EFFICIENT SUPER-RESOLUTION DRIVEN BY SALIENCY SELECTIVITY

Nabil G. Sadaka and Lina J. Karam

School of Electrical, Computer, & Energy Engineering, Arizona State University, Tempe, AZ 85287
{nabil.sadaka, karam}@asu.edu

ABSTRACT

This paper presents a low-complexity saliency detector targeted towards efficient selective Super-Resolution (SR). As a result, an improved efficient ATtentive-SELective Perceptual (AT-SELP) framework is presented. The proposed AT-SELP scheme results in a reduced computational complexity for iterative SR algorithms without any perceptible loss in the desired enhanced image/video quality. A perceptually significant set of active pixels is selected for processing by the SR algorithm based on a local contrast sensitivity threshold model and the proposed low complexity saliency detector. Simulation results show that the proposed AT-SELP scheme results in a 15-40% reduction in computational complexity over an efficient Selective Perceptual (SELP) SR scheme without degradation in the visual quality.

Index Terms—Visual attention, Super-resolution, MAP-estimation, Contrast sensitivity, Masking.

1. INTRODUCTION

Due to the vast advancement of High Definition (HD) display technologies and digital multimedia delivery methods, High Resolution (HR) image and video content has become the backbone of many industries and applications. HR digital multimedia content is inevitably gaining presence in many applications such as mobile telephony, HDTV, medical imaging, border surveillance, military reconnaissance, and manufacturing defects detection. Super-Resolution (SR) techniques are widely used resolution enhancement solutions that overcome limitations on quality and cost of image hardware acquisition and content delivery.

Multi-frame SR techniques, as described in [1], exploit extra information from multiple neighboring frames of the same scene in a degraded Low Resolution (LR) sequence to estimate a High-Resolution (HR) unaliased and deblurred image. In [2, 3], Hardie et al. proposed a MAP-based SR solution by imposing a piecewise smoothness assumption on the image prior. In [2], an iterative gradient descent solution is used for simultaneously estimating the registration parameters and the HR image. Additionally, in [3], an iterative conjugate gradient descent solution is presented to estimate the HR image. In order to increase the computational efficiency of SR solutions, the two-stage Fusion-Restoration (FR) SR estimation method was proposed in [4]. A median shift and add operation is used to fuse the LR frames on the HR grid, followed by a regularized $L_1$ error minimization that is performed using an iterative gradient-descent deblurring-interpolation step. In [5], Hardie et al. proposed an SR approach by using subpixel registration to fuse the LR frames on the HR grid and a one-step adaptive Wiener filter to reconstruct the HR image. Although a one-step Wiener filtering approach is inherently more computationally efficient in nature, the quality of the resulting SR image is highly dependent on the statistical modeling and suffers from limited reconstruction accuracy.

Although iterative [2-4] techniques provide a relatively good reconstruction quality, they still suffer from high computational complexity due to the inherent high dimensionality of the SR inverse problem. Towards this issue, a selective approach of SR estimators [6-9] is introduced to reduce the dimensionality or the computational complexity of popular SR algorithms without a noticeable degradation in the desired HR reconstruction quality. These selective algorithms detect only a subset of active pixels to be processed in an iterative SR solution based on a local significance measure. In [6], local gradient thresholds are used to detect the active pixels for SR processing. This technique is impractical due to manually tweaking the gradient thresholds for each image differently in order to attain the best desired SR quality. In [7], a set of significant pixels is determined adaptively using an automated perceptual detection mechanism without any manual tuning. Due to Human Visual Attention, attended regions are processed at high visual acuity, hence details present in these regions should be reconstructed with higher accuracy than those present in non-attended areas. In [8, 9], a visual attention model proposed by Itti et al. [10], is used to further reduce the perceptually detected pixels that are processed while maintaining the desired visual quality.

However, the SR schemes of [8, 9] assume that the saliency-based visual attention (VA) information is already computed and stored offline, thus ignoring the high computational overhead introduced by the adopted complex VA model [10]. Towards an effective SR solution, we propose a low complexity saliency detector designed for efficient attentively selective SR estimators. Consequently, an improved ATtentive-SELective Perceptual (AT-SELP) framework is presented in this work in order to reduce the computational complexity of iterative Super-Resolution (SR) algorithms without any perceptible loss in the desired enhanced image/video quality.

