# Characterizing Feature Matching Performance Over Long Time Periods

Abby Stylianou, Austin Abrams, Robert Pless
Washington University in St. Louis
astylianou@seas.wustl.edu, {abramsa|pless}@cse.wustl.edu

## Abstract

*Many computer vision applications rely on matching features of a query image to reference data sets, but little work has explored how quickly data sets become out of date. In this paper we measure feature matching performance across 5 years of time-lapse data from 20 static cameras to empirically study how feature matching is affected by changing sunlight direction, seasons, weather, and the structural changes over time in outdoor settings.*

*We identify several trends that may be relevant in real-world applications: (1) features are much more likely to match within a few days of the reference data, (2) weather and sun-direction have a large effect on feature matching, and (3) there is a slow decay over time due to physical changes in a scene, but this decay is much smaller than effects of lighting direction and weather.*

*These trends are consistent across standard choices for feature detection (DoG, MSER) and feature description (SIFT, SURF, and DAISY). Across all choices, analysis of the feature detection and matching pipeline highlights that performance decay is mostly due to failures in key point detection rather than feature description.*

## 1. Introduction

Robust approaches to matching features taken at different times and from slightly different viewpoints have made numerous computer vision applications possible. For some of those applications, the problem domain requires matching a current image to older imagery. These applications, including approaches to geo-location, geo-orientation [13], geo-tagging [16], landmark recognition [23], image based localization [14], camera-pose estimation [7], and historical rephotography [1], are all based on a database of reference imagery.

Here we ask the question, "how does this reference imagery age?", or more precisely, "how does feature matching performance change over time?"

This question is relevant at different time scales; over short time periods, the illumination conditions may change, and over longer time periods there may be effects due to weather, seasons, and erosion, plant growth or building construction. How much can these factors change before feature matching fails, and what components of the algorithm cause this failure?

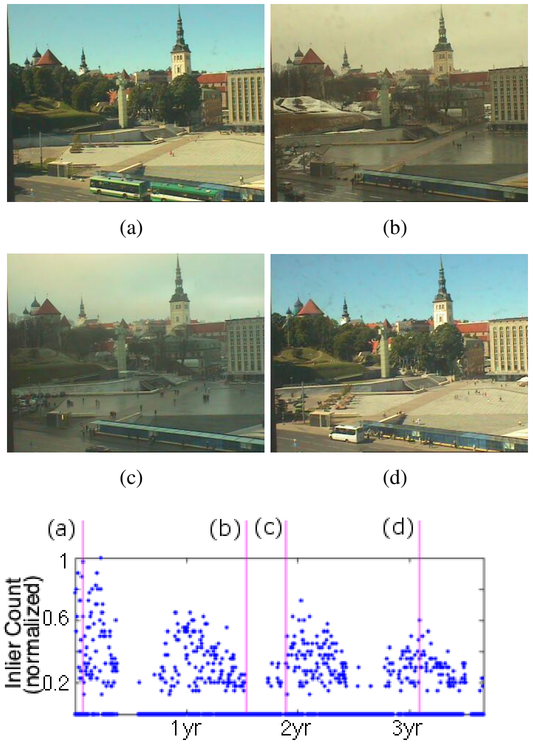

(a) (b)
(c) (d)

Figure 1. Our pipeline starts by extracting key points and computing feature descriptors in a static webcam image from within 24 hours of either the summer or winter solstice. We then extract key points and compute feature descriptors in every subsequent image from that camera at the same time of day, and find the number of inliers between each image pair. This process reveals significant trends over time that are explored throughout this paper.

To answer this question, we empirically study image data captured over five years from 20 static outdoor webcams. This is an ideal data set because we know that feature matches should occur at the same scene location, and the only change in imagery comes from extrinsic changes in the scene. In practice, small camera jitters force us to allow small geometric transformations between frames, but since these are due to camera rotations, those transforms are captured by a homography. Because most real world applications of feature matching arise in conditions with cameras that are not static, we explicitly validate that we see the same trends of matching efficiency over time on several cases where we have images from a different viewpoint.

