# Similarity criteria: evaluating perceptual change for visual localization

Stephanie Lowry[1]

***Abstract*— Visual localization systems may operate in environments that exhibit considerable perceptual change. This paper proposes a method of evaluating the degree of appearance change using a similarity criteria based on comparing the subspaces spanned by the principal components of the observed image descriptors. We propose two criteria – $\theta_{\min}$ measures the minimum angle between subspaces and $S_{\texttt{total}}$ measures the total similarity between the subspaces. These criteria are introspective – they evaluate the performance of the image descriptor using nothing more than the image descriptor itself. Furthermore, we demonstrate that these similarity criteria reflect the ability of the image descriptor to perform visual localization successfully, thus allowing a measure of quality control on the localization output.**

## I. INTRODUCTION

Visual localization – the ability for robots to recognize and identify their location in a large-scale environment – has made rapid progress in recent years [1]. Perceptual change can be challenging for visual localization systems, as places in the environment may not look the same as they did on previous occasions and a visual localization system can fail to correctly identify matching locations due to extreme changes in appearance. Perceptual change is particularly challenging when it happens uniformly over a region – for example, if day turns to night or snow falls – as all places in that region will become difficult to recognize (Fig. 1). However, the spatial relationship between places does not change, so as long some similarity in appearance remains, a weak location hypothesis can be formed and by observing multiple nearby locations and the spatial relationship between them, you can gradually build up confidence in your location belief.

It helps to know whether or not the system is localizing in a perceptually changing environment. Perceptual change requires a permissive matching strategy where places may be matched together even when they do not appear similar at all, and such an approach may cause incorrect matching if applied to a less challenging environment: if the appearance of the environment has not changed, a strict matching strategy where places must be both highly similar and highly distinctive may be more reliable and stop false positive matches (this opposing problem is known as *perceptual aliasing* where different places may look very similar).

Thus an important consideration for a localization system is *context* – is the system in a situation when it should demand highly rigorous matching expectations or is a more permissive strategy necessary? Furthermore, the environment may take on a large range of different conditions: not only may night and day be different, but dawn, noon, and dusk may also exhibit different characteristics. Weather conditions could include sun, rain, and cloud, while seasonal changes may affect whether there are leaves on the trees or snow on the ground. Since all these factors can operate independently, it is not feasible to expect a system to be able to predict all the potential scenarios it may see. Thus we propose a simplified single-valued criteria based on *similarity*. This similarity criteria is intended to subsume all the possible condition changes and evaluate if the change is sufficiently major to pose challenges to the visual localization system.

(a) Perceptual change

(b) Perceptual aliasing

Fig. 1. Perceptual change and perceptual aliasing are two challenges for visual localization systems. (a) Perceptual change occurs when the same location looks different at different times. (b) Perceptual aliasing occurs when two different locations look similar.

This paper presents a method of quantifying environmental similarity and perceptual change. It uses principal component subspace comparison to identify the perceptual context within the environment, and therefore will provide information to a visual localization regarding the level of perceptual change. These methods require no external information and only using the image descriptors included in the visual localization system. We present two *similarity criteria*: one that measures the total similarity between the perceptual contexts and one the measures the extent to which the subspaces differ.

This paper demonstrates that both these criteria are correlated with the ability of the selected image descriptor to perform visual localization. This result demonstrates that these similarity criteria provide information about the likely performance of the visual localization system, without any additional information about the environment.

## II. PRIOR WORK

Visual localization has developed rapidly as a field in recent years. One driving motivation is the observation that

*This work was supported by the Swedish Research Council (grant no. 2018-03807).

[1]The author is with the Centre for Applied Autonomous Sensor Systems (AASS), Örebro University, 70281 Örebro, Sweden stephanie.lowry@oru.se

978-1-7281-3605-9/19/$31.00 

changing appearance in an environment, including (most commonly) lighting change, is a major cause of failure in location matching experiments [2]. Like many other vision-based fields, visual localization has been inspired by the astonishing advances in machine learning for computer vision and learning methods are utilized to improve location representations [3], [4], [5], [1]. The results have shown impressive perceptual change invariance, allowing localization systems to operate successfully in highly perceptually changing environments. Furthermore, methods such as NetVLAD [5] and DenseVLAD [6] can perform localization that is robust to both perceptual change and considerable viewpoint variation [1].

