QoE-Based Multi-Exposure Fusion in Hierarchical Multivariate Gaussian CRF

Rui Shen, Member, IEEE, Irene Cheng, Senior Member, IEEE, and Anup Basu, Senior Member, IEEE

Abstract—Many state-of-the-art fusion methods, combining details in images taken under different exposures into one well-exposed image, can be found in the literature. However, insufficient study has been conducted to explore how perceptual factors can provide viewers better Quality of Experience (QoE) on fused images. We propose two perceptual quality measures: perceived local contrast and color saturation, which are embedded in our novel hierarchical multivariate Gaussian Conditional Random Field (CRF) model, to illustrate improved performance for multi-exposure fusion. We show that our method generates images with better quality than existing methods for a variety of scenes.

Index Terms—Multi-exposure fusion, human perception, QoE, conditional random field, MAP estimation

I. INTRODUCTION

HUMAN perceptual factors have attracted increasing attention in research on visual communication techniques [1], [2]. The rationale behind this trend is to appeal to human observers with high visual quality images, videos, and graphics. Thus, it is important for applications to take the human visual system into consideration when designing image processing algorithms. Contrast and color are generally recognized to be important parameters [3], [4] in image quality. Motivated by these research findings, we study the visual impact of perceived local contrast and color saturation on fused images.

Multi-Exposure Fusion (MEF) is necessary because a conventional digital camera often produces images with insufficient details on a natural scene due to the incompatibility of its Low Dynamic Range (LDR) relative to the High Dynamic Range (HDR) of the scene. As shown in Figure 1(a), neither of the source images captured under different exposures is able to present all of the details in the scene, although the individual images combined contain complementary high-quality details of the scene like in Figure 1(b). This composition of local details can be achieved using MEF techniques [5], [6] or HDR imaging techniques [7]. In the HDR imaging approach, a radiance map of the scene is constructed, which allows a physical interpretation of the pixel values, but the process needs to adapt the reconstructed HDR image on consumer displays for viewing by applying Tone Mapping (TM) methods [8], [9]. In contrast, the MEF approach bypasses the HDR generation process and directly builds a visually appealing result based on certain perceptual criteria requiring minimal user intervention.

We propose a novel MEF method based on perceptual quality measures that exploit both contrast and color information. In order to deliver maximum image details, we model the probability for the human visual system to detect local contrast based on physiological findings. Incorporating these perceptual measures, the optimal fusion weights are then derived using Maximum A Posteriori (MAP) estimation in our Hierarchical Multivariate Gaussian Conditional Random Field (HMGCRF) model. The remainder of this paper is organized as follows. Section II presents the background of MEF. Section III explains our proposed Quality of Experience (QoE) based MAP-HMGCRF fusion method. Experimental results are summarized in Section IV. Conclusions and future work are given in Section V.

II. BACKGROUND OF MULTI-EXPOSURE FUSION

The history of image fusion research dates back to 1984 when Burt [10] proposed Laplacian pyramid-based fusion for binocular grayscale images. In 1993, Burt and Kolczynski [11] applied this method to fuse multi-exposure grayscale images. Mertens et al. [5] used local variation, saturation and well-exposedness based measures in a Laplacian pyramid-based fusion scheme for color images. Raman and Chaudhuri [12] estimated the fused pixel values by solving an optimization problem. Local variation and gradient based measures were used in [13] to infer the luminance components of the fused pixels in a Bayesian model. A gradient-directed MEF method [14] was proposed for dynamic scenes. In contrast, our focus is the study of perceptual factors on static scenes. In our previous work [6], a probabilistic fusion method was proposed, which applies local variation and neighborhood consistency computation. Although the technique generates high-quality results, we believe that images delivered by MEF techniques can be more visually appealing by considering human perceptual parameters. To address some common issues, which include loss of local details and poor color scheme leading to the loss of vividness, we propose using two perceptual quality measures (i.e., perceived local contrast and color saturation) to give a more accurate evaluation of pixel contributions, in order to achieve higher-quality fused images. After validating the effectiveness of the perceptual measures using our previous model [6], we then propose a more flexible fusion model, where the fusion weights are computed as the MAP estimate in a hierarchical multivariate Gaussian CRF model to further illustrate the effectiveness of the perceptual parameters.
III. QOE-BASED MULTI-EXPOSURE FUSION