Given the importance of feature matching, there has been surprisingly little work to evaluate its performance over time. This is because it is challenging to structure an experiment where the correct feature matches are known even if current feature matching algorithms fail — especially in the context of long term, real-world data sets. Related work includes tests on synthetic dataset from a photorealistic visual world, lit to simulate different times of day [5]. The only study we know of that uses real data tests location recognition for autonomous driving. They capturing data by driving around a neighborhood several times over a nine-month period [20]. Both studies report that feature matching decreases substantially as a function of changes in illumination direction, and the real-world data also highlights weather as an important determiner of feature matching efficiency.

Relative to these previous works, our contributions are fourfold:

First, we define an experimental protocol that uses long term static webcam archives as a structured data source, allowing the first comparative evaluation of different feature detectors and feature descriptors over time in many real-world settings over time scales of many years.

Second, we show feature matching over time in fixed viewpoint images has the same trends as matching features from different viewpoints, so the experimental results will hold across standard application domains that require feature matching.

Third, we show that the effects of lighting and weather variation are the dominant cause of matching failures and dominate accumulated physical scene changes even after many years. These trends are consistent across all tested feature detectors and matchers.

Fourth, we characterize where in the feature matching pipeline the failure occurs and highlight that the vast majority of failures occur at the feature detection stage, not the feature description stage.

We choose to use widely accessible implementations of standard feature detectors (DoG, MSER) and descriptors (SIFT, SURF, DAISY). To support future comparisons with alternative implementations, the features, datasets and code for all parts of this paper will be released.

## 2. Background and Related Work

There have been a number of works comparing choices in the feature detection/description pipeline. Prior works have a large variation in the way that they create or find data with known ground truth matches. Early work evaluating interest point detection [17, 10] explored performance with respect to image transformations (affine geometric transformations and intensity variations), and viewpoint variability from moving video data of a location. Viewpoint variability was evaluated both in terms of the stability of the key point detection and matching the description, as a function of the angular viewpoint change [11], based on 144 calibrated views of objects, such as telephones, pineapples, and statues. In the context of video streams, combinations of detectors and descriptors have been evaluated with respect to changes in lighting conditions, geometric changes and motion blur [3].

Fewer works explicitly consider the effect of long-term time variations on feature matching. The matching of SIFT features in historical imagery has been used to to sort them by date, based on the construction and demolition of buildings [15]. Matching of World War II era aerial images to modern aerial images was used to detect possible buried unexploded ordnance based on short line features [12]. [22] explored matching modern images to a few historical images for each of 10 specific landmarks and found that SIFT and SURF are similar and work better than Harris or KLT feature matching. Work with photorealistic rendering of scenes [5] over the course of a day, and outdoor robotic driving around a neighborhood [20] also found minimal difference between standard detector and descriptor combinations. To our knowledge, there is no large scale systematic study of the impact that years long time differences have on the effectiveness of feature matching in varied, real-world environments, nor has anyone studied where failures occur in the feature matching pipeline for features imaged at different times.

## 3. Implementation Details & Method

We assess the robustness of commonly used key point extraction methods and local feature descriptors over time by extracting key points and computing feature descriptors on a large number of stable webcam images. In this section, we describe our experimental design and implementation details.

### 3.1. Image Selection

We collected images from 20 particularly stable, long-lived webcams from the Archive of Many Outdoor

Scenes [4]. For each of these 20 cameras, we select a single clear daytime image, from within 48 hours of either the winter or summer solstice, as a reference image. We then find one image per day captured by that camera within 30 minutes of the reference image. Once we have this set of images, we manually remove any images that substantially moved relative to the reference image (i.e., a pan-tilt-roll camera that turned to a different view of the scene).

### 3.2. Key Point Extraction & Feature Description

We extract key points from each one of these images. We use both scale-space local maxima of Difference of Gaussian (DoG) filters as implemented by [21] and the MATLAB Computer Vision Toolbox implementation of Maximally Stable Extremal Regions (MSER) [9].

We then find SIFT [8], SURF [2], and DAISY [18, 19] descriptors at the locations identified by the key point extraction methods listed above. We use the VLFeat [21] implementation of SIFT, the MATLAB Computer Vision Toolbox implementation of SURF, and the author provided implementation of DAISY.

