Natural Scene Statistics of Long Wave Infrared and Visible Images

D.E. Moreno-Villamarín\textsuperscript{a,}\textsuperscript{*}, H.D. Benítez-Restrepo\textsuperscript{a}, A.C. Bovik\textsuperscript{b}

\textsuperscript{a}Pontificia Universidad Javeriana, Seccional Cali, Departamento de Electrónica y Ciencias de la Computación, Calle 18 No 118-250, Cali, Colombia
\textsuperscript{b}The University of Texas at Austin, Department of Electrical and Computer Engineering, TX 78712, Austin, USA

Abstract

Quality evaluation of long wave infrared (LWIR) and visible images plays an important role in the quality of a resulting fused image. Extensive work has been conducted on studying statistics of natural scenes in LWIR and visible images. Nonetheless, there has been little work done on analyzing the statistics of fused LWIR and visible images’ distortions. In this paper, we analyze five multi-resolution based image fusion methods in the presence of several common distortions, including blur, white noise, JPEG compression, and non-uniformity (NU) on LWIR images and how they affect natural scene statistics (NSS) models in resulting fused images. Furthermore, we conducted a human study on the subjective quality of pristine and degraded LWIR and visible fused images and created an opinion-distortion-unaware (ODU) fused image quality analyzer. In this test, 27 subjects evaluated 750 images in five sessions. In addition, we propose an opinion-aware (OA) fused image quality analyzer, whose relative predictions with respect to other state-of-the-art metrics correlated better to human perceptual evaluations. An implementation of the proposed fused image quality measures can be found at \url{https://github.com/ujemd/NSS-of-LWIR-and-Vissible-Images}.

Keywords: NSS, LWIR, multi-resolution image fusion, fusion performance

\textsuperscript{*}Corresponding author

Email addresses: david.moreno@javerianacali.edu.co (D.E. Moreno-Villamarín), benitez@ieee.org (H.D. Benítez-Restrepo), bovik@ece.utexas.edu (A.C. Bovik)
1. Introduction

The increasing level of uncertain security in the last years in our world, along with the availability of cheaper and more intelligent digital cameras technology is encouraging great interest on the development of video-based systems capable of observing and interpreting specific scenes, with the aim of detecting anomalies or events that may affect the safety, security, economics, and vital parts of human activity [1]. Outdoor surveillance systems that rely on electro-optical sensors such as CCD cameras are prone to failures due to illumination changes, and weather conditions [2, 3]. Nonetheless, due to the decreasing costs and increasing miniaturization of infrared sensors, the use of this technology has become an interesting alternative in surveillance systems [1, 5, 6, 7].

Although Long Wave Infrared (LWIR) sensors can accurately capture video in low-light and night-vision applications, they lack the color and relative luminances of visible spectrum sensors, whereas RGB sensors do capture color and correct relative luminances, but are sensitive to illumination variations, noisy, and lack fine features due to short video exposure times [8]. Therefore, two main benefits of the joint use of thermal and visible sensors are first the complementary nature of different modalities that provides the thermal and color information from the scene and second, the redundancy of information captured by the sensors, which increases the reliability and robustness of a surveillance system. These advantages have motivated the computer vision community to study and investigate algorithms for the fusing of infrared and visible videos for surveillance applications [5].

Due to the growing interest in LWIR and visible image fusion, considerable efforts have been made to develop objective quality measures of fused images. The performance of different image fusion algorithms can be evaluated by existing image fusion quality metrics based on information theory [9, 10, 11], image features [12, 13, 14, 15], image structural similarity [16, 17, 18], and human perception [19, 20]. Chen and Blum [20] investigated the performance of fusion metrics based on human vision system models under several levels of additive white Gaussian noise (AWGN). Liu et al [21] analyzed the impact of AWGN and blur on fused images. They found that the quality of fused images degrades with the decrease of image quality. When the AWGN level is severe, all the fused images are almost of the same quality, regardless of the fusion scheme. These studies lack the analysis of specific distortions in LWIR sensors such as non uniformity (NU) and “Halo
Effect.” Even though extensive work has been conducted on studying statistics of natural scenes in the visible spectrum [22, 23, 24] and some studies have been done on LWIR images [25, 26], there has been very little work done on analyzing the statistics of fused LWIR and visible images afflicted by any of multiple impairments.

The objective of this work is to analyze how image distortions such as AWGN, blur, JPEG compression, and non-uniformity noise in LWIR and visible images affect natural scene statistics (NSS) on fused LWIR-visible images. We deploy previous models proposed in [25, 27, 28] as a starting point to task on NSS, and create ‘opinion-distortion-unaware’ (ODU) and ‘opinion-aware’ (OA) image quality measures. An image quality measure is ODU if it does not require training on databases of human judgments of distorted images and it does not rely on training and tuning on specific distortions. In contrast, a model is OA if it has been trained on a database(s) of human rated distorted images and associated subjective opinion scores. The comparison of the results obtained with the proposed OA and ODU measures with state-of-the-art metrics shows outstanding results.

This paper is organized as follows: Section 2 outlines the image databases and methods used throughout our analysis of NSS. Section 3 describes the processing and feature models we employed. Section 4 describes the subjective human study carried out in this work, and also presents two fused image quality measures. Sections 5 and 6 discuss the results and present the conclusions and further work, respectively.

