Evaluation of Synthetically Generated Airborne Image Datasets using Feature Detectors as Performance Metric

Georg Hummel[1], **and Peter Stütz**[1]
[1]Institute of Flight Systems, University of the Bundeswehr, Munich, Germany

Abstract - *The use of synthetic datasets to develop, prototype and qualify new computer vision algorithms is currently not widely accepted, though highly sought after by the industry. This is due to lack of knowledge on how the results acquired with such datasets will transfer to real live performance. Therefore, this paper introduces an approach to evaluate modelled synthetic datasets against their real counterparts. In a use case, the performance of common feature detectors is evaluated using the repeatability metric against real and synthetic datasets. Based on resulting performances; general usability, rendering techniques and modelling efforts for generation of synthetic datasets are discussed.*

Keywords: synthetic datasets; dataset evaluation; feature detectors; homography; performance analysis;

1 Introduction

Today the development of new CV-algorithms often depends on the quality of design, training or test datasets. However, when it comes to applications striving to process data from aircraft mounted sensors, public availability of datasets is rare and available data are homogeneous or fragmented. Therefore, the resulting algorithms are often limited to the operational conditions available in the used datasets to perform as intended. Datasets such as VIVID[1] or NGSIM[2] are providing good means for prototyping of specific algorithm but cannot cover the complexity of weather and lighting influences in aerial imagery due to their recording at one specific date and location.

In 1995 [3] already discussed the concept of using a synthetic environment to develop CV-algorithms. In the last 20 years computer graphic technologies experienced a technology leap allowing to model weather conditions, illumination or shadowing in photo-realistic qualities. In [4] an airborne object algorithm designed on real datasets was evaluated on its performance with synthetic data. It has become a common procedure to use abstract synthetic datasets for initial development of new computer vision algorithms [5] followed by further steps using real datasets. Several image processing benchmarks [6]–[8] use synthetic data due to the easy to access ground truth for quantitative performance measurements. Still, in the final stage, the computer vision domain seeks to extract information from real (recorded) imagery, which is much more complex than its synthetic representations. Thus the acceptance using synthetic data for evaluation of algorithms is low, since while the content of the scene can be the same, the image structure may be fundamentally different (e.g. texture, color, contrast, etc.) [9].

This paper details the CV-algorithm evaluation step suggested in the concept presented in [10] using geo-referenced airborne image datasets of real and synthetic nature. Therefore, the general concept is briefly introduced in the following section.

2 General concept

The general evaluation concept is intended to allow investigation of essential image properties and influencing rendering technologies and to identify a trade-off between modelling detail and algorithm performance. It further aims to provide suggestions and design guidelines towards a benchmark simulation system. This shall be achieved by evaluating basic CV-algorithms against datasets consisting of sequential images by varying rendering techniques and to compare their performance. Thus, we can derive conclusions on the suitability of conducted modelling and rendering efforts of synthetic datasets.

The multi-level concept consists of four different levels as depicted in Fig. 1. The bottom layer (layer 1) contains the datasets and their corresponding ground truth. These datasets consist of aerial imagery and aircraft telemetry derived from test flights performed in either the real ("real datasets") or the synthetic environment ("synthetic datasets"). The ground truth contains the camera movement between compared images as geometric transformation. Level two analyzes the image structure of evaluated datasets using image descriptors (MPEG7) usually deployed for image queries to search engines or image databases. This mechanism is explained in detail in [10]. It allows the direct comparison of image properties. Level three uses computer vision algorithms as test

Table 1: Aircraft parameters provided by telemetry

Recorded Telemetry			
Parameter	***Accuracy***	***Unit***	***Update Frequency***
Latitude (WGS84)	0.06"	degree	5 Hz
Longitude (WGS84)	0.06"	degree	5 Hz
Altitude (AGL)	0.1	meter	100 Hz
Yaw (Euler)	1°	degree	100 Hz
Pitch (Euler)	1°	degree	100 Hz
Bank (Euler)	1°	degree	100 Hz