A. Overview

Given a source image sequence, the contributions from individual pixels to the fused image are perceptually tuned by two locally-defined quality measures, i.e., perceived local contrast (Section III-B) and color saturation (Section III-C). First, physical contrasts are calculated for each pixel in the luminance channel. Visual responses to these physical contrasts, i.e., perceived contrasts, are then modeled using a transducer function followed by a psychometric function. These perceived contrasts, together with color saturation, are used in our MAP-HMGCRF model for pixel contribution evaluation, where the contributions are modeled as the MAP estimate of a multidimensional potential field (Section III-E).

B. Perceived Local Contrast

Earlier physiological studies of contrast sensitivity in three opponent color channels, i.e., black-white (luminance), red-green, and yellow-blue, show that luminance sensitivity is normally higher than chromatic sensitivity [15], which inspires us to employ a luminance contrast measure to help preserve details. Local contrast represents the perception of local luminance variations with respect to the surrounding luminance, and different measures of local contrast exist in the literature. Simple definitions like Weber contrast normally assume small targets on a large uniform background [16]. In order to deal with complex images of natural scenes in MEF, we modified the local band-limited contrast proposed by Peli [16], which defines local contrast as the ratio between the band-pass filtered image and the low-pass filtered image. We perform contrast calculation in the luminance channel of the LHS color space.

If we directly use Peli’s contrast in MEF, under-exposed regions, which are normally noisy, may produce stronger responses than well-exposed regions. This makes under-exposed regions contribute more to the fused image and reduce the overall brightness. Thus, if the local background luminance at a pixel is below a threshold \( \theta \), we weight its contrast by the background luminance to suppress noise. When \( \theta \) is no less than 0.2, the fused image is brighter and preserves more details. When \( \theta \) is above 0.4, the image shows less vivid colors. Therefore, we suggest using \( \theta \in [0.2, 0.4] \).

When combined with the Gaussian pyramid representation of a luminance image, we can construct a contrast pyramid.

Let \( C_{i,k}^n \) denote the weighted contrast at the i-th pixel location at level \( n \) of the Gaussian pyramid, where \( n \in [0, N_c - 1] \). Then, \( C_{i,k}^n \) can be calculated as:

\[
C_{i,k}^n = \begin{cases} 
G^n_{i,k} - [\phi \ast G^n_{i,k}], & [\phi \ast G^n_{i,k}]_2 < \theta; \\
(G^n_{i,k} - [\phi \ast G^n_{i,k}])/[\phi \ast G^n_{i,k}], & \text{otherwise}.
\end{cases}
\]

(1)

where \( G^n \) denotes the n-th level of the Gaussian pyramid and \( C_{i,k}^n \), the i-th coefficient in \( G^n \), and we take \( \phi \) as a \( 5 \times 5 \) Gaussian filter with variance 1. Figure 2 gives a comparison between our weighted and Peli’s contrasts. Peli’s contrast produces noisy responses in the under-exposed regions of the low-exposure image, which reduces the brightness of the fused image. This issue is resolved using our weighted contrast.

Furthermore, in order to obtain the best representative information from lower levels, the contrast magnitude \( C_{i,k}^n \) at a higher-level coefficient is determined as the maximum contrast magnitude among those associated with that coefficient and its corresponding lower-level coefficients.

1) Transducer and Psychometric Functions: The nonlinearity of human perception of contrast has been studied by many researchers [17]–[19]. According to [17], contrast perception can be considered as a two-stage procedure. In the first stage, the stimulus contrast is mapped to the internal/physiological response of the sensory system via a transducer function \( \mu \). In the second stage, the probability of correctly discriminating a stimulus with certain contrast from the standard stimulus with a fixed contrast \( C_s \) is expressed by a psychometric function \( \Psi \), where we take \( C_s = 0 \). A formal relationship between the psychometric function \( \Psi \) and the transducer function \( \mu \) was developed in [18], where \( \Psi \) is determined by \( \mu \) and the distribution of internal responses. We normalize \( \hat{C}_{i,k}^n \)'s to \([0, 1]\) before using it in \( \mu \) to fulfill the assumption of the stimulus contrast range in [18].

