CHAPTER 7

Ground Truth Data, Content, Metrics, and Analysis

Buy the truth and do not sell it.

—Proverbs 23:23

This chapter discusses several topics pertaining to *ground truth data*, the basis for computer vision metric analysis. We look at examples to illustrate the importance of ground truth data design and use, including manual and automated methods. We then propose a method and corresponding ground truth dataset for measuring interest point detector response as compared to human visual system response and human expectations. Also included here are example applications of the general robustness criteria and the general vision taxonomy developed in Chapter 5 as applied to the preparation of hypothetical ground truth data. Lastly, we look at the current state of the art, its best practices, and a survey of available ground truth datasets.

Key topics include:

- Creating and collecting ground truth data: manual vs. synthetic methods
- Labeling and describing ground truth data: automated vs. human annotated
- Selected ground truth datasets
- Metrics paired with ground truth data
- Over-fitting, under-fitting, and measuring quality
- Publically available datasets
- An example scenario that compares the human visual system to machine vision detectors, using a synthetic ground truth dataset

Ground truth data may not be a cutting-edge research area, however it is as important as the algorithms for machine vision. Let's explore some of the best-known methods and consider some open questions.

What Is Ground Truth Data?

In the context of computer vision, ground truth data includes a set of images, and a set of labels on the images, and defining a modelfor object recognition as discussed in Chapter 4, including the count, location, and relationships of key features. The labels are added either by a human or automatically by image analysis, depending on the complexity of the problem. The collection of labels, such as interest points, corners, feature descriptors, shapes, and histograms, form a model.

A model may be trained using a variety of machine learning methods. At run-time, the detected features are fed into a classifier to measure the correspondence between detected features and modeled features. Modeling, classification, and training are statistical and machine learning problems, however, that are outside the scope of this book. Instead, we are concerned here with the content and design of the ground truth images.

Creating a ground truth dataset, then, may include condieration of the following major tasks:

- **Model design.** The model defines the composition of the objects—for example, the count, strength, and location relationship of a set of SIFT features. The model should be correctly fitted to the problem and image data so as to yield meaningful results.
- **Training set.** This set is collected and labeled to work with the model, and it contains both positive and negative images and features. Negatives contain images and features intended to generate false matches; see Figure 7-1.

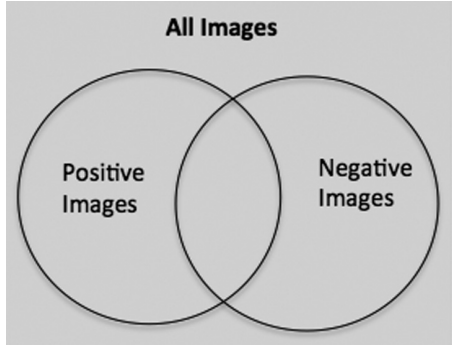


Figure 7-1. *Set of all ground truth data, composed of both positive and negative training examples*

- **Test set.** A set of images is collected for testing against the training set to verify the accuracy of the model to predict the correct matches.

- **Classifier design.** This is constructed to meet the application goals for speed and accuracy, including data organization and searching optimizations for the model.
- **Training and testing.** This work is done using several sets of images to check against ground truth.

Unless the ground truth data contains carefully selected and prepared image content, the algorithms cannot be measured effectively. Thus, *ground-truthing* is closely related to root-causing: there is no way to improve what we cannot measure and do not understand. Being able to root-cause algorithm problems and understand performance and accuracy are primary purposes for establishing ground truth data. Better ground truth data will enable better analysis.

Ground truth data varies by task. For example, in 3D image reconstruction or face recognition, different attributes of the ground truth data must be recognized for each task. Some tasks, such as face recognition, require segmentation and labeling to define the known objects, such as face locations, position and orientation of faces, size of faces, and attributes of the face, such as emotion, gender, and age. Other tasks, such as 3D reconstruction, need the raw pixels in the images and a reference 3D mesh or point cloud as their ground truth.

Ground truth datasets fall into several categories:

- **Synthetic produced**: images are generated from computer models or renderings.
- **Real produced**: a video or image sequence is designed and produced.
- **Real Selected**: real images are selected from existing sources.
- **Machine-automated annotation:** feature analysis and learning method are used to extract features from the data.
- **Human annotated**: an expert defines the location of features and objects.
- **Combined**: any mixture of the above.

Many practitioners are firmly against using synthetic datasets and insist on using real datasets. In some cases, random ground truth images are required; in other cases, carefully scripted and designed ground truth images need to be produced, similar to creating a movie with scenes and actors.

Random and natural ground truth data with unpredictable artifacts, such as poor lighting, motion blur, and geometric transformation, is often preferred. Many computer problems demand real images for ground truth, and random variations in the images are important. Real images are often easy to obtain and/or easy to generate using a video camera or even a cell phone camera. But creating synthetic datasets is not as clear; it requires knowledge of appropriate computer graphics rendering systems and tools, so the time investment to learn and use those tools may outweigh their benefits.

However, synthetic computer-generated datasets can be a way to avoid legal and privacy issues concerning the use of real images.

Previous Work on Ground Truth Data: Art vs. Science

In this section, we survey some literature on ground truth data. We also highlight several examples of automatic ground truth data labeling, as well as other research on metrics for establishing if, in fact, the ground truth data is effective. Other research surveyed here includes how closely ground truth features agree with human perception and expectations, for example, whether or not the edges that humans detect in the ground truth data are, in fact, found by the chosen detector algorithms.

General Measures of Quality Performance

Compared to other topics in computer vision, little formal or analytic work has been published to guide the *creation* of ground truth data. However, the machine learning community provides a wealth of guidance for measuring the *quality* of visual recognition between ground truth data used for training and test datasets. In general, the size of the training set or ground truth data is key to its accuracy [336–338] and the larger the better, assuming the right data is used.

Key journals to dig deeper into machine learning and testing against ground truth data include the journal IEEE PAMI for Pattern Analysis and Machine Intelligence, whose articles on the subject go back to 1979. While the majority of ground truth datasets contain real images and video sequences, some practitioners have chosen to create synthetic ground truth datasets for various application domains, such as the standard Middlebury dataset with synthetic 3D images. See Appendix B for available real ground truth datasets, along with a few synthetic datasets.

One noteworthy example framework for ground truth data, detector, and descriptor evaluation is the Mikolajczyk and Schmidt methodology (M&S), discussed later in this chapter. Many computer vision research projects follow the M&S methodology using a variety of datasets.