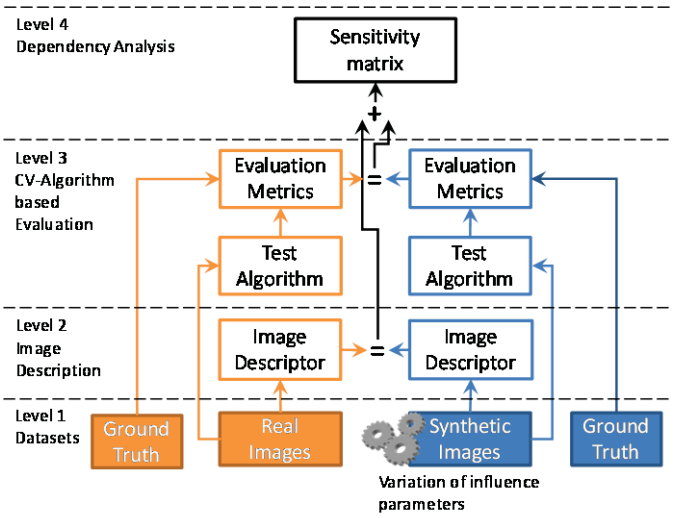

Fig. 1. Evaluation concept of datasets against known computer vision algorithms. This concept is part of the more general concept presented in [10].

algorithm to extract the performance differences among datasets. For clarity, only widely used metrics measuring the quality of algorithms are selected (time based metrics are not considered). The last level performs dependency analysis using the results from level two and three, which allows correlation and weighing between resulting performance and identified image properties. Thus, conclusions in level four shall allow to identify rendering techniques suitable for computer vision algorithm testing and evaluation. This paper, discusses level three (CV-algorithm based evaluation) and level one (dataset generation).

3 Dataset generation

Special interest in this work has been laid on the generation of datasets. First, we had to ensure the scenic correlation between synthetic and real datasets. Therefore, the test flight area was modelled in a virtual environment to create snapshots with identical scenery. Secondly, we had to record telemetry data representing the sensors pose and location (e.g. location, attitude and altitude of the aircraft at which images are taken in flight). This enables us to position the camera in the virtual environment equivalently. The following subsection explains the dataset in detail.

3.1 Test flight dataset "Real"

The taxiway of a former airport on the premises of the University of the Bundeswehr Munich was selected as test site, because it was easy to access, free from unauthorized persons, allows small aircraft operation and had changing terrain (e.g. field, woods, buildings, etc.). As sensor platform, a Multicopter equipped with eight 350W motors and 13" propellers was selected due to its payload, in air stability and low vibrations. This platform has a maximum take-off weight (MTOW) of 6kg allowing 2.2kg payload at max. The aircraft can be navigated via waypoints at a fixed above ground altitude. The camera has been mounted perpendicular to the aircraft frame using a fixed rigid non-stabilized mount. The deployed camera, a XIMEA MQ042CG-CM has been configured to a resolution of 2048x2048 at 30 Hz. A detachable C-Mount lens from Myutron, achieving a total field of view of 25.4° has been deployed. The telemetry was received directly from the flight control system via serial interface at 100 Hz and containing several aircraft parameters detailed in Table 1. Telemetry and image data were recorded on-board in sync using Linux based distributed data services [11], running on a Commel LS-37B Single-board computer.

The actual test flight was conducted on March 18, 2015 at noon on sunny weather leading to crisp shadows and some reflections on buildings. The altitude has been fixed to 75 meters above ground. During the flight, 1000 meter of terrain have been covered that were categorized in nine classes of which three are presented in this paper later on. Each category was reduced to 11 sequentially taken images at 1 Hz to reduce data while retaining sufficient overlap. The images have been resized and cropped to 1024x768 pixels for comparison with synthetic datasets. Due to automatic white balancing the images of the real dataset had a slightly green tint. In future tests manual color calibration may minimize this effect.

3.2 Synthetic dataset

At first, a virtual environment (engine) suitable for geo-referenced dataset generation was selected. VBS3 from Bohemia Systems was preferred as it is widely used in tactical military simulation, capable to reproduce high ground detail, wide terrain areas, providing a resource database and tools for geo-referenced map generation. The virtual database was modelled in four different quality levels as can be seen in Table 2. The raw data used to model the database variants comprised satellite images, elevation data, geo-spatial vector data and 3D-Objects. The department of geo-information of the Bundeswehr provided orthographic satellite images used in various resolutions and a digital surface model (3D altitude mesh) in 15 meters per pixel (mpp) resolution. Rasterized Shapefiles are used as masks populating the area with different detail maps (e.g. concrete, grass high, etc.) for high-resolution texture details at low altitude. Finally, the terrain was populated using geo-referenced and geo-specific 3D-Objects either provided by VBS3 or created using Blender. All Buildings were modelled after their blueprints to ensure accurate dimensions.