### 3.3. Feature Matching & Assessment of Robustness

For every image, we use the VLFeat implementation of the matching algorithm described by [8], to find the set of features which are considered matches in feature space between the reference image and the subsequent image. We then solve for the optimal homography and subset of the feature matches that are geometrically consistent using [6]'s RANSAC implementation. This allows for the slight camera jitter that can be present in even very stable webcams (less than 10 pixels of movement). It also allows us to check that the $3 \times 3$ homography, $M_0$, computed through this matching process is close to the identity matrix, using the following protocol:

1. Normalize the homography by its bottom right element: $M_1 = M_0 / M_0(3,3)$.
2. If $|M_1(1,1) - M_1(2,2)| > 0.1$, the homography is not consistent with small camera pans and tilts.

Our dataset is derived from very stable webcams, that had occasional small pan, tilt and zoom variations. This simple rule above was more consistent than any threshold comparing $M_1$ to the identity matrix. Visual inspection of many cases where $M_1$ failed this test found that all arose in situations where all inliers after RANSAC were in a nearly degenerate configuration with all matches nearly colinear.

When $M_1$ passes this test, we consider all inliers to be geometrically consistent feature matches between this image and the reference image for this camera. Because different cameras have reference images that may have dramatically different numbers of features, we normalize the inlier counts by the maximum number of inliers ever seen between that reference image and any other image.

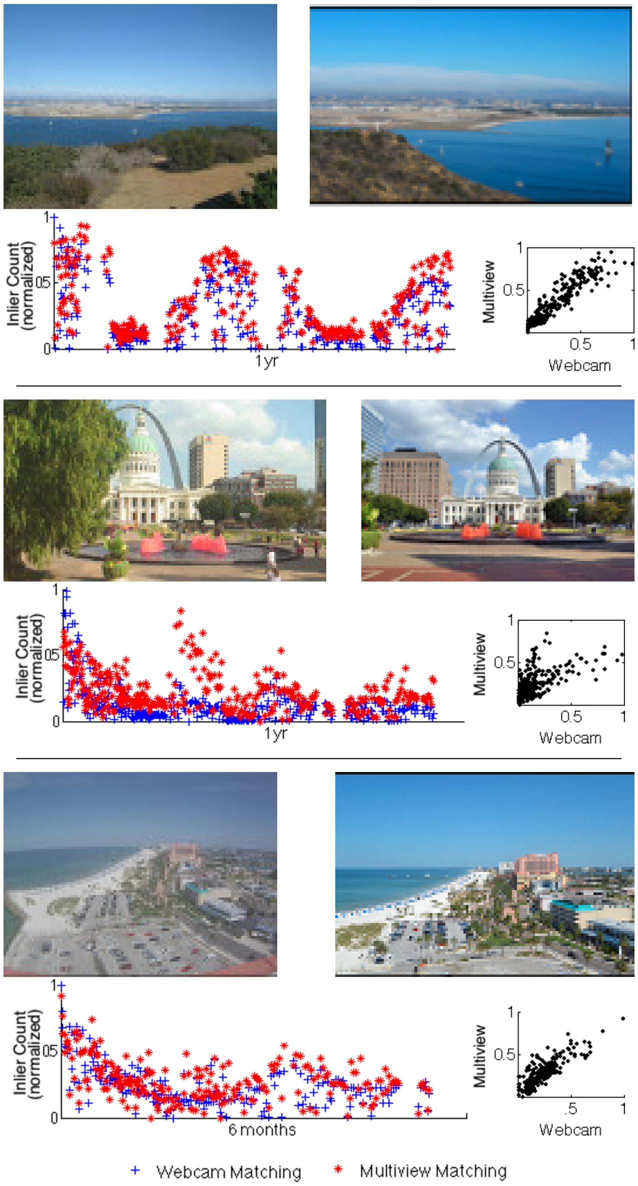


Figure 2. For three scenes we compare match features back through time when using an image from the webcam (left images) and when using an image captured at the same date and time from a different viewpoint (right images). The plots over time show the normalized inlier count as a function of time-difference from the query image, (+) indicating matching to the webcam image and (*) the image from a different viewpoint. The black plots on the right show the correlation in feature matching efficiency as a function of time, highlighting that matching webcam images across time is a good proxy for multiview matching performance.

