Efficient Perceptual Super-Resolution

by

Nabil Sadaka

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2011 by the
Graduate Supervisory Committee:

Lina J. Karam, Chair
Andreas S. Spanias
Antonia Papandreou-Suppappola
Glen P. Abousleman
Michael Goryll

ARIZONA STATE UNIVERSITY
December 2011
ABSTRACT

Super-Resolution (SR) techniques are widely developed to increase image resolution by fusing several Low-Resolution (LR) images of the same scene to overcome sensor hardware limitations and reduce media impairments in a cost-effective manner. When choosing a solution for the SR problem, there is always a trade-off between computational efficiency and High-Resolution (HR) image quality. Existing SR approaches suffer from extremely high computational requirements due to the high number of unknowns to be estimated in the solution of the SR inverse problem. This thesis proposes efficient iterative SR techniques based on Visual Attention (VA) and perceptual modeling of the human visual system.

In the first part of this thesis, an efficient ATtentive-SELEctive Perceptual-based (AT-SELP) SR framework is presented, where only a subset of perceptually significant active pixels is selected for processing by the SR algorithm based on a local contrast sensitivity threshold model and a proposed low complexity saliency detector. The proposed saliency detector utilizes a probability of detection rule inspired by concepts of luminance masking and visual attention. The second part of this thesis further enhances on the efficiency of selective SR approaches by presenting an ATtentive (AT) SR framework that is completely driven by VA region detectors. Additionally, different VA techniques that combine several low-level features, such as center-surround differences in intensity and orientation, patch luminance and contrast, bandpass outputs of patch luminance and contrast, and difference of Gaussians of luminance intensity are integrated and analyzed to illustrate the effectiveness of the proposed selective SR frameworks. The proposed AT-SELP SR and AT-SR frameworks proved to be flexible by integrating a Maximum A Posteriori (MAP)-based SR algorithm as well as a fast two-stage Fusion-Restoration (FR) SR estimator. By adopting the proposed selective
SR frameworks, simulation results show significant reduction on average in computational complexity with comparable visual quality in terms of quantitative metrics such as PSNR, SNR or MAE gains, and subjective assessment. The third part of this thesis proposes a Perceptually Weighted (WP) SR technique that incorporates unequal weighting parameters in the cost function of iterative SR problems. The proposed approach is inspired by the unequal processing of the Human Visual System (HVS) to different local image features in an image. Simulation results show an enhanced reconstruction quality and faster convergence rates when applied to the MAP-based and FR-based SR schemes.
To my parents, George and Samiyya
I am deeply indebted to my family for their continued support and unwavering faith in me. To my parents, George and Samiyya Sadaka, thank you for setting me on a path of knowledge and persistence. It truly had profound effects on who I am today. My deepest gratitude to my sister and brother, Mariam and Michel Sadaka, for their infinite support and understanding.

I would like to acknowledge my dissertation advisor Dr. Lina Karam. Thank you for your guidance, enthusiasm, and inspiration during my PhD years and for your great effort for meetings and discussions.

I am also grateful to Dr. Glen Abousleman for providing constant funding and support for many years. Thank you for giving me the opportunity to fill the gap between research and product development.

I want to thank Dr. Joseph Palais for giving me the opportunity to work as a teaching assistant. This has been one of the more rewarding experiences I have had during my time at ASU.

I also wish to thank the other members of my committee, Dr. Andreas Spanias, Dr. Antonia Papandreou, and Dr. Michael Goryll for serving on my committee.

I would like to acknowledge all the students in the IVU lab for being great friends and making the lab such a pleasant working environment.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Difficulties of High-Resolution Imaging Systems</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Super-Resolution Applications</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Overview of Efficient Super-Resolution Problem-Solution</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Thesis Organization</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>GENERAL SUPER-RESOLUTION BACKGROUND AND FORMULATION</td>
<td>14</td>
</tr>
<tr>
<td>2.1</td>
<td>Single-Frame Super-Resolution</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Multi-Frame Super-Resolution</td>
<td>17</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Bayesian Super-Resolution Methods</td>
<td>19</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Regularized-Norm Minimization Super-Resolution Methods</td>
<td>19</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Fusion-Restoration Super-Resolution Methods</td>
<td>20</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Non-Iterative Fusion-Interpolation Super-Resolution Methods</td>
<td>20</td>
</tr>
<tr>
<td>2.2.5</td>
<td>Learning-based Super-Resolution Methods</td>
<td>21</td>
</tr>
<tr>
<td>2.3</td>
<td>General Super-Resolution Observation Model</td>
<td>22</td>
</tr>
<tr>
<td>2.4</td>
<td>Bayesian Super-Resolution Formulation</td>
<td>24</td>
</tr>
<tr>
<td>2.5</td>
<td>Regularized-Norm Minimization Formulation</td>
<td>27</td>
</tr>
<tr>
<td>2.6</td>
<td>Non-Iterative Kernel-Based Formulation</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>HUMAN VISUAL PERCEPTION</td>
<td>31</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview of the HVS</td>
<td>31</td>
</tr>
<tr>
<td>3.2</td>
<td>Light Adaptation</td>
<td>37</td>
</tr>
<tr>
<td>3.3</td>
<td>Contrast Sensitivity</td>
<td>38</td>
</tr>
<tr>
<td>3.4</td>
<td>Contrast Masking</td>
<td>40</td>
</tr>
</tbody>
</table>
CHAPTER 3.5 Human Visual Attention . . . . . . . . . . . . . . . . . . . . 41
3.6 General Framework of Visual Attention Systems . . . . . . . . . . . 43
3.7 Overview of Existing Computational Visual Attention Methods . . . 46
3.8 Hierarchical VA Model . . . . . . . . . . . . . . . . . . . . . . . . . . 49
3.9 Foveated VA Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
3.10 Frequency-Tuned VA Model . . . . . . . . . . . . . . . . . . . . . . . 61

4 ATTENTIVE-SELECTIVE SUPER-RESOLUTION ESTIMATOR BASED ON PERCEPTUALLY SALIENT FEATURES . . . . . . . . . . . . . . . 63
4.1 ATtentive-SELective Perceptual Super-Resolution Framework . . . . 67
4.2 Perceptual Contrast Sensitivity Threshold Model . . . . . . . . . . . 71
4.3 Proposed ATtentive-SELective Perceptual Detection . . . . . . . . . 73
4.4 ATtentive-SELective Perceptual MAP-Based SR Estimation . . . . . 76
4.5 ATtentive-SELective Perceptual Fusion-Restoration SR Estimation . . 79
4.6 Simulation Results . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
4.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98

5 EFFICIENT SUPER-RESOLUTION ESTIMATOR DRIVEN BY VISUALLY ATTENDED REGIONS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100
5.1 ATtentive Super-Resolution Framework . . . . . . . . . . . . . . . . . 102
5.2 Visual Attention Mask Detection . . . . . . . . . . . . . . . . . . . . . 104
5.3 Simulation Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
5.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 128

6 PERCEPTUALLY WEIGHTED SUPER-RESOLUTION . . . . . . . . . . 131
6.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 131
6.2 Proposed Perceptually Weighted SR Estimator . . . . . . . . . . . . 133
6.3 Perceptual Weights Generation . . . . . . . . . . . . . . . . . . . . . . 135
6.4 Simulation Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137
<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5 Conclusion</td>
<td>149</td>
</tr>
<tr>
<td>7 CONCLUSION AND FUTURE RESEARCH</td>
<td>150</td>
</tr>
<tr>
<td>7.1 Contributions</td>
<td>150</td>
</tr>
<tr>
<td>7.2 Future Research</td>
<td>152</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>155</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.</td>
</tr>
<tr>
<td>4.2</td>
<td>PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.</td>
</tr>
<tr>
<td>4.3</td>
<td>Percentage of Pixel Savings for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.</td>
</tr>
<tr>
<td>4.4</td>
<td>Percentage of Pixel Savings for FR-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.</td>
</tr>
<tr>
<td>4.5</td>
<td>Percentage of Operations Savings in computations for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.</td>
</tr>
<tr>
<td>4.6</td>
<td>Percentage of Operations Savings in computations for FR-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.</td>
</tr>
<tr>
<td>4.7</td>
<td>Savings in computations when applying the proposed AT-SELP-MAP SR method on Frame 9 of the $96 \times 128$ Alpaca real video sequence for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 0$.</td>
</tr>
<tr>
<td>4.8</td>
<td>Savings in computations when applying the proposed AT-SELP-FR SR method on Frame 9 of the $96 \times 128$ Alpaca real video sequence for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 0$.</td>
</tr>
<tr>
<td>4.9</td>
<td>Mean Opinion Scores (MOS): scores 1,2,3,4,&amp;5 correspond to AT-SELP-SR(JND) is “Worse”, “Slightly Worse”, “Same”, “Slightly Better”, &amp; “Better”, respectively, than the baseline SR scheme.</td>
</tr>
<tr>
<td>5.1</td>
<td>PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma^2_\eta = 16$, and $\tau = 20%$.</td>
</tr>
</tbody>
</table>
TABLE

