Matchable Image Transformations for Long-term Metric Visual Localization

Lee Clement, Mona Gridseth, Justin Tomasi and Jonathan Kelly
University of Toronto Institute for Aerospace Studies
{lee.clement,mona.gridseth,justin.tomasi,jonathan.kelly}@robotics.utias.utoronto.ca

1. Introduction

Long-term metric localization is an essential capability of autonomous mobile robots, but remains challenging for vision-based systems in the presence of appearance change caused by lighting, weather or seasonal variations. While experience-based mapping [1, 7] has proven to be an effective technique for enabling visual localization across appearance change [9, 10], the number of experiences required for reliable long-term localization can be large, and methods for reducing the necessary number of experiences are desired. Taking inspiration from physics-based models of color constancy [12], we propose a method for learning a nonlinear mapping from RGB to grayscale colorspaces that maximizes the number of feature matches for images captured under varying lighting and weather conditions. Our key insight is that useful image transformations can be learned by approximating conventional non-differentiable feature matching algorithms with a differentiable learned model. Moreover, we find that the generalization of appearance-robust RGB-to-grayscale mappings can be improved by incorporating a learned low-dimensional context feature computed for a specific image pair. Using synthetic and real-world datasets, we show that our method substantially improves feature matching across day-night cycles and presents a viable strategy for improving the efficiency of experience-based visual localization.

2. Technical Approach

Our goal in this work is to learn a nonlinear transformation $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ mapping the RGB colorspace onto a grayscale colorspace that explicitly maximizes the number of feature matches for a given image pair. We investigate two approaches to formulating such a mapping: 1) a single function to be applied to all incoming images, similarly to [2, 8, 11]; and 2) a parametrized function tailored to the specific image pair to be used for localization, where the parameters are derived from the images themselves.

Ideally we would like the learned colorspace transformation to be tied to the performance of the target feature detector/matcher. However, the most commonly used feature detectors/matchers in robotics rely on non-differentiable components such as nearest-neighbors search and RANSAC [4], which makes gradient-based optimization infeasible. In this work we *learn* an objective function by training a deep convolutional neural network (CNN) to act as a *differentiable proxy* to the localization front-end. Specifically, we train a siamese CNN $\mathcal{M}_{\boldsymbol{\theta}}$ to predict the number of feature matches for a large set of image pairs, where the training targets are generated using a conventional non-differentiable feature detector/matcher algorithm D such as `libviso2` [6]. We train $\mathcal{M}_{\boldsymbol{\theta}}$ to approximate D by minimizing the squared error of the prediction for overlapping image pairs $(\mathbf{I}_1, \mathbf{I}_2)$:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{I}_1^i, \mathbf{I}_2^i) - D(\mathbf{I}_1^i, \mathbf{I}_2^i) \right)^2 \tag{1}$$

This formulation naturally admits a self-supervised training approach as training targets can be generated on the fly by D. We then use the trained proxy network to define a fully differentiable objective function, which we can use to train a nonlinear colorspace mapping using gradient-based methods. Figure 1 summarizes our full data pipeline visually.

We compare two approaches to formulating the colorspace transformation. The first is a generalization of the color constancy transformation proposed in [12]:

$$\mathbf{F} = \alpha \log \mathbf{R} + \beta \log \mathbf{G} + \gamma \log \mathbf{B}, \tag{2}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha & \beta & \gamma \end{bmatrix}^T = \mathcal{E}_{\boldsymbol{\phi}}(\mathbf{I}_1, \mathbf{I}_2). \tag{3}$$

where $\mathcal{E}_{\boldsymbol{\phi}}$ is a pairwise encoder network. The second is to replace Equation (2) with a learned transformation $\mathcal{T}_{\boldsymbol{\psi}}$ parametrized as a multilayer perceptron (MLP). In both cases, we optimize the joint model by minimizing

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\psi}) = \frac{1}{N} \sum_{i=1}^{N} \left| \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{I}_1^i, \mathbf{I}_1^i) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{F}_1^i, \mathbf{F}_2^i) \right|, \tag{4}$$

alternating between optimizing Equations (1) and (4) at each iteration. We refer to the physics-based model as "SumLog" and "SumLog-E", where the latter uses the encoder network $\mathcal{E}_{\boldsymbol{\phi}}$ to derive the $\boldsymbol{\beta}$ and the former uses a constant value. We refer to the equivalent MLP-based models

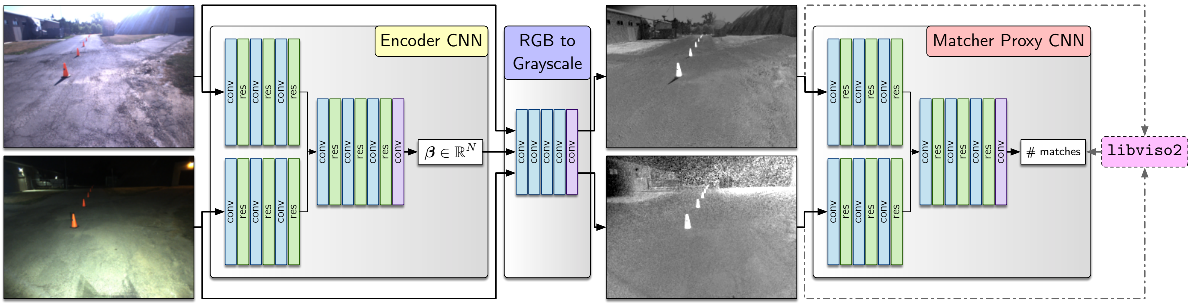
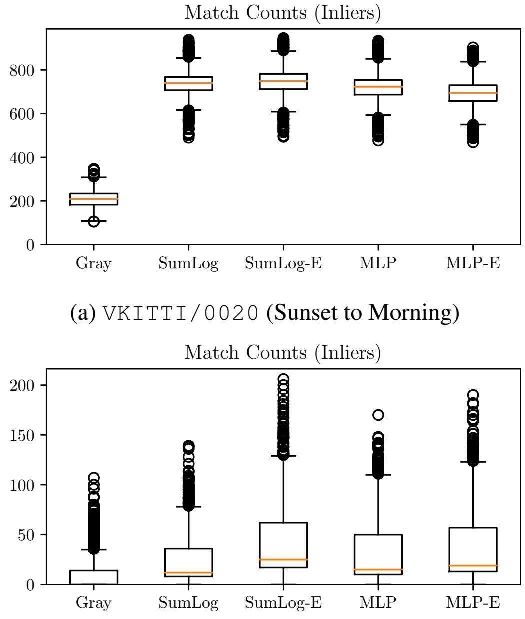