### 3.4. Webcams as a Proxy for Multiple Viewpoints

Applications of feature matching usually do not involve images from exactly the same viewpoint, so in this section we test whether the feature matching experiment described in the previous section is consistent with feature matching from different viewpoints.

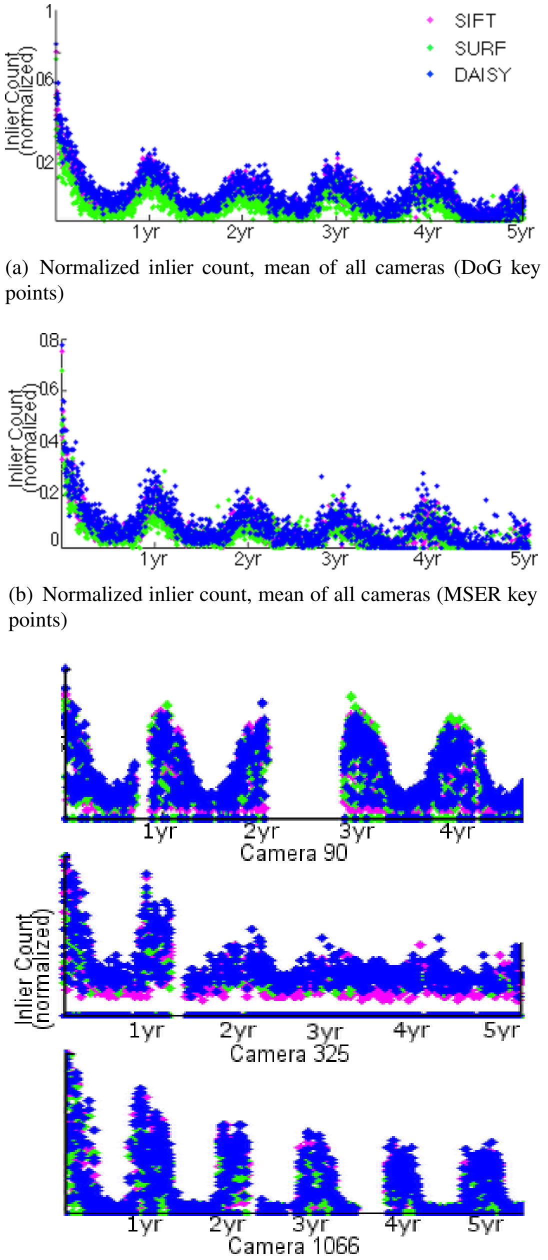
To assess this, we adapt the experiment described in Sections 3.1 - 3.3. We find images of the same scene as several of our webcams, but from taken from different viewpoints. We use these as our multiview reference image. We select a webcam image taken within 24 hours of the multiview reference image to use as the webcam reference image. We then run the matching protocol once using the multiview reference image and once using the webcam reference image. In the multiview case, we use RANSAC to fit a fundamental matrix rather than a homography, again using [6].

In Figure 2, we observe that the same trends seen in webcams over time and described later in Section 4, are also observed in this multiview approach. The correlation between the single view and multiview approaches is very strong, and we visually observe similar trends over time in the number of features that are matched. This justifies our use of feature matching in static webcam images over time as a proxy for the more general feature matching that is required for many applications. This is fortunate because the static camera matching problem can be evaluated more rigorously, and we can discover more easily where in the pipeline failures occur.

## 4. Analysis

We apply the procedure from Sections 3.1 - 3.3 to compute feature matching performance across many years for many cameras. In the data collected through this procedure, we observe four persistent trends, shown in the plots in Figure 3. First, there are day-to-day fluctuations in feature matching performance due to factors such as weather or temporary occlusions like people walking through a scene. Second, there is an annual trend due to changes in illumination as a function of sun position that can be seen clearly in Figure 3, where feature matching performance peaks cyclically every year. Third, feature matching performance is extremely good across the first few images taken in similar weather conditions, due to the similar scene illumination and lack of long term changes that can occur over longer periods of time. Fourth, there is an overall decrease in feature matching performance over time due to factors like plant growth in natural scenes or larger scale construction activities in urban scenes, such as a building being put up or street lanes being painted.