<table>
<thead>
<tr>
<th></th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 20% ).</td>
</tr>
<tr>
<td>5.3</td>
<td>PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 40% ).</td>
</tr>
<tr>
<td>5.4</td>
<td>PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 40% ).</td>
</tr>
<tr>
<td>5.5</td>
<td>Percentage of Operations Savings in computations for MAP-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 20% ).</td>
</tr>
<tr>
<td>5.6</td>
<td>Percentage of Operations Savings in computations for FR-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 20% ).</td>
</tr>
<tr>
<td>5.7</td>
<td>Percentage of Operations Savings in computations for MAP-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 40% ).</td>
</tr>
<tr>
<td>5.8</td>
<td>Percentage of Operations Savings in computations for FR-based SR methods for all test sequences for a magnification factor of ( L = 4 ), noise variance ( \sigma^2_\eta = 16 ), and ( \tau = 40% ).</td>
</tr>
<tr>
<td>5.9</td>
<td>Mean Opinion Scores (MOS) for ( \tau = 20% ): scores 1, 2, 3, 4, &amp; 5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, &amp; “Excellent”, respectively.</td>
</tr>
<tr>
<td>5.10</td>
<td>Mean Opinion Scores (MOS) for ( \tau = 20% ): scores 1, 2, 3, 4, &amp; 5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, &amp; “Excellent”, respectively.</td>
</tr>
</tbody>
</table>
5.11 Mean Opinion Scores (MOS) for $\tau = 40\%$: scores 1, 2, 3, 4, & 5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, & “Excellent”, respectively. ........................................ 128

5.12 Mean Opinion Scores (MOS) for $\tau = 40\%$: scores 1, 2, 3, 4, & 5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, & “Excellent”, respectively. ........................................ 128

6.1 PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$. ........................................ 141

6.2 PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$. ........................................ 141

6.3 SNR gain measures in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$. ........................................ 142

6.4 SNR gain measures in dBs for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$. ........................................ 142

6.5 Percentage of operation savings in computations for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$. ........................................ 148

6.6 Percentage of operation savings in computations for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$. ........................................ 148

6.7 Mean Opinion Scores (MOS): scores 1, 2, 3, 4, & 5 correspond to PW-SR is “Worse”, “Slightly Worse”, “Same”, “Slightly Better”, & “Better”, respectively, than the baseline SR scheme. ........................................ 149

6.8 Mean Opinion Scores (MOS): scores 1, 2, 3, 4, & 5 correspond to PW-SR is “Worse”, “Slightly Worse”, “Same”, “Slightly Better”, & “Better”, respectively, than the SELP-SR scheme. ........................................ 149
LIST OF FIGURES

FIGURE

1.1 Artifacts introduced by a typical HR imaging system. (a) Original 256 × 256 Lena image. (b) Sensor averaging blur, 5 × 5 kernel size. (c) Camera-to-object motion blur, shift = 10 pixels, angle = 45°. (d) System Gaussian noise, zero mean, variance = 16. (e) System saturation or salt-and-pepper noise, density = 10%. (f) JPEG2000 compression artifacts, 40 : 1 compression ratio. ........................................... 3

1.2 Super-resolved 256 × 256 HR Cameraman image obtained using sixteen 64 × 64 low-resolution images with magnification factor L = 4, average blur of size 4 × 4, and noise standard deviation σn = 4. (a) Original image; (b) Bicubic interpolation; (c) Baseline MAP-SR; (d) Baseline FR-SR; (e) Non-iterative FI-SR. ...................................................... 9

2.1 Single-frame SR block diagram. The forward model, f{z}, is a mathematical description of the image degradation process exploiting the relationship between HR neighboring pixels. The inverse problem or backward model, f−1{x}, is estimating the HR from the low-quality captured image. ........ 15

2.2 Multi-frame SR block diagram. ................................................. 18

2.3 General multi-frame SR observation model. ............................. 24

2.4 The blur-warp observation model for multi-frame SR. ................. 28

2.5 The non-iterative kernel-based SR approach. ............................. 29

3.1 Primary visual pathway of the HVS. ......................................... 33

3.2 Anatomy of the retina. ...................................................... 34

3.3 Ganglion cell receptive field. ............................................... 34

3.4 PSF of the human eyeball. The x and y axis are in degrees of visual angle and the foveated area corresponds to the high acuity of the fovea. ........... 35
<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>Receptive field profile of a retinal ganglion cell modeled as a DoG. The $x$ and $y$ axis are in degrees of visual angle.</td>
</tr>
<tr>
<td>3.6</td>
<td>Receptive field profile of a simple cell in the primary visual cortex modeled by Gabor filters.</td>
</tr>
<tr>
<td>3.7</td>
<td>Small squares in the middle have equal luminance levels but are perceived with different brightness.</td>
</tr>
<tr>
<td>3.8</td>
<td>Just noticeable difference luminance, $\Delta L$, experiment.</td>
</tr>
<tr>
<td>3.9</td>
<td>Contrast sensitivity function.</td>
</tr>
<tr>
<td>3.10</td>
<td>Demonstration of noise masking by adding a $30 \times 30$ pixels of a zero-mean Gaussian noise with variance 150, indicated by the arrow. Left image is the original image, Middle image with noise added on the sky, Right image with noise added on the grass.</td>
</tr>
<tr>
<td>3.11</td>
<td>General computational visual attention structure.</td>
</tr>
<tr>
<td>3.12</td>
<td>Calculated conspicuity maps of the hierarchical VA model.</td>
</tr>
<tr>
<td>3.13</td>
<td>Saliency map and attended regions of the hierarchical VA model.</td>
</tr>
<tr>
<td>3.14</td>
<td>Generated feature maps of the foveated VA model.</td>
</tr>
<tr>
<td>3.15</td>
<td>Saliency map and attended regions of the foveated VA model.</td>
</tr>
<tr>
<td>3.16</td>
<td>Saliency map and attended regions detected by keeping the highest 20% of the saliency map of the frequency-tuned VA model.</td>
</tr>
<tr>
<td>4.1</td>
<td>Super-resolved $256 \times 256$ HR Cameraman image obtained using sixteen $64 \times 64$ low-resolution images with a $4 \times 4$ average blur and a zero mean Gaussian noise with standard deviation $\sigma_n = 10$.</td>
</tr>
<tr>
<td>4.2</td>
<td>SNR comparison among the baseline MAP SR, the SELP-MAP SR, G-MAP SR with block mean thresholds, E-MAP SR with block mean thresholds, using sixteen $64 \times 64$ low-resolution images with noise standard deviation $\sigma_n = 10$ for the $256 \times 256$ Cameraman image.</td>
</tr>
<tr>
<td>FIGURE</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>4.3</td>
<td>Flowchart of the proposed AT-SELP SR framework.</td>
</tr>
<tr>
<td>4.4</td>
<td>Generation process of the SELP mask.</td>
</tr>
<tr>
<td>4.5</td>
<td>Illustration of the obtained set of active pixels that are obtained by generating and applying (a) the perceptual mask ($M_p$) and (b) the attentive mask ($M_a$) for the blurred and noisy version of the $256 \times 256$ ‘Lena’ image. ‘White’ intensities correspond to selected (active) pixels.</td>
</tr>
<tr>
<td>4.6</td>
<td>Motion vectors on the HR grid.</td>
</tr>
<tr>
<td>4.7</td>
<td>Original images of size $256 \times 256$ used to test the proposed AT-SELP SR framework.</td>
</tr>
<tr>
<td>4.8</td>
<td>Detected salient active pixels, $M_a$, for the $256 \times 256$ bicubicly interpolated Frame 9 of the Cameraman image with a $4 \times 4$ average blur, a Gaussian noise with $\sigma^2 = 16$, and $\tau = 20%$.</td>
</tr>
<tr>
<td>4.9</td>
<td>Comparison between the baseline MAP, SELP-MAP, and the proposed AT-SELP-MAP SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, and a noise variance $\sigma^2 = 16$ for Frame number 9 of the $256 \times 256$ Cameraman sequence.</td>
</tr>
<tr>
<td>4.10</td>
<td>Comparison between the baseline FR, SELP-FR, and the proposed AT-SELP-FR SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, and a noise variance $\sigma^2 = 16$ for Frame number 9 of the $256 \times 256$ Cameraman sequence.</td>
</tr>
<tr>
<td>4.11</td>
<td>Super-resolved Frame 9 of the $256 \times 256$ HR Cameraman image obtained using MAP-based SR methods from sixteen $64 \times 64$ LR images with $\sigma^2 = 16$ and $\tau = 20%$.</td>
</tr>
<tr>
<td>4.12</td>
<td>Super-resolved Frame 9 of the $256 \times 256$ HR Cameraman image obtained using FR-based SR methods from sixteen $64 \times 64$ LR images with $\sigma^2 = 16$ and $\tau = 20%$.</td>
</tr>
</tbody>
</table>
FIGURE

