# Image Based Quantitative Mosaic Evaluation with Artificial Video

Pekka Paalanen, Joni-Kristian Kämäräinen*, and Heikki Kälviäinen

Machine Vision and Pattern Recognition Research Group (MVPR)
*MVPR/Computational Vision Group, Kouvola
Lappeenranta University of Technology

**Abstract.** Interest towards image mosaicing has existed since the dawn of photography. Many automatic digital mosaicing methods have been developed, but unfortunately their evaluation has been only qualitative. Lack of generally approved measures and standard test data sets impedes comparison of the works by different research groups. For scientific evaluation, mosaic quality should be quantitatively measured, and standard protocols established. In this paper the authors propose a method for creating artificial video images with virtual camera parameters and properties for testing mosaicing performance. Important evaluation issues are addressed, especially mosaic coverage. The authors present a measuring method for evaluating mosaicing performance of different algorithms, and showcase it with the root-mean-squared error. Three artificial test videos are presented, ran through real-time mosaicing method as an example, and published in the Web to facilitate future performance comparisons.

## 1 Introduction

Many automatic digital mosaicing (stitching, panorama) methods have been developed [1,2,3,4,5], but unfortunately their evaluation has been only qualitative. There seems to exist some generally used image sets for mosaicing, for instance the "S. Zeno" (e.g. in [4]), but being real world data, they lack proper ground truth information for basis of objective evaluation, especially intensity and color ground truth. Evaluations have been mostly based on human judgment, while others use ad hoc computational measures such as image blurriness [4]. The ad hoc measures are usually tailored for specific image registration and blending algorithms, possibly giving meaningless results for other mosaicing methods and failing in many simple cases. On the other hand, comparison to any reference mosaic is misleading, if the reference method does not generate an ideal reference mosaic. The very definition of ideal mosaic is ill-posed in most real world scenarios. Ground truth information is crucial for evaluating mosaicing methods on an absolute level and an important research question remains how the ground truth can be formed.

In this paper we propose a method for creating artificial video images for testing mosaicing performance. The problem with real world data is that ground truth information is nearly impossible to gather at sufficient accuracy. Yet ground

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 470–479, 2009.

truth must be the foundation for quantitative analysis. Defining the ground truth ourselves and from it generating the video images (frames) allows to use whatever error measures required. Issues with mosaic coverage are addressed, what to do when a mosaic covers areas it should not cover and vice versa. Finally, we propose an evaluation method, or more precisely, a visualization method which can be used with different error metrics (e.g. root-mean-squared error).

The terminology is used as follows. Base image is the large high resolution image that is decided to be the ground truth. Video frames, small sub-images that represent (virtual) camera output, are generated from the base image. An intermediate step between the base image and the video frame is an optical image, which covers the area the camera sees at a time, and has a higher resolution than the base image. Sequence of video frames, or the video, is fed to a mosaicing algorithm producing a mosaic image. Depending on the camera scanning path (location and orientation of the visible area at each video frame), even the ideal mosaic would not cover the whole base image. The area of the base image, that would be covered by the ideal mosaic, is called the base area.

The main contributions of this work are 1) a method for generating artificial video sequences, as seen by a virtual camera with the most significant camera parameters implemented, and photometric and geometric ground truth, 2) a method for evaluating mosaicing performance (photometric error representation) and 3) publicly available video sequences and ground truth facilitating future comparisons for other research groups.

### 1.1 Related Work

The work by Boutellier et al. [6] is in essence very similar to ours. They also have the basic idea of creating artificial image sequences and then comparing generated mosaics to the base image. The generator applies perspective and radial geometric distortions, vignetting, changes in exposure, and motion blur. Apparently they assume that a camera mainly rotates when imaging different parts of a scene. Boutellier uses an interest point based registration and a warping method to align the mosaic to the base image for pixel-wise comparison. Due to additional registration steps this evaluation scheme will likely be too inaccurate for superresolution methods. It also presents mosaic quality as a single number, which cannot provide sufficient information.

Möller et al. [7] present a taxonomy of image differences and classify error types into registration errors and visual errors. Registration errors are due to incorrect geometric registration and visual errors appear because of vignetting, illumination and small moving objects in images. Based on pixel-wise intensity and gradient magnitude differences and edge preservation score, they have composed a voting scheme for assigning small image blocks labels depicting present error types. Another voting scheme then suggests what kind of errors an image pair as a whole has, including radial lens distortion and vignetting. Möller's evaluation method is aimed to evaluate mosaics as such, but ranking mosaicing algorithms by performance is more difficult.