Measures of Algorithm Performance

Ericsson and Karlsson[102] developed a ground truth correspondence measure (GCM) for benchmarking and ranking algorithm performance across seven real datasets and one synthetic dataset. Their work focused on statistical shape models and boundaries, referred to as *polygon shape descriptors* in the vision taxonomy in Chapter 5. The goal was to automate the correspondence between shape models in the database and detected shapes from the ground truth data using their GCM. Since shape models can be fairly complex, the goal of automating model comparisons and generating quality metrics specific to shape description is novel.

Dutagaci et al.[91] developed a framework and method, including ground truth data, to measure the *perceptual* agreement between humans and 3D interest point detectors—in other words, do the 3D interest point detectors find the same interest points as the humans expect? The ground truth data includes a known set of human-labeled interest points within a set of images, which were collected automatically by an Internet

scraper application. The human-labeled interest points were sorted toward a consensus set, and outliers were rejected. The consensus criterion was a radius region counting the number of humans who labeled interest points within the radius. A set of 3D interest point detectors was ran against the data and compared using simple metrics such as false positives, false negatives, and a weighted miss error. The ground truth data was used to test the agreement between humans and machine vision algorithms for 3D interest point detectors. The conclusions included observations that humans are indecisive and widely divergent about choosing interest points, and also that interest point detection algorithms are a fuzzy problem in computer vision.

Hamameh et al.[88] develop a method of automatically generating ground truth data for medical applications from a reference dataset with known landmarks, such as segmentation boundaries and interest points. The lack of experts trained to annotate the medical images and generate the ground truth data motivated the research. In this work, the data was created by generating synthetic images simulating object motion, vibrations, and other considerations, such as noise. Prestawa et al.[89] developed a similar approach for medical ground truth generation. Haltakov et al.[510] developed synthetic ground truth data from an automobile-driving simulator for testing driver assistance algorithms, which provided situation awareness using computer vision methods.

Vedaldi et al.[90] devised a framework for characterizing affine co-variant detectors, using synthetically generated ground truth as 3D scenes employing raytracing, including simulated natural and man-made environments; a depth map was provided with each scene. The goal was to characterize co-variant detector performance under affine deformations, and to design better covariant detectors as a result. A set of parameterized features were defined for modeling the detectors, including points, disks and oriented disks, and various ellipses and oriented ellipses. A large number of 3D scenes were generated, with up to 1,000 perspective views, including depth maps and camera calibration information. In this work, the metrics and ground truth data were designed together to focus on the analysis of geometric variations. Feature region shapes were analyzed with emphasis on disks and warped elliptical disks to discover any correspondence and robustness over different orientations, occlusion, folding, translation, and scaling. (The source code developed for this work is available.[1])

Rosin's Work on Corners

Research by Rosin[61,92] involved the development of an analytical taxonomy for gray scale corner properties, as illustrated in Figure 7-2. Rosin developed a methodology and case study to generate both the ground truth dataset and the metric basis for evaluating the performance and accuracy of a few well-known corner detectors. The metric is based on the receiver operating characteristic (ROC) to measure the accuracy of detectors to assess corners vs. noncorners. The work was carried out over 13,000 synthetic corner images with variations on the synthetic corners to span different orientations, subtended angles, noise, and scale. The synthetic ground truth dataset was specifically designed to enable the detection and analysis of a set of chosen corner properties, including bluntness or shape of apex, boundary shape of cusps, contrast, orientation, and subtended angle of the corner.

[1]See the "*VLFeat*" open-source project online (`http://www.vlfeat.org`").

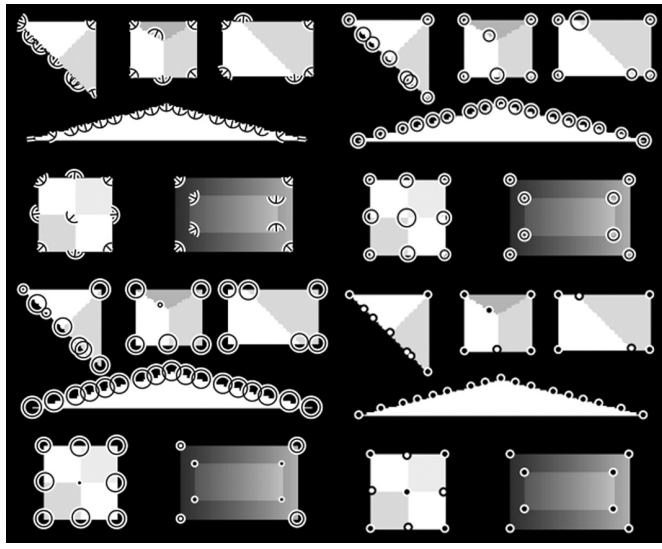


Figure 7-2. *Images illustrating the Rosin corner metrics: (Top left) Corner orientation and subtended angle. (Top right) Bluntness. (Bottom left) Contrast. (Bottom right) Black/white corner color. (Images © Paul Rosin and used by permission[61])*

A novel aspect of Rosin's work was the generation of explicit types of synthetic interest points such as corners, nonobvious corners, and noncorners into the dataset, with the goal of creating a statistically interesting set of features for evaluation that diverged from idealized features. The synthetic corners were created and generated in a simulated optical system for realistic rendering to produce corners with parameterized variations including affine transformations, diffraction, sub-sampling, and in some cases, adding noise. Rosin's ground truth dataset is available for research use, and has been used for corner detector evaluation of methods from Kitchen and Rosenfeld, Paler, Foglein, and Illingworth, as well as the Kittler Detector and the Harris & Stephens Detector.

Similar to Rosin, a set of synthetic interest point alphabets are developed later in this chapter snf tested in Appendix A, including edge and corner alphabets, with the goal of comparing human perception of interest points against machine vision methods. The synthetic interest points and corners are designed to test pixel thickness, edge intersections, shape, and complexity. The set diverges significantly from those of Rosin and others, and attempts to fill a void in the analysis of interest point detectors. The alphabets are placed on a regular grid, allowing for detmining position detection count.

Key Questions For Constructing Ground Truth Data

In this section we identify some key questions to answer for creating ground truth data, rather than provideing much specific guidance or answers. The type of work undertaken will dictates the type of guidance, for example, published research usually requires widely accepted ground truth data to allow for peer review and duplication of results. In medical or automobile industries,there may be government regulations, and also legal issues if competitors publish measurement or performance data. For example, if a company publishes any type of benchmark results against a ground truth data set comparing the results with those of competitor systems, all such data and claims should be reviewed by an attorney to avoid the complexities and penalties of commerce regulations, which can be daunting and severe.

For real products and real systems, perhaps the best guidance comes from the requirements, expectations and goals for performance and accuracy.Once a clear set of requirements are in place, then the ground truth selection process can begin.