4.13 Zoomed in super-resolved Frame 9 of the 256 × 256 HR Cameraman image using MAP-based SR methods. .................................................. 92

4.14 Super-resolved Frame 9 of the 96 × 128 Alpaca real video sequence using the first sixteen frames with magnification factor \( L = 4 \), \( \sigma^2_\eta = 0 \) and \( \tau = 20\% \). (a) Bicubicly interpolated Frame; (b) Baseline MAP SR; (c) SELP-MAP SR; (d) proposed AT-SELP-MAP SR. .......................... 96

4.15 Super-resolved Frame 9 of the 96 × 128 Alpaca real video sequence using the first sixteen frames with magnification factor \( L = 4 \), \( \sigma^2_\eta = 0 \) and \( \tau = 20\% \). (a) Bicubicly interpolated frame; (b) Baseline FR SR; (c) SELP-FR SR; (d) proposed AT-SELP-FR SR. .......................... 97

4.16 Snapshot of performed subjective test comparing the baseline MAP-SR with the proposed AT-SELP-MAP(JND) method, and the baseline FR-SR with the proposed AT-SELP-FR(JND) method. ......................... 99

5.1 Flowchart of the proposed AT-SR framework. ......................... 103

5.2 Detected VA regions, \( M_n \), for the 256 × 256 bicubicly interpolated Frame 9 of the Monarch image with a 4 × 4 average blur, a Gaussian noise with \( \sigma^2_\eta = 16 \), and \( \tau = 20\% \). ........................................ 107

5.3 Detected VA regions, \( M_n \), for the 256 × 256 bicubicly interpolated Frame 9 of the Monarch image with a 4 × 4 average blur, a Gaussian noise with \( \sigma^2_\eta = 16 \), and \( \tau = 40\% \). ........................................ 108

5.4 Comparison between the baseline MAP and the proposed AT-MAP SR estimators using sixteen 64 × 64 LR images, a resizing factor \( L = 4 \), a noise variance \( \sigma^2_\eta = 16 \), and \( \tau = 20\% \) for Frame number 9 of the 256 × 256 Monarch sequence. ................................. 109
5.5 Comparison between the baseline MAP and the proposed AT-MAP SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2 = 16$, and $\tau = 40\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.

5.6 Comparison between the baseline FR and the proposed AT-FR SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2 = 16$, and $\tau = 20\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.

5.7 Comparison between the baseline FR and the proposed AT-FR SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2 = 16$, and $\tau = 40\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.

5.8 Super-resolved Frame number 9 of $256 \times 256$ HR Monarch image obtained using MAP-based SR methods from sixteen $64 \times 64$ low-resolution images with $\sigma^2 = 16$ and $\tau = 20\%$.

5.9 Super-resolved Frame number 9 of $256 \times 256$ HR Monarch image obtained using MAP-based SR methods from sixteen $64 \times 64$ low-resolution images with $\sigma^2 = 16$ and $\tau = 20\%$.

5.10 Super-resolved Frame number 9 of $256 \times 256$ HR Monarch image obtained using FR-based SR methods from sixteen $64 \times 64$ low-resolution images with $\sigma^2 = 16$ and $\tau = 20\%$.

5.11 Super-resolved Frame number 9 of $256 \times 256$ HR Monarch image obtained using FR-based SR methods from sixteen $64 \times 64$ low-resolution images with $\sigma^2 = 16$ and $\tau = 20\%$. 
FIGURE

5.12 Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with \( \sigma^2 = 16 \) and \( \tau = 40\% \) . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121

5.13 Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with \( \sigma^2 = 16 \) and \( \tau = 40\% \) . . . . . . . . . . . . . . . . . . . . . . . . . . . 122

5.14 Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using FR-based SR methods from sixteen 64 × 64 low-resolution images with \( \sigma^2 = 16 \) and \( \tau = 40\% \) . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

5.15 Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using FR-based SR methods from sixteen 64 × 64 low-resolution images with \( \sigma^2 = 16 \) and \( \tau = 40\% \) . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

5.16 Snapshot of performed subjective test for rating the SR methods . . . . . . . 129

6.1 Perceptual weights, \( \Psi \), used by the proposed PW-MAP SR method for all test sequences for a magnification factor of \( L = 4 \), an average blur of size \( 4 \times 4 \), and noise variance \( \sigma^2 = 16 \) . . . . . . . . . . . . . . . . . . . . . . . . . . . 137

6.2 Perceptual weights, \( \Psi \), used by the proposed PW-FR SR method for all test sequences for a magnification factor of \( L = 4 \), a Gaussian blur of size \( 4 \times 4 \), and noise variance \( \sigma^2 = 16 \) . . . . . . . . . . . . . . . . . . . . . . . . . . . 138

6.3 SNR gain per iteration comparisons among the baseline MAP, efficient SELP-MAP, and the proposed PW-MAP SR estimator using sixteen 64 × 64 LR images, a resizing factor \( L = 4 \), and a noise variance \( \sigma^2 = 16 \) for frame number 9 of all the 256 × 256 test sequences . . . . . . . . . . . . . . . . . . . . . . . . . . . 139
FIGURE

6.4 SNR gain per iteration comparisons among the baseline FR, efficient SELP-FR, and the proposed PW-FR SR estimator using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, and a noise variance $\sigma^2_{\eta} = 16$ for frame number 9 of all the $256 \times 256$ test sequences. .......................... 140

6.5 Super-resolved Frame 9 of $256 \times 256$ HR Lena image obtained using the baseline MAP SR, SELP-MAP SR, and the proposed PW-MAP SR from sixteen $64 \times 64$ low-resolution images with $\sigma^2_{\eta} = 16$. .......................... 144

6.6 Super-resolved Frame 9 of $256 \times 256$ HR Cameraman image obtained using the baseline MAP SR, SELP-MAP SR, and the proposed PW-MAP SR from sixteen $64 \times 64$ low-resolution images with $\sigma^2_{\eta} = 16$. .......................... 145

6.7 Super-resolved Frame 9 of $256 \times 256$ HR Lena image obtained using the baseline FR SR, SELP-FR SR, and the proposed PW-FR SR from sixteen $64 \times 64$ low-resolution images with $\sigma^2_{\eta} = 16$. .......................... 146

6.8 Super-resolved Frame 9 of $256 \times 256$ HR Cameraman image obtained using the baseline FR SR, SELP-FR SR, and the proposed PW-FR SR from sixteen $64 \times 64$ low-resolution images with $\sigma^2_{\eta} = 16$. .......................... 147
Chapter 1

INTRODUCTION

Nowadays, High Resolution (HR) image/video applications are reaching every home and soon will become a necessity rather than a mere luxury. Whether it is through a High-Definition (HD) TV, computer monitor, HD camera, smart phones, or many such handheld devices, HR multimedia applications are becoming essential components of consumers’ daily lives. Moreover, the demand for HR images transcends the need for offering a better quality picture to the common viewer. HR imagery is continuing to gain popularity and dominate many industries that require accurate image analysis. For example, in surveillance applications, high resolution images are needed for a better performance of target detection, face recognition, and text recognition [1, 2]. Furthermore, medical imaging applications require HR images for accurate assessment and detection of small lesions [3, 4].