However, even the most powerful visual description methods can struggle with drastic perceptual change, especially if combined with other challenges such as low light, blur, and general poor image quality. It is useful for a localization system to produce a confidence metric that assesses how reliable the localization system is in the current scenario. Confidence metrics have typically been provided by frameworks such as probabilistic localization systems [7], [8], [9]. However, probabilistic systems depend on prior information, and likelihood models of the environment which in a perceptually changing environment may themselves be unreliable.

A measure of environmental appearance change can be used as a method of selecting an observation likelihood model [9]. This paper extends these results to a more general similarity criteria based on the subspaces formed by the principal components of image descriptors in different perceptual conditions. This approach has some similarities with hierarchical environment selection [10], but in this case there is no assumption that the same environment will maintain the same overall perceptual characteristics.

## III. APPROACH

Suppose a visual localization system is matching a set of images $\mathcal{A}$ and $\mathcal{B}$. The images within each set are captured close together both spatially and temporally, and thus are likely to have similar perceptual characteristics – for example, it might by day-time or nighttime, or snowy, or rainy. However, the particular perceptual characteristics of each cluster are unknown. The goal of the similarity critieria evaluation is to determine whether the perceptual characteristics of $\mathcal{A}$ and $\mathcal{B}$ are *similar* or *different*.

The similarity criteria does not perform visual localization – that is, it does not determine if groups $\mathcal{A}$ and $\mathcal{B}$ are from a similar location, as two spatially distant environments might appear superficially similar if they were viewed at similar times of day under similar conditions, while the same environment might appear very different under very different conditions.

More formally, these groups of images have the same variables measured on them (that is, the chosen image descriptor) and we wish to evaluate how similar the groups are with respect to their overall features. One straightforward and intuitive procedure to analyze relevant factors is to describe each group in terms of its principal components and compare these. We follow the method used in [11].

### A. Notation

Let $\mathcal{A}$ and $\mathcal{B}$ represent two groups of images such as those described above. Let $\mathbf{L}_k$ be the matrix containing the first $k$ principal components of $\mathcal{A}$ and let $\mathbf{M}_k$ be the matrix containing the first $k$ principal components of $\mathcal{B}$. Define

$$\mathbf{S}_k = \mathbf{L}_k \mathbf{M}_k^\top \mathbf{M}_k \mathbf{L}_k^\top. \tag{1}$$

### B. Minimum subspace angle

The first similarity criterion we define is the *minimum subspace angle*, denoted $\theta_{\min}$. This angle is the minimum angle between an arbitrary vector in the space spanned by $\mathbf{L}_k$ and the space spanned by $\mathbf{M}_k$. The size of the minimum subspace angle provides a measure of the extent to which the subspaces differ: a small minimum subspace angle suggests the subspaces are more similar than a large minimum subspace angle.

As shown in [11], the minimum subspace angle is given by

$$\theta_{\min} = \cos^{-1}(\sqrt{\lambda_1}). \tag{2}$$

In this equation, $\lambda_1$ is the largest eigenvalue of $\mathbf{S}_k$.

### C. Total similarity

The second similarity criteria we define is the *total similarity*, denoted $S_{\texttt{total}}$, as the sum of the squares of the cosines $\cos^2(\theta_{ij})$ between each of the $k$ principal components in $\mathbf{L}_k$ and $\mathbf{M}_k$. The sum is then normalized between 0 and 1. It is shown in [11] that if $\lambda_i$ is the $i$-th eigenvalue of $\mathbf{S}_k$ then:

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\cos^2(\theta_{ij}) = \sum_{i=1}^{k}\lambda_i \tag{3}$$

$$= \texttt{trace}(\mathbf{S}_k). \tag{4}$$

Since the sum of the eigenvalues is constrained between 0 for completely orthogonal spaces and $k$ for coincident spaces, a normalized $S_{\texttt{total}}$ can be calculated via:

$$S_{\texttt{total}} = \frac{\texttt{trace}(\mathbf{S}_k)}{k}. \tag{5}$$

### D. Descriptor-dependent similarity criteria

These similarity criteria are descriptor-dependent, as they measure the similarity of the two environments according to a particular descriptor space. This is a reasonable approach, since localization is itself descriptor-dependent – one descriptor may easily perform visual localization on one environment but be outperformed by another descriptor on another environment. However, a deficiency in this formulation is that similarity criteria are not able to be compared between descriptors – because the similarity criteria is defined on the original descriptor space it is not necessarily meaningful to compare the results between two descriptor types. The similarity criteria can be used to compare different environments, as we show in the experiments below.