We now discuss why these patterns emerge by closely studying the effects of weather and lighting variation, and investigating where the feature matching pipeline fails.



(a) Normalized inlier count, mean of all cameras (DoG key points)

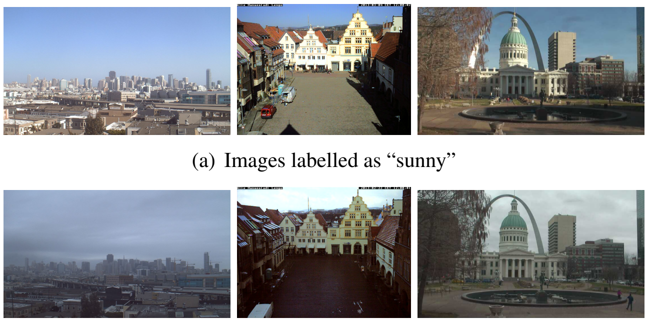
(b) Normalized inlier count, mean of all cameras (MSER key points)

(c) Exemplar individual cameras

Figure 3. Each of the above plots show the (normalized) number of inliers after RANSAC for various feature matching routines through time using stable webcams from the Archive of Many Outdoor Scenes [4]. We show the average plot for all cameras with DoG key points (a) and MSER key points (b), along with three of the 20 cameras (c) using DoG key points. Across all cameras, feature matching performance has a cyclic pattern, where the most likely images to be matched are taken exactly a year (or two, or three...) later. Furthermore, feature matching performance is roughly equivalent for all feature extraction routines.

### 4.1. Weather Changes

To assess the extent to which feature matching performance over time is affected by weather patterns, we ran-



(a) Images labelled as "sunny"

(b) Images labelled as "not sunny"

Figure 4. Example image pairs and their corresponding weather labels from three webcams collected from the AMOS data set.

| | Both Sunny | Both Not Sunny | Sunny-Not Sunny |
|---|---|---|---|
| DoG+SIFT | 0.547 | 0.400 | 0.193 |
| MSER+SIFT | 0.450 | 0.332 | 0.139 |
| DoG+SURF | 0.504 | 0.362 | 0.160 |
| MSER+SURF | 0.459 | 0.343 | 0.139 |
| DoG+DAISY | 0.520 | 0.383 | 0.208 |
| MSER+DAISY | 0.497 | 0.347 | 0.172 |
| Mean | 0.496 | 0.361 | 0.196 |

Table 1. The mean normalized inlier count for different key point extraction and feature description methods when comparing different weather conditions 24 hours apart. Matching between images taken during like weather conditions significantly outperforms matching between images taken during different weather conditions.

domly select 25 images from 13 different webcams. Each image is then paired with an image from 24 hours later. The change in lighting direction between images taken 24 hours apart is minimal, and so the driving force in feature matching performance at this scale is due to weather.

We then hand label each of these 325 images as either sunny or not sunny. We then find feature matches between the image pairs and compute the normalized inlier counts for each method under the different sets of weather conditions, averaged across all 13 cameras.

This data is presented in Table 1, which shows the robustness of each key point extraction and feature descriptor pair when matching between different combinations of weather conditions. We find that the sunny-to-sunny matching performance is best, because images that are taken exactly 24 hours apart have shadows that provide strong features and are in almost exactly the same position. Not sunny-to-not sunny matching performance is better than sunny-to-not sunny matching, indicating that weather plays a large role in the success or failure of the matching protocol. Sunny-to-not-sunny matching is the worst, probably because the strong shadow features often cause the feature detectors to fire in different image locations. Figure 4 shows example image pairs with their labels – the significant difference in these images' appearances demonstrates why feature matching across different weather conditions is a difficult domain.

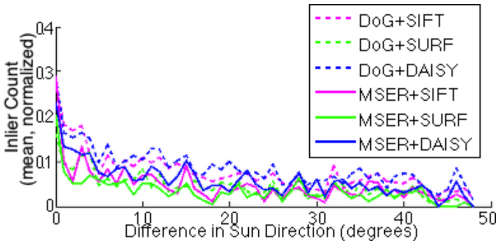


Figure 5. This plot shows the normalized inlier count, over all cameras for all key point detector/descriptor routines, as a function of the difference in lighting direction. This demonstrates that matching performance is strongly correlated when the sun positions are similar, regardless of the feature.