1.1 Difficulties of High-Resolution Imaging Systems

The capture and delivery of HR multimedia content is a complex and problematic process [5]. In any typical imaging system or multimedia delivery chain, the quality of HR media can be impaired due to several processes of acquisition, transmission, and display. The resolution of the imaging systems is physically limited by the pixel density on their sensors (e.g., couple charge device (CCD)). On one hand, increasing the number of pixels on a CCD chip via reducing the pixel size is limited by the existence of shot noise and associated high cost. On the other hand, increasing the chip size is deemed ineffective due to the existence of a large capacitance that slows the charge transfer rate and limits the data transfer [6]. Thus, a major drawback of HR image acquisition in many of the aforementioned applications is the inadequate resolution of the sensors, either because of cost or hardware limitation [7]. In addition to hardware limitations, HR imagery can be blurred during acquisition due to atmospheric turbulence and the
camera-to-object motion [8]. Transmission of HR media requires extremely high bandwidth, which is usually not available in practical scenarios. Thus, high compression rates are imposed on the media content resulting in annoying compression artifacts and high frame dropping rates [9]. Moreover, large displays use interpolation techniques to scale video content to fit a target screen size, thus introducing blurred details and system noise enlargement. Hence, signal processing techniques, more specifically Super-Resolution (SR), provide cost-effective solutions to increase image resolution by overcoming sensor hardware limitations and reducing media impairments.

Fig. 1.1 illustrates some common artifacts that are generated by typical HR imaging systems. Fig. 1.1(b) shows the blurring introduced by the point spread function (PSF) of the sensors which can be modeled by an averaging kernel (5×5 averaging kernel). Motion blur caused by camera-to-object motion is illustrated in Fig. 1.1(c) with a linear shift of 10 pixels and an angle of 45° between the object and the camera. The system noise introduced by sensor readout can be modeled by a Gaussian noise as illustrated in Fig. 1.1(d) with a zero mean Gaussian noise having a variance of 16. Also, system noise in the form of sensor saturation can be modeled by a salt-and-pepper noise as shown in Fig. 1.1(e) with 10% noise saturation density. The transmission artifacts caused by JPEG2000 [10] compression are shown in Fig. 1.1(f) with a compression ratio of 40:1. Generally, single-frame and multi-frame Super-Resolution techniques are effective image enhancement and reconstruction solutions that can overcome the limitations and reduce the artifacts of HR imaging systems. As will be described in this thesis, single-frame SR methods require one under-sampled or degraded image to reconstruct a higher resolution enhanced image. However, multi-frame SR methods combine multiple degraded frames or views of a scene to estimate a higher resolution enhanced image.
Figure 1.1: Artifacts introduced by a typical HR imaging system. (a) Original 256 × 256 Lena image. (b) Sensor averaging blur, 5 × 5 kernel size. (c) Camera-to-object motion blur, shift = 10 pixels, angle = 45°. (d) System Gaussian noise, zero mean, variance = 16. (e) System saturation or salt-and-pepper noise, density = 10%. (f) JPEG2000 compression artifacts, 40 : 1 compression ratio.
1.2 Super-Resolution Applications

Recently, Super-Resolution techniques are at the heart of HR image/video technologies in various industries such as consumer electronics, entertainment, digital communications, military, and medical applications. Mainly, in the consumer electronics sector, SR can play a major role in designing cost-effective digital cameras using cheap sensors as described in [11–13]. Low-cost cameras use a mosaic of color filters to capture only one color component (R, G, or B) at each sensor location on the CCD array. The color samples obtained with such a color filter array (CFA) are spatially subsampled and must then be interpolated to estimate the missing samples. This CFA interpolation problem is also referred to as the demosaicking problem. The most commonly used CFA pattern is the Bayer pattern, where the green channel is more densely sampled than the red and blue channels on a rectangular grid. With efficient SR techniques, spatial resolution can be improved beyond the physical limits of a sensor chip avoiding unpleasant interpolation artifacts.

Surveillance applications for military or civilian use require smart monitoring techniques to cater applications such as face detection, text reconstruction for license plate identification, or target detection after segmenting or even tagging an object of interest in a scene. Super-resolution, as shown in [14], is used to get a higher quality video sequence from low resolution multiple cameras for low-cost surveillance applications. In [15], SR techniques offer a robust and effective solutions for license plate recognition in intelligent transportation systems under various factors that increase the difficulty of the automatic identification; namely, bad ambient illumination, poor weather condition, car motion and griminess or abrasion of the plate, where areas of the plate may be blurred and degraded badly. In [16, 17], SR techniques are required for a more accurate face recognition system with low quality multiple/single cameras under normal
conditions, or high quality cameras under environments with degraded imaging conditions. Automated target tracking and recognition (ATR) is one of the major backbones of military surveillance and reconnaissance in the modern digital world. In [18], an SR technique is used to aid target tracking and recognition under heavy clutter conditions. A major problem for ATR systems under heavy urban clutter is that the high-contrast building corners, roads, and trees will cause a high false detection rate. The work in [18] employ an SR image enhancement process to improve the vehicle resolution and signal-to-noise ratio (SNR) for better automatic target recognition or human/pilot-monitored recognition performance.

Medical imaging has been one of the major means of medical diagnosis and, with the advancement of imaging systems, is continuing to be one of the highly flourishing sectors of medical diagnostics. Robinson et al. in [4] used SR techniques to reconstruct a single higher resolution image from fusing a collection of multiple extremely low-dosage aliased X-ray images. The proposed technique proved to be effective in enhancing digital mammogram images for accurate assessment and detection of small lesions under low exposure of radiation. In [19], SR is applied to respiratory gated Positron Emission Tomography (PET) images for motion compensation. Respiratory motion is a major source of reduced quality in PET imaging. Respiratory synchronized acquisitions techniques are used to minimize these motion effects, which lead to gated frames with low signal-to-noise ratio (SNR) as they contain reduced statistics. In [19], SR was performed on the gated frames to correct the effects of respiratory motion and enhance the overall image quality.

In the scientific and biological research fields, digital microscopy imaging has been widely used for cell migration analysis and studying the behavior of abnormal cells. Tracking fluorescently tagged objects in living cells makes it possible to follow the dynamics of specific cellular structures under physiological conditions. However,
the number of images acquired in live-cell experiments and processing of image data to extract dynamic information is overwhelming. In [20], SR techniques are used to aid computational tools for semi-automatic or automatic tracking of fluorescent markers in biological images.

Super-resolution has played a major role in printing still images from low-resolution video or scanning a noise-free text that can be employed for improving the display of low-resolution scans of archival documents or low-resolution bit-mapped fonts on high-resolution output devices. In [21], SR is used to produce a noiseless high-resolution scanned text given a single image of text scanned at low resolution, thus, avoiding artifacts in the high-resolution image such as blurry edges and rounded corners. It has also been used to recover from quantization noise and grid-alignment effects that introduce errors in the low-resolution image, and to handle documents with very large glyph.

Another application of SR is in multimedia delivery for home entertainment systems or personal mobile devices. A high-quality low-bit-rate compression algorithms should be applied for real-time delivery of such HD multimedia content. In [22], SR is used to remove the visually annoying artifacts of H.264/AVC [23] video compression standard. Visually annoying artifacts such as ringing are reduced, while preserving the sharpness of edges.

Most of the previously mentioned applications require real-time or near real-time processing with limitations on computational power and the essential demand for high image quality requirements. Efficient super-resolution methods are vital solutions to the digital multimedia acquisition, delivery, and display sectors.
1.3 Overview of Efficient Super-Resolution Problem-Solution

When choosing a solution for the super-resolution problem, there is always a trade-off between computational efficiency and HR image quality. Existing SR approaches suffer from extremely high computational requirements due to the high dimensionality in terms of large number of unknowns to be estimated in the solution of the SR inverse problem. Generally, MAP-based SR solutions are iterative in nature and are highly computational, but can converge to a high quality HR estimate. Even for solutions with fast convergence rates, commonly used Bayesian approaches are conditioned on the number of different LR observations and the HR image prior statistical model, which can lead to very high computational requirements even for small size image estimates. To reduce the computations required for the regularized norm minimization SR solutions, the Fusion-Restoration (FR) methods register and merge all the LR observations on one HR grid before using an iterative regularized minimization reconstruction process. However, these good reconstruction quality solutions are still computationally intensive due their iterative nature and the large number of pixels estimated. Moreover, the non-iterative Fusion-Interpolation (FI) SR approaches, also referred to as kernel-based SR, are inherently less computationally intensive but suffer from limited reconstruction quality depending on their assumed statistical model. For example, the MAP-based SR approach described in [24] requires a total of approximately $395 \times 10^6$ multiplication and addition operations to estimate an HR image of size $256 \times 256$ from 16 LR frames of size $64 \times 64$ pixels with translational motion and noise. Also, for the same problem, the FR-based SR method proposed in [25] requires approximately a total of $200 \times 10^6$ multiplication and addition operations. Additionally, the inherently faster non-iterative FI-based SR approach in [26] with parameters set to $W_x = W_y = 12$, $D_x = D_y = 4$, $\rho = 0.75$, requires $72 \times 10^6$ multiplication and addition operations but suffers from a limited reconstruction quality. The visual quality of the HR estimates
using each of the SR solutions described in [24–26] are compared in Fig. 1.2. It can be clearly seen from Fig. 1.2 that the MAP-based [24] and the FR-based [25] schemes result in a noticeably sharper image than the FI-based SR approach [26]. Furthermore, the iterative MAP-based method [24] and the FR-based method [25] result in around 1 dB increase in PSNR over the non-iterative FI-based SR method of [26].