This paper is organized as follows. Section 2 presents the proposed attentive-selective perceptual (AT-SELP) SR framework. The proposed perceptual-based saliency detector is described in Section 3. Section 4 presents an attentive selective MAP-based SR scheme based on the proposed AT-SELP framework. Simulation results are presented in Section 5.

2. ATTENTIVE SELECTIVE SR FRAMEWORK

Consider the SR observation model of $M$ LR frames of size $N_1 \times N_2$ pixels. Then, the values of the pixels in the $k^{th}$ LR frame can be expressed in matrix notation as:

$$y_k = W_k z + n_k$$

(1)
where $n_k$ is the additive noise, $W_k$ is the sub-pixel-accuracy warping, blurring and downsampling matrix, and $z$ is the undegraded HR image of size $l_1N_1 \times l_2N_2$ where $(l_1, l_2)$ are the upsampling factors in each direction. In an underdetermined system, estimating the HR image, $z$, from the LR observations, $y_k$, $k = [1, 2, ... M]$, is commonly formulated as an optimization problem minimizing an error criteria and a regularization term. Thus, the SR problem can be represented as:
\[
\hat{z} = \arg \min \| f(z) \|
\]
where the cost function $f(x)$ is of the form:
\[
f(z) = \sum_{k=1}^{M} E(y_k, W_k z) + \rho R(z)
\]
In (3), $E(.)$ is the error term in function of the LR observations $y_k$ and HR image $z$, $R(.)$ is the regularization term in function of $z$, and $\rho$ is a tuning parameter. In [2-4], the gradient steepest descent solution is adopted for an iterative solution of (2). At each iteration of the SR estimation, all the pixels are processed on an HR grid inclusively, thus, demanding high computational requirements in solving the inverse problem, which makes these methods impractical in real-time environments or under platforms of low computational specifications. Consequently, in order to minimize the computational complexity, the selective SR framework as presented in [7-9], processes only a subset of active pixels at each iteration of the SR solution. In the selective SR framework, the gradient descent iterative solution [2-4] of the SR problem (2) is modified as follows:
\[
\hat{z}_{n+1} = \hat{z}_{n} - \beta_n \cdot M \cdot \nabla f(\hat{z}_n)
\]
where $\beta_n$ is the step size in the direction of the gradient, $\nabla f(\hat{z}_n)$, that is computed at each iteration, $n$, and $M$ is a binary selective mask that signals the active pixel locations that need to be processed at each iteration. In [7], a selective perceptual mask is presented such that active pixels are detected by exploiting the contrast sensitivity of the Human Visual System (HVS). Moreover, in [8, 9], the high-complexity saliency-based visual attention model of [10] is adopted to further reduce the set of active pixels. This work presents an efficient ATtentive-SELPective Perceptual (AT-SELP) SR framework that integrates a new low-complexity perceptual-based saliency detector.

Fig. 1 shows a block diagram of the proposed AT-SELP framework. Initially, an estimate of the HR image, $\hat{z}_0$, is obtained by interpolating the target LR frame to be reconstructed. The first phase of the SR algorithm processes the perceptually active pixels determined by a contrast sensitivity mask, $M_p$. Then, in the second phase of the SR estimation, only the subset of active pixels that is determined to be salient by the selective attention mask, $M_s$, is iterated upon. The process of updating the HR estimates of the perceptual/attentive active pixels continues until the system stabilizes, i.e. until $\|M_p (\hat{z}_{n+1} - \hat{z}_n)\|/\|M_p \hat{z}_n\| < \epsilon$, in the attentive active region and $\|M_p (\hat{z}_{n+1} - \hat{z}_n)\|/\|M_p \hat{z}_n\| < s \cdot \epsilon$ in the perceptual non-attentive active region, where $s$ is a scaling factor greater than 1 and $\epsilon$ is a predetermined threshold which represents the desired accuracy of the SR algorithm.