2. Material and methods

2.1. LWIR and visible image sources

This study of multimodal image fusion uses databases that we denote as OSU [29], TNO [30, 31], and MORRIS [26]. The OSU database contains two scenes from the Ohio State University campus. From each scene, 40 visible and LWIR image pairs are used. The TNO includes three different outside scenes comprising 23, 19, and 32 images, respectively. The MORRIS database contains indoor and outdoor images of urban environments. Before processing the images, they were linearly re-scaled to the range 0 to 1, to apply the different artificial distortions consistently. Example images from these databases can be seen in Figure 1.
2.2. Distortion models

Several studies have characterized and modeled noise in the LWIR spectrum. Images obtained from focal plane arrays (FPA) can present NU fixed pattern noise [32], which produces a grid-like pattern. In [33], Pezoa and Medina describe a model for this type of noise in LWIR images. By using spectral analysis, the model is:

\[ |\tilde{I}(u, v)| = B_u \exp\left(-\frac{(u - u_0)^2}{2\sigma_u^2}\right) + B_v \exp\left(-\frac{(v - v_0)^2}{2\sigma_v^2}\right) \quad (1) \]

\[ \angle \tilde{I}(u, v) \sim U[-\pi, \pi] \quad (2) \]

where \( \tilde{I} \) is the Fourier Transform of the non-uniformity noise. The parameters \( u_0 \) and \( v_0 \) represent the location, \( B_u = B_v = 5.2 \) the amplitude, \( \sigma_u = \)
\[ \sigma_v = 2.5 \] the scale of the respective horizontal or vertical band, and \( U[a, b] \) denotes the uniform distribution over the interval \([a, b]\). The distortion level is controlled using a standard deviation parameter \( \sigma_{NU} \), which scales the dynamic range of the NU noise. Other common types of distortion which could affect both LWIR and visible images are also considered in this study, such as AWGN, blur, and JPEG compression.

Three distortion levels are used throughout the study for each distortion type, which were applied to the LWIR and visible images of the three databases. For AWGN and NU the standard deviation is varied as \( \sigma_{AWGN} = \sigma_{NU} = \{0.0025, 0.01375, 0.025\} \); for blur, a Gaussian blur kernel of size 15 \( \times \) 15 pixels with \( \sigma_{blur} = \{1, 2, 3\} \) is used; and for JPEG compression, the quality is set to 100, 90 and 80 percent. Figure 2 depicts fused images when both sources are affected by the most severe distortion level.

\[ \begin{align*}
\text{(a)} & \quad \text{(b)} & \quad \text{(c)} & \quad \text{(d)} \\
\end{align*} \]

Figure 2: Example of distortions occurring in fused visible and LWIR images. (a) Additive white noise. (b) Non-Uniformity. (c) Blur. (d) JPEG compression. Images from [26].

2.3. Multi-resolution fusion methods

In night-vision, one of the most important techniques is multi-resolution image fusion (MIF), which aims to retain the main features from the source images [34]. This technique focuses on accessible multi-resolution feature representations and an image fusion rule to guide the combination of coefficients in the transform domain. How the fusion algorithm adapts to different object-to-background situations is still not well understood.

In Liu et al. [21], the fusion performance models evaluated six multi-resolution fusion methods, from which we consider the following five: average (AVG), gradient pyramid (GP) [35], Laplacian pyramid (LP) [36], ratio of low-pass pyramid (RP) [37], and shift-invariant discrete wavelet transform with Haar wavelet (SIDWT) [38]. The decomposition level in the algorithms
is set to four, and the fusion rule is the maximum selection for the high-pass combination and average of the low-pass combination. Our work considers a direct heterogeneous image fusion scheme based on a multi-resolution approach at the pixel level.

3. NSS of fused LWIR and visible images

3.1. Processing model

Highly successful IQA models have been based on the early work by Ruderman on ‘natural images’. Natural light images are captured by an optical camera, and contain particular types of structures, contrary to images completely generated by a computer using artificial processing. The work on natural scene statistics has focused towards predicting a response of a biological vision system [28, 39]. Ruderman found that removing the local mean from a natural image and variance normalizing has a Gaussianizing effect on its distribution. This operation produces the Mean-Subtracted Contrast Normalized (MSCN) coefficients of an image, which can be computed as follows:

\[
\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{3}
\]

where \( I \) is a luminance image or image patch with \( i \in 1, 2, ..., M \) and \( j \in 1, 2, ..., N \), where \( M \) and \( N \) are the image height and width, respectively. The constant \( C \) is usually set as 1, preventing a division by zero. Local mean \( \mu \) and standard deviation \( \sigma \) are defined as:

\[
\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_{k,l}(i,j) \tag{4}
\]

\[
\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} (I_{k,l}(i,j) - \mu(i,j))^2} \tag{5}
\]

where \( w \) is a 2D circularly-symmetric weighting function sampled out to 3 standard deviations and normalized to unit volume.

The histograms of the MSCN coefficients for regions of interest (ROI) of fused LWIR and visible images are depicted in Figure 3. A total of 154 ROIs in five scenes were selected from the OSU and TNO databases. In this case,
Figure 3: Comparison of MSCN histograms of 154 ROIs obtained from fused images (80 from OSU and 74 from TNO databases). ROIs sizes are 64 by 64 pixels. The figures show an increasing distortion level for AWGN, blur, JPEG compression, and non uniformity (NU). The terms AVG, GP, LP, RP, SIDWT refer to the fusion methods Average, Gradient Pyramid, Laplacian Pyramid, Ratio Pyramid, and Shift Invariant Discrete Wavelet Transform, respectively.
three distortion levels are used. For AWGN and NU the standard deviation is varied as $\sigma_{\text{AWGN}} = \sigma_{\text{NU}} = \{0.0025, 0.01375, 0.025\}$; for blur, a Gaussian blur kernel of size 15 by 15 pixels with $\sigma_{\text{blur}} = \{1, 2, 3\}$ is used; and for JPEG compression, the quality is set to 100, 90 and 80 percent. We observe that the Laplacian pyramid fusion method appears to affect the naturalness of the resulting fused images. On the other hand, blur distortion produces thinner histograms, in contrast to the effects of AWGN and NU, which produce wider histograms.

