PERFORMANCE EVALUATION OF SIMILARITY MEASURES FOR DENSE MULTI-MODAL STEREO-VISION

Mustafa Yaman*, Sinan Kalkan

Dept. of Computer Engineering
Middle East Technical University
Ankara, Turkey
Email: {mustafa.yaman, skalkan}@ceng.metu.edu.tr

Abstract

Multi-modal imaging systems have recently been drawing attention in fields such as medical imaging, remote sensing and video surveillance systems. In such systems, estimating depth has become possible due to promising progress of multi-modal matching techniques. In this article, we perform a systematic performance evaluation of similarity measures frequently used in the literature for dense multi-modal stereo-vision. The evaluated measures include Mutual Information (MI), Sum of Squared Distances (SSD), Normalized Cross Correlation (NCC), Census Transform (CENSUS), Local Self Similarity (LSS) as well as descriptors adopted to multi-modal settings like SIFT, SURF, HOG, BRIEF and FREAK. We evaluate the measures over datasets we generated and compare the performances using the "Winner Takes All" (WTA) method. The generated datasets are (i) synthetically-modified four popular pairs from the Middlebury Stereo Dataset (namely, tsukuba, venus, cones and teddy), and (ii) our own multi-modal image pairs acquired using the infrared (IR) and the Electro-optical(EO) cameras of a Kinect device. The results show that MI and HOG provides promising results for multi-modal imagery and FREAK, SURF, SIFT and LSS can be considered as alternatives depending on the multi-modality level and the computational complexity requirements of the intended application.

Keywords: dense multi-modal stereo-vision, similarity measures, kinect device

1. INTRODUCTION

Imaging systems of different modalities have been in use for a long time, especially in medical imaging [1, 2, 3, 4, 5, 6, 7] and remote sensing [8, 9, 10, 11, 12, 13]. In such systems, registration and/or fusion of such imagery is a major concern since information from multiple modalities need to be combined for solving a task. However, this is challenging since objects and surfaces look very different in different modalities.

Recently, using multi-modal cameras for video surveillance systems has been growing in popularity [14, 15, 16, 17, 18]. Such multi-modal systems can combine information from multiple sources provided by different types of sensors in order to get a more accurate and robust interpretation of an environment [16]. The type of sensors include audio, thermal, infrared, vibration sensors etc. Especially, using bi-modal setups including visible and infrared/thermal cameras has become quite prevalent in such systems since such combinations enable surveillance during not only daytime but also night-time, under low visibility or lighting conditions [19, 20, 21, 22]. Figure 1 shows sample images taken from such bi-modal systems.

*Corresponding author

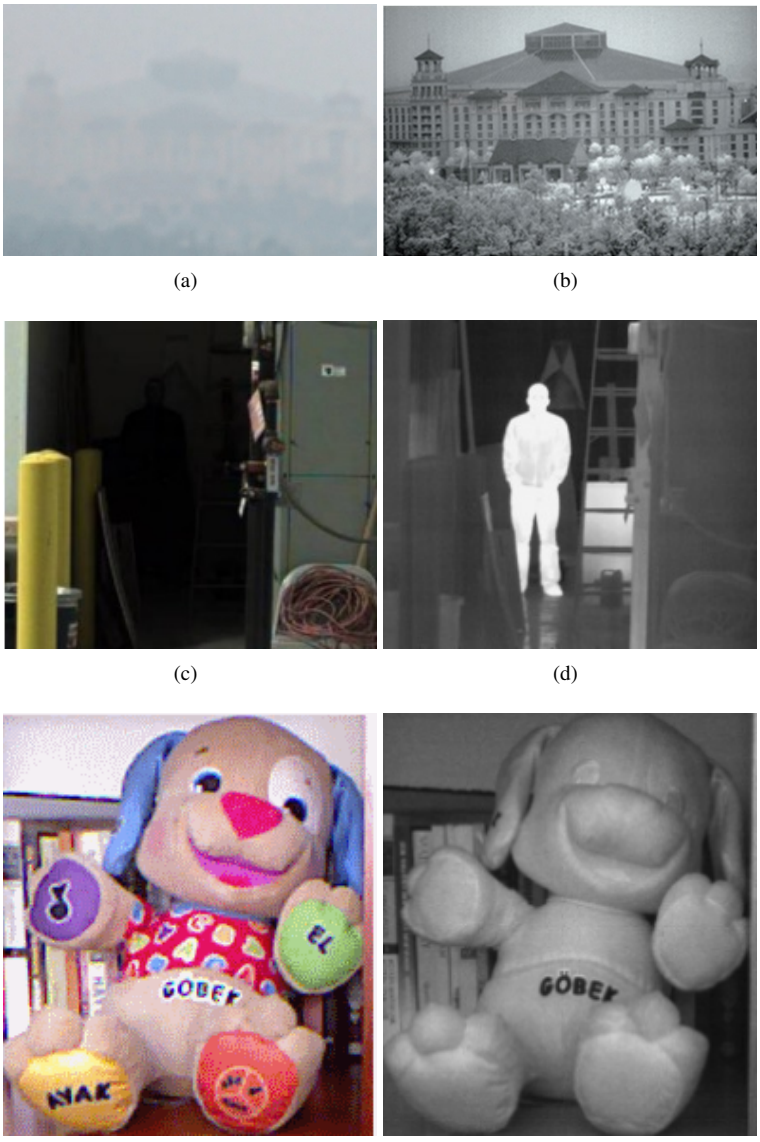


Figure 1: Sample visible-infrared image couples from multi-modal imaging systems for surveillance applications. (a,b) Visible-SWIR image couple (Source: [23]). (c,d) Visible-LWIR image couple (Source: [24]). (e,f) Visible-NIR image couple (Source: [25]).

Depth or the distance in surveillance systems is a very important source of information for the operator to better interpret a scene or a behavior. Currently, this is achieved by an additional active system like a Laser Range Finder (LRF) [19, 20, 21, 22] or photogrammetrically by making an assumption of average height of the target pedestrian, vehicle or even the vehicle tires detected on the sensor [15]. Alternatively, unimodal stereo-vision can be used provided that there is an additional camera of the same modality.

It has been recently shown that a good alternative for acquiring depth in such surveillance systems is stereo-vision from multi-modal camera pairs by several studies in the literature *e.g.* [18, 26, 27, 28, 29] and our previous studies [25, 30]. The classical stereo-vision techniques are not applicable in such a multi-modal setting owing to the differences in corresponding pixels' intensities, making the problem challenging (see Figure 1). However, there have been many successful promising solutions to the problem - see Section 2.

1.1. The Current Study

In this study, we provide a comprehensive analysis of different similarity measures used for dense multi-modal stereo-vision in the literature. Our main contributions are the following:

- A comprehensive analysis: In this study, a systematic performance evaluation of alternative similarity measures frequently used in the literature for dense multi-modal stereo-vision is performed. To the best of our knowledge, there is no such comprehensive study on multi-modal stereo-vision in the literature. On two datasets, the measures are especially compared under different multi-modality and noise settings.

- Multi-modal stereo-vision datasets: MY: For the performance evaluation, two image datasets are compiled and provided online [31] and demonstrated for use as benchmark to similar studies via this study. Since, up to the author's knowledge, there are no dense multi-modal stereo-vision datasets available in the literature with accurate ground truth information, which is considered as another important contribution of the study.

  MY: The first dataset was generated by authors in [25] from four popular images in the Middlebury Stereo Vision Page [32] (tsukuba, venus, cones and teddy) [33, 34]. The left images in this dataset are synthetically altered. The second dataset was generated by our several Kinect camera shootings composed of 24 stereo-rectified image pairs. The Kinect devices [35] have a built-in Infrared (IR) (left) and Electro-optical (EO) (right) camera along with an IR projector which has a built-in depth computation feature (see Section 4 for details).

2. Uni-modal and Multi-modal Stereo-vision

In this section, we first provide a summary of unimodal stereo-vision techniques, and then review the existing multi-modal stereo-vision studies and the similarity measures available in literature.

2.1. Unimodal Stereo-vision

Stereo-vision [36, 37] deals with computing depth by finding the corresponding pixels in different views. The correspondences, which are generally determined by comparing intensities of pixels, are used for computing the 3D positions using simple triangulation.

Stereo-vision is one of the most studied problems of computer vision - for reviews, see, *e.g.*, [38, 33]. Stereo-vision methods are mainly clustered around two main axes: Sparse or feature-based vs. dense methods; and, local vs. global methods. In dense methods (*e.g.*, [33, 39, 40]), stereo correspondences (and therefore the pixel disparities) are computed for all the pixels in the images. On the other hand, sparse methods (*e.g.*, [41, 42]) compute disparity information only for some reliable features extracted from images, such as salient points, edges, corners, curves etc.

Regarding the second grouping; local methods use only the local neighborhood information for finding stereo correspondences. Dense methods performing a window-based matching over pixels in this neighbourhood are grouped as local methods (*e.g.*, [43, 44]). Sparse or feature-based methods can also be local (*e.g.*, [38, 42, 45, 39]). Global methods, on the other hand, use global constraints to correct wrong correspondences that are otherwise not possible locally (*e.g.*, [46, 47, 48, 49]). Although computational complexity is much higher, significant improvements in the accuracy of disparity maps can be achieved.

An important landmark for stereo-vision is the Middlebury benchmark by Scharstein and Szeliski [33] for comparing the current state of the art solutions to the problem. The dataset along with the performance evaluation of many algorithms are provided by a web site [32]. With the Middlebury dataset, which includes scenes of different complexities with ground truth disparities, methods can be compared on different performance criteria. We believe that a similar dataset and an evaluation framework for multi-modal stereo-vision is needed for similar benefits.

2.2. Multi-modal Stereo-vision

Although classical stereo-vision techniques have had tremendous success in terms of both accuracy and running time, they are not directly applicable in a multi-modal setting. The reason is that computing similarities between intensities of pixels or windows will not work using unimodal matching methods simply because the intensities of

the corresponding pixels will be different. For example, an RGB-thermal image pair would have totally different intensities for corresponding pixels (see, *e.g.*, Figure 1).

Multi-modal stereo-vision is defined as performing stereo-vision using two camera pairs of different modalities, for instance an infrared-visible camera pair which has not been studied much until 2000s. The earliest of such studies, to the best of our knowledge, is from Egnal [43], who, affected from Viola's studies of multi-modal registration [50], applied mutual information (MI) as the basic similarity measure for stereo correspondence. Egnal tested his method both on unimodal images and also on red / blue filtered, multimodal (an NIR and Visible/NIR image couple) and differently lighted images and compared to a modified NCC algorithm as a baseline. The results were promising in that they demonstrated MI's power compared to standard methods, based on correlation, especially on images with spectrally distinct characteristics for the same scene. However, using MI still produced low quality results compared to what multi-modal stereo vision techniques can produce.

Fookes *et al.* enhanced the approach based on MI with adaptive windowing [51] and added prior probabilities to the method using a 2D matching surface [52]. A 2D matching surface is simply formed by computing MI costs for every possible combinations of left and right pixels. From the 2D matching surface, first, the maximum of a row is found. The maximum is compared to all the costs on the same column; if it is also the maximum of the column, then it is determined as a valid match. This is claimed to enforce left-right and uniqueness constraints. Prior probability incorporation is performed by computing a joint histogram from all the intensities in the stereo image pair and using this as prior probabilities added to the joint probability of matching windows along with a weighting constant. The results were taken from unimodal images altered synthetically by negative, solarized, posterized or simulated by $(cos(\pi f(I)/255)255)$ versions of one of the image pairs, where

$$f(I) = \begin{cases} 0 & \text{if } I < 0 \\ I & \text{otherwise,} \end{cases} \tag{1}$$

and compared to results from NCC and rank transform where MI outperforms all these methods.

This study is important for showing that stereo-vision results using MI could be significantly enhanced when combined with other state-of-the-art stereo-vision techniques. However, the results were taken only from unimodal images that were altered synthetically, which do not provide distinct characters of segments or edges that multi-modal images may have.

Another notable work on multi-modal stereo vision is by Krotosky and Trivedi [18, 26, 53], who studied pedestrian detection and tracking using mutual information via an infrared-visible camera pair. They established stereo correspondence using mutual information within region of interests (ROI) comprised by human body silhouettes, and proposed a method using disparity voting enabling computation of depth information for the corresponding regions which is a significant restriction. Finally, this depth information is used to accurately register the multi-modal images for the ROIs.

An extension of how MI can be used for stereo-vision was proposed by Barrera *et al.* [54], who integrated the gradient information and also a scale-space analysis into MI calculation and performed an evaluation of the proposed similarity functions. In their study, a multi-modal stereo rig (with thermal and visible cameras) was developed and images taken from this setup were used. However, the evaluation dataset did not contain accurate ground truth data and they used image scenes comprised of planar surfaces to perform the evaluation in the *v-disparity space*. They conclude that MI is a powerful method for multi-modal stereo vision and adding gradient and the scale space analysis increases the accuracy of MI. However, their evaluation method was limited to only the planar surfaces fitted to calculated disparities due to unavailable ground truth data and does not cover scenes with objects of several shapes, sizes, curvature and obscuring objects. In their next study, [27], the IGSS method, which is defined as MI integrated with gradient and scale space analysis is further analysed. The results they presented in this work show that generated 3D depth data are quite sparse considering the tested scenes. Nonetheless, their studies are promising since they show that stereo-vision is possible from images with very distinct spectral bands from thermal and visible range of the electromagnetic spectrum.

Recently, a measure, called Local Self Similarity (LSS), originally proposed for image template matching [55], has been applied as a thermal-visible stereo correspondence measure by Torabi and Bilodeau [28]. In their study, the focus was on ROI-based image matching and they tried tracking people in the scene using their silhouette. They compared LSS against MI-based similarity descriptors. In their first publication [56], they showed that LSS outperformed MI

and HOG (Histogram of Oriented Gradients) measures. Later, they used LSS measure in an energy minimization framework, enhancing the results compared to their previous work [57]. However, in 2014, Bilodeau *et al.* [29] showed for the problem of thermal-visible registration of segmented human silhouettes that MI performs better than LSS and other alternative descriptors. Note that, unlike our article, the focus of the analysis by Bilodeau *et al.* [29] was on thermal-visible registration of human silhouettes. In fact, in our article, we show that, due to the nature of the problem, different performance characteristics are observed in multi-modal stereo-vision. Namely, we show that MI is still better than LSS unlike the results provided by Bilodeau *et al.* for even smaller window sizes. Moreover, HOG also outperforms LSS in multi-modal stereo-vision although more vulnerable to noise.

