 2013.

[17] Roger Y Tsai, "A versatile camera calibration technique for high-accuracy 3d vision metrology using off-the-shelf tv cameras and lenses," *Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[18] Janne Heikkila and Olli Silvén, "A four-step camera calibration procedure with implicit image correction," in *CVPR*, 1997.

[19] Peter F Sturm and Stephen J Maybank, "On plane-based camera calibration: A general algorithm, singularities, applications," in *CVPR*, 1999.

[20] Zhengyou Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[21] Guang Jiang and Long Quan, "Detection of concentric circles for camera calibration," in *ICCV*, 2005.

[22] Roberto Cipolla, Tom Drummond, and Duncan P Robertson, "Camera calibration from vanishing points in image of architectural scenes.," in *BMVC*, 1999.

[23] Jean-Francois Lalonde, Srinivasa G Narasimhan, and Alexei A Efros, "Camera parameters estimation from hand-labelled sun positions in image sequences," Tech. Rep., CMU Robotics Institute, 2008.

[24] Scott Workman, R. Paul Mihail, and Nathan Jacobs, "A Pot of Gold: Rainbows as a Calibration Cue," in *ECCV*, 2014.

[25] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[27] Sergey Karayev, Aaron Hertzmann, Holger Winnemoeller, Aseem Agarwala, and Trevor Darrell, "Recognizing image style," in *BMVC*, 2014.

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

[29] Kyle Wilson and Noah Snavely, "Robust global translations with 1dsfm," in *ECCV*, 2014.

[30] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *NIPS*, 2014.

[31] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACMMM*, 2014.

[32] Patrick Denis, James H Elder, and Francisco J Estrada, "Efficient edge-based methods for estimating manhattan frames in urban imagery," in *ECCV*, 2008.

[33] Olga Barinova, Victor Lempitsky, Elena Tretiak, and Pushmeet Kohli, "Geometric image parsing in man-made environments," in *ECCV*, 2010.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, 2015.

[35] Bo Li, Kun Peng, Xianghua Ying, and Hongbin Zha, "Simultaneous vanishing point detection and camera calibration from single images," in *ISVC*, 2010.

[36] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.