In [27], the authors work with four paired product coefficients calculated as the multiplication of neighboring MSCN coefficients. With these coefficients, it is possible to analyze the directional behavior of both the statistical regularity and perturbations of the images. The paired products are calculated along four orientations: horizontal ($H$), vertical ($V$), main-diagonal ($D_1$), and secondary diagonal ($D_2$):

$$
H(i, i) = \hat{I}(i, j) \hat{I}(i, j + 1) 
$$

$$
V(i, i) = \hat{I}(i, j) \hat{I}(i + 1, j) 
$$

$$
D_1(i, i) = \hat{I}(i, j) \hat{I}(i + 1, j + 1) 
$$

$$
D_2(i, i) = \hat{I}(i, j) \hat{I}(i + 1, j - 1) 
$$

Histograms of these paired product coefficients were generated following the same procedure used for the MSCN images. Figure 4 depicts the $H$ paired product histograms. A remarkable characteristic is the high sensitivity to blur distortions, which produces thinner histograms. In the case of the Laplacian pyramid fusion method, there is a noticeable sensitivity to AWGN, which produces wider histograms, as well as for NU distortion in a lesser degree. Nonetheless, these histograms seem to indicate small differentiation between JPEG compressed and pristine images. Histograms of $V$, $D_1$, and $D_2$ coefficients also present similar characteristics.

In a further analysis presented in [40], the MSCN coefficients are supplemented by a set of log-derivative coefficients, which are intended to provide higher sensitivity to high-frequency noise:

$$
J(i, j) = \log(\|\hat{I}(i, j)\| + K) 
$$

where $K$ is a stabilizing constant, in this case set to $K = 0.1$. The log-derivative coefficients are calculated as:
Figure 4: Comparison of horizontal paired product histograms of 154 ROIs obtained from fused images. ROIs sizes are 64 by 64 pixels. The figures show an increasing distortion level for AWGN, blur, JPEG compression, and NU.

\[ PD1(i, j) = J(i, j + 1) - J(i, j) \]  \hspace{1cm} (11)
\[ PD2(i,j) = J(i + 1, j) - J(i, j) \] (12)
\[ PD3(i,j) = J(i + 1, j + 1) - J(i, j) \] (13)
\[ PD4(i,j) = J(i + 1, j - 1) - J(i, j) \] (14)
\[ PD5(i,j) = J(i - 1, j) + J(i + 1, j) \] (15)
\[ \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad - J(i, j - 1) - J(i, j + 1) \]
\[ PD6(i,j) = J(i, j) + J(i + 1, j + 1) \] (16)
\[ \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad - J(i, j + 1) - J(i + 1, j) \]
\[ PD7(i,j) = J(i - 1, j - 1) + J(i + 1, j + 1) \] (17)
\[ \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad - J(i - 1, j + 1) - J(i + 1, j - 1) \]

Following the described ROI extraction procedure for both pristine and distorted images, histograms of the \( PD6 \) log-derivative coefficients are plotted in [Figure 5]. \( PD1, PD2, PD3, PD4, PD5, \) and \( PD7 \) coefficients have also a high sensitivity to blur, AWGN, and NU distortions. It is interesting to see that in \( PD6 \) and \( PD7 \) histograms, JPEG distortion produces thinner histograms as long as the image quality decreases.

The aforementioned coefficients evaluate only normalized high-pass images; therefore, a steerable decomposition is used to describe band-pass characteristics [25]. The divisively normalized steerable pyramid decomposition is inspired in the decomposition that occurs in area V1 of the primary visual cortex, and has been previously used in FR and RR IQA [23]. In our case, the decomposition is made over six orientations, where each band is denoted \( d_{\theta}^{\alpha} \), where \( \alpha \) indicates the scale and \( \theta \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ\} \). Here the histograms for these coefficients are depicted using only one scale, as it has been found that increasing the scale selectivity does not improve performance [23], however, in the following sections, three scales are analyzed.

Using the same pooled ROI extraction procedure, histograms produced from \( d_{150^\circ}^{30} \) coefficients are plotted in [Figure 6]. When analyzing distortion behavior, NU has a distinctively large standard deviation in the horizontal and vertical subbands, \( d_{150}^{0} \) and \( d_{150}^{90} \), which follows from the striping behavior of NU. \( d_{150}^{0} \) presents also a high sensitivity to AWGN, but the effects on histograms of images affected by blur are not discernible. Regarding \( d_{150}^{30} \) and \( d_{150}^{90} \), the effect of AWGN is in general noticeable, yet not as much as in other coefficients. The effect of NU noise is minimal, contrary to the blur distortion, in which the width of the histograms decreases at higher distortion levels. Histograms of \( d_{150}^{90} \) coefficients show a certain asymmetry, especially for the average fusion method, where the shape for the highest level.
Figure 5: Comparison of log-derivative histograms of 154 ROIs obtained from fused images. ROIs sizes are 64 by 64 pixels. The figures show an increasing distortion level for AWGN, blur, JPEG compression, and NU.

of blur distortion is wider for the other levels. The $d_{1}^{120^0}$ and $d_{1}^{150^0}$ coefficients present an asymmetrical shape when affected by blur distortion, however, they describe well the effect of this distortion type except for the average
In general, each type of coefficient helps capturing the distortion effect on fused images. ROIs sizes are 64 by 64 pixels. The figures show an increasing distortion level for AWGN, blur, JPEG compression, and NU.

Figure 6: Comparison of \(d_1^{30°}\) steerable pyramid histograms of 154 ROIs obtained from fused images. ROIs sizes are 64 by 64 pixels. The figures show an increasing distortion level for AWGN, blur, JPEG compression, and NU.

fusion method. For AWGN and NU, the changes between distortion levels do not become evident as in MSCN and paired log-derivative coefficients.

In general, each type of coefficient helps capturing the distortion effect on
fused images, some types better than others. Therefore, statistical models can fit these histograms to extract perceptual features, as presented in the next section.