In order to further enhance the computational efficiency of iterative SR algorithms, Ivanovski et al. [27, 28] introduced a selective MAP-based SR algorithm in which only pixels with significant spatial activity are super-resolved. The local gradients of the estimated HR image at each iteration are used to detect the pixels with significant spatial activities (i.e., pixels at which the gradient is above a certain threshold). The obtained results in [27, 28] demonstrated a significant reduction in the computational complexity needed for the HR image estimation along with imperceptible loss in visual quality. However, the drawback of the algorithm presented in [27, 28] lies in manually tweaking the gradient threshold for each image differently to attain the best desired SR quality. As a consequence, a new class of selective SR estimators is presented in this thesis to reduce the dimensionality or the computational complexity of popular SR algorithms while maintaining the desired visual quality of the reconstructed HR image. Generally, these selective algorithms detect only a subset of active pixels that are super-resolved iteratively based on the pixel’s local perceptual significance measure to the final SR result.

Unlike the hard gradient thresholding in the previously presented approach [27, 28], a SELective Perceptual-based (SELP) SR framework is presented in [29,30] where the set of significant pixels is determined adaptively using an automated perceptual decision mechanism without any manual tuning. The introduced detection model is based on a perceptual contrast sensitivity threshold model that detects details with perceived contrast over a uniform background. Building on this work, our previous work
Figure 1.2: Super-resolved $256 \times 256$ HR Cameraman image obtained using sixteen $64 \times 64$ low-resolution images with magnification factor $L = 4$, average blur of size $4 \times 4$, and noise standard deviation $\sigma_n = 4$. (a) Original image; (b) Bicubic interpolation; (c) Baseline MAP-SR; (d) Baseline FR-SR; (e) Non-iterative FI-SR.
in [31, 32], showed that not all the detail pixels detected by the SELP algorithm are needed to preserve the overall visual quality of an HR image. Due to Human Visual Attention, attended regions are processed at high visual acuity, hence details present in these regions are better perceived by the HVS than those present in non-attended areas. In consequence, the observer’s perception of image quality is prejudiced by distortions present on signals lying in salient regions [33]. In our previous work [31, 32], a visual attention model proposed by Itti et al. [34], is used to further reduce the processed pixels and thus the computational complexity of the SELP SR technique. However, the previously proposed SR framework assumes that the attention information is already computed and stored offline thus ignoring any computational overhead introduced by the adopted complex visual attention model [35]. This approach was motivated by SR applications incorporated in a quality assessment framework where the visual attention information is computed as part of the quality assessment stage, and is thus available to the SR stage without adding extra overhead [35]. Thus, towards an effective SR solution, this thesis proposes a low complexity saliency detector designed for efficient attentively selective SR estimators. Consequently, an efficient ATtentive-SELectional Perceptual (AT-SELP) SR framework is presented in order to reduce the computational complexity of iterative SR algorithms without any perceptible loss in the desired enhanced image/video quality. Moreover, different low-level features influenced by visual attention models presented in [34,36,37] are studied to illustrate the efficiency and quality of the proposed attentive SR framework. To further reduce the computational complexity of selective SR methods, a highly efficient ATtentive (AT)-SR framework that is completely driven by VA region detectors, is presented as part of this work.

Previous SR methods focus on minimizing non-perceptual error metrics such as the mean squared error (MSE) or signal-to-noise ratio (SNR) that do not necessarily correlate with the perceptual quality. Furthermore, existing SR methods as shown as
part of this work may over-process some pixels leading to enhanced artifacts as seen in the flat areas of Fig. 1.2(c). This is due to controlling the SR iterative process with a global error measure which considers all pixels equally. In the proposed scheme, the set of perceptually significant pixels is determined adaptively based on the local image characteristics and used to control the SR iterations by pooling the error measure locally over a set of perceptually significant regions. Thus, we also demonstrate that the proposed AT-SELP and AT-SR frameworks are locally adaptive by keeping a balance between sharpening edges and denoising smooth regions according to human perception and visual attention. Then, a computational complexity analysis based on the total operation count is presented taking into account the operation overhead introduced by the computation of the attentive features.

In the last part of this thesis, a perceptually weighted approach to the SR inverse problem is presented. This approach is inspired by the HVS unequal processing of the visual stimuli in a scene. The perceptual weighting operator is incorporated in the cost function of the SR minimization problem to bias the balance of sharpening and smoothing of different pixel locations in a perceptual manner. Perceptual weights can be computed adaptively thus better preserving perceptual edges and texture while smoothing noisy regions.

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 provides a general background on multi-frame SR techniques that are addressed in this thesis. The necessary assumptions for the SR solutions to work, are discussed. Then, different observation models are reviewed followed by several SR problem formulations and solutions.

Chapter 3 provides background material on the Human Visual System (HVS) and visual perception in relevance to the proposed attentive selective SR framework. An
overview of the HVS as well as a description of perceptual concepts, namely, contrast sensitivity, luminance and contrast masking, and Visual Attention (VA) background and illustrations are presented. Also, several saliency detection algorithms based on visual attention modeling are reviewed and a detailed computational analysis is provided.

Chapter 4 proposes an efficient ATtentive-SELective Perceptual (AT-SELP) SR framework based on the detection of salient low-level features. After describing the improved selective SR framework, the contrast sensitivity threshold model used for detecting active pixels for selective SR processing is presented. Due to the high complexity of the existing saliency detectors, a new low-complexity saliency detector is proposed based on notions of just noticeable difference thresholding. Finally, the proposed AT-SELP SR framework is integrated in a MAP-based and FR-based SR solutions, and simulation results show the reduced complexity and preserved visual quality of the proposed selective framework.

Chapter 5 proposes an enhanced selective framework that is completely driven by visual attention (VA) information. The proposed efficient ATtentive (AT) SR framework is described, and several VA models are shown to be easily integrated within the proposed framework. Existing VA detectors are shown to be computationally complex or to fail to detect regions essential for efficient SR processing. The proposed low-complexity VA detector based on the Just-Noticeable-Difference (JND) model is adopted for an enhanced performance in quality and computational efficiency of the MAP-based and FR-based selective SR solutions. Finally, simulation results of the proposed AT-SR framework with different detection thresholds show the enhanced efficiency and preserved visual quality of the proposed attentive SR scheme.

Chapter 6 considers a different approach to the SR reconstruction solution mimicking the unequal processing of the human visual system to different image features.
Perceptual weighting parameters are used in minimizing the cost function of the SR problem in order to locally enhance the perceptually relevant image features. Then, a Perceptually Weighted (PW) SR technique is proposed that enhances on the reconstruction quality of iterative SR techniques at a faster convergence rate.

Chapter 7 summarizes the key contributions and discusses possible future research directions in the area of efficient super-resolution techniques.
Chapter 2

GENERAL SUPER-RESOLUTION BACKGROUND AND FORMULATION

This chapter presents a literature survey of single/multi-frame SR techniques and introduces essential SR background material. Then, an overview of SR solution formulations that are relevant to this thesis is presented.

2.1 Single-Frame Super-Resolution

Single-frame SR approaches are widely applied solutions to the resolution enhancement problem, especially, in cases where only one degraded Low-Resolution (LR) observation is available. Single-frame SR techniques, as described in the literature [38], enhance the resolution of an image from a single degraded LR image. Single-frame SR approaches are also referred to as image interpolation or reconstruction and these terms will be used interchangeably in this thesis. Recovering a high-resolution image from an under-sampled (according to Nyquist limits [39]) and noisy observation is a highly ill-posed inverse problem. Thus, prior image models relating neighboring pixels or prior models learnt through similar image patches are needed extra information that can aid in solving the inverse problem. A common approach among these SR techniques is that they take advantage of the relation between neighboring pixels of the same image to estimate the values of missing pixels. Fig. 2.1 shows a general single-frame SR process where \( f\{z\} \) and \( f^{-1}\{x\} \) denote the forward and backward degradation process, respectively.