Content: Adopt, Modify, or Create

It is useful to become familiar with existing ground truth datasets prior to creating a new one. The choices are obvious:

- Adopt an existing dataset.
- Adopt-And-Modify an existing data set.
- Create a new dataset.

Survey Of Available Ground Truth Data

Appendix B has information on several existing ground truth datasets. Take some time to get to know what is already available, and study the research papers coming out of SIGGRAPH, CVPR, IJCV, NIPS in Appendix C, and other research conferences to learn more about new datasets and how they are being used. The available datasets come from a variety of sources, including:

- Academic research organizations, usually available free of charge for academic research.
- Government datasets, sometimes with restricted use.
- Industry datasets, available from major corporations like Microsoft, sometimes can be licensed for commercial use.

Fitting Data to Algorithms

Perhaps the biggest challenge is to determine whether a dataset is a correct fit for the problem at hand. Is the detail in the ground truth data sufficient to find the boundaries and limits of the chosen algorithms and systems? "Fitting" applies to key variables such as the ground truth data, the algorithms used, the object models, classifier, and the intended use-cases. See Figure 7-3, which shows how ground truth data, image pre-processing, detector and descriptor algorithms, and model metrics should be fitted together.

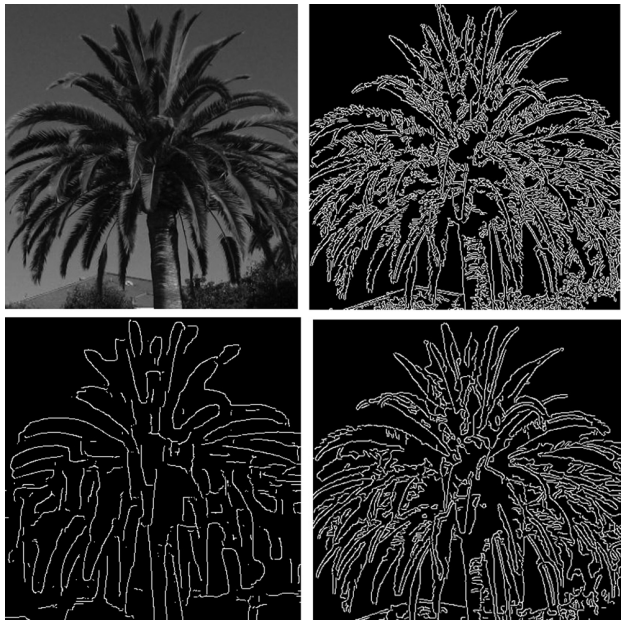


Figure 7-3. *(Top left) Image pre-processing for edges shown using Shen-Castan edge detection against ground truth data. (Top right) Over-fitting detection parameters yield too many small edges. (Bottom left) Under fitting parameters yield too few edges. (Bottom right) Relaxed parameters yield reasonable edges*

Here are a few examples to illustrate the variables.

- **Data fitting:** If the dataset does not provide enough pixel resolution or bit depth, or there are insufficient unique samples in the training set, the model will be incomplete, the matching may suffer, and the data is *under-fitted* to the problem. Or, if the ground truth contains too many different types of features that will never be encoutered in the test set or in real applications. If the model resolution is 16 bits per RGB channel when only 8 bits per color channel are provided in real data, the data and model are *over-fitted* to the problem.
- **Algorithm fitting:** If scale invariance is included in the ground truth data, and the LBP operator being tested is not claimed to be scale invariant, then the algorithm is *under-fitted* to the data. If the SIFT method is used on data with no scale or rotation variations, then the SIFT algithm is *over-fitted* to the data.
- **Use-case fitting:** If the use-cases are not represented in the data and model, the data and model are *under-fitted* to the problem.

Scene Composition and Labeling

Ground truth data is composed of labeled features such as foreground, background, and objects or features to recognize. The labels define exactly what features are present in the images, and these labels may be a combination of on-screen labels, associated label files, or databases. Sometimes a randomly composed scene from the wild is preferred as ground truth data, and then only the required items in the scene are labeled. Other times, ground truth data is scripted and composed the way a scene for a movie would be.

In any case, the appropriate objects and actors in the scene must be labeled, and perhaps the positions of each must be known and recorded as well. A database or file containing the labels must therefore be created and associated with each ground truth image to allow for testing. See Figure 7-4, which shows annotated or labeled ground truth dataset images for a scene analysis of cuboids [62]. See also the Labelme database described in Appendix B, which allows contributors to provide labeled databases.

Figure 7-4. *Annotated or labeled ground-truth dataset images for scene analysis of cuboids (left and center). The labels are annotated manually into the ground- truth dataset, in yellow (light gray in B&W version) marking the cuboid edges and corners. (right) Ground-truth data contains pre-computed 3D corner HOG descriptor sets, which are matched against live detected cuboid HOG feature sets. Successful matches shown in green (dark gray in B&W version). (Images used by permission © Bryan Russel, Jianxiong Xiao, and Antonio Torralba)*

Composition

Establishing the right set of ground truth data is like asssembling a composition; several variables are involved, including:

- **Scene Content**: Designing the visual content, including fixed objects (those that do not move), dynamic objects (those that enter and leave the scene), and dynamic variables (such as position and movement of objects in the scene).
- **Lighting**: Casting appropriate lighting onto the scene.
- **Distance**: Setting and labeling the correct distance for each object to get the pixel resolution needed—too far away means not enough pixels.
- **Motion Scripting**: Determining the appropriate motion of objects in the scene for each frame; for example, how many people are in the scene, what are their positions and distances, number of frames where each person appears, and where each person enters and exits. Also, scripting scenes to enable invariance testing for changes in perspective, scale, affine geometry, occlusion.
- **Labeling**: Creating a formatted file, database, or spreadsheet to describe each labeled ground truth object in the scene for each frame.
- **Intended Algorithms**: Deciding which algorithms for interest point and feature detection will be used, what metrics are to be produced, and which invariance attributes are expected from each algorithm; for example, an LBP by itself does not provide scale invariance, but SIFT does.

- **Intended Use-Cases**: Determining the problem domain or application. Does the ground truth data represent enough real use-cases?
- **Image Channel Bit Depth, Resolution:** Setting these to match requirements.
- **Metrics**: Defining the group of metrics to measure—for example, false positives and false negatives. Creating a test fixture to run the algorithms against the dataset, measuring and recording all necessary results.
- **Analysis**: Interpreting the metrics by understanding the limitations of both the ground truth data and the algorithms, defining the success criteria.
- **Open Rating Systems:** Exploring whether there is an open rating system that can be used to report the results. For example, the Middlebury Dataset provides an open rating system for 3D stereo algorithms, and is described in Appendix B; other rating systems are published as a part of grand challenge contests held by computer vision organizations and governments, and some are reviewed in Appendix B. Open rating systems allow existing and new algorithms to be compared on a uniform scale.