Image fusion is basically very different from mosaicing. Image fusion combines images from different sensors to provide a sum of information in the images. One sensor can see something another cannot, and vice versa, the fused image should contain both modes of information. In mosaicing all images come from the same sensor and all images should provide the same information from a same physical target. It is still interesting to view the paper by Petrović and Xydeas [8]. They propose an objective image fusion performance metric. Based on gradient information they provide models for information conservation and loss, and artificial information (fusion artifacts) due to image fusion.

ISET vCamera [9] is a Matlab software that simulates imaging with a camera to utmost realism and processes spectral data. We did not use this software, because we could not find a direct way to image only a portion of a source image with rotation. Furthermore, the level of realism and spectral processing was mostly unnecessary in our case contributing only excessive computations.

## 2 Generating Video

The high resolution base image is considered as the ground truth, an exact representation of the world. All image discontinuities (pixel borders) belong to the exact representation, i.e. the pixel values are not just samples from the world in the middle of logical pixels but the whole finite pixel area is of that uniform color. This decision makes the base image solid, i.e., there are no gaps in the data and nothing to interpolate. It also means that the source image can be sampled using the nearest pixel method. For simplicity, the mosaic image plane is assumed to be parallel to the base image. To avoid registering the future mosaic to the base image, the pose of the first frame in a video is fixed and provides the coordinate reference. This aligns the mosaic and the base image at sub-pixel accuracy and allows to evaluate also superresolution methods.

The base image is sampled to create an optical image, that spans a virtual sensor array exactly. Resolution of the optical image is $k_{\text{interp}}$ times the base image resolution, and it must be considerably higher than the array resolution. Note, that resolution here means the number of pixels per physical length unit, not the image size. The optical image is formed by accounting the virtual camera location and orientation. The area of view is determined by a magnification factor $k_{\text{magn}}$ and the sensor array size $w_{\text{s}}, h_{\text{s}}$ such that the optical image in terms of base image pixels is of the size $\frac{w_{\text{s}}}{k_{\text{magn}}}, \frac{h_{\text{s}}}{k_{\text{magn}}}$. All pixels are square.

The optical image is integrated to form the sensor output image. Figure 1(a) presents the structure coordinate system of the virtual sensor array element. A "light sensitive" area inside each logical pixel is defined by its location $(x, y) \in ([0, 1], [0, 1])$ and size $w, h$ such that $x + w \leq 1$ and $y + h \leq 1$. The pixel fill ratio, as related to true camera sensor arrays, is $wh$. The value of a pixel in the output image is calculated by averaging the optical image over the light sensitive area. Most color cameras currently use a Bayer mask to reproduce the three color values R, G and B. The Bayer-mask is a per-pixel color mask which transmits only one of the color components. This is simulated by discarding the other two color components for each pixel.

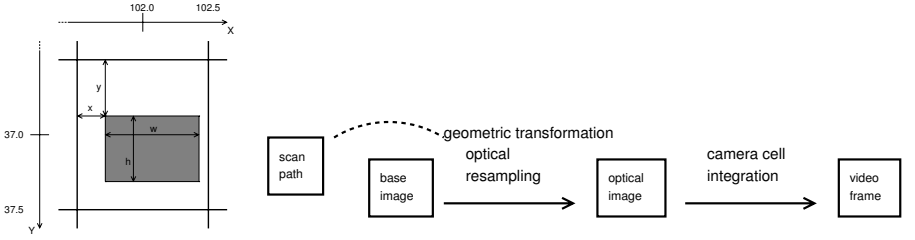


**Fig. 1.** (a) The structure of a logical pixel in the artificial sensor array. Each logical pixel contains a rectangular "light sensitive" area (the gray box) which determines the value of the pixel. (b) Flow of the artificial video frame generation from a base image and a scan path.