It is notable that while Difference of Gaussian key point extraction performs marginally better over all weather conditions and descriptors, there is no feature descriptor that performs significantly better the others. Indeed, this is consistent with the results shown in Figure 3, where no one particular combination of key point extraction or feature description method significantly out performs another.

### 4.2. Illumination Changes

The most noticeable trend over time is the clear annual cycle noticeable in every plot in Figure 3. This trend is due to the differences in the lighting direction of the scene as the tilt of the Earth changes over the course of the year. The impact of sun direction over the course of a year also becomes less significant the closer the scene is to the equator, where the maximum difference in sun direction over the course of a year is less than at the poles.

In order to assess the role illumination changes have on feature matching performance, we require that the starting image for each camera be taken within 48 hours of either the summer or winter solstices. We then compute feature matches between each starting image and that webcam's subsequent images from the same time of day, using the protocol described in Section 3. Using the image time-stamp and the geo-location of the camera, we compute the direction to the sun, and then the angular difference of the sun direction between each pair of images.

Figure 5 summarizes this experiment, showing the normalized inliers counts as a function of difference in illumination angle. The particular combination of key point extraction and feature description has minimal effect on matching performance under different illumination conditions. This relative matching performance of descriptors over different lighting directions is consistent with the relative matching performance of SIFT and DAISY found by [5], which assessed feature matching performance over the course of a single day in synthetic scenes.

All of the key point detection and feature descriptor com-

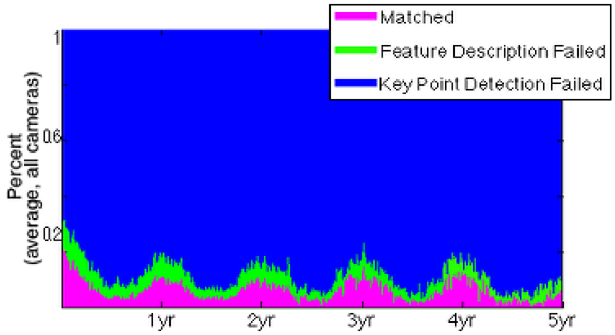


Figure 6. Each vertical slice of this temporal histogram displays the likelihood that a key point from the original image found a match (magenta), found a key point in the right location with a dissimilar descriptor (green), or failed to find a key point (blue), in some new image taken at a later time. The main cause for matching failure is that key points are not extracted in enough common locations from frame to frame.

binations show a significant decline in matching performance as the difference in sun direction between the reference image and each subsequent image increases. Approximately 60% of this decline occurs within the first 5 degrees of lighting direction difference. This corresponds to the change in illumination direction that occurs in about 20 minutes on one day, or the illumination difference between images taken at exactly the same time 40 days apart (depending on time of year and latitude). For any applications reliant on feature matching between outdoor images, this highlights the importance that the images be captured with as minimal a difference in lighting direction as possible.

### 4.3. Key Point Extraction vs. Feature Description

Feature matching can fail because the feature is not detected or because the feature descriptor is not similar enough to be matched. In this section, we evaluate what parts of the feature matching routine break for images captured over time. Due to the similarly in feature matching performance across the different key point extraction methods and feature descriptors discussed previously, and for the purpose of having a simple and clear experimental design, we only explore this using DoG key points and SIFT descriptors.

For each camera, we find the set of SIFT features extracted from that camera's reference image and then from each subsequent image from the same time of day. For each feature from the reference image, we then iterate through every subsequent image and label that feature at each time interval, as one of the following:

- **Successful Match:** There was a feature in the second image that was geometrically consistent with the feature in the first image, and it was sufficiently close in feature space.
- **Feature Description Failure:** There was a feature in the second image that was geometrically consistent with the feature in the first image, but it was insufficiently close in feature space.
- **Key Point Extraction Failure:** There was no key point extracted in the second image that was geometrically consistent with the feature in the first image.