A well-known problem with common kernel-based super-resolution (such as bilinear and bicubic interpolation [40]) is the blurring and blockiness effects that are introduced to sharp edges. The blurring of sharp edges results from the inaccurate kernel resizing to adapt to the edge sharpness. The blockiness, also known as staircase effects, on oriented edges is mainly due to the failure of the filter to adapt to various edge orientations. In [41], an edge-directed interpolation method is proposed that uses
Acquired aliased noisy LR image, \( x \)

Forward Model,
\[
x = f\{z\}
\]

Backward Model,
\[
z = f^{-1}\{x\}
\]

Super-Resolved HR image, \( z \)

**Super-Resolution** using:
- Statistical modeling
- Pixel classification
- Structural properties
- Edge orientations

Figure 2.1: Single-frame SR block diagram. The forward model, \( f\{z\} \), is a mathematical description of the image degradation process exploiting the relationship between HR neighboring pixels. The inverse problem or backward model, \( f^{-1}\{x\} \), is estimating the HR from the low-quality captured image.

Local covariance estimates of the input LR frame to adapt interpolation coefficients to arbitrarily oriented edges of the reconstructed HR image. This method is motivated by the geometric regularity property [42] of an ideal step edge. The geometric regularity property of edges refers to the sharpness constraint across the edge orientation and the smoothness constraint along the edge orientation. Starting with previous findings on edge-directed prediction for lossless image coding [43], Li et al. show that covariance-based adaptation is able to tune the prediction kernel support to match an arbitrarily oriented edge. Then the geometric duality between the low-resolution covariance and the high-resolution covariance, which couples the pair of pixels along the same orientation, is used to determine the coefficients of the edge-directed interpolator [41]. Recently, in [44], a new edge-directed interpolation approach is proposed that uses multiple low resolution training windows to reduce the covariance mismatch be-
tween the LR and HR pixels. The previously described methods use the concepts of kernel-based resizing and orientation that are non-iterative in nature.

A rule-based technique based on a set of simple/heuristic rules and variable thresholds is presented in [45]. The interpolation algorithm is locally adaptive by incrementally filling the missing pixels in a step-by-step approach based on simple rules for spatial dependencies of missing pixels with their surrounding neighbors. Then the last step of the algorithm [45] uses histogram averaging over the bins that correspond to the neighboring pixel values. The values of the bins are represented by the median of each bin. In [46], a smart interpolation method is proposed that uses anisotropic diffusion as introduced in [47]. Anisotropic diffusion, also called Perona-Malik diffusion, is a technique aiming at reducing noise and enhancing contrast without significantly degrading edges and objects in an image. The algorithm proposed in [46], involves enlarging the image beyond the required resolution then performing anisotropic diffusion [48] to sharpen image details by applying. Then, the resulting image is downsampled and low-pass filtered to reduce aliasing and a weighted averaging technique is applied to obtain the final super-resolved target image. A detailed comparison between the previously described methods is also found in [49].

Muresan et al. in [50–54], proposed model-based super-resolution methods based on the optimal recovery principle. The presented methods model the local image regions into ellipsoidal signal classes and adapts the interpolating kernels accordingly. Given the optimized recovery formulation in [50], the challenge is determining the local quadratic signal class of the high resolution image. This is achieved by taking image patches from the local neighborhood of the up-sampled observation. In [55], another model-based SR approach is presented based on multi-resolution analysis in the wavelet domain. The statistical relationships between coefficients at coarser scales are modeled by using hidden Markov trees to predict the coefficients on the finest scale.
An inverse wavelet transform is applied after prediction to obtain the super-resolved image.

Although single-frame SR algorithms may seem relatively efficient in nature, they still suffer from the inability to generate fine details from the limited reconstruction of high frequency components. Super-resolving from a single low-resolution image is known to be a highly ill-posed inverse problem due to the low number of observations relative to the large number of missing pixels/unknowns. Thus, the gain in quality in the single-frame SR approach is limited by the minimal number of information provided to recover missing details in the reconstructed HR signal.

2.2 Multi-Frame Super-Resolution

Multi-frame SR techniques offer a better solution to the resolution enhancement problem by exploiting extra information from several neighboring frames in a video sequence. In a multi-frame acquisition system, the sub-pixel motion between the camera and object allows the frames of the video sequence to contain over-sampled information that make the reconstruction of an HR image possible. Thus, in this thesis, the focus is on multi-frame SR techniques that enhance the resolution of images by combining information from multiple Low-Resolution (LR) frames of the same scene to estimate a High-Resolution (HR) unaliased and sharp/deblurred image under realistic bandwidth requirements and cost-effective hardware [6]. Interest in multi-frame super-resolution re-emerged in the recent years for two main reasons: first, the use of multi-frame image sequences which can take advantage of additional spatio-temporal information available in the video content, and second, the increase of hardware computational power and advancement of display technologies which make SR applications possible. Fig. 2.2, presents a block diagram showing a multi-frame SR estimation process using multiple degraded LR frames with sub-pixel shifts to estimate one HR reference frame. The figure also shows the LR pixels registered relative to a common
HR reference grid to further visualize the number of pixels used in the process of estimating the missing pixels of the SR image. Since Tsai et al. [56] first introduced the multi-frame image restoration and registration problem/solution, several multi-frame SR approaches have been proposed in the past two decades. In the following literature review, several multi-frame SR approaches are described and broadly categorized according to their methods of solution into Bayesian, regularized norm minimization, Fusion-Restoration (FR), non-iterative Fusion-Interpolation (FI), and learning-based approaches.
Bayesian Super-Resolution Methods

Bayesian Maximum A-Posteriori (MAP) solutions have gained great attention and proved to be effective due to the inclusion of a-priori knowledge about the HR estimate and inherent probabilistic formulation of the relation between the LR observations and the HR image. In [57], Shultz et al. proposed a MAP-based SR solution by imposing a piecewise smoothness assumption on data regions separated by discontinuities. An edge preserving Huber-Markov random field is presented to model discontinuities in the image prior. A similar, but improved approach, is proposed in [24] by using a gradient descent solution for the simultaneous estimation of the registration parameters and the HR image. More recently, Gunturk et al. [58] introduced the camera parameters, such as, exposure time and camera response function in the Bayesian SR framework. Also in [59], a transform-domain formulation of the MAP-based SR solution is presented by taking into consideration the quantization errors in the data prior for a better reconstruction of compressed video. The common disadvantage of MAP-based multi-frame SR methods is their high sensitivity to the assumed statistical model of data and noise as part of the Bayesian problem formulation.

Regularized-Norm Minimization Super-Resolution Methods

Elad et al. [60], based on the assumption of additive and mutually independent Gaussian noise, formulated the Maximum-Likelihood (ML) estimation problem into an unconstrained least-square minimization problem. In [61], a more robust $l_2$-norm minimization problem is presented by applying a median estimator to the back-projected error corresponding to each observation in the gradient-descent iterative solution. Farsiu et al., in [25], studied the robustness of the $l_p$-norm type of minimization approaches against outliers in the SR model. It was concluded, theoretically and experimentally, that the $l_1$-norm minimization approach is the most robust to errors in the system.
Also in [25], different regularization approaches are investigated and the bilateral total variation function was shown to give the best performance in terms of robustness and edge-preservation. While MAP-based and constrained $l_p$-norm minimization techniques provide effective solutions for resolution enhancement, they are still iterative in nature and suffer from high computational complexity. Towards that problem, many efficient solutions are proposed to reduce the SR estimation complexity.

**Fusion-Restoration Super-Resolution Methods**

The two-stage Fusion-Restoration (FR) SR estimation methods [25, 60] were devised in order to decrease the computational requirements and to increase the robustness to outliers. The FR methods are composed of a non-iterative fusion step, which includes a one-step registration process, and a restoration step, which simultaneously deblurs and denoises the fused image by minimizing an error function with or without a specific regularization term. In [60], Elad et al. proposed a fast algorithm for the fusion stage by reducing the SR minimization problem to a pixel-wise average of the measurements, then a $15 \times 15$ non-adaptive Wiener filter is manually tweaked for best de-blurring performance in the reconstruction stage. Subsequently, in [25], a median shift and add operation is used to fuse the LR frames on the HR grid, then an iterative $l_1$ error norm minimization with a bilateral total variation regularization term is applied for reconstruction. Furthermore in [62], the FR SR method of [25] is made more robust to motion errors by weighting each pixel in the SR observation by a coefficient determined by both the distance from the estimated pixel and the motion estimation block matching error associated with the observation.