**Table 1.** Parameters and features used in the video generator

| | |
|---|---|
| **Base image.** | The selected ground truth image. Its contents are critical for automatic mosaicing and photometric error scores. |
| **Scan path.** | The locations and orientations of the snapshots from a base image. Determines motion velocities, accelerations, mosaic coverage and video length. Video frames must not cross base image borders. |
| **Optical magnification,** $k_{\text{magn}} = 0.5$**.** | Pixel size relationship between base image and video frames. Must be less than one when evaluating superresolution. |
| **Optical interpolation factor,** $k_{\text{interp}} = 5$**.** | Additional resolution multiplier for producing more accurate projections of the base image, defines the resolution of the optical image. |
| **Camera cell array size,** $400 \times 300$ $pix$**.** | Affects directly the visible area per frame in the base image. The video frame size. |
| **Camera cell structure,** $x = 0.1, y = 0.1, w = 0.8, h = 0.8$**.** | The size and position of the rectangular light sensitive area inside each camera pixel (Figure 1(a)). In reality this approximation is also related to the point spread function (PSF), as we do not handle PSF explicitly. |
| **Camera color filter.** | Either 3CCD (every color channel for each pixel) or Bayer mask. We use 3CCD model. |
| **Video frame color depth.** | The same as we use for the base image: 8 bits per color channel per pixel. |
| **Interpolation method in image trans.** | Due to the definition of the base image we can use nearest pixel interpolation in forming the optical image. |
| **Photometric error measure.** | A pixel-wise error measure scaled to the range $[0, 1]$. Two options: i) root-mean-squared error in RGB space, and ii) root-mean-squared error in L*u*v* space assuming the pixels are in sRGB color space. |
| **Spatial resolution of photometric error.** | The finer resolution of the base image and the mosaic resolutions. |

An artificial video is composed of output images defined by a scan path. The scan path can be manually created by a user plotting ground truth locations with orientation on the base image. For narrow baseline videos cubic interpolation is used to create a denser path. A diagram of the artificial video generation is presented in Figure 1(b).

Instead of describing the artificial video generator in detail we list the parameters which are included in our implementation and summarize their values and meaning in Table 1. The most important parameters we use are the base

image itself and the scan path. Other variables can be fixed to sensible defaults as proposed in the table. Other unimplemented, but still noteworthy, parameters are noise in image acquisition (e.g. in [10]) and photometric and geometric distortions.

## 3 Evaluating Mosaicing Error

Next we formulate a mosaic image quality representation or visualization, referenced to as *coverage–cumulative error score graph*, for comparing mosaicing methods. First we justify the use of solely photometric information in the representation and second we introduce the importance of coverage information.

### 3.1 Geometric vs. Photometric Error

Mosaicing, in principle, is based on two rather separate processing steps: registration of video frames, in which the spatial relations between frames is estimated, and blending the frames into a mosaic image, that is deriving mosaic pixel values from the frame pixel values. Since the blending requires accurate registration of frames, especially in superresolution methods, it sounds reasonable to measure the registration accuracy or the geometric error. However, in the following we describe why measuring the success of a blending result (photometric error) is the correct approach.

Geometric error occurs, and typically also cumulates, due to image registration inaccuracy or failure. The geometric error can be considered as errors in geometric transformation parameters, assuming that the transformation model is sufficient. In the simplest case this is the error in frame pose in reference coordinates. Geometric error is the error in pixel (measurement) location.

Two distinct sources for photometric error exist. The first is due to geometric error, e.g., points detected to overlap are not the same point in reality. The second is due to the imaging process itself. Measurements from the same point are likely to differ because of noise, changing illumination, exposure or other imaging parameters, vignetting, and spatially varying response characteristics of the camera. Photometric error is the error in pixel (measurement) value.

Usually a reasonable assumption is that geometric and photometric errors correlate. This is true for natural, diverse scenes, and constant imaging process. It is easy, however, to show pathological cases, where the correlation does not hold. For example, if all frames (and the world) are of uniform color, the photometric error can be zero, but geometric error can be arbitrarily high. On the other hand, if geometric error is zero, the photometric error can be arbitrary by radically changing the imaging parameters. Moreover, even if the geometric error is zero and photometric information in frames is correct, non-ideal blending process may introduce errors. This is the case especially in superresolution methods (the same world location is swiped several times) and the error certainly belongs to the category of photometric error.

From the practical point of view, common for all mosaicing systems is that they take a set of images as input and the mosaic is the output. Without any further insight into a mosaicing system only the output is measurable and, therefore, a general evaluation framework should be based on photometric error. Geometric error cannot be computed if it is not available. For this reason we concentrate on photometric error, which allows to take any mosaicing system as a black box (including proprietary commercial systems).