Table 2: Generated terrain databases detailed with raw data used for modelling in meters per pixel (mpp).

Terrain Databases				
Surface Detail (Database Name)	***Resolution Satellite Images***	***Resolution Digital Surface Model***	***Resolution Rasterized Shapefiles***	***Objects***
Low	5 mpp	15 mpp	5 mpp	Yes
Mid	1 mpp	15 mpp	1 mpp	Yes
High	0.2 mpp	15 mpp	0.2 mpp	Yes
High no Building	0.2 mpp	15 mpp	0.2 mpp	No

Facade textures have been photographed and applied after rectification using perspective transformation. Roofs are most prominent in aerial images therefore after identification of type, material and color; their textures have been modeled precisely using free texture databases. Each 3D model consists of geometry, texture map, normal map, specular map and material definition (setting the lighting behavior).
Common industrial tool chains and efforts have been employed in generating the virtual database and its 3D objects. However, the additional requirement of a geo-referenced database necessary for real- and synthetic- dataset comparison increased the development time significantly.
To create synthetic imagery correlating to the test flight, the virtual camera had to be positioned according to recorded telemetry data. Thus, the telemetry data were replayed and used as a trigger to synchronize the image extraction of the virtual environment. Lighting was adjusted using a hemi-spherical lighting model to adjust day light color and strength as well as length and orientation of shadows. The implement-ted camera model of VBS3 was employed which enables the parametrization of focus, aperture, field of view and zoom. These were set to comparable values of its real counterpart. The image resolution has been fixed to 1024x768. The focus was set to infinity equivalent to the real camera.

3.3 Specific dataset categories

The test flight route was separated in nine different classes concerning the nature of the scene. For each of these classes synthetic datasets were extracted. In this paper, the dataset classes Field, Woods and Concrete are discussed.
Field designates a regularly mowed meadow on even terrain. There are no objects in the scene and it is homogenous without any sharp edges or specific high contrast textures. This dataset is intended to demonstrate the differences in terrain image quality between real and synthetic datasets.
Woods designates a dense forest, hiding the ground texture almost completely. In the synthetic datasets, trees have been placed approximately and geo-typical tree models have been used. This dataset was used with caution for two reasons: Firstly, the virtual environment has been modelled using aerial images taken in summer, meaning all trees are in full bloom, while during real test flight they were leafless. Secondly, the height of trees (up to 15 meters) violates the homography constraint, which states that all features shall be in a plane. This reduces the overall results of real and synthetic datasets. However, the highly heterogenic, diverse and cluttered textures are demanding for feature detectors.
Concrete designates a concrete area with transport containers, concrete plates, a mobile bridge and a silver car. This dataset provides sharp edges on several man-made objects as well as a highly textured surface. The object heights are not exceeding two meters, which is small compared to the aircraft altitude. Thus, the homography error was considered negligible.

4 CV-algorithm based evaluation

To investigate the usability of synthetic data for CV-algorithm development and prototyping it was important to use well-known algorithms to allow a comprehensible assessment of acquired results. Therefore, the feature detectors SIFT [12], SURF [13] and MSER [14] were selected as test algorithms. Feature detectors in particular are interesting, because they filter the image domain for recognizable locations, which were used to extract information from the image domain to the feature domain. Often, further processing is solely working on the feature domain (e.g. stereo vision, image stitching). Thus, the performance of feature detectors influences the performance of many specific algorithms and implementations.
Performance of these algorithms is measured using evaluation concept and metric presented by Mikolajczyk in [15]. This *repeatability* metric measures the number of detected corresponding regions in image pairs. It assumes that all features found in Image *I*, mapped to a plane, experience a global geometric trans-formation and can again be found on the transformed plane in Image *J*. The homography matrix H^{ij} describes this geometric transformation and allows reprojection of Features R_J to Image *I*. Features are described as regions *R* on the image with a location and a radius. A feature pair is corresponding when the region ${}^{i}R_j = (H^{ij})^T \, {}^{j}R_j$ reprojected into Image *I* overlaps with R_I [15]:

$$1 - \frac{R_I \cap {}^{I}R_J}{R_I \cup {}^{I}R_J} < e_O \qquad (1)$$

This means a pair is accurate when the overlap error e_O smaller is then the intersection of R_I and ${}^{I}R_J$ divided by its union. The overlap error is set to 0.4 ≙ 40%. This metric is scale dependent, thus punishing differences in region size. The resulting number of correspondences against all possible correspondences is the repeatability measure depicted in percentage. The evaluation is performed in MATLAB and based on the benchmark framework "VLBenchmark" [16].
Additionally to the repeatability, the number of resulting correspondences was analyzed to provide an absolute measure that needs to be considered when evaluating the relative repeatability. Thus, it is possible that the performance of a detector, which only detects four regions in an image but identifies them all results in 100% repeatability. The number of correspondences shows that the detector performs poorly on given dataset, since these few features were not sufficient for possible subsequent processing steps. The necessary ground truth (homography matrix between the images) is first calculated using SURF features [13] matched with a brute force SSD matching algorithm as initialization of an iterative RANSAC-Algorithm for optimization [17].
By using homography, it was possible to measure algorithm performance against generated ground truth in real and synthetic datasets. The decision to create the ground truth of the simulation also using homography deems from the intention of having comparable results therefore using the

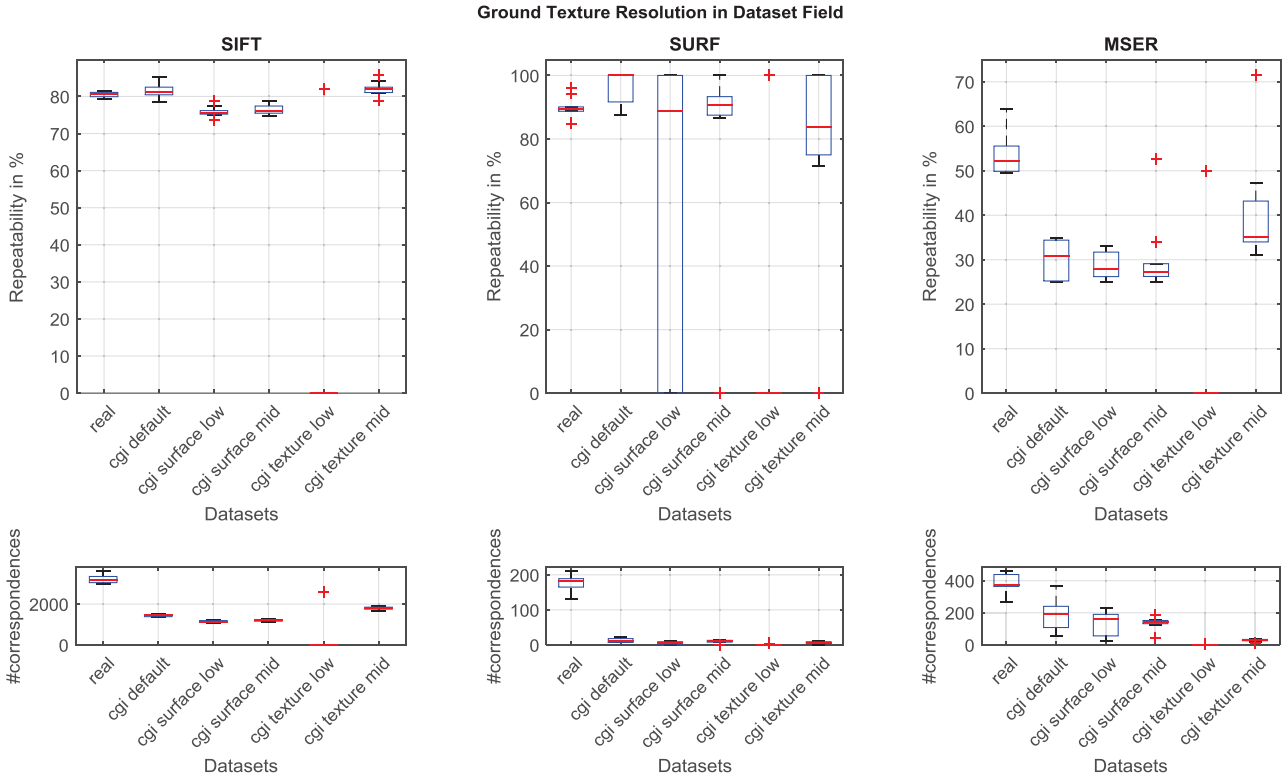

Fig. 2.Evaluation of the dataset class Field using parameter group Ground Texture Resolution. Each box presents the performance of one dataset. The red line inside the box marks the merdian, the upper and lower end mark the 75th and 25th percentile, the black whiskers mark the outmost inliers. Outliers are marked with a red plus.