Any saliency-based VA model, such as [10-12], can be adopted to detect the attentive mask, $M_s$. In [8, 9], the proposed AT-SELP SR methods ignore the extremely high computational overhead introduced by the adopted saliency based VA model of [10] and thus assume that the VA saliency information is already computed and stored offline. Although these methods provided relatively high-quality SR reconstruction, a low complexity saliency detector is needed to deem this framework of any practical value. Thus, in this work, a low complexity saliency-based detector is proposed for attentive selective SR, as described in Section 3.

Masking thresholds are levels above which a human can start distinguishing a stimulus or distortion. The contrast sensitivity threshold is the measure of the smallest contrast, or Just Noticeable Difference (JND), that yields a visible signal over a uniform background [7]. Ferzli et al. [7] proposed a model where the contrast sensitivity thresholds are computed in the spatial domain using a sliding window block of size $N_{bhik} \times N_{bhik}$. This is done by first computing the contrast sensitivity threshold for a uniform block having a mean grayscale value equals to 128, and then calculating the threshold for any block having an arbitrary mean intensity using the approximation model presented in [13]. The contrast sensitivity threshold of a block in the spatial domain is computed as:
\[
t_{\text{JND}} = t_{128} \left( \frac{\gamma_M}{\gamma_{L_{\text{max}} - L_{\text{min}}}} \right)^{\alpha_T}
\]
where $\gamma_M$ is the total number of grey scale levels, and $L_{\text{min}}$ and $L_{\text{max}}$ are the minimum and maximum display luminances, respectively. In (5), $T$ is evaluated based on the parametric model derived by Ahumada and Peterson [14] using a parabolic approximation.

Once the threshold at a grayscale value 128, $t_{128}$, is calculated using (5), the Just Noticeable Difference (JND) thresholds for the other grayscale values are approximated using a power function [13] as follows:
\[
t_{\text{JND}} = t_{128} \left( \frac{\gamma_M}{\gamma_{L_{\text{max}} - L_{\text{min}}}} \right)^{\alpha_T}
\]
where $l$ is the pixel intensity at location $[r_1, r_2]$ and $\alpha_T$ is a correction exponent set to 0.649 [13]. In the proposed AT-SELP scheme, the obtained local $t_{\text{JND}}$ thresholds are used in generating the attentive mask, $M_a$, in addition to the perceptual mask, $M_p$, in order to select the candidate pixels to be super-resolved for each HR estimate.

The perceptual mask, $M_p$, is generated by computing the difference between the center pixel of a 3x3 sliding window with its 4 cardinal neighbors. If any of the 4 absolute differences is greater than the local computed $t_{\text{JND}}$, then the corresponding center
pixel location is flagged as a perceptually active pixel for SR processing. The attentive mask, \( M_a \), is generated by first creating an initial saliency map based on the already computed contrast sensitivity thresholds \( t_{\text{JND}} \), and then finding the most significantly attended region according to a probability of saliency detection. The initial saliency map, \( S_{\text{JND}} \), is computed by weighting the maximum of the computed sliding window differences with the corresponding locally computed \( t_{\text{JND}} \). Then, a saliency detection rule is applied by finding a threshold, \( C \), above which the probability of saliency detection is equal to \( \tau \% \) as follows:

\[
Pr(S_{\text{JND}} > C) = \tau \%
\]

The salient active pixels, corresponding to pixels at which \( S_{\text{JND}} > C \), are then flagged as attentive locations for SR processing in the proposed AT-SELP framework. Note that the computational overhead for generating the perceptual and attentive masks is minimal compared to the complex and computationally demanding saliency-based VA methods proposed in [10-12].

### 4. ATTENTIVE SELECTIVE MAP SR

The proposed AT-SELP SR framework (Sections 2 and 3) is very flexible in that any iterative SR estimation algorithm can be easily integrated in it. Due to space limitation, we present in here the application of the proposed framework to an iterative MAP-based SR method [2]. A similar treatment and performance are obtained when applied to fusion-restoration methods [4].