### 3.2 Quality Computation and Representation

Seemingly straightforward measure is to compute the mean squared error (MSE) between a base image and a corresponding aligned mosaic. However, in many cases the mosaic and the base image are in different resolutions, having different pixel sizes. The mosaic may not cover all of the base area of the base image, and it may cover areas outside the base area. For these reasons it is not trivial to define as what should be computed for MSE. Furthermore, MSE as such does not really tell the "quality" of a mosaic image. If the average pixel-wise error is constant, MSE is unaffected by coverage. The sum of squared error (SSE) suffers from similar problems.

Interpretation of the base image is simple compared to the mosaic. The base image, and also the base area, is defined as a two-dimensional function with complete support. The pixels in a base image are not just point samples but really cover the whole pixel area. How should the mosaic image be interpreted; as point samples, full pixels, or maybe even with a point spread function (PSF)? Using a PSF would imply that the mosaic image is taken with a virtual camera having the PSF. What should the PSF be? Point sample covers an infinitely small area, which is not realistic. Interpreting the mosaic image the same way as the base image seems the only feasible solution, and is justified by the graphical interpretation of an image pixel (a solid rectangle).

Combing the information about SSE and coverage in a graph can better visualize the quality differences between mosaic images. We borrow from the idea of Receiver Operating Characteristic curve and propose to draw the SSE as a function of coverage. SSE here is the smallest possible SSE when selecting $n$ determined pixels from the mosaic image. This makes all graphs monotonically increasing and thus easily comparable. Define $N$ as the number of mosaic image pixels required to cover exactly the base area. Then coverage $a = n/N$. Note that $n$ must be integer to correspond to binary decision on each mosaic pixel whether to include that pixel. Section 4 contains many graphs as examples.

How to account for differences in resolution, i.e., pixel size? Both the base image and the mosaic have been defined as functions having complete support and composing of rectangular or preferably square constant value areas. For error computation each mosaic pixel is always considered as a whole. The error value for the pixel is the squared error integrated over the pixel area. Whether the resolution of the base image is coarser or finer does not make a difference.

How to deal with undetermined or excessive pixels? Undetermined pixels are areas the mosaic should have covered according to the base area but are not

determined. Excessive pixels are pixels in the mosaic covering areas outside the base area. Undetermined pixels do not contribute to the mosaic coverage or error score. If a mosaicing method leaves undetermined pixels, the error curve does not reach 100% coverage. Excessive pixels contribute the theoretical maximum error to the error score, but the effect on coverage is zero. This is justified by the fact that in this case the mosaicing method is giving measurements from an area that is not measured, creating false information.

## 4 Example Cases

As example methods two different mosaicing algorithms are used. The first one, referenced to as the ground truth mosaic, is a mosaic constructed based on the ground truth geometric transformations (no estimated registration), using nearest pixel interpolation in blending video frames into a mosaic one by one. There is also an option to use linear interpolation for resampling. The second mosaicing algorithm is our real-time mosaicing system that estimates geometric transformations from video images using point trackers and random sample consensus, and uses OpenGL for real-time blending of frames into a mosaic. Neither of these algorithms uses a superresolution approach.

Three artificial videos have been created, each from a different base image. The base images are in Figure 2. The bunker image ($2048 \times 3072$ px) contains a natural random texture. The device image ($2430 \times 1936$ px) is a photograph with strong edges and smooth surfaces. The face image ($3797 \times 2762$ px) is scanned from a print at such resolution that the print raster is almost visible and produces interference patterns when further subsampled (we have experienced this situation with our real-time mosaicing system's imaging hardware). As noted in Table 1, $k_{\text{magn}} = 0.5$ so the resulting ground truth mosaic is in half the resolution, and is scaled up by repeating pixel rows and columns. The real-time mosaicing system uses a scale factor 2 in blending to compensate.