## IV. EXPERIMENTAL SETUP

### A. Datasets

The similarity criteria were evaluated using two benchmark datasets: the RobotCar Seasons dataset [1],[12] and the Nordland train dataset. The RobotCar Seasons visual localization dataset[1] is derived from a subset of the larger Oxford RobotCar dataset [12]. The images were recorded in the city of Oxford, UK on a car driving the same route over a period of 12 months. The RobotCar Seasons dataset contains 10 traversals under different conditions (see Fig 2 for sample images from each condition). This paper uses the left camera image from each location, and derives an approximate ground truth using the GPS+Inertial data provided by the original dataset [12][2]. The Nordland train dataset[3] consists of four traversals of a 700-kilometres Norwegian train journey during four different seasons (see Fig 3 for sample images).

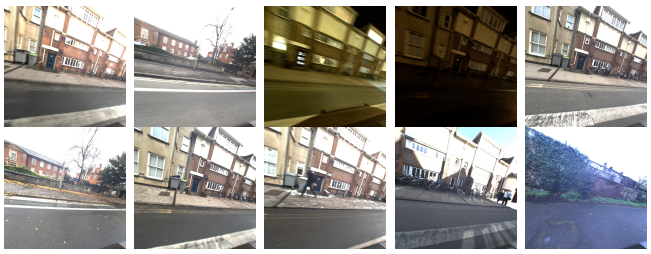


Fig. 2. Sample images from each of the RobotCar Seasons traversals. The night-time traversals are particularly challenging due to motion blur and low image quality.

Fig. 3. Sample images from each of the Nordland traversals.

For each traversal from the datasets, 358 images were extracted. For the RobotCar Seasons traversals, the first 358 images were used while for the Nordland traversals, the GPS-aligned images were sampled and every 100th image was used.

### B. Image Description

The goal of these experiments was to assess whether the similarity criteria indicated whether the conditions were perceptually challenging for a visual localization system. To do so, we used a visual localization system based on two different image descriptors: NetVLAD [5] and downsampled images.

*1) NetVLAD:* NetVLAD is a state-of-the-art image description technique for visual localization, and has shown extremely good performance at place recognition on many datasets. On the RobotCar Seasons dataset specifically, NetVLAD was demonstrated in [1] to have excellent performance at coarse-precision place recognition (matching places within 5 metres and 10 degrees of each other) on the day-time conditions, but struggled on the night-time images. Nonetheless, it was one of the top two performers (along with DenseVLAD [6]) for the night-time scenario.

We used the implementation of NetVLAD provided by the authors[4], using the best performing pre-trained model[5].

*2) Images:* Downsampled images can be used as a simple descriptor for a visual localization system. In this work the images were converted to grayscale and downsampled to $64 \times 64$ pixels to retain the same number of dimensions (4096) as NetVLAD. The small images were whitened using the method presented in [13].

### C. Visual Localization

This evaluation uses an extremely simple visual localization technique. For each image $I_1$ with $N$-dimensional descriptor $d_1$ in dataset $D_1$, to determine the best matching image in dataset $D_2$, we extracted the descriptor $d_i$ for each image $I_i \in D_2$. We then calculated the zero normalized cross correlation (ZNCC) [14] via:

$$z(d_1, d_2) = \frac{\sum_{u \leq N} (d_1[u] - \bar{d_1}) \cdot (d_2[u] - \bar{d_2})}{\sqrt{\sum_{u \leq N} (d_1[u] - \bar{d_1})^2 \cdot \sum_{u \leq N} (d_2[u] - \bar{d_2})^2}}$$

The ZNCC is typically a measure used for comparing images, but also demonstrates good performance on non-image descriptors such as NetVLAD. The best match for $d_1$ was considered the image descriptor $d \in D_2$ with the largest ZNCC value $z(d_1, d_2)$.

### D. Localization metrics

To evaluate whether the similarity criteria provide useful information about the potential success of the visual localization success, a very simple localization metric was used: the Fraction of Correct Matches (FCM). We calculated for how many images $I_1$ its corresponding best matching image $I_2$ was a ground truth match, out of the total number of images for which a ground truth match existed.