In order to illustrate the significant reduction in computations for MAP-SR techniques, the popular algorithm presented by Hardie et al. [2] is integrated into the proposed AT-SELP SR framework. In [2], an iterative gradient descent optimization approach is used to update the HR estimate \( \hat{z} \). Following (4) within the proposed AT-SELP framework, the SR iterative process is modified as follows:

\[
\hat{z}_{n+1} = \hat{z}_n - \beta_n M_a \left( \frac{1}{\sigma_n^2} W^T(Wz_n - y) + \frac{1}{2} C_z^{-1}z_n \right)
\]

where \( \sigma_n^2 \) is the noise variance, \( C_z \) is the covariance matrix of \( z \), \( \beta_n \) is the step size at iteration \( n \). Initially, an estimate of the HR image, \( \hat{z}_0 \), is obtained by interpolating one of the corresponding LR frames. The detection mask, \( M_a \), is calculated as described previously in Sections 2 and 3.

### 5. SIMULATION RESULTS

The proposed JND-based saliency detection map is integrated in an AT-SELP-MAP SR algorithm and tested on a simulated sequence of images. In this case, a single HR image is passed through the SR degradation model to generate a sequence of blurred, shifted, and noisy LR images. This is done by shifting the reference 256x256 HR image, blurring the shifted images with an averaging filter of size 4x4, and sub-sampling these by a factor of 4 in each direction to generate sixteen 64x64 LR frames. Then, an additive Gaussian noise of variance 16 is added to the resulting LR sequence. The images used in this experiment are ‘Clock’, ‘Cameraman’ and ‘Lena’ from the USC-SIPI image database [15], and ‘Monarch’ from the LIVE Image Quality Assessment Database [16].

The simulations parameters of all the compared SR methods are set to \( \lambda=150, \tau=20\% \), \( \epsilon=0.0001 \), \( s=50 \) and a maximum of 20 iterations are performed.

Table 1 presents the significant reduction in the total number of pixels that are processed by the proposed AT-SELP-MAP method versus the baseline MAP [2] and the selective efficient SELP-MAP [7] super-resolution methods. Based on these results (Table 1), around 15-40% increase in computational efficiency is achieved by the proposed method over the efficient SELP-MAP method. A 52-67% increase in computational efficiency is achieved by the proposed scheme when compared to the baseline MAP method. Additionally from Table 1, comparable PSNR values are attained while significantly less active pixels are processed by the proposed method. Fig. 2 clearly shows the preserved desired visual quality among the baseline MAP SR [2], the SELP-MAP SR [7], and the AT-SELP-MAP SR with the proposed low complexity JND-based saliency detector. It can also be seen in Fig. 2 that the selective SR methods (Figs. 2(c) & (d)), as compared to the baseline non-selective results (Fig. 1(b)), result in a better denoising of the smooth areas where noise is easily noticeable, such as the sky area. This is due to the perceptual detector, \( M_a \), and the saliency detector, \( M_s \), locally adapting to the different image content leading to a content adaptive SR reconstruction.

For further comparisons, different saliency maps generated from existing VA models presented in [10-12] are also integrated and tested in the proposed AT-SELP-MAP SR framework. In [10], Itti et al. compute a saliency map using center-surround differences of intensity and orientation between different Gaussian scales. In [11], a saliency map is computed using low-level features of patch luminance and contrast, and bandpass outputs of patch luminance and contrast. In [12], a Frequency Tuned Saliency (FTS) approach using difference of Gaussians of luminance intensity is proposed to generate saliency information. Similar pixel savings as the proposed JND-based saliency detector can be achieved by these existing different VA methods [10, 11, 12]. However, those VA methods are more complex than the proposed JND-based saliency detector scheme, and the saliency detection needs to be computed offline for a real-time application of the AT-SELP SR approach. The proposed JND-based detection method does not add extra computational overhead since it reuses the JND thresholds computed in the first phase of the AT-SELP SR framework. Fig. 3 shows the attention mask, \( M_s(\text{JND}) \), that is generated using the proposed JND-based saliency detection scheme. For comparison, Fig. 3 also shows the attention masks, \( M_i(\text{ITTI}) \), \( M_i(\text{GAFFE}) \), and \( M_i(\text{FTS}) \), that are generated from the computed saliency maps of ITTI [10], GAFFE [11], and FTS [12], respectively. As shown in Fig. 3, the attention mask generated by the proposed low-complexity JND-based detector is better adapting to information essential to SR processing such as edges and perceived noise over smooth areas as compared to the existing schemes.