3.2. Feature models

Previous works have shown that the Generalized Gaussian Distribution (GGD) captures the behavior of natural and distorted versions of the MSCN, paired log-derivative, and steerable pyramid coefficients \[27, 25\]. The fit to the histograms of the coefficients is made using the GGD probability density function:

\[
f(x; \alpha, \sigma) = \frac{\alpha}{2\beta \Gamma(1/\alpha)} \exp \left( -\frac{|x|^{\alpha}}{\beta} \right)
\]

where \(\alpha\) denotes the shape, \(\sigma\) the standard deviation, and \(\Gamma\) is the Gamma function defined as:

\[
\Gamma(t) = \int_0^\infty x^{t-1} \exp^{-x} \, dx
\]

The paired product coefficients are then modeled with an Asymmetric Gaussian Distribution (AGGD), with its probability density function defined as:

\[
f(x; v, \sigma_l, \sigma_r) = \begin{cases} 
\frac{v}{(\beta_l + \beta_r)\Gamma(1/v)} \exp \left( -\frac{|x|^{1/v}}{\beta_l} \right) & x < 0 \\
\frac{v}{(\beta_l + \beta_r)\Gamma(1/v)} \exp \left( -\frac{|x|^{1/v}}{\beta_r} \right) & x \geq 0
\end{cases}
\]

where

\[
\beta_l = \sigma_l \sqrt{\frac{\Gamma(1/v)}{\Gamma(3/v)}} \quad (21)
\]

\[
\beta_r = \sigma_r \sqrt{\frac{\Gamma(1/v)}{\Gamma(3/v)}} \quad (22)
\]

As before, \(v\) denotes the shape, and \(\sigma\) the standard deviation.

Following the work presented by Goodall and Bovik \[25\], we extract the GGD \((\alpha, \sigma)\) and the AGGD parameters \((v, \sigma_l, \sigma_r)\) by using the moment...
matching technique [41]. For the product pair, a mean parameter η is also computed, which is defined as:

\[
\eta = (\beta_r - \beta_l) \frac{\Gamma(2/v)}{\Gamma(1/v)}
\]  

(23)

Table 1: Feature summary for MSCN (f), Paired Products (pp), Paired Log-derivatives (pd), and Steerable Pyramid Coefficients (sp) for one scale

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Feature Description</th>
<th>Computation Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1 - f2</td>
<td>α and σ</td>
<td>GGD fit to I</td>
</tr>
<tr>
<td>f3 - f4</td>
<td>α_r - α_l and σ_r - σ_l</td>
<td>GGD fit to right and left halves of I</td>
</tr>
<tr>
<td>pp1 - pp4</td>
<td>v, η, σ_l, σ_r</td>
<td>AGGD fit to H</td>
</tr>
<tr>
<td>pp5 - pp8</td>
<td>v, η, σ_l, σ_r</td>
<td>AGGD fit to V</td>
</tr>
<tr>
<td>pp9 - pp12</td>
<td>v, η, σ_l, σ_r</td>
<td>AGGD fit to D1</td>
</tr>
<tr>
<td>pp13 - pp16</td>
<td>v, η, σ_l, σ_r</td>
<td>AGGD fit to D2</td>
</tr>
<tr>
<td>pd1 - pd2</td>
<td>α and σ</td>
<td>GGD fit to PD1</td>
</tr>
<tr>
<td>pd3 - pd4</td>
<td>α and σ</td>
<td>GGD fit to PD2</td>
</tr>
<tr>
<td>pd5 - pd6</td>
<td>α and σ</td>
<td>GGD fit to PD3</td>
</tr>
<tr>
<td>pd7 - pd8</td>
<td>α and σ</td>
<td>GGD fit to PD4</td>
</tr>
<tr>
<td>pd9 - pd10</td>
<td>α and σ</td>
<td>GGD fit to PD5</td>
</tr>
<tr>
<td>pd11 - pd12</td>
<td>α and σ</td>
<td>GGD fit to PD6</td>
</tr>
<tr>
<td>pd13 - pd14</td>
<td>α and σ</td>
<td>GGD fit to PD7</td>
</tr>
<tr>
<td>sp1 - sp2</td>
<td>α and σ</td>
<td>GGD fit to d_0^1</td>
</tr>
<tr>
<td>sp3 - sp4</td>
<td>α and σ</td>
<td>GGD fit to d_30^1</td>
</tr>
<tr>
<td>sp5 - sp6</td>
<td>α and σ</td>
<td>GGD fit to d_60^1</td>
</tr>
<tr>
<td>sp7 - sp8</td>
<td>α and σ</td>
<td>GGD fit to d_90^1</td>
</tr>
<tr>
<td>sp9 - sp10</td>
<td>α and σ</td>
<td>GGD fit to d_120^1</td>
</tr>
<tr>
<td>sp11 - sp12</td>
<td>α and σ</td>
<td>GGD fit to d_150^1</td>
</tr>
</tbody>
</table>

To address some asymmetric histograms, negative and positive MSCN coefficients are measured separately, which correspond to the left and right halves of the histograms. Hence, four parameters are extracted (α_l, σ_l, α_r, and σ_r) from the MSCN coefficients, and the differences in value between the left and right halves are used to capture the asymmetry. We then obtain 46 features per image, as summarized in Table 1. These features are computed over three scales, yielding a total of 138.
To have a clear visualization of the features and their clustering, they are projected into a bi-dimensional space using Principal Component Analysis (PCA). Figure 7 depicts the two-dimensional space of features for images from each scene of the image databases, and for each one of the fusion algorithms used. As it can be seen, the positions and variances of the clusters generated by the average, gradient pyramid, and SIDW fusion algorithms remain stable. Nonetheless, for Laplacian and ratio pyramid fusion algorithms the clusters change position significantly.

Figure 8 shows the features for all the images and fusion algorithms. We observe the formation of sub-clusters for each database, presumably due to differences between sensor technologies. OSU database LWIR images were taken with a ferro-electric thermal sensor that follows a non-linear function of luminance, which we believe affects the NSS extracted from the images. Nonetheless, as shown in Figure 8c, features from distorted images still appear to cluster away from the pristine images.