Because the most representative information is passed upwards along the contrast pyramids, the transducer and psychometric functions are only applied to level \( N_c - 1 \), which also reduces the computational cost. We adopt the transducer function proposed in [17]:

\[
\mu(\hat{C}_{i,k}^{N_c-1}) = \frac{(\hat{C}_{i,k}^{N_c-1}S_E)^p}{(\hat{C}_{i,k}^{N_c-1}S_t)^q} + Z,
\]

(2)

where \( S_E = 100 \) is a constant; \( S_t, p, q, Z \) are four free parameters, which we set to the mean values reported from...
the experiments in [17], i.e., $S_f = 75.70, p = 4.03, q = 3.59, Z = 24.87$. Two other forms of $\mu$ were also tested. The three-parameter function in [18] did not produce comparable results. Wilson's [19] transducer function for threshold and suprathreshold vision produced similar results. We adopted the results. Wilson’s [19] transducer function for threshold and three-parameter function in [18] did not produce comparable results.

3

$S$ contrast, i.e., information when the filter’s size and variance parameters are

ntween levels is sufficient to produce satisfactory fusion results in practice, we observe that saturation calculation performed only

r saturation components at every level. To be consistent with the first building a Gaussian pyramid and then calculating the

The results on one scene are presented in Figure 3. Instead of employing local variations in a non-linear function to indicate contrasts [6], we believe modeling the probability of the human visual system to perceive a given contrast will generate better perceptual quality, because this new modeling scheme offers a more accurate estimation of the amount of visual stimuli delivered from each image region, which leads to better detail preservation, as shown in the insets. Together with the color saturation measure, the fused image can exhibit more vivid colors (e.g., for the sky and the street lamp).

E. MAP-HMGCRF Model

In order to illustrate the quality contribution of the proposed perceived local contrast and color saturation measures, we incorporate these two measures in our previously published GRW model [6]. The results on one scene are presented in Figure 3. Instead of employing local variations in a non-linear function to indicate contrasts [6], we believe modeling the probability of the human visual system to perceive a given contrast will generate better perceptual quality, because this new modeling scheme offers a more accurate estimation of the amount of visual stimuli delivered from each image region, which leads to better detail preservation, as shown in the insets. Together with the color saturation measure, the fused image can exhibit more vivid colors (e.g., for the sky and the street lamp).

D. Perceptual Impact of the Proposed Quality Measures

In order to illustrate the quality contribution of the proposed perceived local contrast and color saturation measures, we incorporate these two measures in our previously published GRW model [6]. The results on one scene are presented in Figure 3. Instead of employing local variations in a non-linear function to indicate contrasts [6], we believe modeling the probability of the human visual system to perceive a given contrast will generate better perceptual quality, because this new modeling scheme offers a more accurate estimation of the amount of visual stimuli delivered from each image region, which leads to better detail preservation, as shown in the insets. Together with the color saturation measure, the fused image can exhibit more vivid colors (e.g., for the sky and the street lamp).
that measures the contribution from pixel and the

where

This minimization problem is equivalent to solving the linear

Here,

Then, the posterior density

follows a multivariate Gaussian distribution:

Then, the posterior density

is an element defined as

With local variation measure

With perceived local contrast measure

With both perceptual measures

are a K-D weight vector and a K-D pixel vector, respectively.

If \( \hat{p}_i \)'s obtained in Equation (6) exceed the dynamic range of the target device, they are truncated. If we model \( u_i \)'s as the K-D potential field \( x \) on a lattice graph and \( p_i \)'s the observed data \( D \), then estimating the fusion weights is equivalent to estimating the MAP configuration of an MGCRF. To fully specify the MGCRF, we also need to define the precision matrices and the potential boundary. We assume that the precision matrices \( \Sigma_{ik} \) and \( S_{ij} \) are identity matrices subject to individual scaling factors, i.e., \( \Sigma_{ik} = \gamma_1 \Sigma_{ik}, S_{ij} = \gamma_2 S_{ij} \). Here, \( I \) represents a \( K \times K \) identity matrix; \( \gamma_1 \) and \( \gamma_2 \) are data-independent scaling factors; and \( Y_{ik} \) and \( W_{ij} \) are data-dependent scaling factors defined as:

where \( \cdot \) denotes Euclidean distance; and \( \sigma \) is a free parameter. For the potential boundary, we assume that the maximum allowable potentials in each dimension are equal and so do the minimum allowable potentials, i.e., \( U_{\text{max},k} = \alpha_1, U_{\text{min},k} = \alpha_2, \forall k \). Setting \( \alpha_2 = 0 \) and with the identity precision matrix assumption, the MAP-MGCRF model degenerates to the GRW model in terms of steady-state probability calculation, but here we do not restrict the range of the boundary values.