A second localization metric was evaluated: the recall at 100% precision (Recall@100), but due to space constraints only the FCM results are presented in the paper. Recall@100 is a much more sensitive metric than FCM. A single false match can drastically change the value of Recall@100, while a false match can only change FCM a small amount.

## V. RESULTS

The results section presents an analysis of the impact of the primary parameter in the similarity criteria definition, the number of principal components used. It then evalutes the relationship between the similarity criteria and the selected visual localization metric.

[1]Available for download at `www.visuallocalization.net`

[2]Available for download at `https://robotcar-dataset.robots.ox.ac.uk/`

[3]available for download at `https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/`

[4]Available from `https://www.di.ens.fr/willow/research/netvlad/`.

[5]VGG-16+NetVLAD+whitening, trained on Pittsburgh

### A. Number of principal components to use

The only parameter in the definition of the similarity criteria is $k$, the number of principal components used. Fig. 4 displays the relationship between $k$ and the similarity metrics. There is a visible qualitative difference for both $S_{\texttt{total}}$ and $\theta_{\min}$ between the easy-to-match traversals (the `night` traversal for the `night-rain` example) and the more challenging traversals. This qualitative difference is not significantly impacted by the choice of $k$, and both metrics change smoothly and gradually as $k$ increases. Aside from a peak at low $k$ (around $k = 4$) in $S_{\texttt{total}}$, there appears to be little sensitivity to the choice of $k$. While the peak at $k = 4$ might be interesting to investigate for future work, the remainder of this work uses $k = 20$, as it is large enough to avoid this potential source of noise.

### B. Similarity metric evaluation

This section evaluates whether the similarity criteria is able to evaluate the visual localization ability of a particular descriptor in different perceptual conditions.

Fig. 5 displays the range of results when comparing the Nordland traversals to the same environment under different conditions as well as to a totally different environment (the RobotCar Seasons dataset). $S_{\texttt{total}}$ is much higher between the Nordland traversals than between the Nordland and RobotCar Seasons traversals (where it is close to zero). Similarly, $\theta_{\min}$ is much lower between the Nordland traversals than between the Nordland and RobotCar Seasons traversals (it is close to $80°$, compared to less than $25°$ between the Nordland traversals).

Fig. 6 displays the results for the `dawn`, `dusk`, and `night-rain` datasets. Fig. 6a displays the results for NetVLAD and Fig. 6b displays the results for images. The lines show the linear regression for each base dataset. The results show that the similarity criteria demonstrates a clear relationship between the FCM and both $S_{\texttt{total}}$ and $\theta_{\min}$. This result is true for both NetVLAD and for images, but the correlation is weaker for images. To quantify the results the mean $R^2$ value across all RobotCar Seasons traversals is shown in Table I. The mean $R^2$ value for NetVLAD is 0.89 for $S_{\texttt{total}}$ and 0.92 for $\theta_{\min}$, while it is 0.63 for $S_{\texttt{total}}$ and 0.65 for $\theta_{\min}$ when downsampled images are used as the descriptor.

TABLE I
MEAN $R^2$ VALUES ON ROBOTCAR SEASONS DATASET

| Descriptor | $R^2$ value | |
|---|---|---|
| | $S_{\texttt{total}}$ | $\theta_{\min}$ |
| NetVLAD | 0.89 | 0.92 |
| Images | 0.63 | 0.65 |

These results also show some of the limitations of the similarity criteria. Firstly, they cannot be used to compare the performance of the different descriptors – for example, an $\theta_{\min}$ of $45°$ represents poor performance for an image descriptor (only about $20\%$ of places are correctly matched) but good performance for NetVLAD (close to $100\%$ of places correctly matched).

Fig. 7 displays $S_{\texttt{total}}$ and $\theta_{\min}$ for NetVLAD on the Nordland dataset. As for the RobotCar Seasons datasets, a larger $S_{\texttt{total}}$ or a smaller $\theta_{\min}$ are correlated to a higher FCM.

## VI. CONCLUSIONS

Solving the problem of unconstrained localization in uncontrolled environments is of great practical significance as a fundamental step towards mobile systems that can operate autonomously within real-world environments. Visual localization is an important component of a localization and navigation system as cameras provide rich semantic information about the world unmatched by any other sensor. However, to be effective a visual localization framework must be able to independently identify the nature of the environment in which it is operating. The similarity criteria presented here represent a step towards this goal.