### Table 1. Performance results in terms of PSNR and Pixel Savings.

<table>
<thead>
<tr>
<th>Monarch</th>
<th>PSNR (dB)</th>
<th>Pixel Savings</th>
<th>PSNR (dB)</th>
<th>Pixel Savings</th>
<th>PSNR (dB)</th>
<th>Pixel Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>18.18</td>
<td>0%</td>
<td>24.48</td>
<td>0%</td>
<td>22.50</td>
<td>0%</td>
</tr>
<tr>
<td>MAP [2]</td>
<td>21.14</td>
<td>0%</td>
<td>28.72</td>
<td>0%</td>
<td>25.31</td>
<td>0%</td>
</tr>
<tr>
<td>SELP-MAP [7]</td>
<td>21.87</td>
<td>28.00%</td>
<td>29.37</td>
<td>52.29%</td>
<td>25.65</td>
<td>13.61%</td>
</tr>
<tr>
<td>AT-SELP-MAP</td>
<td>21.81</td>
<td>60.56%</td>
<td>29.25</td>
<td>67.57%</td>
<td>25.55</td>
<td>52.41%</td>
</tr>
</tbody>
</table>

Based on these results (Table 1), around 15-40% increase in computational efficiency is achieved by the proposed method over the efficient SELP-MAP method. A 52-67% increase in computational efficiency is achieved by the proposed scheme when compared to the baseline MAP method. Additionally from Table 1, comparable PSNR values are attained while significantly less active pixels are processed by the proposed method. Fig. 2 clearly shows the preserved desired visual quality among the baseline MAP SR [2], the SELP-MAP SR [7], and the AT-SELP-MAP SR with the proposed low complexity JND-based saliency detector. It can also be seen in Fig. 2 that the selective SR methods (Figs. 2(c) & (d)), as compared to the baseline non-selective results (Fig. 1(b)), result in a better denoising of the smooth areas where noise is easily noticeable, such as the sky area. This is due to the perceptual detector, \( M_a \), and the saliency detector, \( M_s \), locally adapting to the different image content leading to a content adaptive SR reconstruction.

For further comparisons, different saliency maps generated from existing VA models presented in [10-12] are also integrated and tested in the proposed AT-SELP-MAP SR framework. In [10], Itti et al. compute a saliency map using center-surround differences of intensity and orientation between different Gaussian scales. In [11], a saliency map is computed using low-level features of patch luminance and contrast, and bandpass outputs of patch luminance and contrast. In [12], a Frequency Tuned Saliency (FTS) approach using difference of Gaussians of luminance intensity is proposed to generate saliency information. Similar pixel savings as the proposed JND-based saliency detector can be achieved by these existing different VA methods [10, 11, 12]. However, those VA methods are more complex than the proposed JND-based saliency detector scheme, and the saliency detection needs to be computed offline for a real-time application of the AT-SELP SR approach. The proposed JND-based detection method does not add extra computational overhead since it reuses the JND thresholds computed in the first phase of the AT-SELP SR framework. Fig. 3 shows the attention mask, \( M_s(\text{JND}) \), that is generated using the proposed JND-based saliency detection scheme. For comparison, Fig. 3 also shows the attention masks, \( M_i(\text{ITTI}) \), \( M_i(\text{GAFFE}) \), and \( M_i(\text{FTS}) \), that are generated from the computed saliency maps of ITTI [10], GAFFE [11], and FTS [12], respectively. As shown in Fig. 3, the attention mask generated by the proposed low-complexity JND-based detector is better adapting to information essential to SR processing such as edges and perceived noise over smooth areas as compared to the existing schemes.
6. REFERENCES