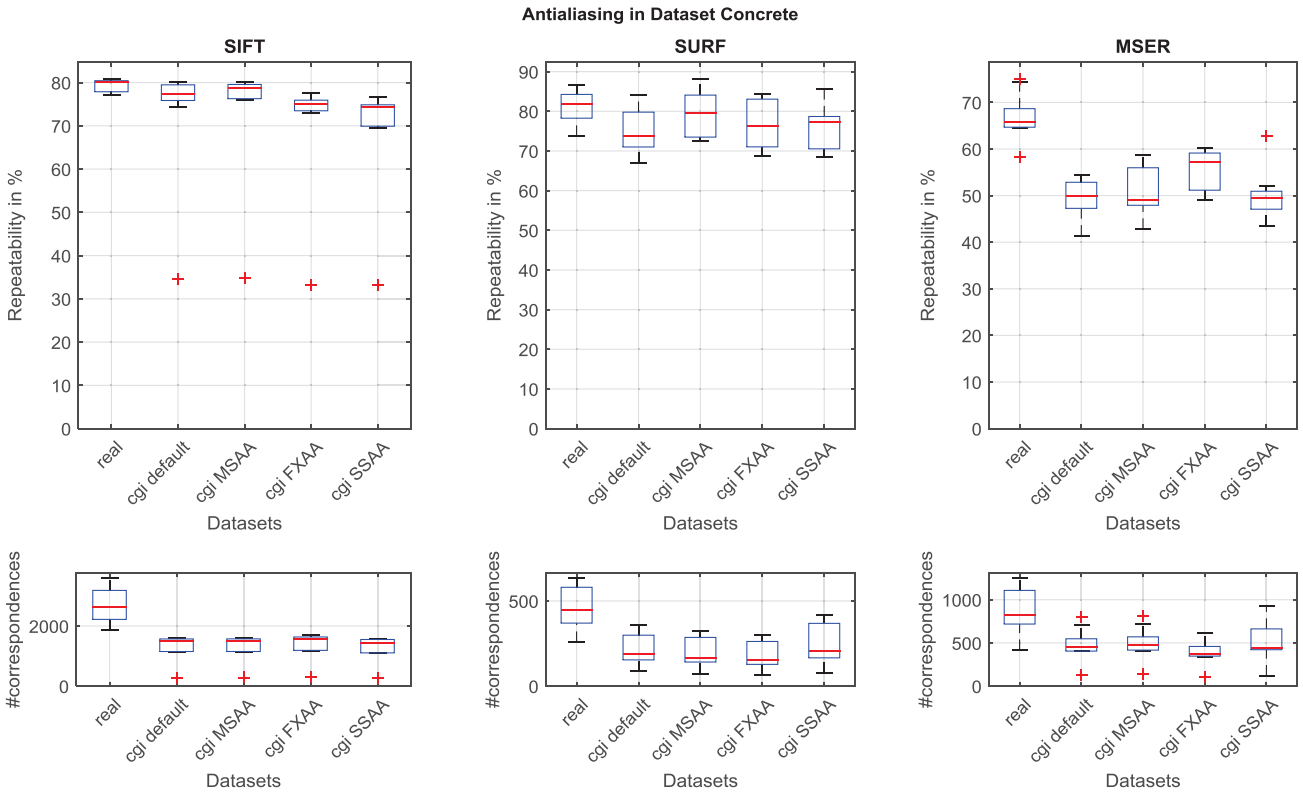

Fig. 3. Evaluation of the dataset class Field using parameter group Ground Texture Resolution. In general synthetic datasets using anti-aliasing increased their performance (except 1.5x SSAA). Using SIFT and SURF synthetic datasets achievied similar performance to real dataset, however finding less corresponding features in total.

same evaluation chain. It needs to be noticed that homography can only be used when either the camera has no or small translation between to images or the displayed surface is planar. Altitude information confirms that the surface of the recorded premises is adequately planar. Due to high aircraft altitudes (75m) and a top down view, small altitude difference of occasional trees and buildings are considered negligible. In the worst case, performance would drop on all datasets and the relative results between tested datasets would not be influenced.