To compute our feature model, 154 pristine image pairs from the OSU
and TNO databases are used, their sizes range from 298 × 217 to 360 × 270. These are later fused (using the five aforementioned fusion algorithms), which yield a total of 770 fused pristine images. 138 features per image are then extracted, from which a pristine model is obtained. This model can be calculated by fitting the features to a multivariate Gaussian model (MVG) to obtain the maximum likelihood estimation of the parameters in a Gaussian mixture model by setting the number of groups $K$ as 1. The MVG consists of a mean vector $\mu$ and a covariance matrix $\Sigma$; with them, it is possible to evaluate the quality of an image by comparing the pristine model to the model of the degraded image. This procedure is further explained in subsection 4.2.

4. Quality Assessment of Fused LWIR and Visible Images

4.1. Subjective Study

In this section, we first describe a human perceptual study, which we use to create an opinion aware IQA model (later described in subsection 4.3), and to assess how well do our quality models for fused LWIR and visible images correlate with subjective judgments. Later, we report the procedure used for the study, and the processing of the opinion scores.

To avoid overloading a subject with too many images to evaluate, we selected a total of 25 pairs of pristine LWIR and visible images, including 11 pairs from the TNO database and 14 pairs from the MORRIS database. Images are studied with three fusion methods, three types of distortion and

Figure 8: A total of 138 features per image are projected in a 2D space using PCA with a cumulative explained variance of 0.9973. (a) Features of Visible, LWIR, and fused images for each scene. (b) Features of fused images for each scene. (c) Features of pristine and distorted images. The terms O1 and O2 refer to the scenes of the OSU database, and the terms DD, TD, UD refer to the scenes of the TNO database.
three levels of distortion. We chose the fusion methods that seem to preserve the clusters that represent each database: average, gradient pyramid, and shift-invariant discrete wavelet transform with Haar wavelet, as seen in Figure 7. The fusions of the image pairs presented are then affected by three levels of distortion and three types of distortion: additive white noise and blur in both LWIR and visible images, and non uniformity in the LWIR images. For AWGN and NU the distortion level is controlled using a standard deviation parameter $\sigma_{\text{AWGN}} = \sigma_{\text{NU}} = \{0.0025, 0.01375, 0.025\}$, and for blur a kernel of 15 by 15 pixels and a standard deviation $\sigma_{\text{Blur}} = \{1, 2, 3\}$.

We conducted this study with 27 voluntaries. Each person was asked to evaluate the images using a procedure we wrote with Matlab and the PsychToolbox \[42\]. A subject evaluates 150 single stimulus images along five testing sessions, meaning a total of 750. The test procedure follows the recommendations mentioned in \[43\], where the authors used a variant of the absolute category rating with hidden reference format (ACR-HR) from ITU-T Rec. P.910, in which each original image is included in the experiment but not identified as such. We also used the same screen resolution of $1024 \times 768$ pixels, while the stimulus images were displayed at their native resolution. The sequence started displaying a single stimulus image for 7 seconds as depicted in Figure 9, then a subject rated the image using a continuous sliding quality bar with labels "Bad," "Poor," "Fair," "Good," or "Excellent," as shown in Figure 10. It is also important to note that stimuli were continuously scored on the scale $[0, 100]$. In addition, we measured the illumination levels during the tests, which varied between 220 and 240 lux, making sure that they did not change significantly between sessions.

![Figure 9: Example stimulus.](image)
The obtained subjective scores are then processed to discount individual preferences of images and differences in image content, as explained in [43, 25]. First, we compute difference scores, which are next converted to Z-scores for each session and combined to build a matrix $z_{ij}$, which corresponds to the Z-score given by subject $i$ to image $j$. After obtaining the matrix, a subject rejection procedure must be done to discard scores from unreliable sources. This procedure is specified in ITU-R BT 500.11 recommendation, which determines that the scores given to a stimulus image are normally distributed if the kurtosis falls between the values of 2 and 4. If scores are normally distributed, a subject is rejected whenever more than 5% of his scores lie outside two standard deviations of the mean score for each image. If scores are not normally distributed, a subject is rejected whenever more than 5% of his scores fall outside the range of $\sqrt{20}$ standard deviations from the mean score for each image [43, 44].

Following this procedure, five outliers were found and removed. Then, the Z-scores are linearly rescaled to $[0, 100]$. The Difference Mean Opinion Scores (DMOS) for each image are computed as the mean of the rescaled Z-scores for the remaining 22 subjects as follows [43, 25].

Histograms for the DMOS are shown in Figure 11, which indicates a fairly broad distribution of the DMOS. Scores before subject rejection lie within the range [45, 80], while scores after subject rejection lie within [31, 74] yielding a wider range of visual quality. For DMOS obtained after subject rejection, it is also to be noted that most evaluations are distributed at half of the quality range.

4.2. Opinion-Distortion-Unaware Image Quality Analyzer

To evaluate the quality of a fused image our procedure extracts 138 features and fits them with a MVG model. The quality is then calculated as the Mahalanobis distance between the pristine feature model and the feature model of the distorted image:
Figure 11: Histograms of the DMOS in 15 equally spaced bins for (a) scores obtained before subject rejection and (b) scores obtained after subject rejection.

\[ Q_D(\mu_1, \mu_2, \Sigma_1, \Sigma_2) = \sqrt{(\mu_1 - \mu_2)^T \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\mu_1 - \mu_2)} \] (24)

where \( \mu_1, \mu_2 \) and \( \Sigma_1, \Sigma_2 \) are the mean vectors and covariance matrices of the models obtained with a standard maximum likelihood estimation procedure \[45\].