3. Similarity Measures for Multi-Modal Stereo-vision

In this section, the similarity measures used as multi-modal stereo correspondence measures in the literature are described in detail. The performances of these similarity measures are evaluated in the article.

3.1. Sum of Square Distances - SSD

SSD is a basic similarity measure used in stereo-vision [38, 33, 58]. It is simply composed of computing the sum of squares of intensity differences of the pixels in the two candidate windows, W_L and W_R. Namely:

$$SSD(W_L, W_R, d) = \sum_{x,y \in W_L; x',y \in W_R;} \left(I_l(x,y) - I_r(x',y)\right)^2, \tag{2}$$

where W_L is the local window around a center pixel (x_c, y_c) in the left image L; W_R is a candidate window from the same row y_c in the right image for disparities $d \in [0, d_{max}]$ and $x' = x - d$ for all $x \in W_L$.

3.2. Normalized Cross Correlation (NCC)

NCC is a classical similarity measure widely used in stereo vision [38, 33, 58]. In NCC, the pixel-wise cross-correlation of the two matching windows are computed and normalized by the overall intensity difference. NCC similarity measure is defined as follows:

$$NCC(W_L, W_R, d) = \frac{\sum_{x,y \in W_L; x',y \in W_R} (I_l(x,y))(I_r(x',y))}{\sqrt{\sum_{x,y \in W_L; x',y \in W_R} (I_l(x,y))^2 (I_r(x',y))^2}}, \tag{3}$$

where W_L is a local window round a center pixel (x_c, y_c) in the left image L; W_R is a candidate window from the same row y_c in the right image for disparities $d \in [0, d_{max}]$ and $x' = x - d$ for all $x \in W_L$. NCC has been tested as the basic similarity measure in several multi-modal stereo-vision studies [29, 43, 56].

3.3. Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) was proposed by Lowe [59] for object detection. The method generates sets of descriptive features from images that are claimed to be invariant to rotation, scaling and partial changes in viewpoint and illumination. The method has drawn significant attention since 2004 and become very popular for detection, recognition and retrieval problems.

The method is composed of four major steps [59]; *Scale-space extrema detection*, *Accurate Keypoint Localization*, *Assignment of Orientation* and *Generating the Local Descriptor*. The method generates local descriptors at the salient points detected by the process that starts by computing image gradients and orientations around the local region of each keypoint. Next, the gradients and orientations are weighted by a Gaussian function with the σ of the Gaussian as the half of the descriptor window determining the 2×2 or 4×4 subregions around the keypoint. The gradients and orientations are accumulated within these descriptor windows where the orientations are binned to 8 directions within each subregion using the gradient values.

In this article, SIFT descriptors are generated for each pixel while evaluating stereo correspondence.

3.4. Speeded-Up Robust Features (SURF)

The Speeded-Up Robust Features (SURF) was proposed by Bay *et al.* [60] as a faster alternative to SIFT with comparable performance. The method uses the Hessian for localization. Scale space analysis is performed by up-sampling the box filters instead of down-sampling the image. Haar-wavelet responses are computed in x and y directions in a circular neighborhood of the interest points for achieving rotational invariance.

In SURF, a descriptor vector is generated by constructing a square region around the interest point oriented in the dominant orientation. This square region is split into 4×4 subregions of 5×5 pixels. The wavelet responses along x and y directions (d_x, d_y) are summed up over each subregion. The absolute values of the responses ($|d_x|$ and $|d_y|$) are also summed. Therefore, the feature vector is comprised of $v = (\sum d_x, \sum |d_x|, \sum d_y, \sum |d_y|)$ for each subregion which yields in total a $4 \times 4 \times 4 = 64$-sized descriptor vector. The descriptor vector is normalized for invariance to contrast changes.

Similar to SIFT, in this article, SURF descriptors are generated for each pixel for evaluating stereo correspondence.

3.5. Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) is a descriptor proposed by Dalal and Triggs [61] for human detection. It was first used by Torabi and Bilodeau [56] in a multi-modal setting where they compared LSS against HOG for human ROI detection in thermal-visible stereo image pairs.

The method first divides the image window at the center pixel q into a grid of *cells*. Each cell accumulates the local 1-D histogram of gradient directions. For robustness to illumination effects, *blocks* composed of several of these cells accumulate the local histograms and they are used for normalization of local cells. The normalized descriptor blocks are called Histogram of Oriented Gradients (HOG). The HOG descriptors of each pixel in a local detection window are then be combined to have a feature vector. Formally, computation of a HOG descriptor can be described as follows:

$$HOG_q(k) = \sum_{(x,y) \in W_q} T\left(\frac{\Theta(x,y)}{\gamma}\right), \tag{4}$$

where $HOG_q(k)$ corresponds to k^{th} bin in the histogram with K bins, $\Theta(x,y)$ is the gradient at pixel (x,y), γ is a scaling constant and $T()$ is defined as:

$$T(u) = \begin{cases} 1 & \text{if } u = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

3.6. Local Self-Similarity (LSS)

Local self similarity (LSS) is a similarity measure proposed initially by Shechtman and Irani [55] for image template matching. The method simply extracts a small patch (*e.g.*, 5×5) from the center pixel q of a larger window (*e.g.*, 40×40) and the sum of squared distances (SSD) between the small patch and the surrounding larger region is computed. Next, the SSD costs are normalized by maximum variance of difference of small image patches ($\sigma_{auto}(q)$)) and a noise term (σ_{noise}) generating a *correlation surface* S_q as follows:

$$S_q(x,y) = \exp\left(-\frac{SSD_q(x,y)}{\max(\sigma_{noise}, \sigma_{auto}(q))}\right). \tag{6}$$

Finally, the LSS descriptor is then the partitioned log-polar representation of this surface $S_q()$ using 20 angles and 4 radial intervals, giving 80 bins.

Torabi and Bilodeau introduced the LSS measure for ROI-based image matching for human tracking and computing the depth of the human ROI in thermal-visible image pairs [28, 57]. In their recent studies on image registration, they compared this measure with other similarity measures like MI, HOG, Census, SIFT, SURF, BRIEF, FREAK etc. [56, 57] where LSS was successful for the smallest window size that was tested but MI was still outperforming LSS for the other larger window sizes.

3.7. Binary Robust Independent Elementary Features (BRIEF)

BRIEF is a descriptor proposed by Calonder *et al.* [62] based on encoding visual information as binary strings over an image patch. The method is composed of three steps:

- A sampling grid of points in a defined pattern is placed around the region of the pixel of interest.
- A list of pairs of points from the sampling grid is constructed.
- A binary string is encoded from the intensities of the sampling pairs. The intensities are first smoothed using Gaussian kernels in order to be robust against noise.

The binary string is encoded using the T function as in Eqn. 7, for an image patch p of size $S\,xS$:

$$T(p; x, y) = \begin{cases} 1 & \text{if } p(x) < p(y) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

For constructing the sampling grid and computing the sampling pairs (x, y), a number of methods were proposed in [62]:

- **(G I)**: Points are evenly distributed, pairs are randomly selected.
- **(G II)**: Points are sampled using a Gaussian distribution and pairs are randomly selected from this distribution of points, which means points near the center are preferred.
- **(G III)**: The first location x is sampled from a Gaussian centered around the origin, the other point is sampled from another Gaussian centered around the x, which creates more local pairs.
- **(G IV)**: A coarse polar grid is used and pairs are randomly selected from this grid.
- **(G V)**: A coarse polar grid is used and pairs are selected as $x = (0, 0)$ at the origin and y is randomly selected.

The *Hamming* distance [63] is used to compare the encoded binary strings which is defined as:

$$Hamming(s_1, s_2) = \sum_{i=0}^{N} T_h(s_1(i), s_2(i)), \tag{8}$$

where T_h function is defined as:

$$T_h(c_1, c_2) = \begin{cases} 1 & \text{if } c_1 \neq c_2, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where s_1 and s_2 are the two binary strings to be compared, $s(i)$ corresponds to the i^{th} bit in the string s.

BRIEF was used as one of the methods to compare against alternative measures in Bilodeau *et al.*, [29], for human ROI detection in thermal-visible stereo image pairs.

3.8. Fast Retina Keypoint (FREAK)

FREAK is yet another descriptor proposed recently by Alahi *et al.* [64]. Similar to BRIEF, FREAK is a binary descriptor. Like BRIEF, the binary strings are generated over a sampling grid, but the difference of FREAK is that the sampling grid is inspired from the human visual system and the spatial arrangement of receptor cells in the retina.

Using a sampling grid similar to receptive fields, a binary descriptor is constructed by first pairing the receptive fields and then thresholding the difference between the receptive fields, as below, as a sequence of one-bit Difference of Gaussians (DoG):

$$FREAK = \sum_{0 \leq a < N} 2^a T(P_a), \tag{10}$$

where function T is defined as:

$$T(P_a) = \begin{cases} 1 & \text{if } I(P_a^{r_1}) - I(P_a^{r_2}) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

and P_a is the pair of receptive fields, N is the size of the descriptor binary string.

In order to select the pairings, a learning phase is performed over the data which yields automatically a coarse-to-fine ordering of DoGs similar to human vision system. Later at the recognition step, matching is performed at several stages: initially first 16 bits are compared, after locating the candidate matchings a more detailed search is performed over the rest of the string. This mechanism is also claimed to mimic how the human visual system performs recognition [64].

The *Hamming* distance [63] is used to compare the encoded binary strings as given in Eqn. 8. FREAK was used as one of the methods to compare against alternative measures in Bilodeau *et al.* [29] for human ROI detection in thermal-visible stereo image pairs.

3.9. Mutual Information (MI)

Mutual information (MI), proposed by Shannon [65], is a measure for the information content of received data over a random variable x having the probability distribution $p(x)$.

In his study, Shannon defined the mutual information using the concepts of entropy ($H(x)$), joint entropy ($H(x, y)$) and conditional entropy ($H(x|y)$) as:

$$MI(x, y) = H(x) - H(y|x) = H(y) - H(x|y) = H(x) + H(y) - H(x, y). \tag{12}$$

Mutual Information (MI) is defined as the subtraction between the information needed to describe x alone and the additional information required to specify y given x, which corresponds to the reduction in the information content of x given that y is known.

An alternative definition of MI uses Kullback-Leibler distance measure [66], which measures the additional information required for defining x using another approximating distribution $q(x)$ instead of $p(x)$ which is assumed to be unknown:

$$KL(p||q) = -\sum_x p(x) \log\left(\frac{q(x)}{p(x)}\right). \tag{13}$$

Egnal [43] is the first to use mutual information for multi-modal stereo correspondence. Later, many others followed using mutual information for multi-modal stereo-vision [27, 51, 52] and related problems such as human ROI tracking on multi-modal stereo image pairs [18, 26, 29, 53].

3.10. Mutual Information with Prior Probabilities (MI(wPR))

Incorporating prior probabilities to MI calculation was proposed by Fookes *et al.* [52] for multi-modal image matching and stereo correspondence. The aim is to increase the statistical discriminability of joint probability calculation of the two local matching windows. In order to accomplish this, joint prior probabilities computed from the whole images are added to the local joint probability calculation.

MI can be calculated from the two local windows W_L and W_R extracted from the left image L and right image R as:

$$MI(W_L, W_R) = \sum_{I_l \in W_L} \sum_{I_r \in W_R} P(I_l, I_r) \log \frac{P(I_l, I_r)}{P(I_l)P(I_r)}, \tag{14}$$

where $P(I_l, I_r)$ corresponds to joint probability for the left and right image patches W_L and W_R, $P(I_l)$ and $P(I_r)$ are the marginal probabilities of the pixel intensities that can be computed by constructing a joint histogram.

In MI(wPR), the MI formulation is changed slightly as follows:

$$MI_{(wPR)}(W_L, W_R) = \sum_{I_l \in W_L} \sum_{I_r \in W_R} P(I_l, I_r) \log \frac{P^*(I_l, I_r)}{P(I_l)P(I_r)}, \tag{15}$$

where $P^*(I_l, I_r)$ can be defined as:

$$P^*(I_l, I_r) = \lambda P(I_l, I_r) + (1 - \lambda) P_{prior}(I_l, I_r), \tag{16}$$

where $P_{prior}(I_l, I_r)$ is the joint prior probability computed from the joint histogram of the whole images as:

$$P_{prior}(I_l, I_r) = \frac{hist(I_l, I_r)}{\sum_{l',r'} hist(I_{l'}, I_{r'})}, \tag{17}$$

for all corresponding pixels I_l in left image L and I_r in right image R. λ in Equation 16 corresponds to the degree of incorporating priors into the joint probability. This modification to MI calculation was shown to increase the performance of MI as a similarity measure in stereo correspondence problem in [52].

3.11. Census Transform (CENSUS)

Census Transform, proposed by Zabih and Woodfill [67], is a non-parametric local transform. The method simply uses relative ordering of pixel intensities in the image patch over the center pixel and transforms it into a binary encoded string as the descriptor. Formally, for a window W_u at a pixel u, the descriptor is determined as follows:

$$C(W_u) = \otimes T(W_u; u, v), \tag{18}$$

where v a neighboring pixel in W_u, $\otimes$ is the concatenation operation and the $T()$ is defined as:

$$T(W_u; u, v) = \begin{cases} 0 & \text{if } W_u(v) < W_u(u) \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

In other words, the neighbor pixels are checked if their value is greater than the center pixel, which leads to a 0 in the binary string and 1 otherwise if smaller. The descriptor is then the composition of these binary values.

Since the feature vector is binary, the *Hamming* distance [63] is used to compare two feature vectors as given in Eqn. 8, like other binary encoded string based similarity measures *i.e.*, FREAK and BRIEF.