The similarity criteria here is introspective – it uses no additional information beyond the image descriptors themselves, and provides some insight into the likely performance of those descriptors on a given environment. This introspective approach could of course also be augmented by including external sources of information like the time of day, or the weather forecast. However, it is interesting that conclusions can be drawn simply from the vector subspace structure of the visual descriptors, without requiring external sources of data.

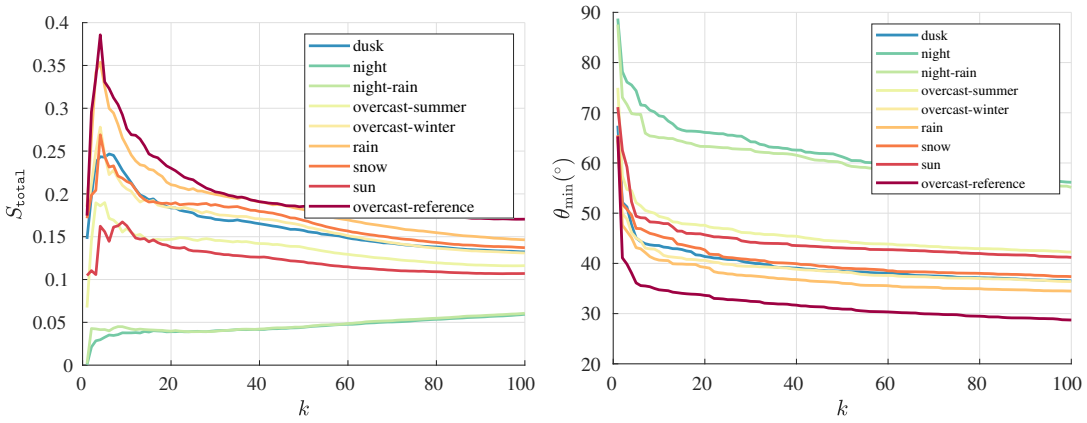
The localization system used for evaluation is very simple and contains no additional filtering based on prior belief or secondary verification. While these sophisticated mechanisms can of course improve localization performance, any localization system is fundamentally dependent on the choice of image descriptors, and the similarity criteria provides information which can then be used to enhance any localization system.

An interesting future research direction is also the interplay between perceptual aliasing and perceptual change. The similarity criteria measures the perceptual change *between* environment states, while perceptual aliasing is a measure of similarity *within* environment states. Future work will investigate the relationship of the similarity criteria to the perceptual aliasing within an environment.

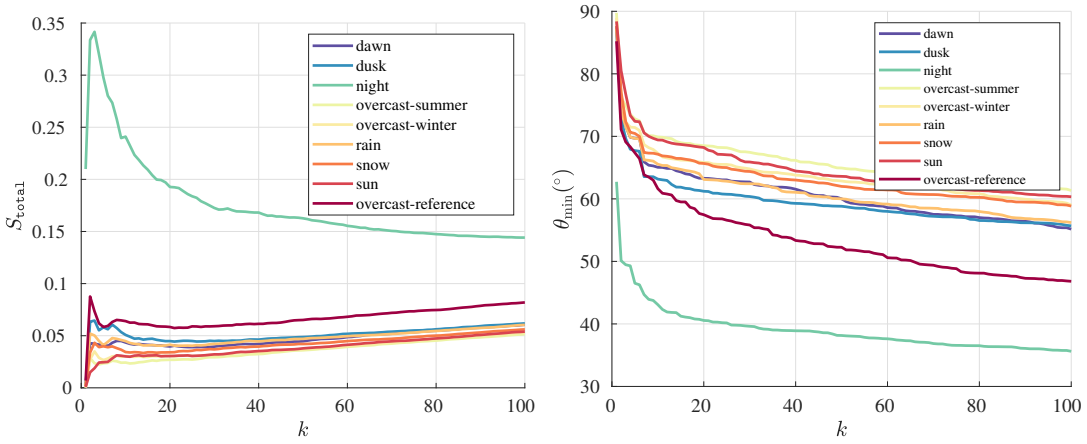
The proposed criteria have limitations. As the criteria are defined according to subspaces within the original descriptor space, they are unable to systematically compare different descriptors defined on different feature spaces. However, similarity criteria across environments have potential to provide valuable contextual information to visual localization systems.