Labeling

Ground truth data may simply be images returned from a search engine, and the label may just be the search engine word or phrase. Figure 7-5 shows a graph of photo connectivity for photo tourism [63–65] that is created from pseudo-random images of a well-known location, the Trevi Fountain in Rome. It is likely that in five to ten years, photo tourism applications will provide high-quality image reconstruction including textures, 3D surfaces, and rerenderings of the same location, rivaling real photographs.

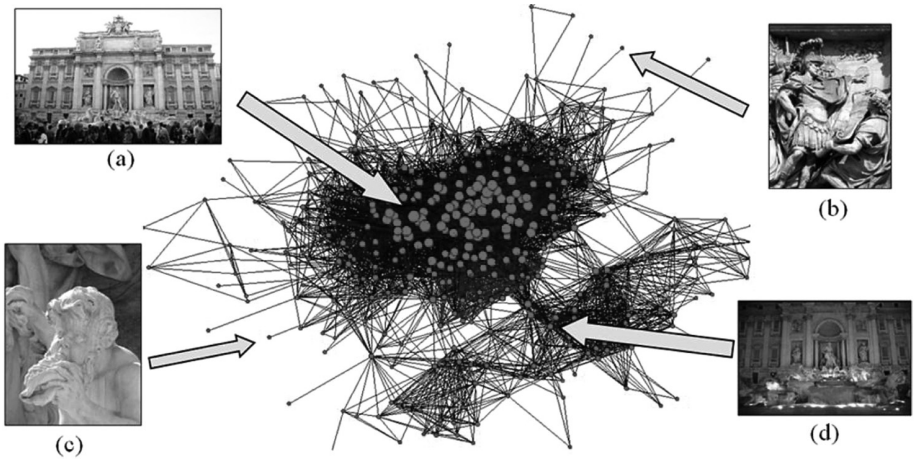


Figure 7-5. *Graph of photo connectivity (center) created from analyzing multiple public images from a search engine of the Trevi Fountain (a). Edges show photos matched and connected to features in the 3D scene, including daytime and nighttime lighting (b)(c)(d). (Images © Noah Snavely and used by permission)*

For some applications, labels and markers are inserted into the ground truth datasets to enable analysis of results, as shown in the 3D scene understanding database for cuboids in Figure 7-4. Another example later in this chapter composes scenes using *synthetic alphabets* of interest points and corners that are superimposed on the images of a regularly spaced grid to enable position verification (see also Appendix A). In some visual tracking applications, *markers* are attached to physical objects (a wrist band, for example) to establish ground truth features.

Another example is ground truth data composed to measure *gaze detection,* using a video sequence containing labels for two human male subjects entering and leaving the scene at a known location and time, walking from left to right at a known speed and depth in the scene. The object they are gazing at would be at a known location and be labeled as well.

Defining the Goals and Expectations

To establish goals for the ground truth data, questions must be asked. For instance, what is the intended use of the application requiring the ground truth data? What decisions must be made from the ground truth data in terms of accuracy and performance? How is quality and success measured? The goals of academic research and commercial systems are quite different.

Mikolajczyk and Schmid Methodology

A set of well-regarded papers by Mikolajczyk, Schmid and others [45,79,82,91,306] provides a good methodology to start with for measuring local interest points and feature detector quality. Of particular interest is the methodology used to measure scale and affine invariant interest point detectors [306] which uses natural images to start, then applies a set of known affine transformations to those images, such as homography, rotation, and scale. Interest point detectors are run against the images, followed by feature extractors, and then the matching recall and precision are measured across the transformed images to yield quality metrics.

Open Rating Systems

The computer vision community is, little by little, developing various open rating systems, which encourage algorithm comparisons and improvements to increase quality. In areas where such open databases exist, there is rapid growth in quality for specific algorithms. Appendix B lists open rating systems such as the Pascal VOC Challenge for object detection. Pascal VOC uses an open ground truth database with associated grand challenge competition problems for measuring the accuracy of the latest algorithms against the dataset.

Another example is the Middlebury Dataset, which provides ground truth datasets covering the 3D stereo algorithm domain, allowing for open comparison of key metrics between new and old algorithms, with the results published online.

Corner Cases and Limits

Finding out where the algorithms fail is valuable. Academic research is often not interested in the rigor required by industry in defining failure modes. One way to find the corner cases and limits is to run the same tests on a wide range of ground truth data, perhaps even data that is outside the scope of the problem at hand. Given the availability of publicly available ground truth databases, using several databases is realistic.

However, once the key ground truth data is gathered, it can also be useful to devise a range of corner cases—for example, by providing noisy data, intensity filtered data, or blurry data to test the limits of performance and accuracy.

Interest Points and Features

Interest points and features are not always detected as expected or predicted. Machine vision algorithms detect a different set of interst points than those humans expect. For example, Figure 7-6 shows obvious interest points missed by the SURF algorithm with a given set of parameters, which uses a method based on determinant of Hessian blob detection. Note that some interest points obvious to humans are not detected at all, some false positives occur, and some identical interest points are not detected consistently.

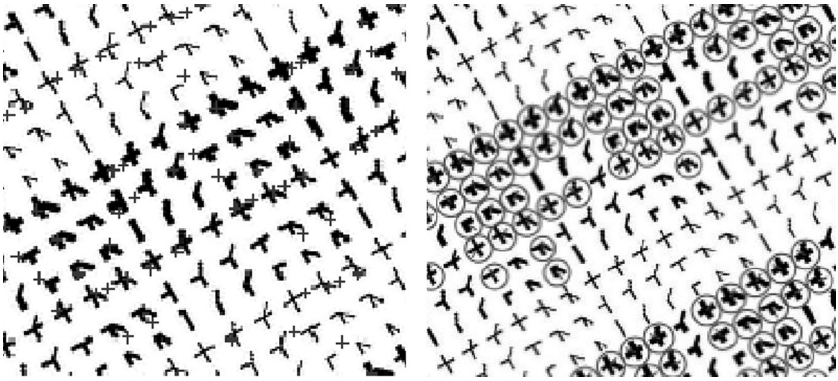


Figure 7-6. *Interest points detected on the same image using different methods: (Left) Shi-Tomasi corners marked with crosses. (Right) SURF interest points marked with circles. Results are not consistent or deterministic*

Also, real interest points change over time—for example, as objects move and rotate—which is a strong agrument for using real ground truth data vs. synthetic data to test a wide range of potential interest points for false positives and false negatives.

Robustness Criteria for Ground Truth Data

In Chapter 5, a robustness criteria was developed listing various invariance attributes, such as rotation and scale. Here, we apply the robustness criteria to the development of ground truth data.