4. DATASETS AND PERFORMANCE EVALUATION

Two types of datasets are collected and used in this study: (1) A dataset with synthetically altered stereo image pairs from the Middlebury Stereo Evaluation Dataset [68], and (2) visible and infrared image pairs captured from a Kinect device [35]. The datasets are available publicly at [31].

4.1. Dataset #1 - The Middlebury Dataset

This dataset contains the four *popular* image pairs (Tsukuba, Venus, Cones and Teddy) from the Middlebury Stereo Evaluation Dataset [68]. The left images in the dataset are altered synthetically by using a cosine transform ($(\cos(\pi f(I)/255)255)$ where $f(I)$ is defined by Equation 1) of pixel intensities just as Fookes *et al.* did [52]. Table 1 provides the list of the image pairs that comprises the dataset along with several properties of the stereo images. Figure 2 presents the image pairs generated and used in the experiments.

Note that, in the left images, important details are lost and some segments are merged due to the cosine transform and the truncation to 8-bit unsigned integer of the result. This alteration enabled us to simulate the challenge of matching IR and EO images. On the other hand, the negative, solarized and posterized versions of the images could be used, however, these transformations was only changing the intensities not the edges and segments of the images.

This dataset is important since (i) it is possible to evaluate the performance of the methods thanks to the ground truth and (ii) it allows us to see the state of the multi-modal stereo-vision methods compared to that of uni-modal stereo-vision methods.

In the experiments, the "all" regions provided by the Middlebury page [68] is used by clipping the specified regions for which matching cannot be performed. In addition, regarding the window-based methods, half of the used window sizes at the borders are also discarded when computing performance statistics for a fair comparison between methods.

Table 1: The Dataset #1 - Synthetically Altered Middlebury Stereo Evaluation Dataset.

Dataset	Image No	Image Name	Resolution	Max. Disparity
Dataset #1	1	Tsukuba	384×288	15
Dataset #1	2	Venus	434×383	19
Dataset #1	3	Teddy	450×375	59
Dataset #1	4	Cones	450×375	59

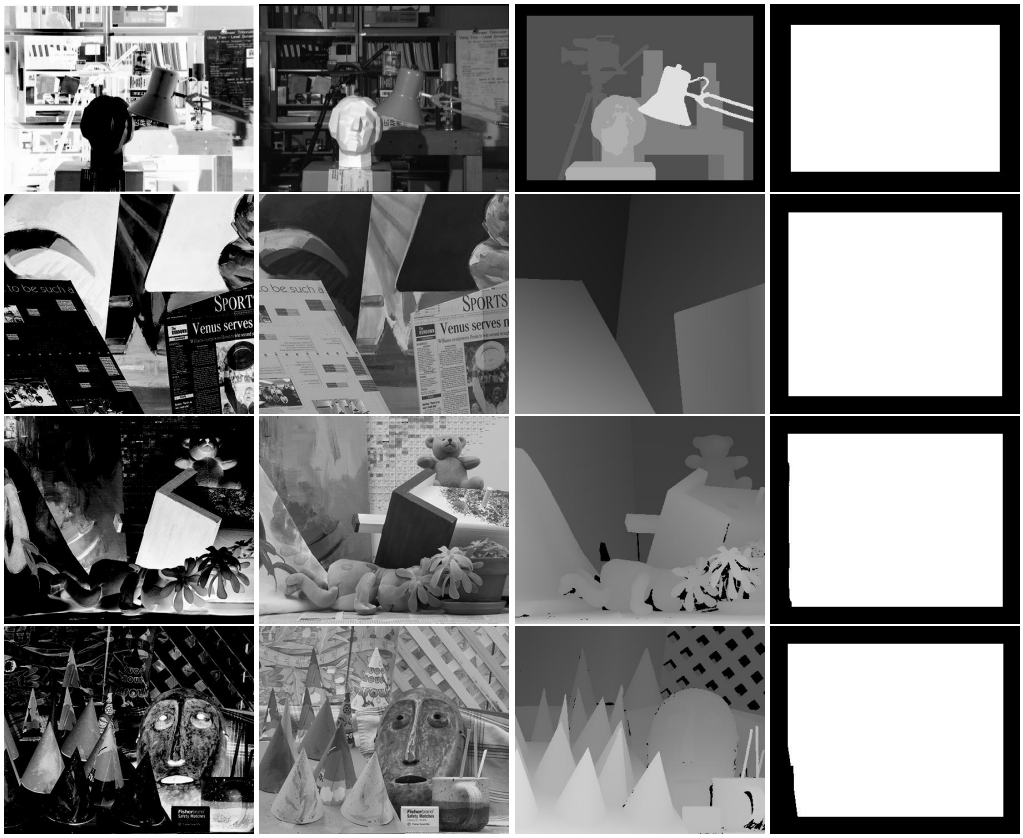


Figure 2: Tsukuba, Venus, Teddy and Cones stereo pairs from the Middlebury Stereo Vision Page - Evaluation Version 2 [68]. *1st column*: Synthetically altered left images. *2nd column*: The right images *(*(grayscale)). *3rd column*: The ground truth disparities. *4th column*: The "all" regions where evaluations are performed, (Only "white" pixels are included in performance evaluation.) Note that, in the left images, important details are lost due to the cosine transformation.

See Figure 2 (rightmost column) for the "all" regions in accordance with the Middlebury page [68], where the white pixels show the region that the performance evaluations are performed.

Performance on Dataset #1 is evaluated using two types of metrics: (i) the Root Mean Square (RMS) distances between estimated and ground-truth disparities, and (ii) the Percentage of Bad Pixels (BAD), which is the percentage of pixels for which the estimation error is greater than a threshold δ_d. As suggested in Middlebury Stereo Vision Evaluation Page, δ_d is set to 1.5. These metrics can be formally defined as follows [33]:

$$RMS = \sqrt{\frac{1}{N}\sum_{(x,y)} |d_C(x,y) - d_T(x,y)|^2}, \tag{20}$$

$$BAD = \frac{1}{N} \sum_{(x,y)} (|d_C(x,y) - d_T(x,y)| > \delta_d)\,, \tag{21}$$

where $d_C(x,y)$ is the computed disparity map, $d_T(x,y)$ is the ground truth disparity map, δ_d is the error threshold (= 1.5), N is the number of pixels being evaluated.

4.2. Dataset #2 - The Kinect Dataset

The Kinect dataset contains infrared (left) and visible (right) images captured from a Kinect device. The Kinect Device was introduced by Microsoft for the Xbox 360 game console [69] which enabled the user use his/her own body as the game controller. As shown in Figure 3, the device has a built-in EO camera and an infrared camera and projector couple. The infrared projector sends beams to the scene and the beams are sensed on the infrared camera which enables the device to generate a 3D depth map of the scene where the intention is the human body.

Figure 3: The Kinect Device having a built-in camera, sensors and features.

To be able to use the built-in infrared and visible camera images for multi-modal stereo-vision, the images need to be rectified so that the epipolar constraint holds. We use the method by Zhang *et al.*, [70], for calibration of the cameras [1].

See Figure 4 for an overview of the process, with sample chessboard images taken by the Kinect infrared and visible cameras, the detected chessboard grid points for the computation of the rectification parameters and the achieved stereo rectification results.

Using the Kinect device and the process described above, a stereo vision evaluation dataset composed of infrared and visible image pairs is constructed. For this purpose, several scenes of indoor environments (office, living room, shelves) with several objects having different reflectance properties are prepared and recorded by the infrared and visible cameras of the device. The images are stereo-rectified. See [31] for the dataset.

For comparing the performances of the similarity measures, all 24 image pairs in this dataset are used in the scope of this study. See Figure 5 for sample four image pairs from the dataset.

Performance evaluation on the Kinect dataset is performed by using depth data that Kinect provides. The accuracy of native Kinect depth data was already evaluated by several studies [71, 72]. Disparity information is computed from the depth data by inversing the disparity to depth computation over epipolar-rectified images [70] as:

$$[x \quad y \quad d_k(x,y) \quad w]^T = Q^{-1} \times [X \quad Y \quad Z_k(X,Y) \quad 1.0]^T \tag{22}$$

where (x,y) represents the column and row of one pixel, d_k is the disparity value at the corresponding pixel along with the scaling constant w. The Kinect depth value at (X,Y) coordinate is at Z_K distance from the camera. The Q matrix

[1] For both calibration and rectification, we use the *calib3d* module in OpenCV.

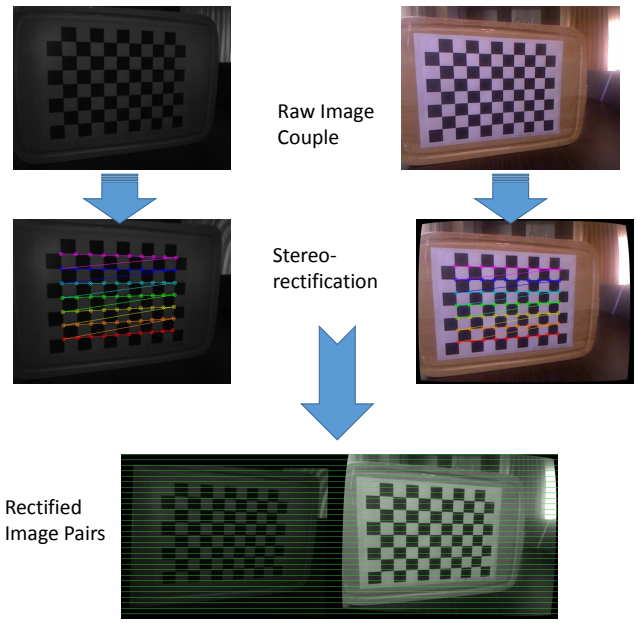


Figure 4: Depiction of the Kinect calibration process

Figure 5: Sample image pairs from Dataset #2 - Kinect Dataset *1st column*: Left (IR) camera images. *2nd column*: Right (EO) camera images. *3rd column*: Kinect's native depth maps (brighter pixels have more depth). *4th column*: Disparity maps

is the perspective transformation matrix from disparity to depth which is constructed as:

$$Q = \begin{bmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/b & (c_x - c'_x)/b \end{bmatrix}, \tag{23}$$

by using the calibration parameters, f the focal length, b the stereo baseline, (c_x, c_y) and (c'_x, c_y) are the principal points.

Performance evaluation on Dataset #2 is also performed using the same two types of metrics, RMS and BAD as defined for Dataset #1. The computation is performed only on the pixels that have a valid depth ($\forall(x, y)\ s.t.\ Z_K(x, y) \in (0.0, 5.0]$) since Kinect native depth estimation may fail on edges and non-fronto-paralel surfaces due to insufficient reflectance of infrared beams on such surfaces (See Figure 5).

5. COMPARISON OF SIMILARITY MEASURES FOR MULTI-MODAL STEREO-VISION

In this section, the descriptors and the similarity measures that are widely used in the literature (see Section 3) are compared using the datasets that were generated in the scope of this study (see Section 4).

The evaluated methods can be grouped into three categories similar to [29]:

1. **Local Window-Based Measures:** These methods calculate the measures by using local windows extracted around the compared pixels from the left and the right images. The SSD, NCC and MI (with and without Prior Probability) measures fall into this category - see Section 5 for details.
2. **Measures based on non-binary feature descriptors:** These methods are composed of initially calculating feature vectors densely for each pixel. The LSS, HOG, SIFT and SURF fall into this category. To compute the stereo correspondences, a similarity measure using the sum of distances of each corresponding feature vector of the pixels within the local windows around the matching left and right image pixels are used:

$$SM(W_L, W_R) = \sqrt{\sum_{x,y} (f_L(x, y) - f_R(x, y))^2}, \tag{24}$$

   where f_L and f_R are the feature vectors of each pixel in the two matching windows W_L and W_R.
3. **Measures based on binary features:** These methods are based on binary descriptors for each pixel. CENSUS,BRIEF, FREAK features are used under this category. *Hamming* distance of the matching windows are applied to the binary descriptors as the similarity measure (see Eqn. 8).

5.1. Performance Evaluation Using Dataset #1 - The Synth. Alt. Middlebury Dataset

In this section, the performance evaluation of the similarity measures are performed using the Dataset #1 - The Synthetically Altered Middlebury Dataset. After computing the similarity measures for the matching pixels, the "WTA" (Winner-Takes-All) disparities are computed by selecting the best disparity having the maximum similarity value over candidate disparities. The performance evaluations are performed as described in Section 4.

Three different experiments are conducted in the scope of this section, as provided in the following subsections: (i) **Effect of window size:** First, the measures are tested using three different window sizes, $9\times9, 21\times21$ and 31×31. This range enabled us to evaluate the performances in small, medium and bigger window sizes which might affect the results for comparison. (ii) **Effect of multi-modality:** Next, the measures are tested for different multi-modality levels of the left image. (iii) **Effect of noise:** Finally, several levels of Gaussian noise are added to the left image and the measures are tested for increasing noise levels.

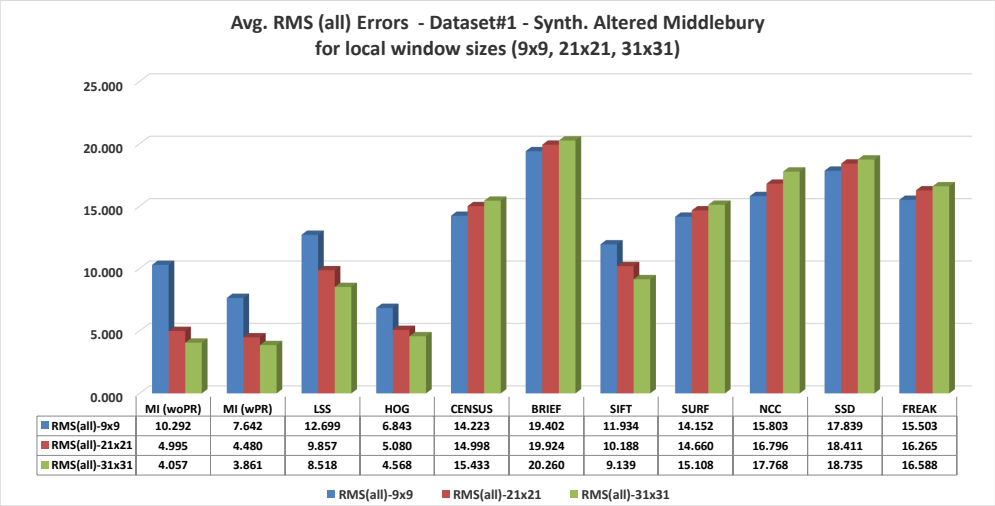
Table 2 provides the parameter settings used in the experiments for each of the method tested where the default parameters are used.

