Automated Calibration of CCTV Cameras

1st Giacomo D'Amicantonio
Technological University of Eindhoven
g.d.amicantonio@tue.nl

2nd Egor Bondarau
Technological University of Eindhoven
e.bondarau@tue.nl

3rd Peter H.N. De With
Technological University of Eindhoven
p.h.n.de.with@tue.nl

I. Introduction

The topic of camera calibration has been of great interest in the Computer Vision community for decades. Extrinsic and intrinsic calibration is required for applications such as sports video broadcasting, object localization and immersive imaging. A multitude of methods and algorithms have been proposed to perform semi-automated calibration in different contexts. Unfortunately, these methods are often impractical in real-world setups such as traffic surveillance cameras, which require frequent and automated re-calibration. We propose a method for automated calibration of traffic cameras that requires only the topview image of an intersection and its semantically segmented map. Our method brings two improvements to the SOTA approaches: a novel loss function called Topological Loss (TL) and a custom implementation of the Spatial Transformer Network (STN) [1].

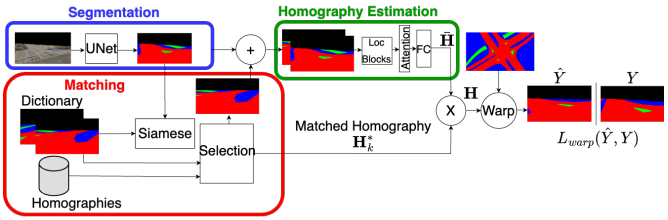

Fig. 1. Architecture of the proposed model. The homography $\hat{\mathbf{H}}$ is estimated by three so-called Localization Blocks (LocBlock) and three fully connected (FC) layers. The two matrices $\hat{\mathbf{H}}$ and $\mathbf{H}^*_K$ are multiplied to produce the final homography $\mathbf{H}$. The model warps the bird's-eye view with $\mathbf{H}$ to generate the image $\hat{Y}$.

II. Method

We generate thousands of homographies by sampling intrinsic parameters, rotation angles and camera translations. We use these homographies to warp both topviews and generate virtual camera views. The camera views are split in training, testing and dictionary splits. The segmentation component of our proposed model, shown in Figure 1, semantically segments the input image and produces a semantic map. The second component of our model, a Siamese network, retrieves the closest match for the semantic map from the dictionary of templates. The two images are concatenated across the channel dimension and passed to the thirs component of our model, the STN. Our implementation of the STN consists of three Localization Blocks, each containing three convolutions connected via skip connect and followed by batch normalization and GELU activation. Finally, a self-attention layer and three fully-connected layers estimate an homography matrix. The homography of the matched image and the estimated one are multiplied to obtain the final homography. The topview is then warped with the resulting homography. The image created by the model and the semantic ground truth used for the segmentation component are compared using a pixel-based loss function. Comparing two semantic images in this way incurs in the pitfall of parts of the images being identical while depicting very different parts of the intersection. To address this problem, our TL splits the two images is patches and computes a score between corresponding patches using a pixel-based loss function such as MSE or Dice Loss as shown in Figure 2. Each patch's score is summed to the scores of its neighbouring patches to enforce the model to generate images consistent with the topology of the intersection.

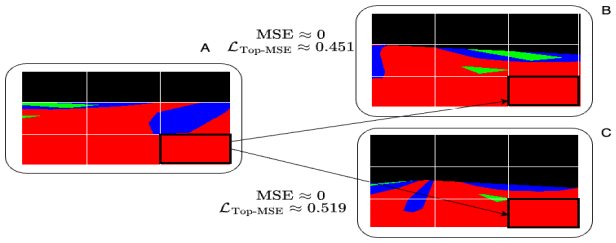

Fig. 2. MSE and $\mathcal{L}_{\text{Top-MSE}}$ scores between two patches. Notice that, using MSE, the patch would be considered almost completely correct while using $\mathcal{L}_{\text{Top-MSE}}$, the error is quite large.

TABLE I
IoU scores.

Method	Measured IoU			
	$\mathcal{L}_{\text{Top-MSE}}$	MSE	$\mathcal{L}_{\text{Top-Dice}}$	Dice
Sha et al.	75.93%	75.15%	**76.18%**	74.77%
Ours	85.12%	83.29%	**87.00%**	84.71%

III. Results

We compare the proposed model and loss function with the previous SOTA model proposed by Sha et al. [2], which used a similar approach. The performance improvement brought by TL can be noticed by comparing adjacent columns in Table I. We implemented TL using both Dice Loss and MSE to show that the idea behind it is sound. The STN improvement can be observed by comparing the rows in the table. The combination of TL and the new STN improves upon the competitor's results by up to 11%.

IV. Conclusion

Our proposed model and loss function proved to be very effective to automatically re-calibrate traffic surveillance cameras. Future work should focus on improving the matching component.

References

[1] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," NIPS, 2015.
[2] L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly, "End-to-end camera calibration for broadcast videos," in *CVPR*, 2020.