## REFERENCES

[1] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF outdoor visual localization in changing conditions," in *CVPR*, 2018.

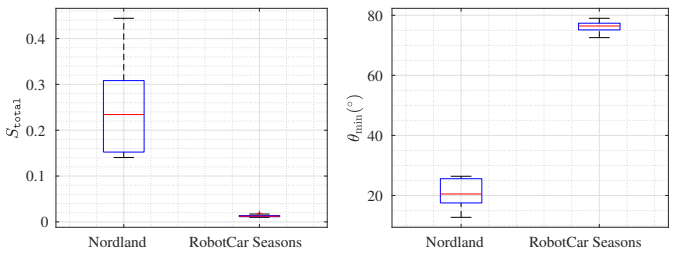



(a) Dawn traversal




(b) Night-rain traversal

Fig. 4. Effect of $k$ (the number of principal components) on $S_{\texttt{total}}$ and $\theta_{\min}$. The value of $k = 20$ was chosen for the evaluation experiments.




(a) $S_{\texttt{total}}$ (b) $\theta_{\min}$

Fig. 5. Similarity criteria $S_{\texttt{total}}$ and $\theta_{\min}$ on the Nordland traversals using NetVLAD compared to the same environment under different conditions (*Nordland*) and a different environment (*RobotCar Seasons*). (a) $S_{\texttt{total}}$ is higher between the Nordland traversals than between the Nordland and RobotCar Seasons traversals. (b) $\theta_{\min}$ is lower between the Nordland traversals than between the Nordland and RobotCar Seasons traversals.


[2] P. Furgale and T. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, vol. 27, no. 5, pp. 534–560, 2010.

[3] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2015, pp. 4297–4304.

[4] N. Carlevaris-Bianco and R. M. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2014, pp. 2769–2776.

[5] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, June 2018.

[6] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015.

[7] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[8] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 1635–1642.

[9] S. Lowry and M. J. Milford, "Building beliefs: Unsupervised generation of observation likelihoods for probabilistic localization in changing environments," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2015, pp. 3071–3078.

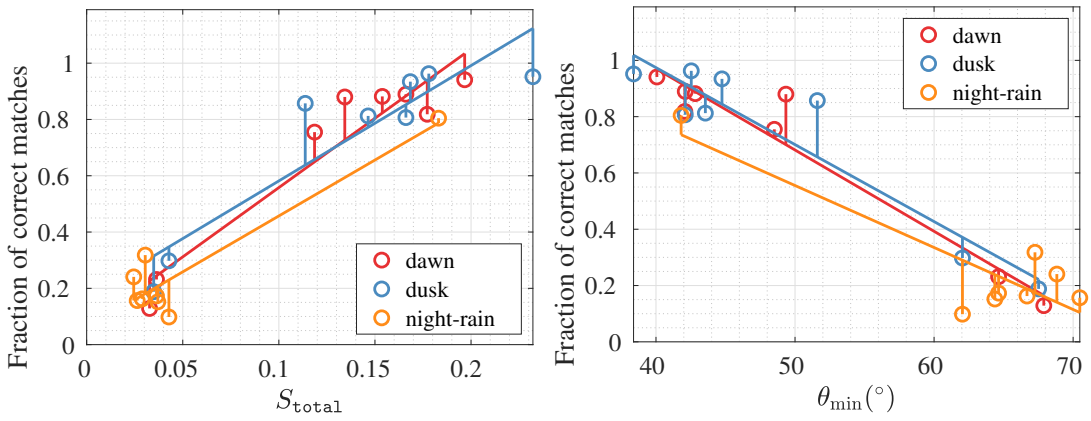
[10] M. Mohan, D. Gálvez-López, C. Monteleoni, and G. Sibley, "Environment selection and hierarchical place recognition," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 5487–5494.

[11] W. Krzanowski, "Between-groups comparison of principal components," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 703–707, 1979.