The quality model was tested using seven distortion levels of blur, JPEG compression, AWGN, and NU. Different set-ups were taken into account: distortion present in both LWIR and visible images, only in LWIR images, and only in visible images. Images with AWGN and NU were generated using standard deviations \( \sigma_{AWGN} = \sigma_{NU} = \{1 \times 10^{-4}, 0.0306, 0.1156, 0.2550, 0.4489, 0.6972, 1\} \), blurred images were generated with a Gaussian kernel with \( \sigma = \{1, 2, 3, 4, 5, 6, 7\} \), and JPEG compressed images have a quality parameter \( Q = \{100, 85, 70, 55, 40, 25, 10\} \). It must be noted that for an increase in the JPEG quality parameter, our quality measure \( Q_D \) should decrease.

Figure 12 shows that when both LWIR and visible images are distorted, the model shows a monotonic result for AWGN and NU, but for JPEG compression, it appears to decrease in a short range, presumably due to how little this distortion type affects the NSS of the images in comparison to other types. In addition, the resulting quality values for blur do not seem to
Figure 12: IQA model calculated from fused images, where both the LWIR and visible spectrum images are distorted with seven levels (for NU, only LWIR images are distorted). The set contains 15 images from the TNO database, different from the ones used to create the pristine model. Figures are ordered by fusion algorithm row-wise, and by type of distortion column-wise. The terms AVG, GP, LP, RP, SIDWT refer to the fusion methods Average, Gradient Pyramid, Laplacian Pyramid, Ratio Pyramid, and Shift Invariant Discrete Wavelet Transform, respectively.

depend on the level of degradation, perhaps due to the low spatial frequency of LWIR images. However, in the case of fusion by ratio pyramid, the quality estimate takes its highest value at early distortion levels and then decreases again.

To verify the performance of our model, we use the subjective scores ob-
Table 2: Description of the fusion performance models for night vision studied in [21]

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Theory</td>
<td>$Q_{MI}$</td>
<td>Normalized Mutual Information [9]</td>
</tr>
<tr>
<td></td>
<td>$Q_{TE}$</td>
<td>Fusion Metric-Based on Tsallis Entropy [10]</td>
</tr>
<tr>
<td></td>
<td>$Q_{NCIE}$</td>
<td>Nonlinear Correlation Information Entropy [11]</td>
</tr>
<tr>
<td>Image Features</td>
<td>$Q_G$</td>
<td>Gradient-Based Fusion Performance [12]</td>
</tr>
<tr>
<td></td>
<td>$Q_M$</td>
<td>Image Fusion Metric-Based on a Multiscale Scheme [13]</td>
</tr>
<tr>
<td></td>
<td>$Q_{SF}$</td>
<td>Image Fusion Metric-Based on Spatial Frequency [14]</td>
</tr>
<tr>
<td></td>
<td>$Q_P$</td>
<td>Image Fusion Metric-Based on Phase Congruency [15]</td>
</tr>
<tr>
<td>Structural Similarity</td>
<td>$Q_S$</td>
<td>Piella’s Metric [16]</td>
</tr>
<tr>
<td></td>
<td>$Q_C$</td>
<td>Cvejic’s Metric [18]</td>
</tr>
<tr>
<td></td>
<td>$Q_Y$</td>
<td>Yang’s Metric [17]</td>
</tr>
<tr>
<td>Human Perception</td>
<td>$Q_{CV}$</td>
<td>Chen-Varshney Metric [19]</td>
</tr>
<tr>
<td></td>
<td>$Q_{CB}$</td>
<td>Chen-Blum Metric [20]</td>
</tr>
</tbody>
</table>

Obtained in the previous section and compare it to the fusion quality models shown in Table 2 which were previously studied by Liu et al in [21]. We also evaluate the performance of individual feature groups ($f$, $pp$, $pd$, and $sp$) using the same approach of $Q_D$, obtaining $Q_f$, $Q_{pp}$, $Q_{pd}$, $Q_{sp}$ quality measures. First, we compute the Z-scores for each raw opinion score $s$, and then rescale it to fill the range from 0 to 1. We proceed to calculate Mean Opinion Scores (MOS) after removing outliers. On the other hand, the values of the quality estimates $Q$ that are inversely proportional to the MOS are adjusted as $Q' = 1/|Q|$, where the new values $Q'$ are used to compute the Spearman’s Rank Correlation Coefficient (SRCC), Pearson’s Linear Correlation Coefficient (LCC), and Root Mean Squared Error (RMSE). Models which show this inverse behavior include the spatial frequency metric $Q_{SF}$ [16], the Chen-Varshney metric $Q_{CV}$ [19], and our models based on NSS fea-
Table 3: SRCC, LCC, and RMSE between MOS and predicted MOS by different quality measures for the fused images evaluated

<table>
<thead>
<tr>
<th>Model</th>
<th>SRCC</th>
<th>LCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{MI}$</td>
<td>0.561</td>
<td>0.533</td>
<td>0.054</td>
</tr>
<tr>
<td>$Q_{TE}$</td>
<td>0.257</td>
<td>0.252</td>
<td>0.069</td>
</tr>
<tr>
<td>$Q_{NC1E}$</td>
<td>0.586</td>
<td>0.567</td>
<td>0.073</td>
</tr>
<tr>
<td>$Q_{G}$</td>
<td>0.398</td>
<td>0.411</td>
<td>0.068</td>
</tr>
<tr>
<td>$Q_{M}$</td>
<td>0.214</td>
<td>0.251</td>
<td>0.076</td>
</tr>
<tr>
<td>$Q_{SF}$</td>
<td>-0.271</td>
<td>-0.261</td>
<td>0.087</td>
</tr>
<tr>
<td>$Q_{P}$</td>
<td>0.585</td>
<td>0.585</td>
<td>0.059</td>
</tr>
<tr>
<td>$Q_{S}$</td>
<td>0.532</td>
<td>0.455</td>
<td>0.081</td>
</tr>
<tr>
<td>$Q_{C}$</td>
<td>0.331</td>
<td>0.289</td>
<td>0.098</td>
</tr>
<tr>
<td>$Q_{Y}$</td>
<td>0.525</td>
<td>0.463</td>
<td>0.084</td>
</tr>
<tr>
<td>$Q_{CV}$</td>
<td>-0.325</td>
<td>-0.353</td>
<td>0.123</td>
</tr>
<tr>
<td>$Q_{CB}$</td>
<td>-0.302</td>
<td>-0.328</td>
<td>0.095</td>
</tr>
</tbody>
</table>