Figure 3 contains coverage–cumulative error score curves of four mosaics created from the same video of the bunker image. In Figure 3(a) it is clear that the real-time methods getting larger error and slightly less coverage are inferior to the ground truth mosaics. The real-time method with sub-pixel accuracy point

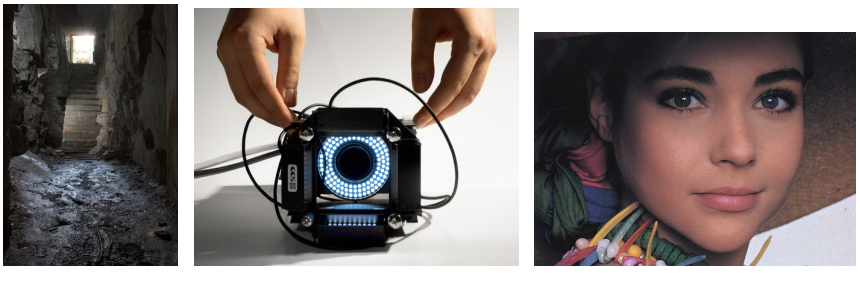

**Fig. 2.** The base images. (a) Bunker. (b) Device. (c) Face.

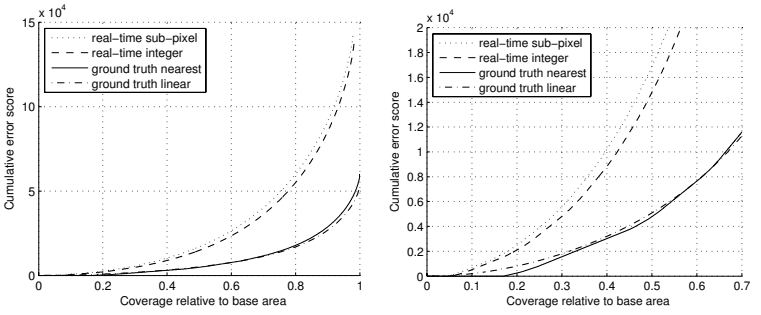


**Fig. 3.** Quality curves for the Bunker mosaics. (a) Full curves. (b) Zoomed in curves.

**Table 2.** Coverage–cumulative error score curve end values for the bunker video

| mosaicing method | max coverage | error at max coverage | total error |
|---|---|---|---|
| real-time sub-pixel | 0.980 | 143282 | 143282 |
| real-time integer | 0.982 | 137119 | 137119 |
| ground truth nearest | 1.000 | 58113 | 60141 |
| ground truth linear | 0.997 | 50941 | 50941 |

tracking is noticeably worse than integer accuracy point tracking, suggesting that the sub-pixel estimates are erroneous. The ground truth mosaic with linear interpolation of frames in blending phase seems to be a little better than using nearest pixel method. However, when looking at the magnified graph in Figure 3(b) the case is not so simple anymore. The nearest pixel method gets some pixel values more correct than linear interpolation, which appears to always make some error. But, when more and more pixels of the mosaics are considered, the nearest pixel method starts to accumulate error faster. If there would be way to select the 50% of the most correct pixels of a mosaic, then in this case the nearest pixel method would be better. A single image quality number, or even coverage and quality together, cannot express this situation. Table 2 shows the maximum coverage values and cumulative error scores without (at max coverage) and with (total) excessive pixels.

To more clearly demonstrate the effect of coverage and excessive pixels, an artificial case is shown in Figure 4. Here the video from the device image is processed with the real-time mosaicing system (integer version). An additional mosaic scale factor was set to 0.85, 1.0 and 1.1. Figure 4(b) presents the resulting graphs along with the ground truth mosaic. When the mosaic scale is too small by factor 0.85, the curve reaches only 0.708 coverage and due to a particular scan path there are no excessive pixels. Too large scale by factor 1.1 introduces a great amount of excessive pixels, which are seen in the coverage–cumulative error score curve as a vertical spike at the end.

The face video is the most controversial because it should have been low-pass filtered to smooth interferences. The non-zero pixel fill ratio in creating the video

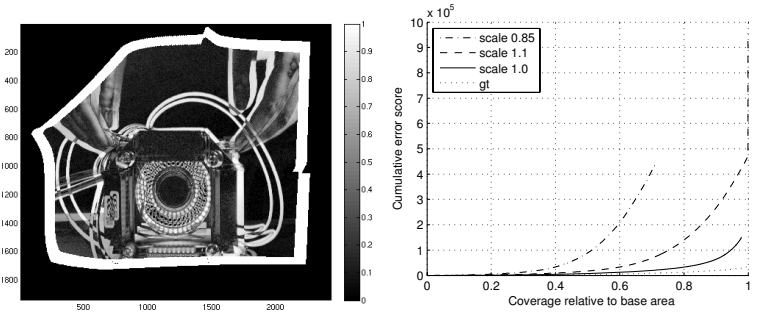


**Fig. 4.** Effect of mosaic coverage. (a) error image with mosaic scale 1.1. (b) Quality curves for different scales in the real-time mosaicing, and the ground truth mosaic `gt`.