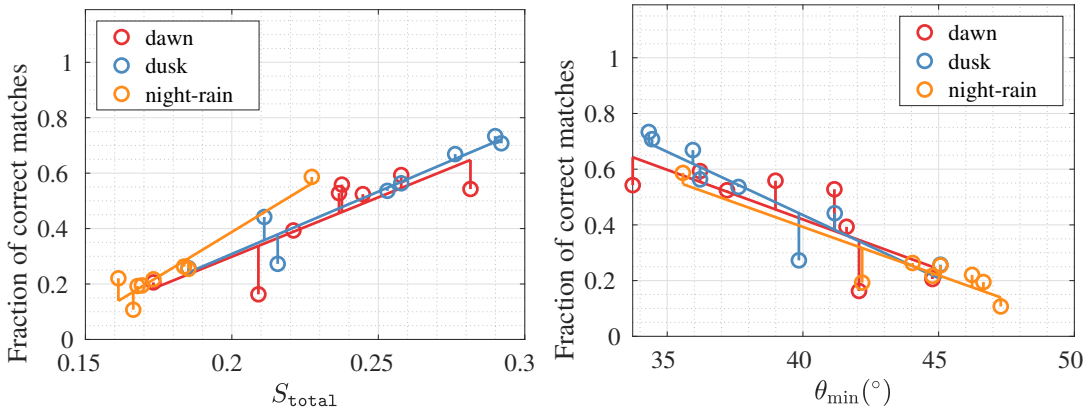
[12] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.

[13] S. Lowry and M. J. Milford, "Supervised and unsupervised linear learning techniques for visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 32, no. 3, pp. 600–613, June 2016.

[14] P. I. Corke, *Robotics, Vision & Control: Fundamental Algorithms in Matlab*. Springer, 2011.




(a) NetVLAD




(b) Images

Fig. 6. Similarity criteria $S_{\texttt{total}}$ and $\theta_{\min}$ against FCM on the RobotCar Seasons datasets. For a given base dataset and descriptor, a greater $S_{\texttt{total}}$ or a smaller $\theta_{\min}$ implies a greater FCM.

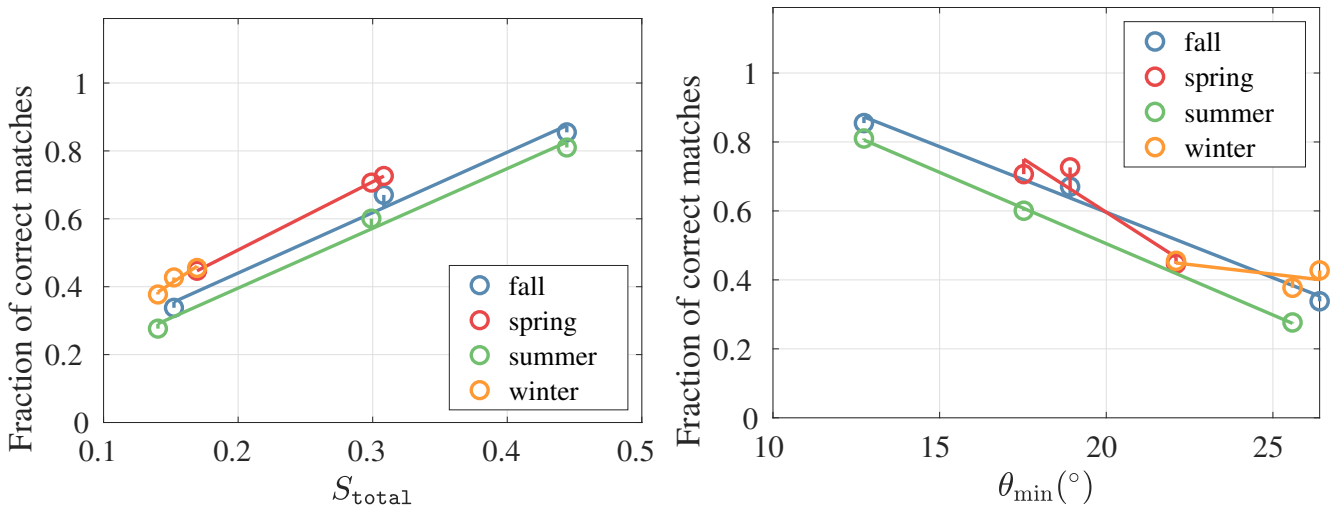



Fig. 7. Similarity criteria $S_{\texttt{total}}$ and $\theta_{\min}$ against FCM on the Nordland datasets using NetVLAD. As for the RobotCar Seasons dataset, a greater $S_{\texttt{total}}$ or a smaller $\theta_{\min}$ implies a greater FCM.