Illustrated Robustness Criteria

Table 7-1 discusses various robustness criteria attributes, not all attributes are needed for a given application. For example, if radial distortion might be present in an optical system, then the best algorithms and corresponding metrics will be devised that are robust to radial distortion, or as mitigation, the vision pipeline must be designed with a pre-processing section to remove or compensate for the radial distortion prior to determining the metrics.

Table 7-1. *Robustness Criteria for Ground Truth Data*

Attribute	Discussion
Uneven illumination	Define range of acceptable illumination for the application; uneven illumination may degrade certain algorithms, some algorithms are more tolerant.
Brightness	Define expected brightness range of key features, and prepare ground-truth data accordingly.
Contrast	Define range of acceptable contrast for the application; some algorithms are more tolerant.
Vignette	Optical systems may degrade light and manifest as dim illumination at the edges. Smaller the features are localized better and may be able to overcome this situation; large features that span areas of uneven light are affected more.
Color accuracy	Inaccurate color space treatment may result in poor color performance. Colorimetry is important; consider choosing the right color space (RGB, YIQ, Lab, Jab, etc.) and use the right level of bit precision for each color, whether 8/16 bits is best.
Clutter	Some algorithms are not tolerant of clutter in images and rely on the scene to be constructed with a minimal number of subjects. Descriptor pixel size may be an issue for block search methods—too much extraneous detail in a region may be a problem for the algorithm.
Occlusion and clipping	Objects may be occluded or hidden or clipped. Algorithms may or may not tolerate such occlusion. Some occlusion artifacts can be eliminated or compensated for using image pre-processing and segmentation methods.
Outliers and proximity	Sometimes groups of objects within a region are the subject, and outliers are to be ignored. Also, proximity of objects or features may guide classification, so varying the arrangement of features or objects in the scene may be critical.
Noise	Noise may take on regular or random patterns, such as snow, rain, single-pixel spot nose, line noise, random electrical noise affecting pixel bit resolution, etc.
Motion blur	Motion blur is an important problem for almost all real-time applications. This can be overcome by using faster frame rates and employing image pre-processing to remove the motion blur, if possible.
Jitter and judder	Common problem in video images taken from moving cameras, where each scan line may be offset from the regular 2D grid.

(*continued*)

Table 7-1. (*continued*)

Attribute	Discussion
Focal plane or depth	If the application or use-case for the algorithm assumes all depths of the image to be in focus, then using ground truth data with out-of-focus depth planes may be a good way to test the limits.
Pixel depth Resolution	If features are matched based on the value of pixels, such as gray scale intensity or color intensity, pixel resolution is an issue. For example, if a feature descriptor uses 16 bits of effective gray scale intensity but the actual use-case and ground truth data provide only 8 bits of resolution, the descriptor may be over-fitted to the data, or the data may be unrealistic for the application.
Geometric distortion	Complex warping may occur due to combinations of geometric errors from optics or distance to subject. On deformable surfaces such as the human face, surface and feature shape may change in ways difficult to geometrically describe.
Scale, projection	Near and far objects will be represented by more or less pixels, thus a multi-scale dataset may be required for a given application, as well as multi-scale feature descriptors. Algorithm sensitivity to feature scale and intended use case also dictate ground truth data scale.
Affine transforms and rotation	In some applications like panoramic image stitching, very little rotation is expected between adjacent frames—perhaps up to 15 degrees may be tolerated. However, in other applications like object analysis and tracking of parts on an industrial conveyor belt, rotation between 0 and 360 degrees is expected.
Feature mirroring, translation	In stereo correspondence, L/R pair matching is done using the assumption that features can be matched within a limited range of translation difference between L/R pairs. If the translation is extreme between points, the stereo algorithm may fail, resulting in holes in the depth map, which must be filled.
Reflection	Some applications, like recognizing automobiles in traffic, require a feature model, which incorporates a reflective representation and a corresponding ground truth dataset. Automobiles may come and go from different directions, and have a reflected right/left feature pair.
Radial distortion	Optics may introduce radial distortion around the fringes; usually this is corrected by a camera system using digital signal processors or fixed-function hardware prior to delivering the image.

Using Robustness Criteria for Real Applications

Each application requires a different set of robustness criteria to be developed into the ground truth data. Table 7-2 illustrates how the robustness criteria may be applied to a few real and diverse applications.

Table 7-2. *Robustness Criteria Applied to Sample Applications (each application with different requirements for robustness)*

General Objective Criteria Attributes	Industrial inspection of apples on a conveyor belt, fixed distance, fixed speed, fixed illumination	Automobile identification on roadway, day and night, all road conditions	Multi-view stereo reconstruction bundle adjustment
Uneven illumination	-	Important	Useful
Brightness	Useful	Important	Useful
Contrast	Useful	Important	Useful
Vignette	Important	Useful	Useful
Color accuracy	Important	Important	Useful
Clutter	-	Important	Important
Occlusion	-	Important	Important
Outliers	-	Important	Important
Noise	-	Important	Useful
Motion blur	Useful	Important	Useful
Focal plane or depth	-	Important	Useful
Pixel depth resolution	Useful	Important	important
Subpixel resolution	-	-	important
Geometric distortion (warp)	-	Useful	Important
Affine transforms	-	Important	Important
Scale	-	Important	Important
Skew	-	-	-
Rotation	Important	Useful	Useful
Translation	Important	Useful	Useful

(*continued*)

Table 7-2. (*continued*)

General Objective Criteria Attributes	Industrial inspection of apples on a conveyor belt, fixed distance, fixed speed, fixed illumination	Automobile identification on roadway, day and night, all road conditions	Multi-view stereo reconstruction bundle adjustment
Projective transformations	Important	Important	-
Reflection	Important	Important	-
Radial distortion	-	-	Important
Polar distortion	-	-	Important
Discrimination or uniqueness	-	Useful	-
Location accuracy	-	Useful	-
Shape and thickness distortion	-	Useful	-

As illustrated in Table 7-2, a multi-view stereo (MVS) application will hold certain geometric criteria as very important, since accurate depth maps require accurate geometry assumptions as a basis for disparity calculations. For algorithm accuracy tuning, corresponding ground truth data should be created using a well-calibrated camera system for positional accuracy of the 3D scene to allow for effective comparisons.

Another example in Table 7-2 with many variables in an uncontrolled environment is that of automobile identification on roadways—which may be concerned with distance, shape, color, and noise. For example, identifying automobiles may require ground truth images of several vehicles from a wide range of natural conditions, such as dawn, dusk, cloudy day, and full sun, and including conditions such as rainfall and snowfall, motion blur, occlusion, and perspective views. An example automobile recognition pipeline is developed in Chapter 8.