Features. The ideal subjective quality measure should have a SRCC and LCC close to one and RMSE near to zero. The results from the evaluation are shown in Table 3. We observe that our model $Q_{D}$, and the feature groups $Q_{pd}$ and $Q_{pp}$ provide the highest correlations with subjective image quality judgments, whereas some models yield negative correlation coefficients.

4.3. Opinion Aware Fused Image Quality Analyzer

We employed a Support Vector Regression (SVR) algorithm to fit the NSS features to the DMOS and to obtain an opinion aware quality model $Q_{SVM}$. This method has been previously applied to IQA using NSS-based features [25, 27]. We utilized the LIBSVM package [46] to implement an $\epsilon$-SVR with a Radial Basis Function (RBF) kernel, and found the best-fitting parameters $C$ and $\gamma$ using 5-fold cross-validation. Figure 13 shows that the best-fitting parameters in this case were $C = 2^{11.5}$ and $\gamma = 2^{-5}$, which yielded a mean squared error $MSE = 7.59$. 

22
In order to test our model $Q_{SVM}$, we compare it with state-of-the-art fusion performance models and evaluate how it correlates to human judgments. Since $Q_{SVM}$ requires a training procedure to calibrate, we divide the data from the subjective study into random subsets, where 80% is used for training and 20% for testing, and take care not to overlap between train and test content. This is done to ensure that results do not depend on features extracted from content, rather than distortion. In order to account for a possible non linear relationship, the algorithm scores pass through the following logistic function to fit the objective models to DMOS:

$$Q'_j = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp^{-\frac{(Q_j - \beta_3)}{|\beta_4|}}} \quad (25)$$

where $Q_j$ is the objective quality value for stimulus image $j$. Each $\beta$ parameter is estimated with nonlinear least squares optimization using the Matlab function "nlinfit," to minimize the least squares error between DMOS$_j$ and the fitted scores $Q_j$. To facilitate numerical convergence, quality predictions are first linearly rescaled before performing optimization. We chose the initial $\beta$ parameters following the recommendation in [47]:

$$\beta_1 = \max(\text{DMOS}) \quad (26)$$
$$\beta_2 = \min(\text{DMOS}) \quad (27)$$
$$\beta_3 = \bar{Q} \quad (28)$$
$$\beta_4 = 1 \quad (29)$$
Table 4: Median SRCC, LCC, and RMSE between DMOS and predicted DMOS measured over 1000 iterations

<table>
<thead>
<tr>
<th>Model</th>
<th>Median SRCC</th>
<th>Median LCC</th>
<th>Median RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q_{MI}</td>
<td>0.185</td>
<td>0.241</td>
<td>9.785</td>
</tr>
<tr>
<td>Q_{TE}</td>
<td>0.042</td>
<td>0.023</td>
<td>10.082</td>
</tr>
<tr>
<td>Q_{NCIE}</td>
<td>0.194</td>
<td>0.270</td>
<td>9.713</td>
</tr>
<tr>
<td>Q_G</td>
<td>0.216</td>
<td>0.252</td>
<td>9.754</td>
</tr>
<tr>
<td>Q_M</td>
<td>0.585</td>
<td>0.798</td>
<td>6.085</td>
</tr>
<tr>
<td>Q_SF</td>
<td>0.159</td>
<td>0.194</td>
<td>9.891</td>
</tr>
<tr>
<td>Q_P</td>
<td>0.066</td>
<td>0.209</td>
<td>9.854</td>
</tr>
<tr>
<td>Q_S</td>
<td>0.267</td>
<td>0.303</td>
<td>9.614</td>
</tr>
<tr>
<td>Q_C</td>
<td>0.274</td>
<td>0.317</td>
<td>9.547</td>
</tr>
<tr>
<td>Q_Y</td>
<td>0.172</td>
<td>0.278</td>
<td>9.681</td>
</tr>
<tr>
<td>Q_{CV}</td>
<td>0.042</td>
<td>0.050</td>
<td>10.067</td>
</tr>
<tr>
<td>Q_{CB}</td>
<td>0.070</td>
<td>0.079</td>
<td>10.048</td>
</tr>
<tr>
<td>Q_D</td>
<td>0.264</td>
<td>0.405</td>
<td>10.032</td>
</tr>
<tr>
<td>Q_{SP}</td>
<td>0.095</td>
<td>0.095</td>
<td>10.041</td>
</tr>
<tr>
<td>Q_{PD}</td>
<td>0.224</td>
<td>0.259</td>
<td>9.742</td>
</tr>
<tr>
<td>Q_{PP}</td>
<td>0.107</td>
<td>0.164</td>
<td>9.949</td>
</tr>
<tr>
<td>Q_F</td>
<td>0.077</td>
<td>0.126</td>
<td>9.999</td>
</tr>
<tr>
<td>SVM</td>
<td>Q_{SMV}</td>
<td>0.932</td>
<td>0.961</td>
</tr>
</tbody>
</table>

We repeat this process across 1000 iterations, and compute the Spearman’s Rank Correlation Coefficient (SRCC), the Pearson’s Linear Correlation Coefficient (LCC), and the Root Mean Squared Error (RMSE) for all models, and present their median values in Table 4. Figure 14 depicts the scatter plot of the predicted scores given by our quality model Q_{SMV} versus DMOS for all the images evaluated in the subjective study described in subsection 4.1 along with the best-fitting logistic function. We observe that the model Q_{SMV} has the highest correlation to human scores, while other models provided poor correlation.