4.1 Evaluated parameter groups

After generation of dataset *Real* and its ground truth, several different synthetic datasets have been created. The first dataset generated was the *Default* dataset, defining the default parameter settings of the rendering engine. Afterwards each additional synthetic dataset was created by modifying a parameter of the engine to identify its influence. Because of space limitations, only a selection of parameters is evaluated in this paper. The parameters have been clustered in two groups. These groups were evaluated against a specific dataset class and results were discussed in detail. Additional significant findings were reviewed without full presentation of the evaluation results.

The first group of parameters was named *ground texture resolution* consisting of surface texture resolution (*surface*) and detail texture resolution (*texture*). The used engine creates ground surfaces by overlaying the geo-specific surface texture with a procedural detail texture. This detail texture emulates a higher resolution of the ground surface but does not provide much contrast due to texture blending. The recorded flight imagery in the real dataset has a ground resolution of 0.03mpp, which corresponded to the highest detail texture resolution setting (see Table 3). This group was tested against dataset class Field that depicts only the ground texture with a repetitive detail texture (meadow).

The second parameter group called *Antialiasing (AA)* embraces three anti-aliasing techniques, namely *Multi Sampling* (*MSAA*), *Fast Approximate* (*FXAA*) and *Super Sampling* (*SSAA*). These techniques all had the goal to reduce jagged nature of sharp edges or lines, which were introduced during rasterization. The reason for different techniques to exist is mostly due the different computation effort necessary. The dataset demonstrating eightfold *MSAA* shows selective sampling depending on polygon-pixel coverage, simple sprites (i.e. tree leaves) are unaffected. The *FXAA* dataset, a post processing antialiasing method, used a high pass filter to detect edges followed by a blur only along those edges. The *SSAA* method simply renders the whole scene in 1.5x of the output resolution and resizes it to its original resolution by averaging. This group was tested against dataset class *Concrete* in detail showing its capabilities on objects in the scene. Each group was additionally tested against *Default* and *Real* dataset to allow absolute comparison.

Table 3: Resolution of parameters in group *ground texture resolution* given in meter per pixel (mpp)

Ground Texture Resolution		
Level	***Surface Texture Resolution***	***Detail Texture Resolution***
default (high)	0.2 mpp	0.03 mpp
mid	1 mpp	0.06 mpp
low	5 mpp	0.12 mpp

4.2 Evaluation and discussion

The first group evaluates the *Field* dataset (empty meadow) against the detectors (SIFT, SURF, MSER) using the datasets *real, computer generated imagery (CGI) default, CGI surface texture low, CGI surface texture mid, CGI texture low and CGI texture mid*. The results are depicted in Fig. 2 using boxplots. Each detector has a separate plot providing its results for each dataset. A red line inside the box marks the median. The upper and lower edges mark the 75th and 25th percentile of the dataset and black lines outside the box mark the maximum and minimum value still considered as inlier. Outliers are marked as a red plus. The results are depicted using the relative repeatability measure indicating the amount of corresponding regions that align with less ten 40% difference of their area. The metric is supported by boxplots (lower row) depicting the absolute number of successfully matched correspondences.

4.2.1 Ground texture resolution

The SIFT detector as shown in Fig. 2 performed quite well (SIFT: 80% repeatability with more than 3000 correspondences on dataset *Real*) on the *Field* dataset class. Comparing the results of dataset *Real* and *Default* results in almost equal performance while the number of correspondences, however was halved. This can be explained by the low contrast of edges due to texture blending. Reducing the resolution of the surface texture results in a drop of repeatability accompanied by a drastically increased standard deviation as indicated by the size of the box. Low surface resolution reduces colored edges since it is smoothing the transition (between background pixels) heavily. This reduces the quality of regions resulting in a lower repeatability rate as shown in the dataset *CGI surface low* but also already indicated in dataset *CGI surface mid*. Reducing the ground surface detail to 0.12 mpp (*CGI Texture Low*) actually disables all detectors. No tested detector was capable to cope the blurring effect of downscaling the detail texture. Interestingly, downscaling the detail texture only once, to a ground resolution of 0.06 mpp was not only allowing the detectors to provide correspondences but to perform even slightly better than dataset *Default*.