Also shown Table 7-2 is an example with a controlled environment: industrial inspection. In industrial settings, the environment can be carefully controlled using known lighting, controlling the speed of a conveyor belt, and limiting the set of objects in the scenes. Accurate models and metrics for each object can be devised, perhaps taking color samples and so forth—all of which can be done a priori. Ground truth data could be easily created from the actual factory location.

Pairing Metrics with Ground Truth

Metrics and ground truth data should go together. Each application will have design goals for robustness and accuracy, and each algorithm will also have different intended uses and capabilities. For example, the SUSAN detector discussed in Chapter 6 is often applied to wide baseline stereo applications, and stereo applications typically are not

concerned much with rotational invariance because the image features are computed on corresponding stereo pair frames that have been affine rectified to align line by line. Feature correspondence between image pairs is expected within a small window, with some minor translation on the *x* axis.

Pairing and Tuning Interest Points, Features, and Ground Truth

Pairing the right interest point detectors and feature descriptors can enhance results, and many interest point methods are available and were discussed in Chapter 6. When preparing ground truth data, the method used for interest point detection should be considered for guidance.

For example, interest point methods using derivatives, such as the Laplace and Hessian style detectors, will not do very well without sufficient contrast in the local pixel regions of the images, since contrast accentuates maxima, minima and local region changes. However, a method such as FAST9 is much more suited to low-contrast images, uses local binary patterns, and is simple to tune the compare threshold and region size to detect corners and edges; but the tradeoff in using FAST9 is that scale invariance is sacrificed.

A method using edge gradients and direction, such as eigen methods, would require ground truth containing sufficient oriented edges at the right contrast levels. A method using morphological interest points would likewise require image data that can be properly thresholded and processed to yield the desired shapes.

Interest point methods also must be tuned for various parameters like strength of thresholds for accepting and rejecting candidate interest points, as well as and region size. Choosing the right interest point detector, tuning, and pairing with appropriate ground truth data are critical. The effect of tuning interest point detector parameters is illustrated in Figures 7-6 and 7-7.

Figure 7-7. *Machine corner detection using the Shi-Tomasi method marked with crosses; results are shown using different parameter settings and thresholds for the strength and pixel size of the corners*

Examples Using The General Vision Taxonomy

As a guideline for pairing metrics and ground truth data, we use the vision taxonomy developed in Chapter 5 to illustrate how feature metrics and ground truth data can be considered together.

Table 7-3 presents a sample taxonomy and classification for SIFT and FREAK descriptors, which can be used to guide selection of ground truth data and also show several similarities in algorithm capabilities. In this example, the invariance attributes built into the data can be about the same— namely scale and rotation invariance. Note that the compute performance claimed by FREAK is orders of magnitude faster than SIFT, so perhaps the ground truth data should contain a sufficient minimum and maximum number of features per frame for good performance measurements.

Table 7-3. *General Vision Taxonomy for Describing FREAK and SIFT*

Visual Metric Taxonomy Comparison		
Attribute	**SIFT**	**FREAK**
Feature Category Family	Spectra Descriptor	Local Binary Descriptor
Spectra Dimensions	Multivariate	Single Variate
Spectra Value	Orientation Vector Gradient Magnitude Gradient Direction HOG, Cartesian Bins	Orientation Vector Bit Vector Of values Cascade of 4 Saccadic Descriptors
Interest Point	SIFT DOG over 3D Scale Pyramid	Multi-scale AGAST
Storage Format	Spectra Vector	Bit Vector Orientation Vector
Data Types	Float	Integer
Descriptor Memory	512 bytes, 128 floats	64 Bytes, 4 16-byte Cascades
Feature Shape	Rectangle	Circular
Feature Search Method	Coarse to Fine Image Pyramid Scale Space Image Pyramid Double-scale First Pyramid Level Sparse at Interest Points	Sparse at interest points
Pattern Pair Sampling	*n.a.*	Foveal Centered Trained Pairs
Pattern Region Size	41x41 Bounding Box	31x31 Bounding Box (may vary)
Distance Function	Euclidean Distance	Hamming Distance
Run-Time Compute	100% (SIFT is the baseline)	.1% of SIFT

(*continued*)

Table 7-3. (*continued*)

Visual Metric Taxonomy Comparison		
Attribute	**SIFT**	**FREAK**
Feature Density	Sparse	Sparse
Feature Pattern	Rectangular kernel Sample Weighting Pattern	Binary compare pattern
Claimed Robustness ***Final robustness is a combination of interest point method, descriptor method, and classifier**	Scale Rotation Noise Affine Distortion Illumination	Scale Rotation Noise

Synthetic Feature Alphabets

In this section, we create synthetic ground truth datasets for interest point algorithm analysis. We create alphabets of *synthetic interest points* and *synthetic corner points*. The alphabets are *synthetic*, meaning that each element is designed to perfectly represent chosen binary patterns, including points, lines, contours, and edges.

Various pixel widths or thickness are used for the alphabet characters to measure fine and coarse feature detection. Each pattern is registered at known pixel coordinates on a grid in the images to allow for detection accuracy to be measured. The datasets are designed to enable comparison between human interest point perception and machine vision interest point detectors.

Here is a high-level description of each synthetic alphabet dataset:

- **Synthetic Interest Point Alphabet.** Contains points such as boxes, triangles, circle, half boxes, half triangles, half circles, edges, and contours.
- **Synthetic Corner Point Alphabet.** Contains several types of corners and multi-corners at different pixel thickness.
- **Natural images overlaid with synthetic alphabets**. Contains both black and white versions of the interest points and corners overlaid on natural images.

■ **Note** The complete set of ground truth data is available in Appendix A.

Analysis is provided in Appendix A, which includes running ten detectors against the datasets. The detectors are implemented in OpenCV, including SIFT, SURF, ORB, BRISK, HARRIS, GFFT, FAST9, SIMPLE BLOB, MSER, and STAR. Note that the methods such as SIFT, SURF, and ORB provide both an interest point detector and a feature descriptor implementation. We are only concerned with the interest point detector portion of each method for the analysis, not the feature descriptor.

The idea of using synthetic image alphabets is not new. As shown in Figure 7-2, Rosin[61] devised a synthetic set of gray corner points and corresponding measurement methods for the purpose of quantifying corner properties via attributes such as bluntness or shape of apex, boundary shape of cusps, contrast, orientation, and subtended angle of the corner. However, the synthetic interest point and corner alphabets in this work are developed to address a different set of goals, discussed next.

Goals for the Synthetic Dataset

The goals and expectations for this synthetic dataset are listed in Table 7-4. They center on enabling analysis to determine which synthetic interest points and corners are found, so the exact count and position of each interest point is a key requirement.