Our criteria for geometric consistency were:

1. Difference in $x, y$ position of SIFT features is less than one of the first feature's bin sizes,
2. The scale of each SIFT feature differs by less than $20\%$.
3. Difference in orientation between features is less than 10 degrees.

We calculate the percent of features from the reference image that fit into each of these three categories for every pair of images. Figure 6 shows these statistics, averaged by day across all 20 cameras. We find that key point detection is the dominant source of failure in feature matching between two images with different lighting or weather patterns. If a match fails, it is almost always because the key point detection algorithm did not find feature points at the same locations and orientation.

## 5. Conclusions

In this paper, we offer the first empirical assessment of feature matching performance in varied real-outdoor scenes over the course of years. We analyze many-year sequences across 20 webcams and find the following strong, common patterns in feature matching performance over time:

1. There is day-to-day variation in feature matching due to weather conditions and short term physical changes in scenes.
2. Feature matching performance declines significantly with seasonal effects and small variations in lighting direction.
3. Changes in physical scene appearance over time, such as vegetation growth, road re-painting or construction projects cause small changes in feature matching performance over time, but at the scale of 5 years, these effects are dwarfed by changes in weather and lighting.
4. In the feature detection and matching pipeline, different detectors and descriptors all show the same trends, and the dominant cause of failure is in the feature detection stage.

This leads to two conclusions. First, when creating datasets to optimize the potential for good feature matching over time, it is important to have imagery from many times of day and times of year, as well as both sunny images and cloudy day images. Second, when doing research to improve feature detection performance over time, effort should be concentrated on creating robust feature detectors, or, potentially, working with dense descriptors.

## References

[1] S. Bae, A. Agarwala, and F. Durand. Computational rephotography. *ACM Trans. Graph.*, 29(3):24:1–24:15, July 2010.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[3] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *Int. J. Comput. Vision*, 94(3):335–360, Sept. 2011.

[4] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *CVPR*, June 2007.

[5] B. Kaneva, A. Torralba, and W. T. Freeman. Evaluating image feaures using a photorealistic virtual world. In *IEEE International Conference on Computer Vision*, 2011.

[6] P. D. Kovesi. MATLAB and Octave functions for computer vision and image processing. Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[7] A. Le Bris and N. Paparoditis. Matching terrestrial images captured by a nomad system to images of a reference database for pose estimation purpose. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS)*, 1:133–138, 2010.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, pages 36.1–36.10. BMVA Press, 2002. doi:10.5244/C.16.36.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[11] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *Int. J. Comput. Vision*, 73(3):263–284, July 2007.

[12] V. Murino, U. Castellani, A. Etrari, and A. Fusiello. Registration of very time-distant aerial images. In *International Conference on Image Processing*, volume 3, pages 989–992. IEEE, 2002.

[13] M. Park, J. Luo, R. T. Collins, and Y. Liu. Beyond GPS: determining the camera viewing direction of a geotagged image. In *Proceedings of the International Conference on Multimedia*, pages 631–634. ACM, 2010.

[14] D. Picard, M. Cord, and E. Valle. Study of SIFT descriptors for image matching based localization in urban street view context. In *Proccedings of City Models, Roads and Traffic (CMRT) ISPRS Workshop*, volume 9, pages 193–198. Citeseer, 2009.

[15] G. Schindler, F. Dellaert, and S. B. Kang. Inferring temporal order of images from 3d structure. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

[16] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert. Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In *Computer Vision and Pattern Recognition*, pages 1–7, June 2008.

[17] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int. J. Comput. Vision*, 37(2):151–172, June 2000.

[18] E. Tola, V. Lepetit, and P. Fua. A Fast Local Descriptor for Dense Matching. In *Proceedings of Computer Vision and Pattern Recognition*, Alaska, USA, 2008.

[19] E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010.

[20] C. Valgren and A. J. Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *EMCR*, 2007.

[21] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM.

[22] R. Wolfe. Modern to historical image feature matching. *Published online: http://robbiewolfe.ca/programming/honoursproject/report.pdf*, 2013.

[23] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. In *CVPR*, pages 1085–1092. IEEE, 2009.