**Non-Iterative Fusion-Interpolation Super-Resolution Methods**

Hardie et al. [26] proposed a SR approach by using sub-pixel registration to fuse the LR frames on the HR grid and an adaptive Wiener filter to reconstruct the HR image. The Wiener filter coefficients are designed locally depending on the relative motion
between the frames. Similarly in [63], a partition-based weighted sum of filters are applied after the fusion step to fill in the missing pixels and deblur the final SR image. These SR approaches are composed of a fusion step and a non-iterative reconstruction step using kernel-based interpolation, which is inherently efficient in nature. However, the methods proposed in [26, 63] are limited by the statistical assumptions imposed on the observations and suffers from limited reconstruction accuracy.

**Learning-based Super-Resolution Methods**

All the previous methods consider the sub-pixel motion between the observations to recover missing information by solving the SR problem. Learning-based SR approaches recover missing high frequency information in an image through matching using a large image database. In [64], an example-based SR approach is proposed where the database is composed of interpolated patches of the LR images. The high-frequency information is recovered by a nearest neighbor search to a given LR matching patch in the dataset. The disadvantage of this approach is that the SR image is composed of interpolated versions of the LR matching patches. In [65], another learning-based SR approach is proposed where the initial estimate of the high-frequency content of the SR image is learned from a large LR-HR image pair database. The learning process is in the DWT domain and the SR observation model parameters are learned from the relationship between the initial LR-HR estimates. These learned parameters are used to estimate the decimation model and the Inhomogeneous Gaussian Markov Random Field (IGMRF) image prior model. A MAP-based estimation is then used for the SR reconstruction with the IGMRF prior image model. The learning-based SR algorithm in [66] uses partial differential equations (PDE) based regularization for an artifact-relieved LR image. Then, this resulting artifact-relieved LR image is bicubically interpolated to be used for pair matching with the SR database. The primitive-based pair matching method is adopted for learning the SR image, where matching is applied only to primitive compo-
nents such as edges, ridges and corners. Learning-based SR approaches require a large set of training LR-HR image pairs. In [67], a color-based single-image SR approach is proposed based on an iterative backprojection estimation with a multi-scale tensor voting for perceptual grouping. In this latter method, only blur and downsampling are considered with no noise.

2.3 General Super-Resolution Observation Model

Multi-frame reconstruction techniques, as described earlier in Section 2.2, reduced the sensitivity of the SR inverse problem solution to some given LR observation measures by exploiting extra information from several neighboring frames in a video sequence. A necessary assumption for multi-frame SR solutions to work requires the existence of sub-pixel shifts between the observed LR frames. Due to these fractional pixel shifts, registered LR samples will not always fall on a uniformly spaced HR grid, thus providing over-sampled information necessary for solving the SR inverse problem. The LR pixels in the data acquisition model are defined as a weighted sum of appropriate HR pixels. The weighting function, also known as the system degradation matrix, models the blurring caused by the point spread function (PSF) of the optics. An additive noise term can be added to compensate for any random errors and reading sensor noise in the acquisition model. Assuming that the resolution enhancement factor is constant and the LR frames are acquired by the same camera, it is logical to consider the same PSF and statistical noise for all the LR observations. Taking into consideration all these assumptions, a common observation model for the SR problem is formulated.

Consider $K$ low-resolution frames, $y_k$, $k = 1, 2, \ldots, K$, each arranged in lexicographical form of size $N_1 N_2 \times 1$ pixels. Let $L_1$ and $L_2$ be the resolution enhancement factors in the horizontal and vertical directions, respectively. For simplicity of derivation, we assume $L = L_1 = L_2$. The values of the pixels in the $k^{th}$ low-resolution frame
The formulation of the existing solutions for the SR problem usually falls into three main categories. The Bayesian Maximum Likelihood (ML) methods that solves for a super-resolution image that maximizes the probability of the observed LR input images under a given model, or Maximum a Posteriori (MAP) methods that stabilizes the ML solution under noisy conditions by making explicit use of prior HR information. The regularized-norm minimization methods that solve for an SR image by minimizing an error criteria ($l_p$-norm) with a regularization term. Efficient methods emerged from this category by swapping the order of the warping and blurring operators in the observation model to fuse the images into one HR grid followed by an iterative regularized
optimization solution; thus, the Fusion-Restoration class of SR methods as discussed in Section 4.5. This swapping is feasible under the assumption of a circularly symmetric blurring matrix and a translation or rotation type of motion. The non-iterative kernel-based, also referred to as Fusion-Interpolation (FI), solutions merge all the observations on a common HR grid and solve for the best interpolation through adaptive kernel design techniques. This category is inherently computationally efficient since it is non-iterative in nature. At this point, we assume that all the parameters required by the model in generating the LR observations from the HR images are known, except for the parameters of the warping model.

2.4 Bayesian Super-Resolution Formulation

Bayesian MAP-based estimators are popular solutions for the SR problem, that offer fast convergence and high quality performance [24, 57–59]. In Bayesian SR solutions, all parameters or unknowns (i.e., HR image, motion parameters, and noise) and observable variables (i.e., the LR observations) are assumed to be unknown stochastic quantities with assumed probability distributions based on subjective beliefs or exper-
In the following formulation of the MAP solution, the motion parameters are assumed to be known for simplicity. For example, in cases of compressed video content, the motion vectors can be retrieved from the headers of the bitstream or can be computed by means of motion estimation techniques. In order to estimate the HR image, $z$, a Bayesian MAP estimator is formed given the low-resolution frames, $y_k$, where $k = 1, 2, \ldots K$, and appropriate prior. The HR estimate, $\hat{z}$, can be computed by maximizing the a posteriori probability, $\Pr(z|\{y_k\})$, or by minimizing the log-likelihood function [24]

$$\hat{z} = \arg\min_z \log[\Pr(z|\{y_k\})].$$

(2.3)

Using Bayes’ rule and assuming that the LR observations, $y_k$, are statistically independent of $z$, the problem reduces to:

$$\hat{z} = \arg\min_z \{-\log[\Pr(\{y_k\}|z)] - \log[\Pr(z)]\}.$$

(2.4)

Now solving for an accurate HR estimate in (2.4) is highly dependent on the prior HR image density $\Pr(z)$ and the conditional LR density $\Pr(\{y_k\}|z)$ models. Note that, when dropping the prior HR probability model in (2.4), the MAP optimization problem reduces to an ML estimation problem that is highly unstable under small errors in the parameters of the acquisition model and under noisy conditions [24]. In the Bayesian SR formulation literature [24, 57–59, 68–75], a zero-mean Gaussian distribution for the noise model is commonly assumed. From (2.4), and given that the elements of $n_k$ are i.i.d. Gaussian random variables, the conditional probability distribution can be modeled as follows:

$$\Pr(y_k|z) = \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^2_{\eta}} \exp\left\{-\frac{1}{2\sigma^2_{\eta}} \|y_k - W_k z\|^2\right\},$$

(2.5)

where $\sigma^2_{\eta}$ is the noise variance.