Figure 1: We train a CNN to estimate the number of feature matches for a given feature detector and image pair, and use the trained model as a differentiable loss function to learn a colorspace transformation that maximizes the number of matches.



(a) `VKITTI/0020` (Sunset to Morning)

(b) `InTheDark/0041` (Night to Day)

Figure 2: Box-and-whiskers plots of inlier feature matches for corresponding image pairs using each transformation.

as “MLP” and “MLP-E”, and the standard weighted average grayscale transformation as “Gray”.

3. Experiments

We conduct experiments on the Virtual KITTI (`VKITTI`) dataset [5] and the UTIAS In The Dark (`InTheDark`) dataset [10], both of which exhibit substantial variation in illumination conditions. Figure 2 shows the distributions of inlier `libviso2` feature matches using each RGB-to-grayscale transformation for nearby image pairs from sequences `VKITTI/0020` (Sunset to Morning) and `InTheDark/0041` (Night to Day). We observe that all four transformations more than double the median number of feature matches for `VKITTI/0020`, and that the gains are only slightly higher using the pairwise encoder. These results are consistent with the findings of [2, 3, 8, 11], where one or two sets of parameter values were sufficient to achieve good performance across varying daytime conditions. For `InTheDark/0041`, we note that while both the SumLog and MLP transformations increase the median number of feature matches, the use of the pairwise encoder network provides a substantial performance boost to both methods. We attribute this difference to a wider variety of illumination conditions in the data, where a single transformation is unlikely to perform well under all conditions.

We also note that the MLP-E transformation generally performs similarly to the SumLog-E transformation on both datasets. Qualitatively, we observed that the MLP and MLP-E methods produce output images that are visually similar to their SumLog counterparts. This suggests that a weighted sum of log-responses may in fact be an optimal solution for this problem, and that a careful choice of weights is key to good cross-appearance feature matching.

4. Conclusions

This paper presents a method for learning RGB-to-grayscale colorspace mappings that explicitly maximize the number of feature matches for a given image pair, feature detector/matcher and operating environment. By training a CNN to approximate the behavior of a conventional non-differentiable feature detector/matcher, we learn a fully differentiable loss function that can be used to train a useful image transformation. We evaluate our approach using both physically motivated colorspace transformations and trainable transformations and demonstrate substantially improved feature matching performance at test time on both synthetic and real long-term vision datasets exhibiting severe illumination change. We find that the best performance is consistently achieved using a physically motivated weighted sum of log-responses with the weights derived from a pairwise context encoder network.

References

[1] W Churchill and P Newman. Experience-based navigation for long-term localisation. *International Journal of Robotics Research*, 32(14):1645–1661, Dec. 2013. 1

[2] L Clement, J Kelly, and TD Barfoot. Robust monocular visual teach and repeat aided by local ground planarity and color-constant imagery. *Journal of Field Robotics*, 34(1):74–97, 1 Jan. 2017. 1, 2

[3] P Corke, R Paul, W Churchill, and P Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2085–2092, Nov. 2013. 2

[4] MA Fischler and RC Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automatated cartography. *Communications of the ACM*, 24(6):381–395, June 1981. 1

[5] A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Compututer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, June 2016. 2

[6] A Geiger, J Ziegler, and C Stiller. StereoScan: Dense 3D reconstruction in real-time. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968. IEEE, June 2011. 1

[7] C Linegar, W Churchill, and P Newman. Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 90–97, May 2015. 1

[8] C McManus, W Churchill, W Maddern, A D Stewart, and P Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 901–906, May 2014. 1, 2

[9] M Paton, K MacTavish, LP Berczi, SK van Es, and TD Barfoot. I can see for miles and miles: An extended field test of visual teach and repeat 2.0. In *Proceedings of the Conference on Field and Service Robotics (FSR)*, pages 415–431, 2018. 1

[10] M Paton, K MacTavish, M Warren, and TD Barfoot. Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1918–1925, Oct. 2016. 1, 2

[11] M Paton, F Pomerleau, K MacTavish, CJ Ostafew, and TD Barfoot. Expanding the limits of vision-based localization for long-term route-following autonomy. *Journal of Field Robotics*, 34(1):98–122, Jan. 2017. 1, 2

[12] S Ratnasingam and S Collins. Study of the photodetector characteristics of a camera for color constancy in natural scenes. *Journal of the Optical Society of America A*, 27(2):286–294, Feb. 2010. 1