Table 2: Parameter setting used for the evaluated similarity measures

Method Name	Parameter Name	Value
MI(woPR)	$binsize(hist)$	40
MI(wPR)	λ	0.3
LSS	$size(smallpatch)$	5
	$size(largepatch)$	41
	$num(angles)$	20
HOG	K	9
	γ	1
	$size(detectionwindow)$	16
	$size(cells)$	8
SIFT	σ	3
	$num(octaves)$	4
	$num(octaveintervals)$	3
	$size(descriptor)$	128
SURF	σ	3.3
	$num(octaves)$	4
	$num(octaveintervals)$	2
	$size(descriptor)$	64
CENSUS	$size(window)$	3
	$size(descriptorstrbits)$	8
BRIEF	$size(descriptorbits)$	32
	$size(patch)$	48
	$size(kernel)$	9
	$gridtype$	GIV
FREAK	$num(octaves)$	4
	$num(scales)$	64
	$num(pairs)$	512
	$num(oripairs)$	45

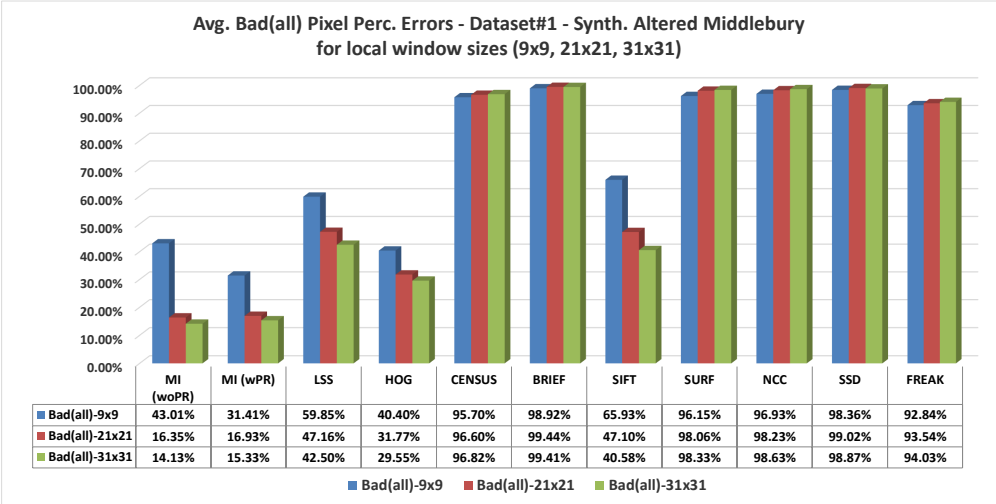
5.1.1. Effect of Window Size

In this part, we evaluate the effect of the window size on the performance of the methods. We use three different window sizes, 9×9, 21×21 and 31×31 (corresponding to a small, medium and a large window, respectively), for the similarity measure computations to perform the evaluation. The results are provided in Figures 6(a) and 6(b) which show the average RMS and BAD metrics for the "all" regions respectively. See Figure 7 for sample visual results on the Tsukuba image for the leading similarity measures.



	MI (woPR)	MI (wPR)	LSS	HOG	CENSUS	BRIEF	SIFT	SURF	NCC	SSD	FREAK
RMS(all)-9x9	10.292	7.642	12.699	6.843	14.223	19.402	11.934	14.152	15.803	17.839	15.503
RMS(all)-21x21	4.995	4.480	9.857	5.080	14.998	19.924	10.188	14.660	16.796	18.411	16.265
RMS(all)-31x31	4.057	3.861	8.518	4.568	15.433	20.260	9.139	15.108	17.768	18.735	16.588

(a)



	MI (woPR)	MI (wPR)	LSS	HOG	CENSUS	BRIEF	SIFT	SURF	NCC	SSD	FREAK
Bad(all)-9x9	43.01%	31.41%	59.85%	40.40%	95.70%	98.92%	65.93%	96.15%	96.93%	98.36%	92.84%
Bad(all)-21x21	16.35%	16.93%	47.16%	31.77%	96.60%	99.44%	47.10%	98.06%	98.23%	99.02%	93.54%
Bad(all)-31x31	14.13%	15.33%	42.50%	29.55%	96.82%	99.41%	40.58%	98.33%	98.63%	98.87%	94.03%

(b)

Figure 6: Average RMS **(a)** and BAD **(b)** errors of the methods. The "WTA" performances in "all" regions for three different window sizes for Dataset #1 are considered.

It can be observed from the results that the average RMS and BAD pixel percentage errors are smallest for the MI(wPR) and MI(woPR) similarity measures, which shows that incorporation of the prior probabilities enhanced the results of MI measure. HOG is ranked right after the MI measures, which indicates that using gradient information in a multi-modal setting is a good alternative. LSS and SIFT are following these measures. On the other hand, the similarity measures SURF, CENSUS, BRIEF, FREAK, NCC and SSD totally perform worst in this dataset since they fail to capture similarity in the multi-modal intensities.

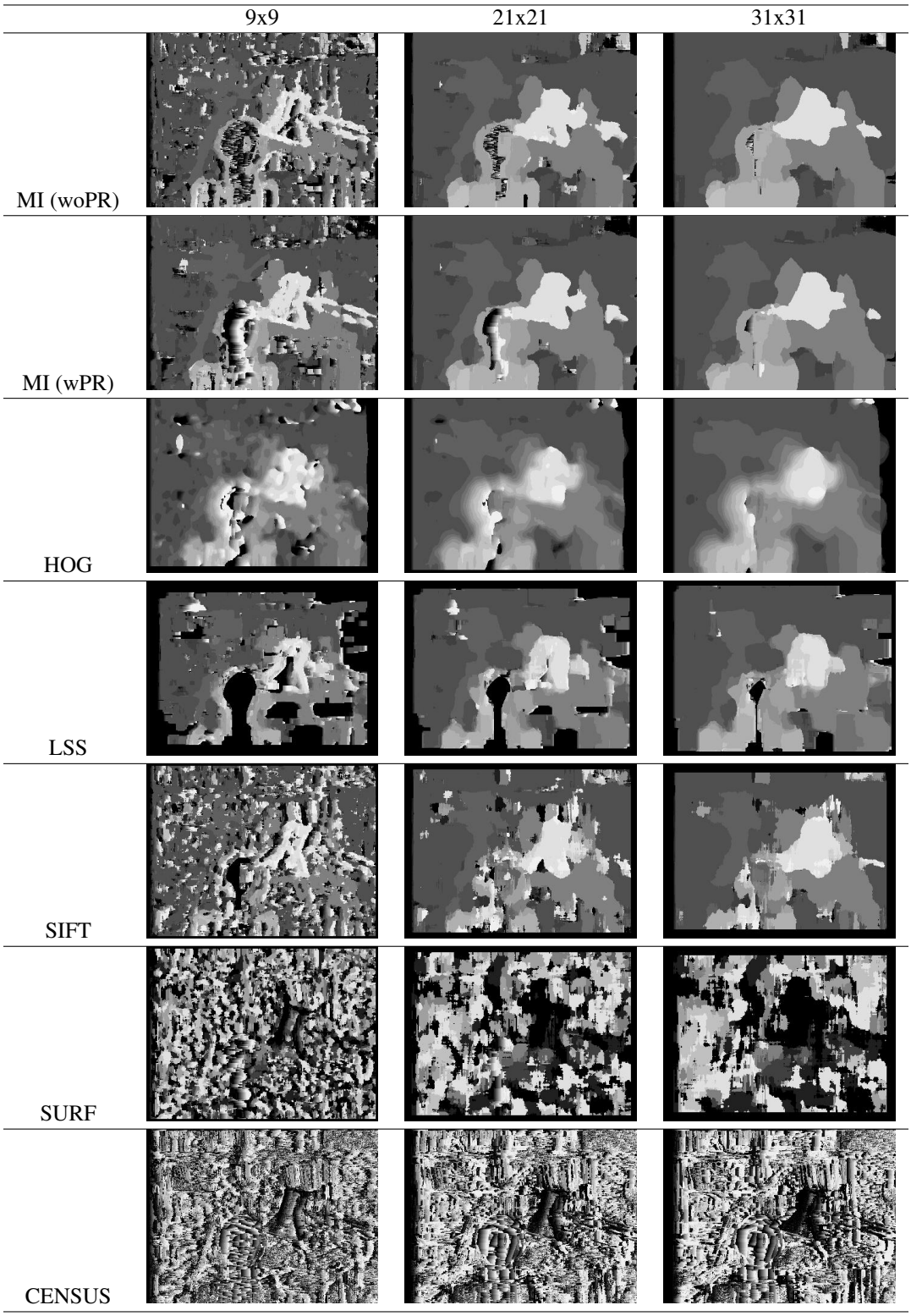


Figure 7: Sample visual results of the leading similarity measures for the synthetically altered Tsukuba image pair in Dataset #1, for the different window sizes 9x9, 21x21 and 31x31 pixels.

The results are analysed for the effect of local window sizes used for computing the similarity measures. It is observed that, as the size of the window increases, the performances of the similarity measures also increase except for the ones that totally fail (i.e. SURF, CENSUS, BRIEF, FREAK, NCC and SSD). However, for the measures that do not fail and can represent multi-modal image patches to some extent (i.e. MI, HOG, LSS, SIFT), the rate of increase in the performance degrades at each enlarged window size. On the other hand, especially for teddy and cones images, having bigger and more curved objects, the methods are affected more by the local window size (results not provided here). HOG is affected less than all other measures (excluding the failed measures) by the window size where it provides best result in average for the smallest window size 9×9. This is because of the similarity measure computing method in which an inner window is used to compute each feature vector for each pixel in the compared local window yielding a greater window in total. MI measures are affected significantly for the window size, as all the computation is performed within the extracted window of pixels.

5.1.2. Effect of Multi-Modality

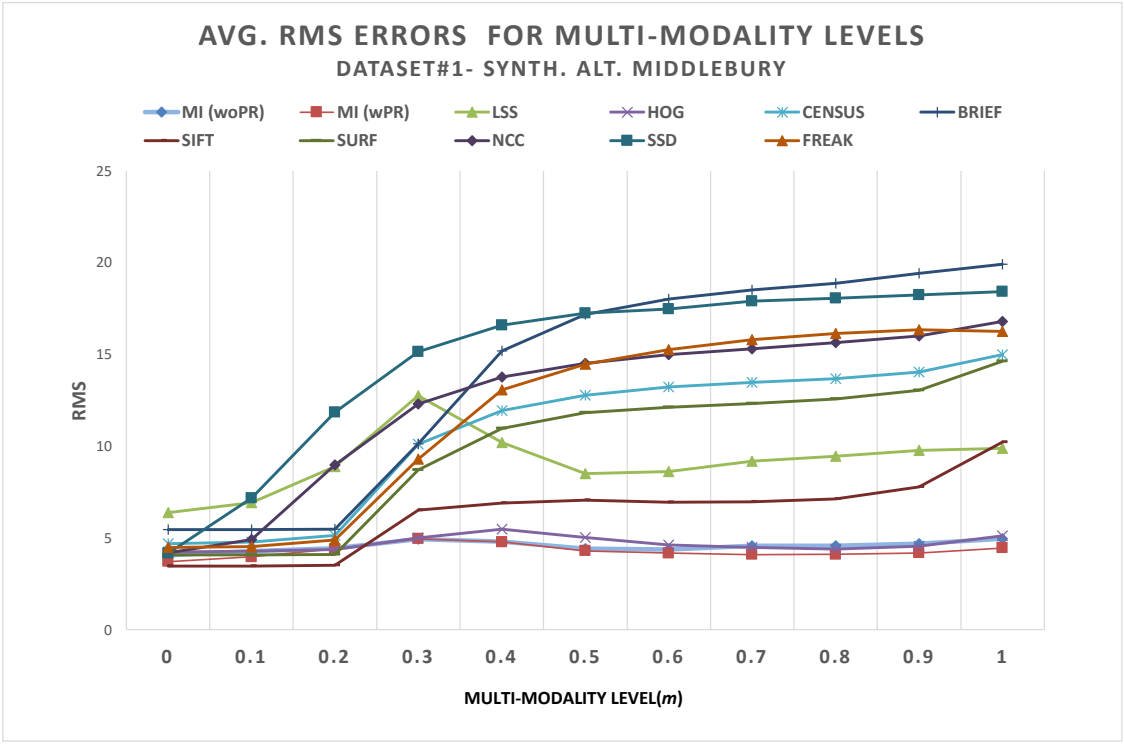
In this part, we evaluate the methods for their resilience against multi-modality. To accomplish this task, we generate stereo pairs of different multi-modalities as follows:

$$I_m(x, y) = (1 - m) \times I_{orig}(x, y) + m \times I_{cos}(x, y), \tag{25}$$

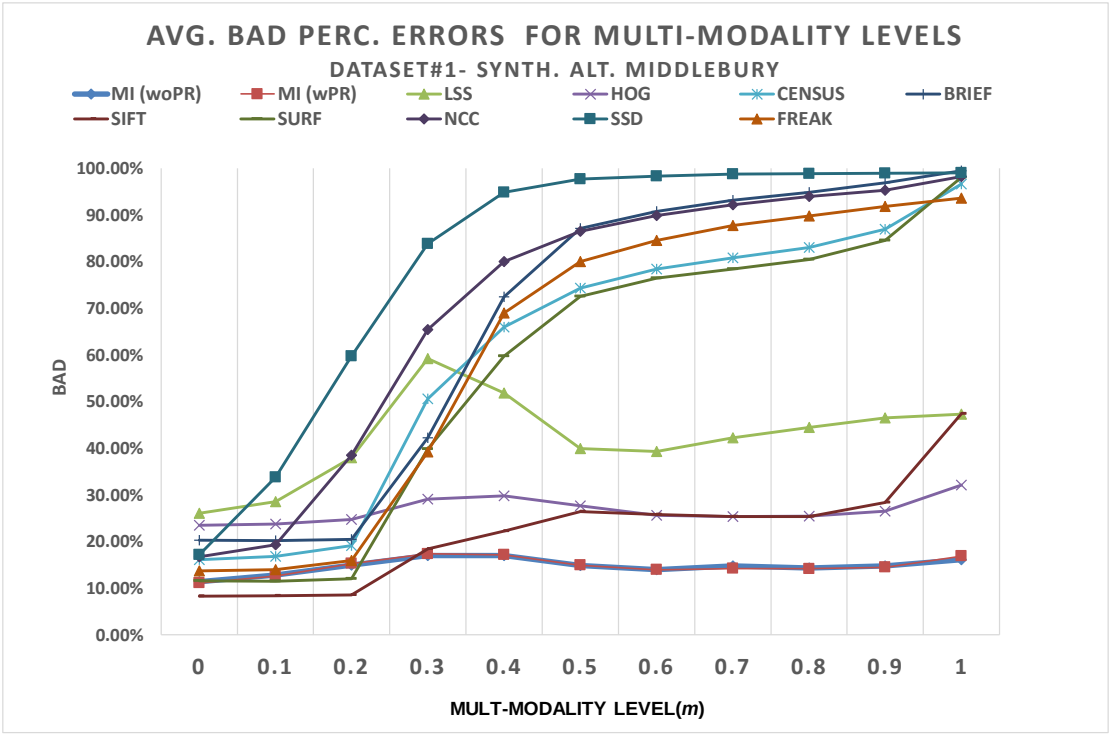
where I_{orig} is the original image from the Middlebury Image Database, I_{cos} is the cosine transformed image ($I_{cos} = (\cos(\pi f(I_{orig})/255)255)$, $f(I)$ is defined in Equation 1) and $m \in [0, 1]$ is the multi-modality level of I_m. Therefore, when $m = 1$ (full multi-modal), I_m is equal to I_{cos} and when $m = 0$ (no multi-modal), I_m is equal to I_{orig}. See Figure 8 for an illustration.

Figure 8: Figure illustrating the method for generating images of different multi-modality. m: the multi-modality level scale ($m = 0.5$ in this case); *Left image*: Original Tsukuba image from Middlebury image database; *Middle image*: Cosine transformed image; *Right image*: Generated image of multi-modality level $m = 0.5$.

The experiments presented in this section performed by testing the similarity measures for 10 multi-modality levels. The local window size is set to 21×21 pixels. Figures 9(a) and 9(b) show the average RMS and BAD pixel percentage errors corresponding to the 10 multi-modality levels of the stereo image pairs in the Dataset #1, where $m = 0$ stands for the original left image in the Middlebury image database and $m = 1$ stands for the cosine transformed left image in Dataset #1. See Figure 10 for sample visual results corresponding to the disparity maps generated for different multi-modality levels by leading similarity measures.



(a)



(b)

Figure 9: Average RMS **(a)** and BAD percentage **(b)** errors of all methods for 10 multi-modality levels for the Dataset #1 image pairs.

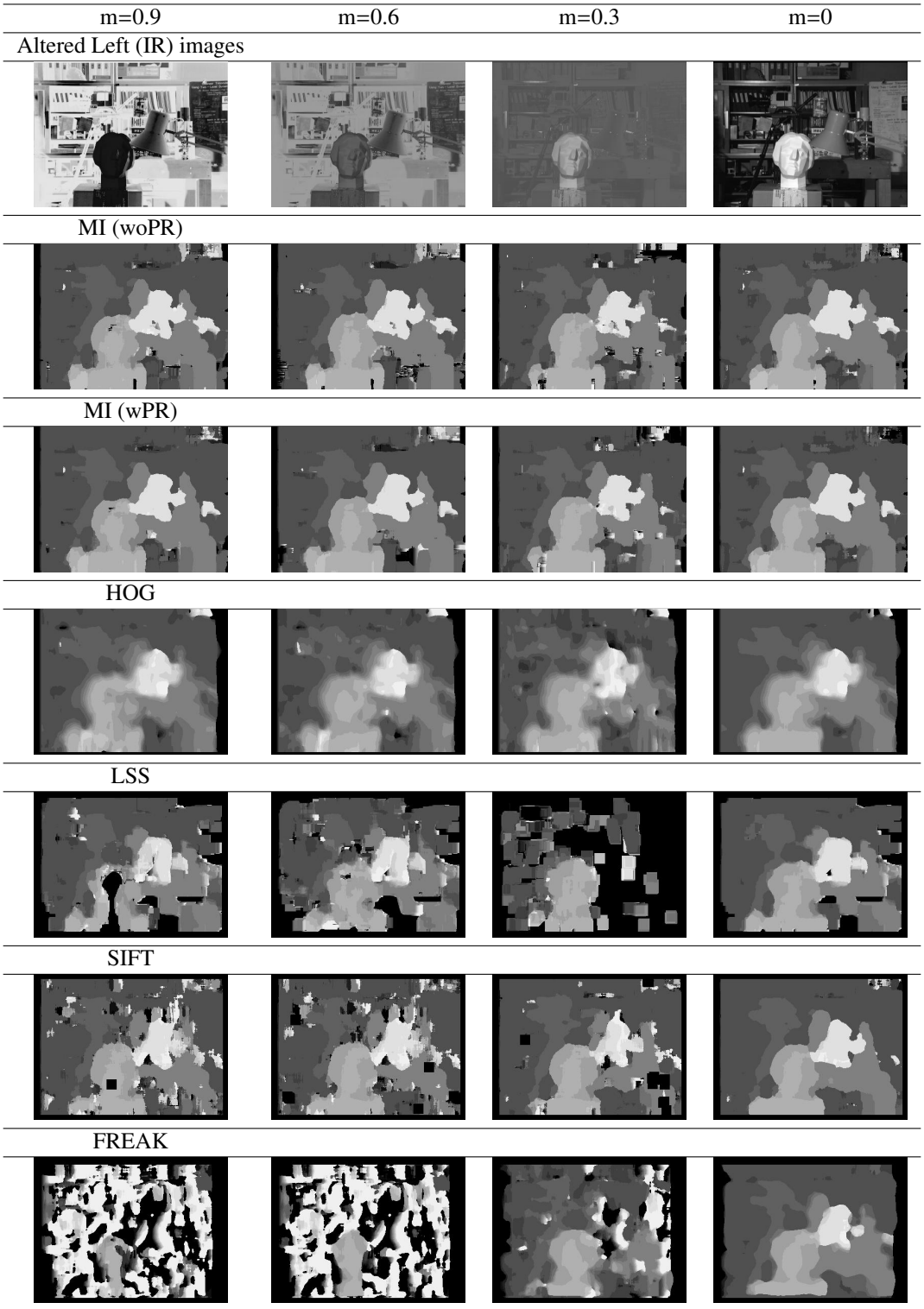


Figure 10: Sample visual results of selected similarity measures for given multi-modality levels (m=0.9, m=0.6, m=0.3 and m=0) of the Tsukuba image pair (local window size=21x21). 1st row shows altered left images of given multi-modality levels).

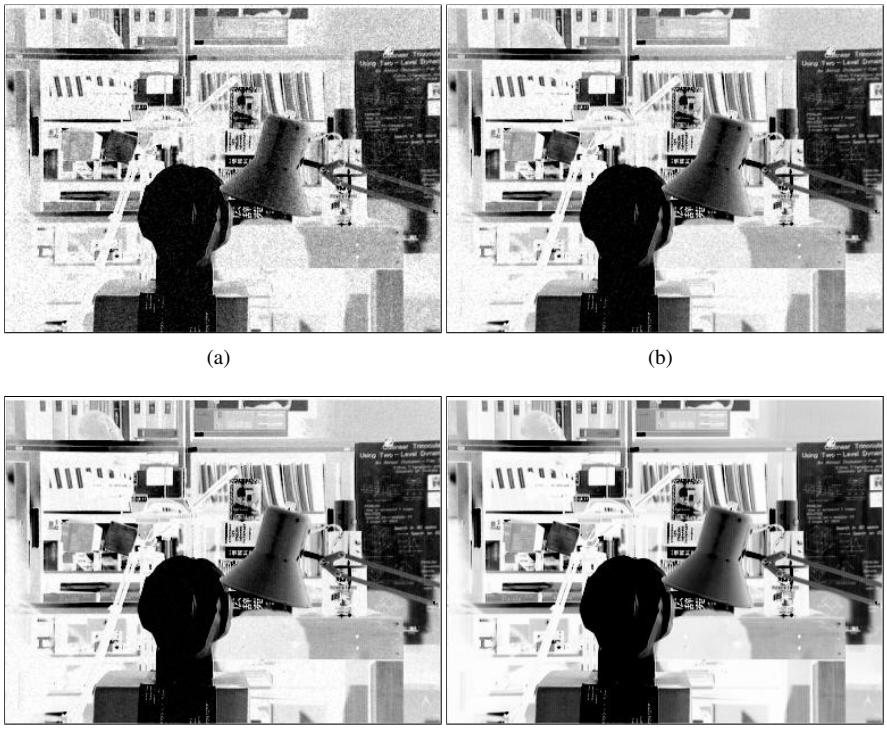
As can be observed from the results, the similarity measures are clustered into three groups. The 1st group is MI(woPR), MI(wPR) and HOG where multi-modality does not have a significant effect. The 2nd group includes LSS and SIFT, which present moderate effect for the multi-modality. LSS makes a peak at the $m = 0.3$ level, which is due to the disappearance of some of the spatial features at this level. The 3rd group includes the other methods (SURF, CENSUS, BRIEF, FREAK, NCC and SSD). Among these, SURF, FREAK, CENSUS and BRIEF have good results only before $m = 0.3$ level, and NCC before the $m = 0.2$ level. As expected, SSD is affected the most by multi-modality but it yields reasonable results at and before $m = 0$.

5.1.3. Effect of Noise

In this part, we evaluate how the methods are affected by noise. To accomplish this, a noisy image is generated using:

$$I_{noisy}(x, y) = I_{cos}(x, y) + N(x, y; \mu, \sigma). \tag{26}$$

where $N()$ is the Gaussian noise with mean $\mu = 0$ and standard deviation σ.

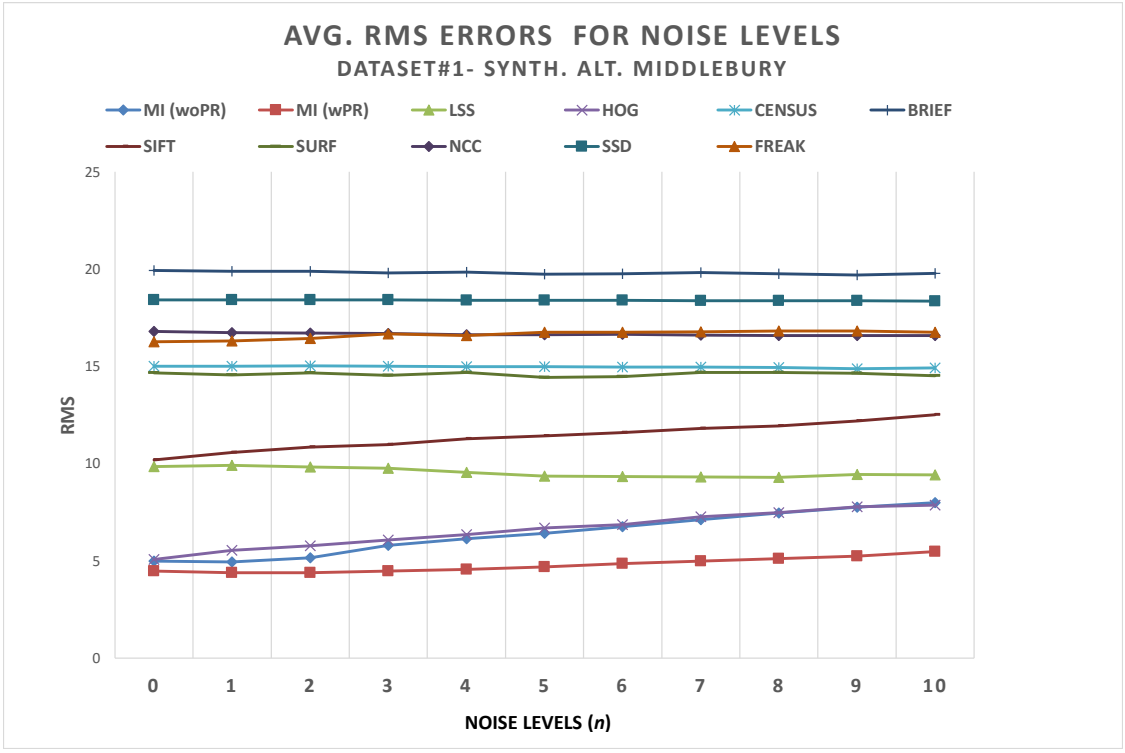


(a) (b)

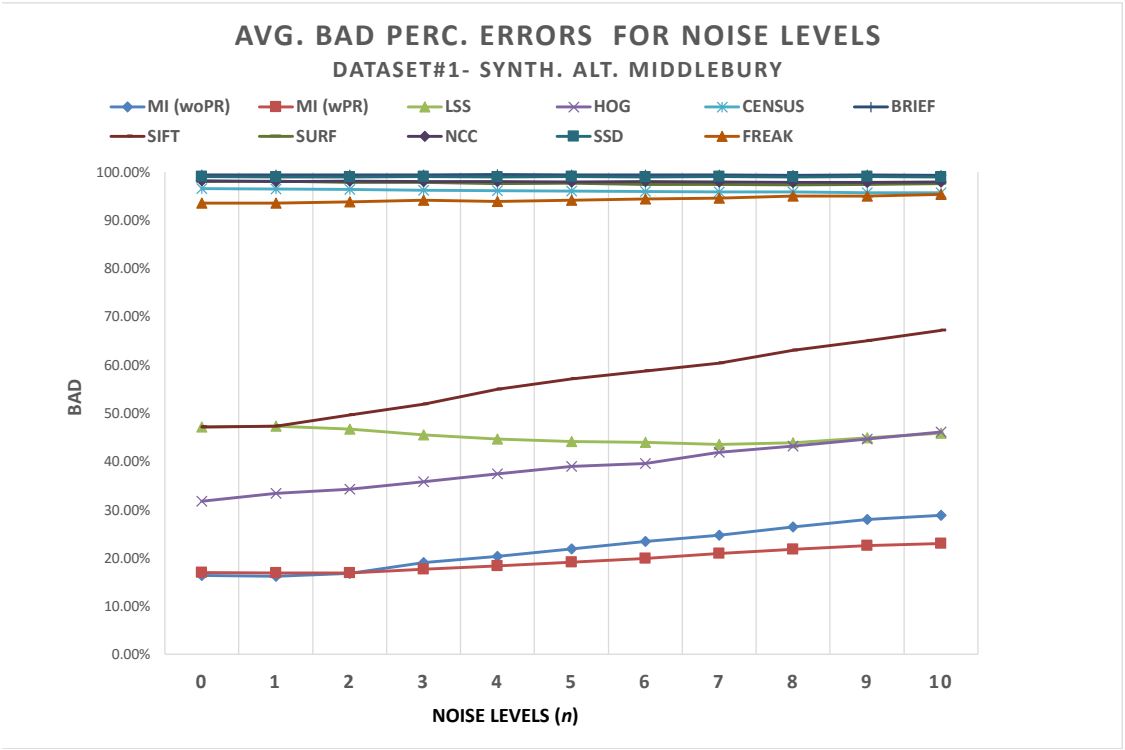
(c) (d)

Figure 11: Different noise levels applied to left Tsukuba image in Dataset #1. (a) Noise level $n = 10$ ($\sigma = 20.0$) (b) Noise level $n = 6$ ($\sigma = 12.0$) (c) Noise level $n = 3$ ($\sigma = 6.0$) (d) Noise level $n = 0$ ($\sigma = 0.0$) the noiseless cosine transformed left image.