To evaluate whether these results are statistically relevant, we perform a Kruskal-Wallis test on each median value of SRCC between the DMOS and the quality measures (after nonlinear mapping). Table 5 tabulates the
Figure 14: Scatter plot of $Q_{SVM}$ prediction scores versus the DMOS for all images assessed in the subjective human study and the best fitting logistic function.

Table 5: Statistical Significance Matrix based on SRCC between DMOS and Quality Measures. A value of "1" indicates that the measure for the row is statistically better than the measure for the column, "0" means that it is statistically worse, and "-" means that it is statistically indistinguishable.

```
<table>
<thead>
<tr>
<th></th>
<th>$Q_{MI}$</th>
<th>$Q_{TE}$</th>
<th>$Q_{NCIE}$</th>
<th>$Q_{G}$</th>
<th>$Q_{M}$</th>
<th>$Q_{SF}$</th>
<th>$Q_{P}$</th>
<th>$Q_{S}$</th>
<th>$Q_{C}$</th>
<th>$Q_{Y}$</th>
<th>$Q_{CV}$</th>
<th>$Q_{CB}$</th>
<th>$Q_{D}$</th>
<th>$Q_{SP}$</th>
<th>$Q_{PD}$</th>
<th>$Q_{PP}$</th>
<th>$Q_{F}$</th>
<th>$Q_{SVM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{MI}$</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{TE}$</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{NCIE}$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{G}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{M}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{SF}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{P}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{S}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{C}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{Y}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{CV}$</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{CB}$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{D}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{SP}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Q_{PD}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{PP}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{F}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$Q_{SVM}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```

results of the statistical significance test. The null hypothesis is that the median correlation for the (row) algorithm is equal to median correlation for the (column) algorithm with a confidence of 95%. The alternate hypothesis is that the median correlation of the row is greater than or lesser than the mean correlation of the column. From Table 5, we conclude that $Q_{SVM}$ is highly competitive and statistically better with respect to all the other
quality algorithms tested. Although the second best model is $Q_M$ [13], our model is still statistically better, standing out the advantage of quality aware features to assess fusion algorithms.

5. Results and Discussion

We found that fused LWIR-visible images created with multi-resolution fusion algorithms such as Average, Gradient Pyramid, Laplacian Pyramid, Ratio Pyramid, and Shift Invariant Discrete Wavelet Transform, also possess statistical regularities, which can be used to describe distortions and estimate fused image quality. As shown through the histogram analysis, some groups of NSS allow to detect better some types of distortion than the rest: $PD$ coefficients detect JPEG compression and blur distortions, while $d_{10}^{o}$ and $d_{10}^{o}$ help measuring NU distortion. Furthermore, we developed a completely blind and an opinion aware image quality analyzer, which predicts human quality evaluations of fused LWIR and visible images more reliably than other state-of-the-art measurements.

Although our $Q_D$ quality model made a good predictor when assessing the preference of a fusion algorithm, it would be possible to improve it by selecting feature groups that are more suitable for quality description. The histogram analysis demonstrated that steerable pyramid coefficients do not capture well the effect of blur in images, possibly due to the low spatial frequency of LWIR images. Even though our experiments reveal that the Laplacian pyramid fusion method affects the naturalness, as shown by the histogram of the MSCN coefficients, this did not affect the results from the $Q_D$ quality model when evaluating a wide range of distortions.

One limitation of our research was the reduced availability of aligned LWIR and visible image pairs. The OSU database had little image content diversity, which did not make it suitable for the inclusion in the subjective human study. In addition, other consulted databases did not provide registered visible and infrared images. It was also found that the DMOS did not cover uniformly a broad range of perceptual quality; therefore, it is recommended to use a broader range of distortion for the subjective study. Nonetheless, the results provided by our opinion-aware quality analyzer in this database outperform other fusion quality algorithms.

Previous studies on fused image quality have not accounted for the presence of distortion in the source images, or even LWIR-specific distortions. Moreover, in some works the authors evaluate their proposed models with
a limited set of image pairs \cite{9, 10, 11, 12, 19}. Although the work in \cite{21} assessed AWGN and blur distortions, our approach considers the effects of NU, and proposes a fusion quality model that analyzes image degradation.

To the extent of our knowledge, there are no previous works in the analysis of NSS extracted from pristine and distorted fused LWIR and visible images. We believe that this work can serve as a solid starting point for further development of perceptual quality aware fusion algorithms.

6. Conclusion

NSS play an important role when analyzing distortions present in image fusion of LWIR and visible images, as they have previously helped modeling degradation in the visible and infrared spectrum. NSS demonstrated being potent descriptors, especially when assessing fused images affected by AWGN and NU. Therefore, we proposed an ODU and an OA fused image quality analyzers that outperformed current fusion quality indexes, correlating better to human subjective evaluations. Even though a broader range of distortion in images would have allowed a deeper insight, the lengthy human study we carried out allowed us to evaluate the impact of image degradation on judgments of perceptual quality, including that present only in LWIR spectrum. Our work shows a different paradigm when assessing the presence of distortions in the input source images of fusion algorithms, and future studies might be able to use the proposed models to evaluate other distortions present in infrared images, and scene statistics of fused images from other image sensors. Furthermore, fused LWIR-visible videos used in surveillance applications are of great interest. These videos could be modeled and studied with the aid of NSS to improve tracking algorithms.

Acknowledgments

The authors would like to thank Z. Liu for providing the codes of fused image quality evaluation algorithms. This research has been supported by COLCIENCIAS and its international mobility program ‘Convocatoria para el apoyo a proyectos con Norteamérica 2014’.

References


[4] M. Freebody, Consumers and cost are driving infrared imagers into new markets, Photonics spectra.