3) \( \text{HMGCRF} \): In order to efficiently estimate the MAP configuration on a lattice graph, we construct an \( N_h \)-level hierarchical MGCRF and perform the calculation in a coarse-to-fine fashion. A coarser-level lattice graph is obtained by downsampling the finer-level graph by a factor of 2 in each dimension. A precision matrix \( \Sigma_{ik} \) between variable/node \( x_s \) and boundary vector \( b_k \) at a coarser-level MGCRF is obtained as a weighted average of the precision matrices of variables in the second-order neighborhood of \( x_s \)'s projection at the finer-level MGCRF. A precision matrix \( S_{st} \) between two adjacent variables \( x_s \) and \( x_t \) at a coarser-level MGCRF is obtained as the precision matrix with the minimum determinant among those defined in the common neighborhood of the projections.

The minimization problem is equivalent to solving the linear system \( Pf^* = -Qb \).

2) MGCRF for MEF: We consider that the fused image is derived as a pixel-wise weighted composition of the source images:

where \( \bar{p}_i \) and \( p_{i,k} \) denote the \( i \)-th pixels in the fused image and the \( k \)-th source image, respectively; \( u_{i,k} \) is a fusion weight that measures the contribution from pixel \( p_{i,k} \); and \( u_i \) and \( p_i \)


<table>
<thead>
<tr>
<th>Input</th>
<th>Size</th>
<th>Time (sec)</th>
<th>Memory (MB)</th>
<th>RMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>House</td>
<td>752 × 500 × 4</td>
<td>1.176</td>
<td>2.398</td>
<td>43</td>
</tr>
<tr>
<td>Chateau</td>
<td>1500 × 644 × 5</td>
<td>3.995</td>
<td>8.418</td>
<td>152</td>
</tr>
<tr>
<td>Belgium House</td>
<td>1025 × 769 × 9</td>
<td>5.362</td>
<td>7.449</td>
<td>148</td>
</tr>
<tr>
<td>Lamp</td>
<td>1600 × 1200 × 15</td>
<td>23.08</td>
<td>130.7</td>
<td>757</td>
</tr>
</tbody>
</table>

Table I: Comparison between solving an HMGCRF and directly solving an MGCRF.
of $x_s$ and $x_t$ at the finer-level MGCRF. At the coarsest level, the MAP estimate is obtained using a direct linear system solver, and then the solution is interpolated downwards along the hierarchy to the finest level.

We evaluated the performance of this HMGCRF with $N_h = 5$ on four multi-exposure sequences of increasing size. Compared with directly solving an MGCRF, solving an HMGCRF requires less time and memory but with good solution accuracy, as shown in Table I. With all the other settings the same, the hierarchical version took 44.081% of the time and 21.085% of the memory needed by directly solving an MGCRF and produced an average root mean squared error of only 1.027%.

IV. EXPERIMENTAL RESULTS

In this section, we summarize the evaluation results. Please refer to the supplementary material for more details and high-resolution images. Our method has eight free parameters, i.e., $\theta, N_c, N_h, \sigma, \gamma_1, \gamma_2, \alpha_1, \alpha_2$. We take $\theta = 0.3$, $N_c = 2$, $\gamma_1 = 1$. Let a source sequence contain $K M \times N$ images. Then, we compute $\sigma = 0.1K$, $\gamma = 0.2K\sqrt{MN}/\kappa$, $N_h = \lceil \log_2(\min(M,N)/\kappa) \rceil$, where $\kappa = 32$ is the maximum number of nodes allowed along the shorter dimension of the coarsest-level lattice. Depending on the size of the source sequence, the value of $N_h$ ranged between 4 and 6 in our experiments. We tested two sets of $\alpha_1$ and $\alpha_2$. In the first set, $\alpha_1 = 1$, $\alpha_2 = 0$, and we denote this algorithm as QBF-1. In the second set, we compute $\alpha_1 = 0.6 + \exp(-\bar{L})$, $\alpha_2 = 0.02\exp(-\bar{L})$, where $\bar{L}$ is the average luminance of $D$, and we denote this algorithm as QBF-2. These parameter settings were used in all experiments.