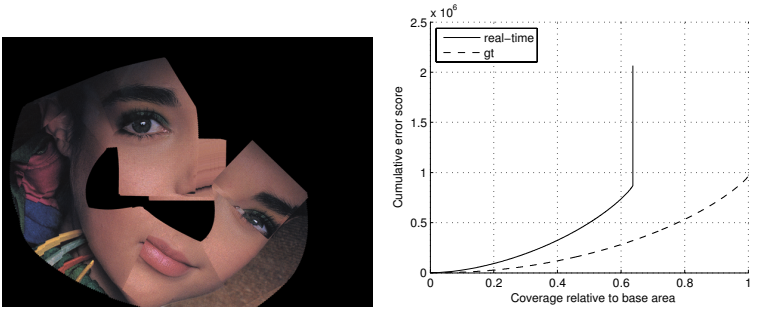


**Fig. 5.** The real-time mosaicing fails. (a) Produced mosaic image. (b) Quality curves for the real-time mosaicing, and the ground truth mosaic `gt`.

removed the worst interference patterns. This is still a usable example, for the real-time mosaicing system fails to properly track the motion. This results in excessive and undetermined pixels as seen in Figure 5, where the curve does not reach full coverage and exhibits the spike at the end. The relatively high error score of ground truth mosaic compared to the failed mosaic is explained by the difficult nature of the source image.

## 5 Discussion

In this paper we have proposed the idea of creating artificial videos from a high resolution ground truth image (base image). The idea of artificial video is not new, but combined with our novel way of representing the errors between a base image and a mosaic image it unfolds new views into comparing the performance of different mosaicing methods. Instead of inspecting the registration errors we consider the photometric or intensity and color value error. Using well-chosen base images the photometric error cannot be small if registration accuracy is lacking. Photometric error also takes into account the effect of blending video frames into a mosaic, giving a full view of the final product quality.

The novel representation is the coverage–cumulative error score graph, which connects the area covered by a mosaic to the photometric error. It must be noted, that the graphs are only comparable when they are based on the same artificial video. To demonstrate the graph, we used a real-time mosaicing method and a ground truth transformations based mosaicing method to create different mosaics. The pixel-wise error metric for computing photometric error was selected to be the simplest possible: length of the normalized error vector in RGB color space. This is likely not the best metric and for instance Structural Similarity Index [11] could be considered.

The base images and artificial videos used in this paper are available at `http://www.it.lut.fi/project/rtmosaic` along with additional related images. Ground truth transformations are provided as Matlab data files and text files.

## References

1. Brown, M., Lowe, D.: Recognizing panoramas. In: ICCV, vol. 2 (2003)
2. Heikkilä, M., Pietikäinen, M.: An image mosaicing module for wide-area surveillance. In: ACM international workshop on Video Surveillance & Sensor Networks (2005)
3. Jia, J., Tang, C.K.: Image registration with global and local luminance alignment. In: ICCV, vol. 1, pp. 156–163 (2003)
4. Marzotto, R., Fusiello, A., Murino, V.: High resolution video mosaicing with global alignment. In: CVPR, vol. 1, pp. I–692–I–698 (2004)
5. Tian, G., Gledhill, D., Taylor, D.: Comprehensive interest points based imaging mosaic. Pattern Recognition Letters 24(9–10), 1171–1179 (2003)
6. Boutellier, J., Silvén, O., Korhonen, L., Tico, M.: Evaluating stitching quality. In: VISAPP (March 2007)
7. Möller, B., Garcia, R., Posch, S.: Towards objective quality assessment of image registration results. In: VISAPP (March 2007)
8. Petrović, V., Xydeas, C.: Objective image fusion performance characterisation. In: ICCV, vol. 2, pp. 1866–1871 (2005)
9. ISET vcamera, `http://www.imageval.com/public/Products/ISET/ISET_vCamera/vCamera_main.htm`
10. Ortiz, A., Oliver, G.: Radiometric calibration of CCD sensors: Dark current and fixed pattern noise estimation. In: ICRA, vol. 5, pp. 4730–4735 (2004)
11. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. Image Processing 13(4), 600–612 (2004)