Evaluating the SURF detector results showed very high relative repeatability for almost all datasets, but at a very low number of correct correspondences. While the *Default* dataset had 177 correspondences, all CGI datasets only had 12 correspondences on average. Thus, the detector was not providing enough significant features, revealing that the box

filter approximation used to find SURF features could not handle homogenous areas of low structure well. Similar to SIFT, on dataset *texture low* no features could be found.
MSER features tend to be small since they are robustly detected extremal regions by thresholding the image at several thresholds. These extremes however are sparse in the dataset class *Field*, therefore leading to repeatability rates of 53% for dataset *Real* and 28-40% on *CGI* datasets. In addition, the number of features is low in respect to usual values of the MSER detector. The performance between dataset *Default*, *CGI surface mid* and *CGI surface low* dropped by 1% each, demonstrating the robustness of MSER features against surface texture changes on larger scale. However, *CGI texture low* and *CGI texture mid* depict that MSER was strongly influenced by reduction of high frequency details of the image. While *CGI texture mid* provides a median repeatability of 35% it only finds 27 correspondences in total, revealing the low absolute performance of the detector.

4.2.2 Antialiasing

In Fig. 3 the results of group *Antialiasing* in dataset class *Concrete* are depicted. Firstly, all datasets perform well using the SIFT detector. Differences between datasets are only small, since image changes were only minor and mostly limited to edges. Similar to aforementioned evaluation, here the detector found more correspondences on *Real* compared to *CGI* datasets (by factor two). However, repeatability as well as the total number of correspondences on synthetic data demonstrated acceptable performance of the detector. Using *MSAA* increased the repeatability, getting closer to Real performance values. Since this dataset contained no trees (sprites), this technique could perform to its fullest. Using *FXAA* reduces the repeatability by 3% showing a visible blurring effect on object edges. *SSAA* lead to repetitions of irregular aliasing patterns along edges leading to displacement errors of skewed lines, which results in a 5% repeatability drop compared to *Default*.
The SURF detector detected five times fewer features than SIFT but achieved a slightly higher repeatability. With this detector, the *MSAA* dataset performed even better in comparison to dataset *Default* closing in to a performance difference of 3% to the *Real* dataset. *FXAA* also performed better on SURF, showing that box filters could take advantage of Antialiasing. Nonetheless, it should be noted that SURF repeatability rates dispersed much more than on SIFT. *Super Sampling* also showed a slight increase, which however can be considered to be within the error of measurement.
Evaluating the repeatability of MSER on dataset *Real* against SIFT and SURF, displayed a drop in repeatability of 13-15%, making it the least suited region detector for aerial images of this nature. Here, all synthetic datasets heavily dropped in repeatability performance. However, *MSAA* and especially *FXAA* could slightly close the resulting gap. *SSAA* decreased the performance even further.

4.2.3 Objects, Shadows and Lighting

In every dataset class, containing man-made or natural objects the CGI dataset *No Objects* performed best, even better than corresponding *Real* datasets. This is due to the prominent 0.03 mpp pattern together with the 0.2 mpp surface texture along the ground surface. This is mainly due to the homography assumption of a planer surface was fully fulfilled.
During evaluation of dataset class *Woods* the number of correspondences raised extensively for MSER (avg. 6000) and SURF (avg. 1700) even higher than evaluated real datasets (MSER: 2000 and SURF: 900). SIFT however behaved similar to its results in dataset class *Field* or *Concrete*. The difference between real and synthetic datasets lay in the absence of leaves in dataset *Real*. Leaves created a large number of extremes, because of their cluttered and overlapping distribution. Repeatability rates using *Woods* datasets ranged from 34 to 67% indicating violation of the homography ground truth constraint.
Dataset classes *Woods* and *Concrete* where used to evaluate the effect of shadows in CGI on detector repeatability and number of correspondences leading to differences of less than 1% in repeatability and number of correspondences. Even the removal of shadows did not change this effect. Thus, shadow generation is not influencing the performance of feature detectors.

5 Conclusion

In this paper, the next step of a concept to evaluate synthetic datasets using computer vision algorithms has been presented. Here, feature detectors were used to evaluate their performance on real and scene-wise corresponding synthetic datasets depicting airborne reconnaissance imagery. In addition, differences in behavior between the detectors have been discussed. The objective was to investigate the use of synthetic environments for CV-algorithm prototyping and evaluation. Additionally the influence of specific rendering techniques has been investigated.
Therefore, a test flight has been conducted recording airborne imagery and position of the aircraft, which was reproduced in a synthetic environment. To achieve correspondence, the terrain has been modelled in geo-referenced detail (textures, terrain and man-made objects). The recorded images have been separated into three terrain classes. The performance was evaluated using the repeatability metric, which used a homography-based ground truth.
In general, *Real* datasets performed roughly equal for SIFT and SURF detectors and 20% better for the MSER detector then *Default* synthetic datasets. Additionally in total, more feature correspondences have been found in real datasets, due to more extremes in the images (e.g. intensity, edges). It has been identified that a high quality ground texture (at least half of the cameras ground resolution) was mandatory.
These textures could however be procedural and repetitive. For increased performance, a high-resolution satellite image (0.2 mpp) was blended with the procedural texture. Additional