Six standard test sequences were used. The fusion results of the proposed QBF-1 and QBF-2 were compared with two other MEF methods, i.e., Probabilistic Fusion (PF) [6] and Exposure Fusion (EF) [5], which have previously demonstrated better performance than many other methods. In addition, two TM methods were also considered: the Photographic Tone Reproduction (PTR) local operator [8] and the iCAM06 operator [9], which have demonstrated good performance in various evaluations. The default parameter settings in PF, EF, and iCAM06 were used. The parameters in PTR were estimated by the method in [23]. The results by EF, PTR, and iCAM06, were generated by the programs provided by their respective authors. The HDR images for PTR and iCAM06 were generated using HDR reconstruction [7]. Both the objective and subjective evaluations were performed in a reference-free manner, where no ideal fused images were available to serve as references.
A. Objective Evaluation

1) Evaluation Using $Q_{AB/F}$: Two objective evaluation metrics were employed to assess the fusion quality. The first one is the $Q_{AB/F}$ metric [24], which is widely used in the image fusion literature to measure correctly transferred edge information in the luminance channel from source images to a fused image. This metric gives a performance score between [0, 1], where a higher score indicates better performance. Traditional fusion quality metrics, including $Q_{AB/F}$, were not designed for cases with more than two source images in MEF. Nevertheless, $Q_{AB/F}$ has been shown to be one of the most robust and consistent metrics [25]. Most evaluation metrics that are designed for the case of two source images, including the other two metrics recommended in [25] (i.e., Cvejic’s metric [26] and Yang’s metric [27]), largely depend on the calculation and manipulation of covariance (or similar statistics) between the two source images and/or between the two source images and the fused image. Therefore, it is relatively difficult to extend such metrics to cases with multiple source images. The advantage of $Q_{AB/F}$ is that it does not rely on calculating statistical score between two source images, and thus it can be directly extended to processes involving multiple source images, such as MEF. This metric has also been proven to correspond best with subjective tests among several other popular metrics [28]. Therefore, we adapted $Q_{AB/F}$ in our evaluation. To the best of our knowledge, our evaluation is the first attempt of extending a traditional fusion quality metric to MEF.

The fusion results of QBF-1 and QBF-2 on the six test scenes, along with objective evaluation results using $Q_{AB/F}$, are shown in Figure 4 and Figure 5. Although PTR and iCAM06 are not MEF methods, they are included in this evaluation for reference purposes only. All of the compared MEF methods successfully transferred most of the edge information, and they have very close performance according to $Q_{AB/F}$. On the average, QBF-2 has slightly better performance than the others.

2) Evaluation Using DRIVDP: To strengthen the evaluation capability of $Q_{AB/F}$, we incorporate the Dynamic Range Independent Visible Difference Predictor (DRIVDP) [29] to assess per-pixel fusion quality. DRIVDP evaluates visual local contrast distortions (i.e., loss of visible contrast, amplification of invisible contrast, and reversal of visible contrast) between images under a specific viewing condition and is widely used in the TM literature. Here, we use it to assess the visual distortions between a test image and each source image. We assume that the images were viewed on a typical LCD with a maximum luminance equivalent to $100\text{cd/m}^2$, a gamma value of 2.2, and a visual resolution of 30 pixels per degree at a viewing distance of 0.5 meter and that the peak contrast sensitivity of the viewer is $0.25\%$.

We chose two images from each of the six source sequences for this evaluation. The evaluation result on one sequence is given in Figure 6. The two source images with good exposures respectively for the bulb and the books are given in Figure 6(a). The distortion maps for each method are given in Figure 6(b)-(g), along with the fused images. In a distortion map, green, blue, red, and gray pixels indicate contrast loss, amplification, reversal, and no distortion, respectively. QBF-1 and QBF-2 are more effective in preserving local details and color schemes than the other methods. QBF-2 performs best in preventing contrast distortions for this sequence. Please note that contrast amplification is normally considered as one of the objectives in image fusion.