The left images are added noise by changing the σ in the range $[20.0, 0.0]$), which is analyzed in 10 noise levels $n = 0, .., 10$, with increments of 2.0: $n = 0$ means no noise. For the experiments, window sizes are fixed to 21×21 pixels. See Figure 11 for the sample images with decreasing noise levels. Figures 12(a) and 12(b) show the average RMS and BAD errors for the 10 noise levels on the left images of Tsukuba, Venus, Teddy and Cones pairs where $n = 0$ stands for the noiseless left image. See Figure 13 for sample visual results for the disparities generated by some similarity measures.



(a)



(b)

Figure 12: Average RMS **(a)** and BAD **(b)** errors of all methods for 10 noise levels for the Dataset #1 image pairs.

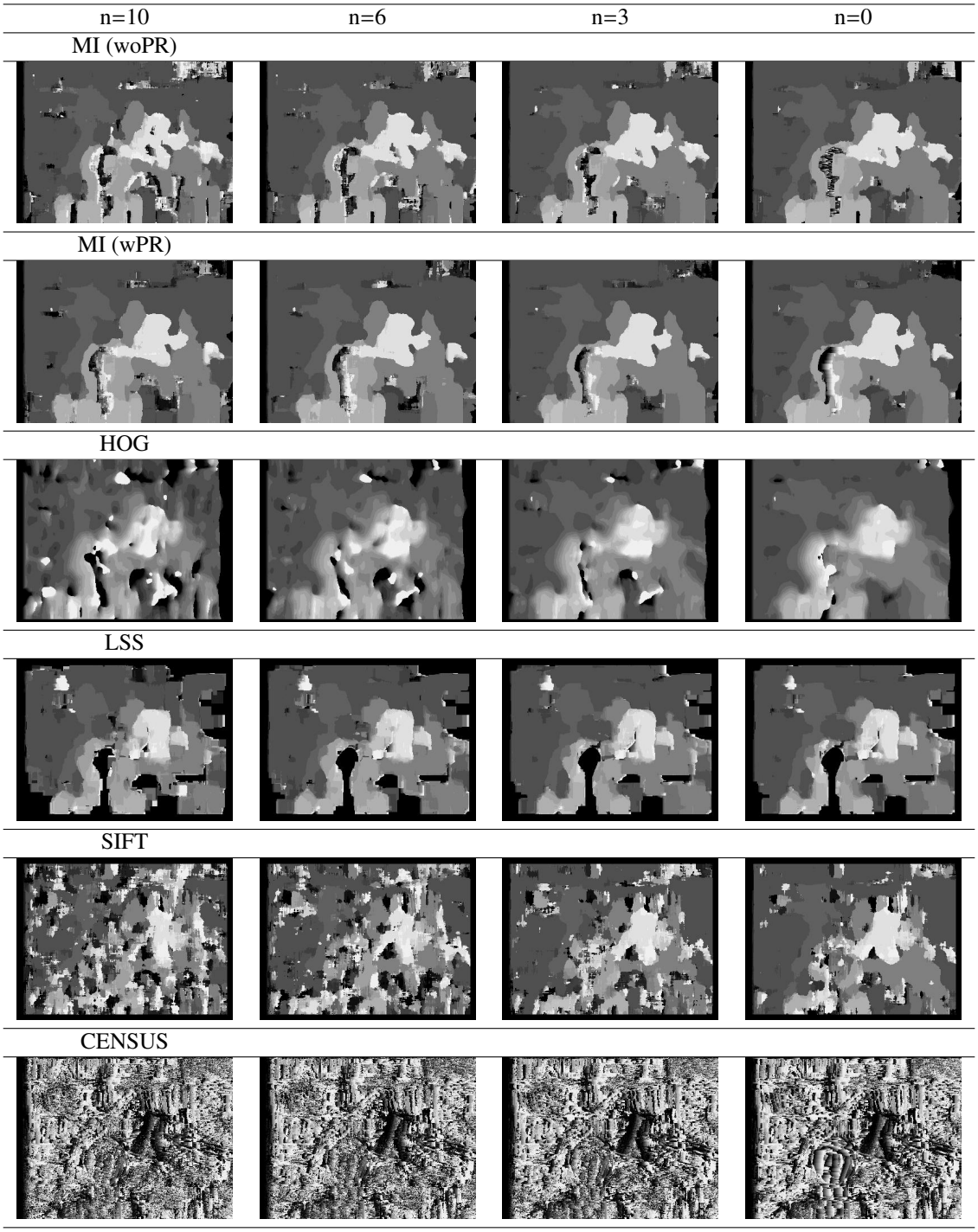


Figure 13: Sample visual results of some similarity measures for the added noise levels to Tsukuba left image in Dataset #1 (local window size=21x21) (noise levels: n = 10, n = 6,n = 3 and noiseless n = 0)

As can be observed from the obtained results, SIFT and HOG are concluded as the most vulnerable measures to noise. MI (woPR) also increase the performance as noise is decreased. MI(wPR) is concluded as the most robust method to noise. On the other hand, LSS is not affected by noise and even has a small shift in error upwards. It is

concluded that this behavior was due to the small increase in spatial correlation of homogeneous segments due to added noise. Other measures which are already shown to fail for multi-modal image pairs does not respond well to added noise as well.

5.2. Performance Evaluation Using Dataset #2 - The Kinect Dataset

In this section, the methods are compared using Dataset #2 - The Kinect Dataset. Same as the Dataset #1, the experiments are performed using the "WTA" (Winner Takes All) disparities and RMS and BAD metrics are applied using ground truth disparity maps computed from the Kinect native depth maps as described in Section 4.2. Since the Kinect images have higher resolution, the three window sizes of 31×31, 41×41 and 51×51 are used. Regarding the similarity measures, the same parameters that was used in Dataset #1 are used (see Table 2).

Figure 14, Figure 15(a) and Figure 15(b) show the resultant RMS and the BAD metrics computed for similarity measures for all 24 image pairs in Dataset #2 (See 5.2 for the description of the metrics). The BAD metrics are computed for two different disparity error thresholds $\delta = 1.5$ and $\delta = 2$ disparities.

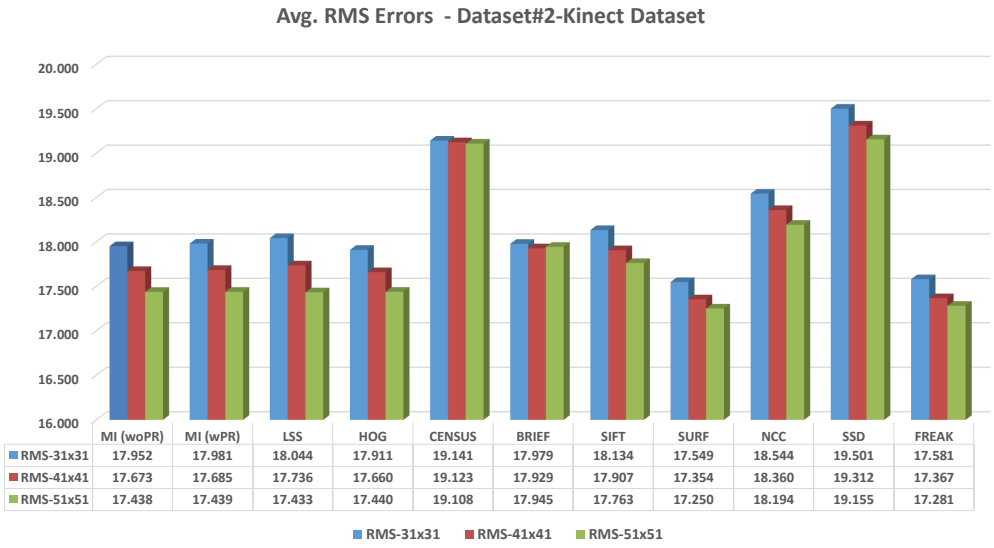
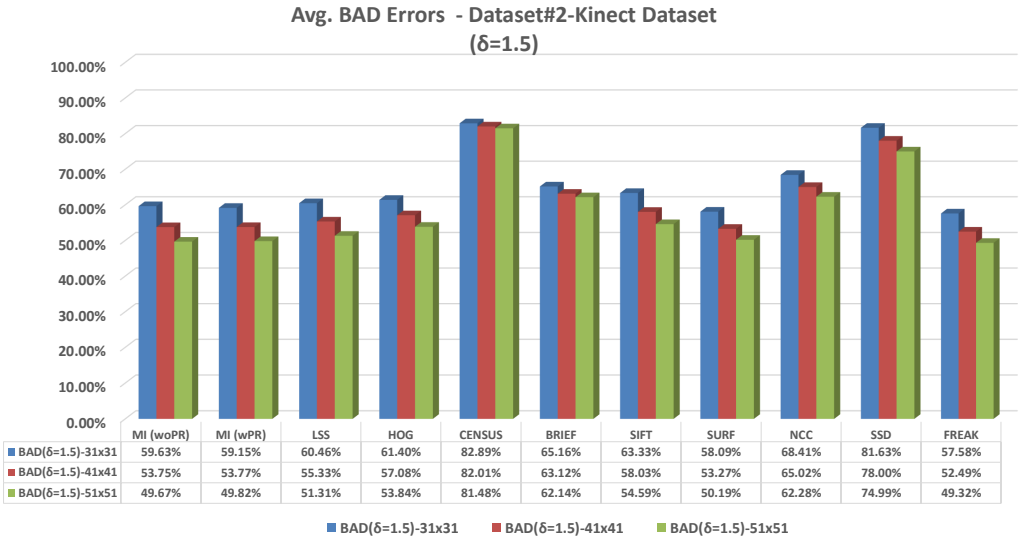
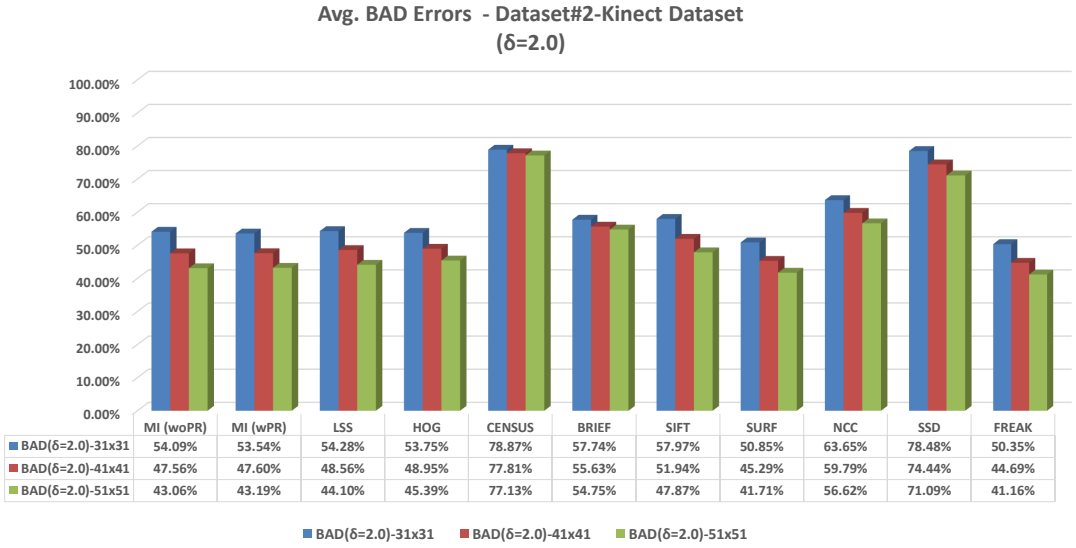


Figure 14: Average RMS errors of all methods for the Dataset #2 image pairs (using three different window sizes, 31x31, 41x41, 51x51 pixels.)