Table 7-4. *Goals and Expectations for the Ground Truth Data Examples: Comparison of Human Expectations with Machine Vision Results*

Goals	Approach
Interest point and corner detectors, stress testing	Provide synthetic features easily recognized by a human; measure how well various detectors perform.
Human recognizable synthetic interest point sets	Synthetic features recognized by humans are developed spanning shapes and sizes of edges and line segments, contours and curved lines, and corners and multi-corners.
Grid positioning of interest points	Each interest point will be placed on a regular grid at a known position for detection accuracy checking.
Scale invariance	Synthetic interest points to be created with the same general shape but using different pixel thickness for scale.
Rotation invariance	Interest points will be created, then rotated in subsequent frames.
Noise invariance	Noise will be added to some interest point sets.
Duplicate interest points, known count	Interest points will be created and duplicated in each frame for determining detection and performance.
Hybrid synthetic interest points overlaid on real images	Synthetic interest points on a grid are overlaid onto real images to allow for hybrid testing.
Interest point detectors,determinism and repeatability	Detectors will include SIFT, SURF, ORB, BRISK, HARRIS, GFFT, FAST9, SIMPLE BLOB, MSER, and STAR. By locating synthetic interest points on a grid, we can compute detection counts.

The human visual system does not work like an interest point detector, since detectors can accept features which humans may not recognize. The human visual system discriminates and responds to gradient information [248] in a scale and rotationally invariant manner across the retina, and tends to look for learned features relationships among gradients and color.

Humans learn about features by observations and experience, so learned expectations play a key role interpreting visual features. *People see what they believe and what they are looking for, and may not believe what they see if they are not looking for it.* For example, Figure 7-7 shows examples of machine corner detection; a human would likely not choose all the same corner features. Note that the results are not what a human might expect, and also the algorithm parameters must be tuned to the ground truth data to get the best results.

Accuracy of Feature Detection via Location Grid

The goal of detector accuracy for this synthetic ground truth is addressed by placing synthetic features at a known position on a regular spaced grid, then after detection, the count and position are analyzed. Some of the detectors will find multiple features for a single synthetic interest point or corner. The feature grid size chosen is 14x14 pixels, and the grid extends across the entire image. See Figures 7-9 and 7-10.

Rotational Invariance via Rotated Image Set

For each ground truth set, rotated versions of each image are created in the range 0 to 90 degrees at 10 degree increments.Since the synthetic features are placed on a regularly spaced grid at known positions, the new positions under rotation are easily computed. The detected synthetic features can be counted and analyzed. See Appendix A for results.

Scale Invariance via Thickness and Bounding Box Size

The synthetic corner point features are rendered into the ground truth data with feature edge thickness ranging from 1 to 3 pixels for simulated scale variation. Some of the interest point features, such as boxes, triangles, and circles, are scaled in a bounding box ranging from 1x1 pixels to 10x10 pixels to allow for scale invariance testing.

Noise and Blur Invariance

A set of synthetic alphabets is rendered using Gaussian noise, and another set using salt-and-pepper noise to add distortion and uncertainty to the images. In addirion, by rotating the interest point alphabet at varying angles between 0 and 90 degrees, digital blur is introduced to the synthetic patterns as they are rendered, owing to the anti-aliasing interpolations introduced in the affine transform algorithms.

Repeatabilty

Each ground truth set contains a known count of synthetic features to enable detection rates to be analyzed. To enable measurement of the repeatability of each detector, there are multiple duplicate copies of each interest point feature in each image. A human would expect identical features to be detected in an identical manner; however, results in Appendix A show that some interest point detectors do not behave in a predictable manner, and some are more predictable than others.

As shown in Figure 7-6, detectors do not always find the same identical features. For example, the synthetic alphabets are provided in three versions— black on white, white on black, and light gray on dark gray—for the purpose of testing each detector on the same pattern with different gray levels and polarity. See Appendix A showing the how the detectors provide different results based on the polarity and gray level factors.

Real Image Overlays of Synthetic Features

A set of images composed of synthetic interest points and corners overlayed on top of real images is provided, sort of like markers. Why overlay interest point markers, since the state of the art has moved beyond markers to markerless tracking? The goal is to understand the limitations and behavior of the detectors themselves, so that analyzing their performance in the presence of natural and synthetic features will provide some insight.

Synthetic Interest Point Alphabet

As shown in Figures 7-8 and 7-9, an alphabet of synthetic interest points is defined across a range of pixel resolutions or thicknesses to include the following features:

- POINT / SQUARE, 1–10 PIXELS SIZE
- POINT / TRIANGLE HALF-SQUARE, 3–1 PIXELS SIZE
- CIRCLE, 3–10 PIXELS SIZE
- CIRCLE / HALF-CIRCLE, 3–10 PIXELS SIZE
- CONTOUR, 3–10 PIXELS SIZE
- CONTOUR / HALF-CONTOUR, 3–10 PIXELS SIZE
- CONNECTED EDGES
- DOUBLE CORNER, 3–10 PIXELS SIZE
- CORNER, 3–10 PIXELS SIZE
- EDGE, 3–10 PIXELS SIZE

Figure 7-8. *Portion of the synthetic interest point alphabet: points, edges, edges, and contours. (Top to bottom) White on black, black on white, light gray on dark gray, added salt and pepper noise, added Gaussian noise*

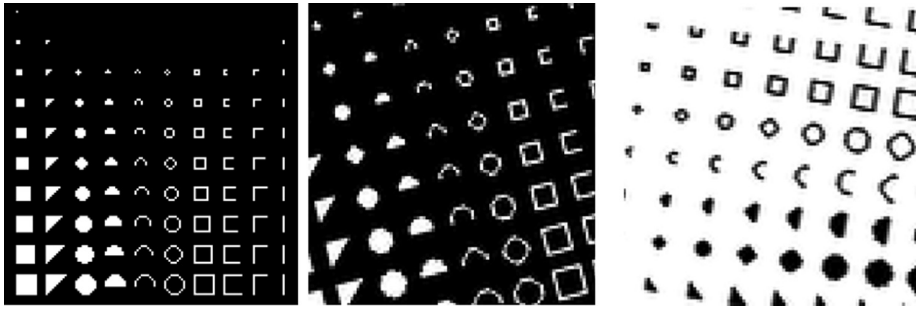


Figure 7-9. *Scaled and rotated examples of the synthetic interest point alphabet. Notice the artifacts introduced by the affine rotation, which distorts the synthetic binary patterns via anti-aliasing and sub-sampling artifacts*

The synthetic interest point alphabet contains 83 unique elements composed on a 14x14 grid, as shown in Figure 7-9. A total of seven rows and seven columns of the complete alphabet can fit inside a 1024x1024 image, yielding a total of 7x7x83=4067 total interest points.