Fig. 4. Example patches of datasets (from left to right). *Field: Real, CGI Default* and *CGI Texure Low. Concrete: Real, CGI Default* and *CGI MSAA*.

rendering methods, such as *Multi Sampling* (for SURF, SIFT) or *Fast Approximate* (for MSER) *Antialiasing* improved the repeatability of synthetic datasets. Synthetic datasets with and without objects have been evaluated resulting in too high performance when objects are missing, due to its planar surface. Shadow generation techniques were also tested showing no influence on repeatability measures.

Aforementioned results lead to the conclusion that used setup demonstrated the usability of synthetic environments. Therefore, feature-based algorithms can be prototyped or evaluated in synthetic environments when mentioned constraints are considered and can be improved using anti-aliasing methods.

The next steps will be a dependency analysis weighing acquired results against numerical distance measures of MPEG7 image retrieval descriptors, intended to identify image parameters influencing the performance of synthetic datasets. Furthermore, a metric allowing evaluation on perspective datasets or without the planar level constraint would increase the range of possible datasets. Moreover, the study could be extended with additional rendering techniques, image descriptors and metrics. In addition, the evaluation of computer vision algorithms could be extended to CV-algorithms that are more complex such as object detectors or trackers.

6 References

[1] R. Collins, X. Zhou, and S. K. Teh, "An open source tracking testbed and evaluation web site," IEEE Int. Work. Perform. Eval. Track. Surveill., pp. 17–24, 2005.

[2] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," ITE J. (Institute Transp. Eng., vol. 74, no. August, pp. 22–26, 2004.

[3] W. Burger and M. Barth, "Virtual Reality for enhanced computer vision," Virtual Prototyp. Virtual Environ. …, 1995.

[4] G. Hummel, L. Kovács, P. Stütz, and T. Szirányi, "Data Simulation and Testing of Visual Algorithms in Synthetic Environments for Security Sensor Networks," in Future Security, 2012, vol. 318, pp. 212–215.

[5] R. Szeliski, Computer vision: algorithms and applications. 2011.

[6] K. Martull, S., Peris, M., & Fukui, "Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps," Int. Conf. Pattern Recognit., pp. 117–118, 2012.

[7] H. Tamura and H. Kato, "Proposal of international voluntary activities on establishing benchmark test schemes for AR/MR geometric registration and tracking methods," in ISMAR. 8th IEEE Int. Symp. on, 2009, pp. 233–236.

[8] M. Berger, J. a. Levine, L. G. Nonato, G. Taubin, and C. T. Silva, "A Benchmark for Surface Reconstruction," ACM Trans. Graph., vol. 32, no. 2, pp. 20:1–20:17, 2013.

[9] J. Ferwerda, "Three varieties of realism in computer graphics," Proc. SPIE Hum. Vis. Electron. …, vol. SPIE 5007, pp. 290–297, 2003.

[10]G. Hummel and P. Stuetz, "Using Virtual Simulation Environments for Development and Qualification of UAV Perceptive Capabilities: Comparison of Real and Rendered Imagery with MPEG7 Image Descriptors," in MESAS: First International Workshop, Rome, Italy, May, 2014, vol. 8906 LNCS, pp. 27-43.

[11]F. Boehm and A. Schulte, "Scalable COTS Based Data Processing and Distribution Architecture for UAV Technology Demonstrators," Eur. Telem. Test Conf. etc …, 2012.

[12]D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[13]H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," Lect. Notes Comput. Sci., vol. 3951 LNCS, pp. 404–417, 2006.

[14]J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *Br. Mach. Vis. Conf.*, pp. 384–393, 2002.

[15]K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J.* ..., 2005.

[16]K. Lenc, V. Gulshan, and A. Vedaldi, "VLBenchmarks," 2012. [Online]. Available: http://www.vlfeat.org/benchmarks/. [Accessed: 25-Mar-2015].

[17]M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24. pp. 381–395, 1981.