	MI (woPR)	MI (wPR)	LSS	HOG	CENSUS	BRIEF	SIFT	SURF	NCC	SSD	FREAK
BAD(δ=1.5)-31x31	59.63%	59.15%	60.46%	61.40%	82.89%	65.16%	63.33%	58.09%	68.41%	81.63%	57.58%
BAD(δ=1.5)-41x41	53.75%	53.77%	55.33%	57.08%	82.01%	63.12%	58.03%	53.27%	65.02%	78.00%	52.49%
BAD(δ=1.5)-51x51	49.67%	49.82%	51.31%	53.84%	81.48%	62.14%	54.59%	50.19%	62.28%	74.99%	49.32%

(a)



	MI (woPR)	MI (wPR)	LSS	HOG	CENSUS	BRIEF	SIFT	SURF	NCC	SSD	FREAK
BAD(δ=2.0)-31x31	54.09%	53.54%	54.28%	53.75%	78.87%	57.74%	57.97%	50.85%	63.65%	78.48%	50.35%
BAD(δ=2.0)-41x41	47.56%	47.60%	48.56%	48.95%	77.81%	55.63%	51.94%	45.29%	59.79%	74.44%	44.69%
BAD(δ=2.0)-51x51	43.06%	43.19%	44.10%	45.39%	77.13%	54.75%	47.87%	41.71%	56.62%	71.09%	41.16%

(b)

Figure 15: Average BAD errors of all methods for disparity error threshold **(a)** $\delta = 1.5$ **(b)** $\delta = 2.0$ for the Dataset #2 image pairs.

Figure 16 and Figure 17 show sample visual results of generated disparities for the dataset images for all the similarity measures tested.

Experiments over the Kinect infrared-visible image pairs show that the leading measures MI, HOG, SIFT and LSS over the Dataset #1 still provide good results for Dataset#2. However, SURF and FREAK perform even slightly better than these measures for this dataset. This is due to the low multi-modality between Near-Infrared and Visible image pairs where cloth textures and monitors behave most different than the rest of the scene objects. Kinect Near-Infrared images correspond to reflection properties of objects around 830nm wavelength of electro-magnetic spectrum where visible images include reflection properties up to 700nm wavelength which is close. This also can be observed by inspecting the multi-modality experiment results on Dataset#1 (see Figure 9(b)) where almost all the measures perform well until $m = 0.3$ level except for the NCC and SDD as expected. On the other hand, when the visual

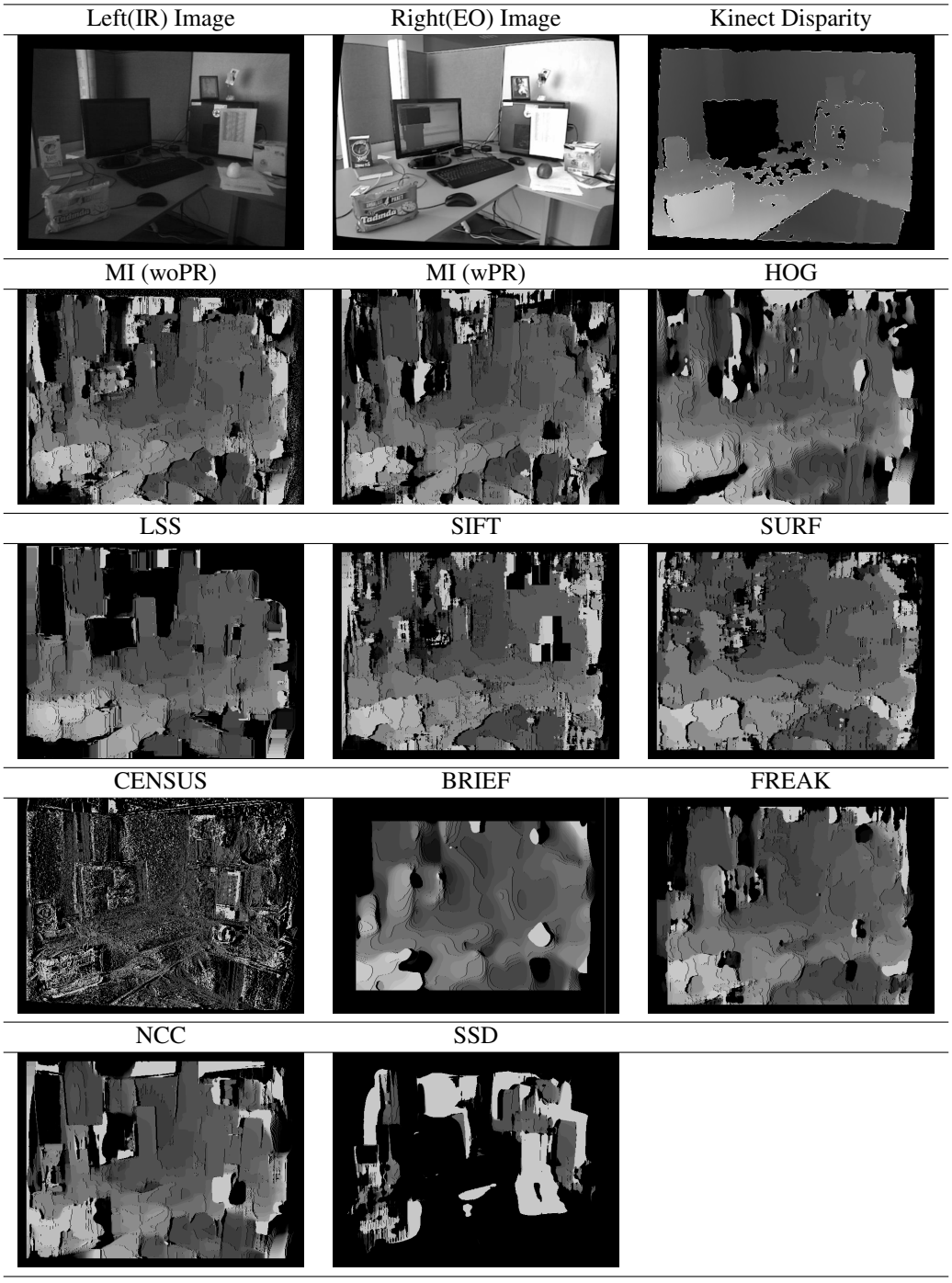


Figure 16: Visual results for Dataset#2-Img#2 showing the computed WTA disparities of the similarity measures tested for the kinect image pair and kinect disparity given in 1st row and results in the following rows.

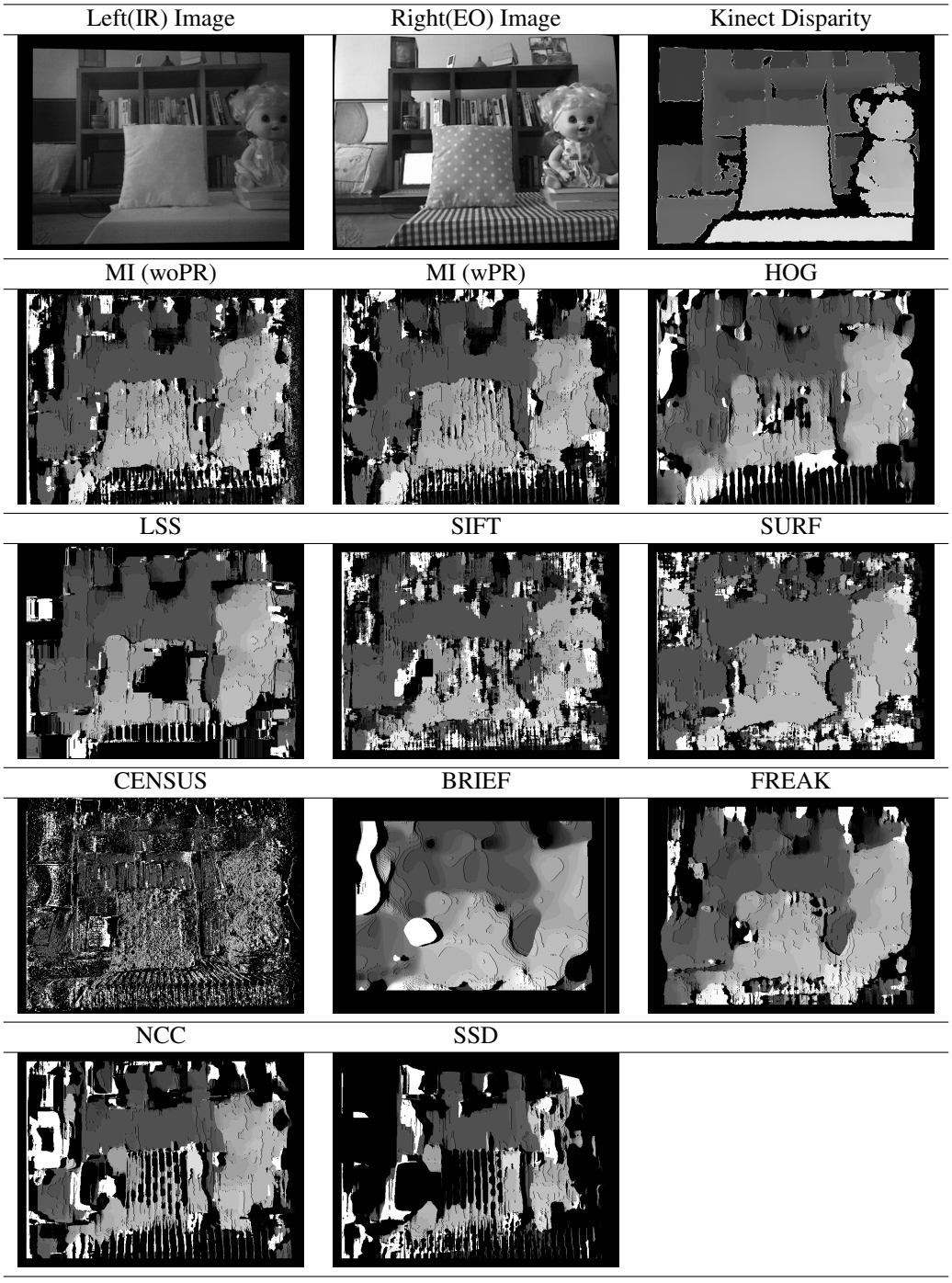


Figure 17: Visual results for Dataset#2-Img#10 showing the computed WTA disparities of the similarity measures tested for the kinect image pair and kinect disparity given in 1st row and results in the following rows.).

results are analysed, the MI measures can be taken as still more preferable to the rest of the measures since they capture the shape of the objects in the scene with better edges and surfaces. Finally, it should be noted that none of the measures are better than 40% regarding the BAD metric, which makes us conclude that there is still significant space for improvement of WTA results of these measures over Near-Infrared and Visible image pairs.

5.3. Comparison of Computational Costs

In this section, the similarity measures are compared on their computational costs which is one of the important concerns in choosing a similarity measure. As mentioned in the beginning of this section, the evaluated measures are grouped into three categories, (i) Local Window Based Measures i.e. SSD, NCC, MI with and without Prior Probability, (ii) Measures based on non-binary feature descriptors *i.e.*, LSS, HOG, SIFT and SURF and (iii) Measures based on binary features *i.e.*, CENSUS, BRIEF and FREAK.

From the computational complexity point of view, the measures can first be regrouped as the ones that need a *Precomputation* phase before the *Similarity Value Computation* phase and the ones that do not. All the measures that use binary or non-binary feature descriptors need a precomputation phase to compute the features to later compute similarity values by the extracted windows of these feature maps generated from left and right image couples for a candidate disparity value. CENSUS is an exception since its implementation requires computing the census information from the extracted window of each image couple corresponding to a candidate disparity similar to MI measure. Besides, MI with Prior Probability method needs first to compute joint probability of intensities for the image pairs globally for all the pixels.

The similarity value computation phase include computing similarity values for each corresponding left and right image pixels $I_l(x, y)$ and $I_r(x - d, y)$ for the candidate disparities $d \in [0, d_{max}]$. This enables the final "WTA - Winner Takes All" selection of best disparities which have the maximum similarity value to generate a dense disparity map for the image scene. At this phase, the Local Window Based Measures directly compute the similarity value using the extracted windows by nature. The computational complexity of these measures are bound by $O(Ndw)$, where N is the number of pixels, d is the maximum disparity and w is the window size. Regarding the Measures Based on Non-Binary and Binary Feature Descriptors, computational complexity of these measures is bound by $O(Ndwf)$ where f is the size of the feature descriptor that determines the computional time of computing the similarity value. However, it should be noted that a non-naive implementation of Hamming distance computation of binary features can be significantly efficient.

Table 3 provides the *average computational time* measured in milliseconds for the Precomputation phase for one pixel and Similarity Value Computation phase for one pixel-disparity pair of $I_l(x, y)$ and $I_r(x - d, y)$ respectively, for all similarity measures. The running times are measured on a PC with Intel Core i5 - 4200U CPU@1.6Ghz with 8.0 GB of RAM over Dataset#2 image pairs using window size of 41×41 pixels.

Table 3: Average Computational Time Measured for Precomputation and Similarity Value (SV) Computation phase for the evaluated similarity measures

Method Name	Method Type	Size (f)	Precomp. (msec)	SV Comp.(msec)
SSD	Local Window Based	n/a	n/a	0.00405
NCC	Local Window Based	n/a	n/a	0.00410
MI(woPR)	Local Window Based	n/a	n/a	0.0763
MI(wPR)	Local Window Based	n/a	0.00006	0.0991
LSS	Non-Binary Features	80	0.0896	0.6506
HOG	Non-Binary Features	36	0.0040	0.1865
SIFT	Non-Binary Features	128	0.1417	0.8492
SURF	Non-Binary Features	64	0.0021	0.3682
CENSUS	Binary Features	8	n/a	0.0196
BRIEF	Binary Features	32	0.0032	0.0848
FREAK	Binary Features	64	0.1375	0.1517

The measurements show that the local window based measures are more advantageous in terms of computational cost. Especially the MI measures can be a good selection since they both provide good performance on disparity

computation and smaller computational cost. Among the measures based on non-binary features, SIFT and LSS need significant amount of computational time for precomputation of feature vectors. However, HOG and SURF have much better computational complexity and smaller feature descriptor size. Considering the measures based on binary features, FREAK has a significant precomputation time and also a bigger binary feature size. On the other hand, CENSUS an BRIEF have much better computational time but their performances in disparity computation are not promising.

6. Conclusion

In this study, a list of similarity measures that are widely used in the literature are evaluated using two datasets generated in the scope of this study: a dataset with synthetically altered images from the Middlebury Dataset, and another dataset collected from a Kinect device.