B. Subjective Evaluation

We also conducted a subjective evaluation, where thirteen subjects (8 males and 5 females) aged between 25 and 35 participated. All of the subjects had normal or corrected-to-normal vision and were non-experts in the field of MEF or TM. The test was performed under normal lighting conditions. For each scene, the results of different methods were anonymized and placed side by side in different orders, along with the source sequence. No other reference image, either manually or automatically fused or tone-mapped, was provided to guide/influence a subject’s judgement. Subjects were asked to rank the results on a scale of 5 (best) to 0 (worst) in four categories: global contrast, details, colors, and overall appearance. These four criteria were also considered in [30]. The global contrast criterion measures the global luminance variations. The details and the colors criteria measure the local details and colors reproduced and/or enhanced from the source images, respectively. The overall appearance criterion measures the overall impression of a fused or tone-mapped image.

The average ranking scores of different algorithms under each criterion are reported in Figure 7. QBF-2 performs consistently well compared to the other methods. It has the best performance under all criteria on 5 out of 6 scenes, and shows similar performance to QBF-1 on the National Cathedral sequence. QBF-1 and EF have similar performance on the average though QBF-1 offers clearly better detail reproduction, followed by PF, iCAM06, and PTR. The evaluation results demonstrate that our proposed fusion method is capable of producing high-quality fused images with quality comparable, or even better quality in many cases, to tone-mapped HDR images and images by other fusion methods.

Even though photo-realistic appearance may not be always achieved due to detail maximization (e.g., the appearance of
Fig. 6. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Lamp sequence using DRIVDP. The two source images give good exposures for the bulb and the books, respectively. In a distortion map, green, blue, red, and gray pixels indicate contrast loss, amplification, reversal, and no distortion, respectively. QBF-1 and QBF-2 are more effective in preserving local details and color schemes than the others. For the bulb, QBF-2 shows the least distortion, followed by QBF-1, PTR, PF, EF, and iCAM06. For the books, QBF-2 shows the least distortion, followed by EF, PF, QBF-1, iCAM06, and PTR.

Fig. 7. Average ranking scores of different algorithms in the subjective evaluation. The global contrast criterion measures the global luminance variations. The details and the colors criteria measure the local details and colors reproduced and/or enhanced from the source images, respectively. The overall appearance criterion measures the overall impression of a fused or tone-mapped image. Our QBF-2 has the best performance under all four criteria for 5 out of 6 scenes, and shows similar performance to our QBF-1 on the other scene. QBF-1 and EF have similar performance on the average though QBF-1 gives better detail reproduction, followed by PF, iCAM06, and PTR.
Subjective details QAB/F metric Cvejac’s metric Yang’s metric
Quality score
QBF-1 QBF−2 PF EF PTR iCAM06

and the $Q^{AB/F}$ scores in Figure 9. From this plot, we can see that $Q^{AB/F}$ provides relatively better correspondence with the subjective study. We then applied Cvejic’s metric, Yang’s metric, and $Q^{AB/F}$ metric to the other five subsequences used in the DRIVDP-based evaluation. $Q^{AB/F}$ metric produced better correspondence with the details criterion in the subjective evaluation than the other two metrics.

V. Conclusion

In this paper, we proposed a novel fusion algorithm based on perceptual quality measures, i.e., perceived local contrast and color saturation. To the best of our knowledge, this is the first time that the modeling of the probability for human eyes to detect local contrast is introduced to multi-exposure fusion, which helps us achieve maximum local detail preservation. A hierarchical multivariate Gaussian conditional random field model was proposed to effectively integrate the perceptual quality measures. Experiments demonstrated better performance of our algorithm compared to other methods. In future work, we will analyze the applicability of other perceptual quality measures in multi-exposure fusion.

ACKNOWLEDGMENT

The authors would like to thank Dr. Z. Wang, University of Waterloo, for providing the code for universal image quality index and SSIM image quality index, which was used in implementing Cvejic’s and Yang’s fusion quality metrics. The authors would also like to thank the reviewers for their constructive comments and suggestions.

REFERENCES