Synthetic Corner Alphabet

The synthetic corner alphabet is shown in Figure 7-10. The alphabet contains the following types of corners and attributes:

- 2-SEGMENT CORNERS, 1,2,3 PIXELS WIDE
- 3-SEGMENT CORNERS, 1,2,3 PIXELS WIDE
- 4-SEGMENT CORNERS, 1,2,3 PIXELS WIDE

Figure 7-10. *Portion of the synthetic corner alphabet, features include 2-,3-, and 4-segment corners. (Top to bottom) White on black, black on white, light gray on dark gray, added salt and pepper noise, added Gaussian noise*

As shown in Figure 7-11, the corner alphabet contains patterns with multiple types of corners composed of two-line segments, three-line segments, and four-line segments, with pixel widths of 1,2, and 3. The synthetic corner alphabet contains 54 unique elements composed on a 14x14 pixel grid.

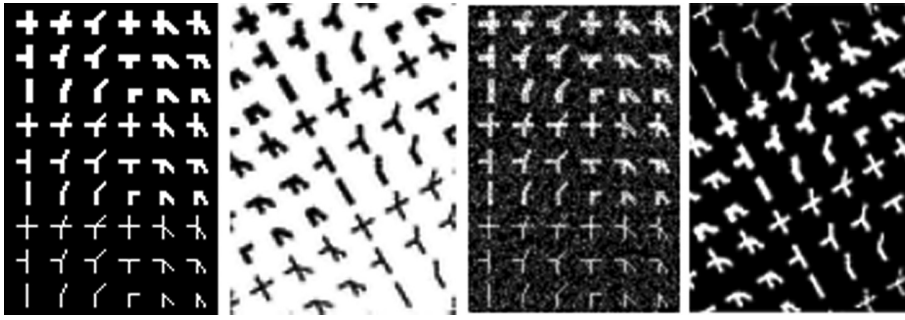


Figure 7-11. *Synthetic corner points image portions*

Each 1024x1024 pixel image contains 8x12 complete alphabets composed of 6x9 unique elements each, yielding 6x9x12x8=5184 total corner points per image. The full dataset includes rotated versions of each image from 0 to 90 degrees at 10 degree intervals.

Hybrid Synthetic Overlays on Real Images

We combine the synthetic interest points and corners as overlays with real images to develop a *hybrid ground truth dataset* as a more complex case.

The merging of synthetic interest points over real data will provide new challenges for the interest point algorithms and corner detectors, as well as illustrate how each detector works. Using hybrid synthetic feature overlays on real images is a new approach for ground truth data (as far as the author is aware), and the benefits are not obvious outside of curiosity. One reason the synthetic overlay approach was chosen here is to fill the gap in the literature and research, since synthetic features overlays are not normally used. See Figure 7-12.

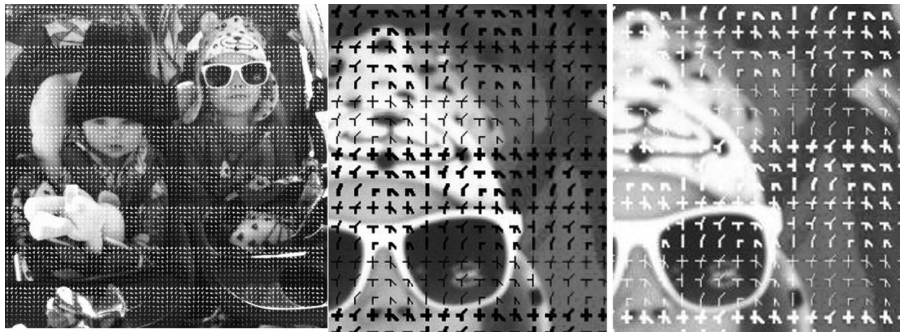


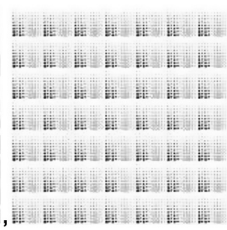
Figure 7-12. *Synthetic interest points combined with real images, used for stress testing interest point and corner detectors with unusual pixel patterns*

The hybrid synthetic and real ground truth datasets are designed with the following goals:

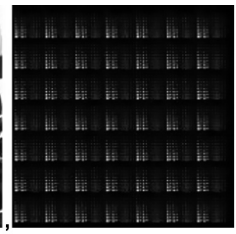
- Separate ground truth sets for interest points and corners, using the full synthetic alphabets overlaid on real images, to provide a range of pixel detail surrounding each interest point and corner.
- Display known positions and counts of interest points on a 14x14 grid.
- Provide color and gray scale images of the same data.
- Provide rotated versions of the same data 0 to 90 degrees at 10 degree intervals.

Method for Creating the Overlays

The alphabet can be used as a *binary mask* of 8-bit pixel values of black 0x00 and white 0xff for composing the image overlays. The following Boolean masking example is performed using Mathematica code ImageMultiply and ImageAd*d* operators.

ImageMultiply [ , ]

ImageMultiply is used to get the negatives, and then followed by ImageAdd to get the positives. Note that in other image processing tool systems, a Boolean ImageAND, ImageOR, and ImageNOT may be provided as alternatives.

ImageAdd [ , ]

Summary

We have surveyed manual and automated approaches to creating ground truth data, have identified some best practices and guidelines, have applied the robustness criteria and vision taxonomy developed in Chapter 5, and have worked through examples to create a ground truth dataset for evaluation of human perceptions compared to machine vision methods for keypoint detectors.

Here are some final thoughts and key questions for perparing ground truth data:

- **Appropriateness:** How appropriate is the ground truth dataset for the analysis and intended application? Are the use-cases and application goals built into the ground truth data and model? Is the dataset under-fitted or over-fitted to the algorithms and use-cases?
- **Public vs. proprietary:** Proprietary ground truth data is a barrier to independent evaluation of metrics and algorithms. It must be possible for interested parties to duplicate the metrics produced by various types of algorithms so they can be compared against the ground truth data. Open rating systems may be preferred, if they exist for the problem domain. But there are credibility and legal hurdles for open-sourcing any proprietary ground truth data.

- **Privacy and legal concerns:** There are privacy concerns for individuals in any images chosen to be used; images of people should not be used without their permission, and prohibitions against the taking of pictures at restricted locations should be observed. Legal concerns are very real.
- **Real data vs. synthetic data**: In some cases it is possible to use computer graphics and animations to create synthetic ground datasets. Synthetic datasets should be considered especially when privacy and legal concerns are involved, as well as be viewed as a way of gaining more control over the data itself.