The experiments conducted using Dataset #1 - Synthetically Altered Middlebury Images provide a good evaluation of the measures regarding the effect of multi-modality, noise and the local window sizes. MI and HOG are shown to have better performance than the rest of the measures for multi-modal imagery where LSS and SIFT yields moderate results. Among these, SIFT and HOG are shown to have vulnerability to noise. Finally, as the local window sizes increase, the performance figures also increase, however, at the expense of blurring in the resultant disparity image as expected. The upward shift in performance is smaller than the difference in the growing window sizes.

Experiments over Kinect image pairs show that the leading measures over the Dataset #1, which are MI, HOG, SIFT and LSS, still provide good results. However, SURF and FREAK perform slightly better on this dataset. The reason is that the Near-Infrared and Visible image pairs indeed provide low multi-modality except for cloth textures and monitor screens where similar results are obtained on multi-modality experiments over Dataset#1 for FREAK and SURF regarding low multi-modality levels. On the other hand, when the visual results are analysed the MI measures are concluded to be more preferable to the rest of the measures since they capture the edge and surface characteristics of the objects better.

Another advantage of MI measures is also on the computational complexity. The local window based measures can directly compute the similarity values before performing the best disparity selection. MI measures are the only measures that perform well in this category. On the other hand, the measures based on non-binary and binary features need a precomputation time to compute the feature descriptors for each pixel which requires a significant time especially for LSS and SIFT methods. CENSUS does not require this time but its performance for multi-modal imagery is not promising.

Finally, our results indicate that there is still a significant space for improvement on WTA results of these measures. However, MI and HOG provides promising results for multi-modal imagery and FREAK, SURF, SIFT and LSS can be considered as alternatives depending on the multi-modality level and the computational complexity requirements of the intended application.

References

[1] J. P. Pluim, J. A. Maintz, M. A. Viergever, Mutual-information-based registration of medical images: a survey, IEEE Transactions on Medical Imaging 22 (8) (2003) 986–1004.

[2] P. Van den Elsen, E.-J. Pol, M. A. Viergever, Medical image matching a review with classification, IEEE Engineering in Medicine and Biology 12 (1) (1993) 26–39.

[3] A. Roche, G. Malandain, X. Pennec, N. Ayache, The correlation ratio as a new similarity measure for multimodal image registration, in: Medical Image Computing and Computer-Assisted Interventation, MICCAI-98, Springer, 1998, pp. 1115–1124.

[4] M. Mellor, M. Brady, Non-rigid multimodal image registration using local phase, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004, Springer, 2004, pp. 789–796.

[5] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality image registration by maximization of mutual information, IEEE Transactions on Medical Imaging 16 (2) (1997) 187–198.

[6] C. B. Fookes (Ed.), Medical Image Registration and Stereo Vision using Mutual Information, Ph.D. Thesis, Queensland University of Technology, 2007.

[7] S. Periaswamy, H. Farid, Medical image registration with partial data, Medical Image Analysis 10 (3) (2006) 452–464.

[8] M. T. Eismann, Hyperspectral remote sensing, Vol. PM210, SPIE Press Monograph, 2012.

[9] J. A. Richards, Remote Sensing Digital Image Analysis, 5th Edition, Vol. PM210, Springer-Verlag, 2013.

[10] P. E. Anuta, Spatial registration of multispectral and multitemporal digital imagery using fast fourier transform techniques, IEEE Transactions on Geoscience Electronics 8 (4) (1970) 353–368.

[11] E. Rignot, R. Kwok, J. Curlander, S. Pang, Automated multisensor registration: Requirements and techniques, in: IEEE International Geoscience and Remote Sensing Symposium, IEEE, 1990, pp. 945–948.
[12] M. A. Ali, D. A. Clausi, Automatic registration of sar and visible band remote sensing images, in: IEEE International Geoscience and Remote Sensing Symposium, Vol. 3, IEEE, 2002, pp. 1331–1333.
[13] A. Wong, D. A. Clausi, Arrsi: automatic registration of remote-sensing images, IEEE Transactions on Geoscience and Remote Sensing 45 (5) (2007) 1483–1493.
[14] A. A. Richards, Alien Vision: Exploring the Electromagnetic Spectrum with Imaging Technology, Vol. PM205, SPIE Press Monograph, 2011.
[15] T. P. Breckon, A. Gaszczak, J. Han, M. L. Eichner, S. E. Barnes, Multi-modal target detection for autonomous wide area search and surveillance, in: Proceedings of SPIE: Emerging Technologies in Security and Defence; and Quantum Security II; and Unmanned Sensor Systems X, Vol. 8899, SPIE, 2013.
[16] Z. Zhu, T. S. Huang (Eds.), Multimodal Surveillance: Sensors, Algorithms and Systems, Artech House, 2007.
[17] C. Beyan, A. Temizel, Mean-shift tracking for surveillance applications using thermal and visible band data fusion, in: Proceedings of SPIE: Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications VIII, Vol. 8020, SPIE, 2011, pp. 1–5.
[18] S. Krotosky, M. Trivedi, Mutual information based registration of multimodal stereo videos for person tracking, Computer Vision and Image Understanding 106 (2) (2007) 270–287.
[19] Ranger hrc: Portable, long range thermal imaging surveillance system with multi-sensor option, http://www.flir.com/uploadedFiles/flirGS/Surveillance/Products/Ranger/HRC/flir-ranger-hrc-datasheet.pdf, last accessed: 22 Feb 2015.
[20] Mx-rsta: A multi-sensor, multi-spectral imaging system, http://www.wescam.com/index.php/products-services/ground-market/mx-rsta/, last accessed: 22 Feb 2015.
[21] Mx-25d: Fully digital, high definition, ultra long-range multi-sensor, multi-spectral imaging and targeting systems, http://www.wescam.com/index.php/products-services/airborne-targeting/mx-25d/, last accessed: 22 Feb 2015.
[22] Seaflir 380hd: Single lru eo/ir imaging system, http://www.flir.com/uploadedFiles/flirGS/Surveillance/Products/SeaFLIR/380-HD/flir-seaflir-380-hd-datasheet-ltr.pdf, last accessed: 22 Feb 2015.
[23] Sensors unlimited: Swir image gallery, http://www.sensorsinc.com/gallery/images, last accessed: 22 Feb 2015.
[24] J. M. Kriesel, N. Gat, True-color night vision (tcnv) fusion system using a vnir emccd and a lwir microbolometer camera, in: SPIE Proceedings, Vol. 7697, 2010.
[25] M. Yaman, S. Kalkan, An iterative adaptive multi-modal stereo-vision method using mutual information, Journal of Visual Communication and Image Representation 26 (2015) 115–131.
[26] S. Krotosky, M. Trivedi, Multimodal stereo image registration for pedestrian detection, in: IEEE Intelligent Transportation Systems Conference, IEEE, 2006, pp. 109–114.
[27] F. Barrera Campo, F. Lumbreras Ruiz, A. D. Sappa, Multimodal stereo vision system: 3d data extraction and algorithm evaluation, IEEE Journal of Selected Topics in Signal Processing 6 (5) (2012) 437–446.
[28] A. Torabi, G.-A. Bilodeau, Local self-similarity as a dense stereo correspondence measure for themal-visible video registration, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2011, pp. 61–67.
[29] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, D. Riahi, Thermal-visible registration of human silhouettes: a similarity measure performance evaluation, Infrared Physics & Technology 64 (2014) 79–86.
[30] M. Yaman, S. Kalkan, Multimodal stereo vision using mutual information with adaptive windowing, in: 13th IAPR International Conference on Machine Vision Applications, IAPR, 2013.
[31] Benchmark datasets for multi-modal stereo-vision : Metu multi-modal stereo datasets, http://kovan.ceng.metu.edu.tr/MMStereoDataset/, last accessed: 10 March 2015.
[32] The middlebury stereo vision page, http://vision.middlebury.edu/stereo/, accessed: 22 Feb 2015.
[33] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (1-3) (2002) 7–42.
[34] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2013, pp. 195–202.
[35] Microsoft's kinect for windows, http://www.microsoft.com/en-us/kinectforwindows/, last accessed: 22 Feb 2015.
[36] R. Hartley, A. Zisserman, Multiple view geometry in computer vision, Cambridge Univ Press, 2000.
[37] R. Szeliski, Computer vision: algorithms and applications, Springer, 2011.
[38] M. Z. Brown, D. Burschka, G. D. Hager, Advances in computational stereo, IEEE Transactions On Pattern Analysis and Machine Intelligence 25 (8) (2003) 993–1008.
[39] E. Tola, V. Lepetit, F. P., Daisy: An efficient dense descriptor applied to wide-baseline stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5) (2010) 815–830.
[40] H. Hirschmueller, Stereo processing by semiglobal matching and mutual information, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 328–341.
[41] U. R. Dhond, J. K. Aggarwal, Structure from stereo-a review, IEEE Transactions on Systems, Man and Cybernetics 19 (6) (1989) 1489–1510.
[42] V. Venkateswar, R. Chellappa, Hierarchical stereo and motion correspondence using feature groupings, International Journal of Computer Vision 15 (3) (1995) 245–269.
[43] G. Egnal, Mutual information as a stereo correspondence measure, Technical Report MS-CIS-00-20, University of Pennsylvania (2000) 113.
[44] K. Ambrosch, W. Kubinger, M. Humenberger, A. Steininger, Flexible hardware-based stereo matching, EURASIP Journal on Embedded Systems 2008 (2).
[45] N. Pugeault, N. Krger, Multi-modal matching applied to stereo, in: Proceedings of the British Machine Vision Conference, 2003, pp. 271–280.
[46] C. Cassisa, Local vs global energy minimization methods: application to stereo matching, in: IEEE International Conference on Progress in Informatics and Computing (PIC), Vol. 2, IEEE, 2010, pp. 678–683.
[47] V. Kolmogorov, R. Zabih, Computing visual correspondence with occlusions using graph cuts, in: IEEE International Conference on Com-

puter Vision, Vol. 2, IEEE, 2001, pp. 508–515.
[48] J. Sun, N. Zheng, H. Shum, Stereo matching using belief propagation, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (7) (2003) 787–800.
[49] J. Kim, V. Kolmogorov, R. Zabih, Visual correspondence using energy minimization and mutual infomation., in: CVPR, IEEE, 2003.
[50] P. Viola, W. M. Wells III, Alignment by maximization of mutual information, International Journal of Computer Vision 24 (2) (1997) 137–154.
[51] C. Fookes, A. Lamanna, M. Bennamoun, A new stereo image matching technique using mutual information, in: International Conference on Computer, Graphics and Imaging, 2001.
[52] C. Fookes, A. Maeder, S. Sridharan, J. Cook, Multi-spectral stereo image matching using mutual information, in: International Symposium on 3D Data Processing, Visualization and Transmission, IEEE, 2004, pp. 961–968.
[53] S. Krotosky, M. Trivedi, Registration of multimodal stereo images using disparity voting from correspondence windows, in: IEEE International Conference on Video and Signal Based Surveillance, IEEE, 2006, pp. 91–91.
[54] Evaluation of similarity functions in multimodal stereo.
[55] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[56] A. Torabi, M. Najafianrazavi, G.-A. Bilodeau, A comparative evaluation of multimodal dense stereo correspondence measures, in: IEEE International Symposium on Robotic and Sensors Environments (ROSE), IEEE, 2011, pp. 143–148.
[57] A. Torabi, G.-A. Bilodeau, A lss-based registration of stereo thermalvisible videos of multiple people using belief propagation, Computer Vision and Image Understanding 117 (12) (2013) 1736–1747.
[58] H. Hirschmueller, D. Scharstein, Evaluation of stereo matching costs on images with radiometric differences, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (9) (September 2009) 1582–1599.
[59] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[60] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision–ECCV, Springer, 2006, pp. 404–417.
[61] B. T. N. Dalal, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05), Vol. 1, IEEE, 2005, p. 886893.
[62] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: binary robust independent elementary features, in: Computer Vision ECCV 2010, Lecture Notes in Computer Science, Vol. 6314, Springer, 2010, p. 778792.
[63] R. W. Hamming, Error detecting and error correcting codes, Bell System Technical Journal 297 (2) (1950) 147160.
[64] A. Alahi, R. Ortiz, P. Vandergheynst, Freak: fast retina keypoint, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, p. 510517.
[65] C. E. Shannon, A mathematical theory of communication, Bell Systems Technical Journal 27 (1948) 379423.
[66] S. Kullback, R. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1) (1951) 7986.
[67] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, in: Computer Vision ECCV 94, Lecture Notes in Computer Science, Vol. 801, Springer, 1994, p. 151158.
[68] Middlebury stereo evaluation - version 2, `http://vision.middlebury.edu/stereo/eval/`, accessed: 22 Feb 2015.
[69] The xbox 360 video game console, `http://www.xbox.com/tr-TR/xbox-360`, last accessed: 22 Feb 2015.
[70] Z. Zhang, A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (11) (2000) 1330–1334.
[71] K. Khoshelham, S. J. Oude Elberink, Accuracy and resolution of kinect depth data for indoor mapping applications, Sensors 12 (2) (2012) 1437–1454.
[72] Precision of the kinect sensor, `http://wiki.ros.org/openni_kinect/kinect_accuracy`, last accessed: 21 Jan 2016 (June 2011).

List of Figures

Authors

Dr. Mustafa Yaman received his M.Sc. degree in Computer Engineering from Middle East Technical University, Turkey in 2003, and his Ph.D. degree in the same department in 2014. He is currently working in a private company in Turkey as Chief Engineer and Research Team Leader, participating in several research projects in Computer Vision, Image Processing and Target Recognition using EO/IR Imaging Systems.

Dr. Sinan Kalkan received his M.Sc. degree in Computer Engineering from Middle East Technical University, Turkey in 2003, and his Ph.D. degree in Informatics from the University of Gttingen, Germany in 2008. After working as a postdoctoral researcher at the University of Gottingen and at Middle East Technical University, he is an assistant professor at Middle East Technical University since 2010. Sinan Kalkan's research interests include biologically motivated Computer Vision and Image Processing and Developmental Robotics.