The problem of determining which HR image prior model is the best for a particular HR reconstruction is still an open problem widely targeted by various existing
literature [6, 76]. However, a common approach followed by existing MAP-based SR solutions is the assumption of smoothness constraints on the HR priors within homogeneous regions [24, 57, 77, 78]. These priors can generally be modeled as:

$$\Pr(z) \propto \exp \left( -\frac{\lambda}{2} \| Qz \|^2 \right)$$  \hspace{1cm} (2.6)

where $Q$ represents a linear high-pass operator that penalizes the estimates that are not smooth and $\lambda$ controls the variance of the prior distribution. In [24, 57], these piecewise smoothness priors in (2.6) takes the form of Huber-Markov random fields that are modeled as Gibbs prior functionals according to [79]. Then the prior model can be written as follows:

$$\Pr(z) = \frac{1}{2} z^T C^{-1} z$$  \hspace{1cm} (2.7)

where $C_z$ is the covariance of the HR image prior model imposing piecewise smoothness constraints between neighboring pixels. Thus, with the smoothness prior model (2.7) and a mutually-independent additive Gaussian noise on the prior error model (2.5), the MAP SR estimation problem can be formulated by minimizing the following convex cost function with a unique global minimum:

$$f(z) = \frac{1}{2} \sigma^2 \eta \sum_{k=1}^{K} (y_k - W_k z^T) (y_k - W_k z) + \frac{1}{2} z^T C^{-1} z$$  \hspace{1cm} (2.8)

where $W_k$ is the degradation matrix for frame $k$, $\sigma^2 \eta$ is the noise variance, $z$ is the HR frame in lexicographical vector form, and $y_k$ are the observed LR frames also in vector form. Thus, the MAP estimator can be reformulated as a least-squares error minimization problem, which in matrix notation is the $l_2$-norm square of the error vector, with a smoothness regularization constraint, given by:

$$\hat{z} = \arg \min_z \left\{ \frac{1}{\sigma^2 \eta} \sum_{k=1}^{K} \| y_k - W_k z \|^2_2 + \lambda \Gamma(z) \right\}$$  \hspace{1cm} (2.9)

where $\| . \|^2_2$ is the square of the $l_2$-norm and $\sigma^2 \eta$ is the variance of the Gaussian noise. The $\lambda$ in the second term is a regularization weighting factor, and $\Gamma(z)$ is a smoothness regularization constraint in function of the SR image prior.
2.5 Regularized-Norm Minimization Formulation

In this category of SR approaches, such as [25, 60–62], an SR image is estimated following a regularized-norm minimization paradigm. Then, in an underdetermined system of equations (2.1), estimating the HR image \( z \) given a sequence of LR observations \( y_k, k = 1, 2, \ldots K \), can be formulated as an optimization problem minimizing an error criteria and a regularization term. Thus, the SR optimization problem can be presented as follows:

\[
\hat{z} = \arg \min_z \{ f(z) \} \tag{2.10}
\]

where the cost function \( f(z) \) is of the form:

\[
f(z) = \frac{1}{\gamma} \sum_{k=1}^{K} E(y_k, W_k z) + \lambda \Gamma(z) \tag{2.11}
\]

In (2.11), \( E(., .) \) is the error term in function of \( y_k \) and \( z \). The weighting factors \( \gamma \) and \( \lambda \) are constant tuning parameters. The regularization term \( \Gamma(.) \), which is in function of the HR image only, is designed to preserve important image content or structures such as edges and objects and also to increase the robustness of the solution to outliers and errors in the system. Unlike MAP-based algorithms, such as the one presented in Section 2.4, there is no a priori assumption made about the distribution function of the reconstructed HR image. Farsiu et al. [25], proved that an \( l_1 \)-norm imposed on the error residual is the most robust solution against outliers. Also in [25], different regularization terms are considered for best performance in terms of robustness and edge preservation. Therefore, a general cost function formulation for the SR problem can be presented as follows:

\[
f(z) = \frac{1}{\gamma} \sum_{k=1}^{K} \| y_k - W_k z \|_p^p + \lambda \Gamma(z) \tag{2.12}
\]

where \( \| . \|_p^p \) is the \( l_p \)-norm raised to the power \( p \). In the methods of [24] and [25], the weighting factor \( \gamma \) is set to \( \{ \sigma_n^2, 1 \} \) for \( p = \{ 2, 1 \} \), respectively.
Previous SR solutions are based on the warp-blur observation model following (2.1). Assuming a circularly symmetric blurring matrix and a translation or rotation type of motion, then the motion and blur matrices in (2.1) are block circulant and can be swapped [80]. Consequently, the observation model, referred to by the blur-warp model, can be deduced from (2.1) by defining the degradation matrix as $W_k = DF_k H$, where $F_k$ is the warping matrix of size $N \times N$, $H$ is the blurring matrix of size $N \times N$ representing the common PSF function, and $D$ is the decimation matrix of size $N_1N_2 \times N$. The question as to which of the two models (blur-warp or warp-blur) should be used in SR solutions is addressed in [81]. Following this blur-warp observation model, a fast implementation of the regularized-norm minimization solution referred to as the Fusion-Restoration (FR) approach, can be achieved by solving for a blurred estimate, $z_b = H\hat{z}$, of the HR image followed by an interpolation and deblurring iterative step. Often, the blurred HR estimate is a non-iterative approach composed of registering all the LR observations relative to the HR grid and estimating the HR pixel by using an average or median operator of the LR pixels at each HR location [25, 60, 61]. The formulation of this FR approach will be further discussed in Section 4.5.
2.6 Non-Iterative Kernel-Based Formulation

In this category of non-iterative kernel-based solutions for the SR problem, following the blur-warp observation model, all the LR observations are registered and merged on a common HR grid and non-iterative kernel-based solutions solve for the best estimate through adaptive kernel design techniques. The blur-warp acquisition model can be visualized as a non-uniform sampler ($U = DF_k$) applied on the blurred HR estimate $z_b$ as shown in Fig. 2.4. A locally adaptive approach of this form of SR estimators is described in [26,63], where the fused LR samples are processed locally by blocks using a moving observation window to estimate the interpolation kernel and an estimation window to apply the designed kernel on the spanned LR observed samples to estimate the missing HR pixels. Fig. 2.5 shows the non-iterative kernel-based estimation approach using a locally adaptive convolution kernel processing.
From Fig. 2.5, assuming that the observation window is of size $W_x \times W_y$ pixels on the HR grid and spans $P = K W_x W_y / L^2$ LR pixels denoted by vector $G_i$, and that the estimation window is of size $D_x \times D_y$ pixels on the HR grid and spans $D_x D_y$ estimated HR pixels denoted by vector $D_i$, where $i$ is the location of the respective window index. Then, estimating a local set of HR pixels can be achieved by simply filtering the vector $G_i$ by its locally designed kernel coefficients $W_i$ following $D_i = W_i G_i$. In [26], the interpolation kernel coefficients are designed by minimizing the mean square error of $(D_i - W_i G_i)$. Solving for the optimal weights of the matrix $W_i$ for each partition reduces to an adaptive Wiener filter solution for the considered observed LR pixels in the window. In [63, 82], the weights of the Wiener filter are shown to be given by:

$$W_i = R_i^{-1} P_i$$

(2.13)

where $R_i = E \{ G_i G_i^T \}$ and $P_i = E \{ G_i D_i^T \}$. Thus, the determination of the weighting coefficients in (2.13) requires the unknown HR image that can be either modeled parametrically or by training data. To avoid training, a parametric modeling approach can be adopted as described in [26]. This category of SR estimation is inherently computationally efficient since it is non-iterative in nature.
Chapter 3

HUMAN VISUAL PERCEPTION

This chapter provides background material on the Human Visual System (HVS) and visual perception that are of particular relevance to the masking properties and visual attention modeling utilized later in this work. A detailed description of the HVS and visual perception can be found in [83, 84]. An overview of the HVS as well as a description of perceptual concepts, namely, contrast sensitivity, luminance and contrast masking, and Visual Attention (VA) background are introduced in this chapter.

3.1 Overview of the HVS

The HVS is an enormously complex system that approximately utilize 80 – 90% of all neurons in the human brain [85]. Young in [85] estimated 70% of the total number of neurons involved in vision are in the V1 area of the brain. However, it is estimated that half of the remaining 30% of neurons outside the V1 area can also be utilized for visual perception processing such as motion and eye movement, visual stability with eye movement, and fusion of visual, auditory and tactile information into a coherent percept [85]. Fig. 3.1 shows an overview of the major components of the HVS that can be subdivided into two major components: the eyes, which focus the incident light on the retina, located at the back of the eye; the retina, which converts the incident light energy into neural signals interpreted by the brain; and the visual pathway, which leads to the primary visual cortex, and along which the signals are transmitted and processed.

The optical system of the eye relies on the principles of refraction where light traverses several layers of transparent media with different refractive indices namely the cornea, the aqueous humor, the lens, and the vitreous humor. These different optical layers of the eye collaborate to focus images of the outside world onto the retina, the neural tissue at the back of the eye (Fig. 3.1). This entire process of adjusting the focus to different distances is called *accommodation*. The retina, as shown in Fig. 3.2,
is composed of different layers of five cell types - photoreceptors (cones, rods), horizontal, bipolar, amacrine, and ganglion cells. Note that, in Fig. 3.2, light traverses from left to right while neural processing traverses from right to left leading to the visual pathway. The actual light photoreceptors are the rods and cones, but the neural cells that transmit to the brain are the ganglion cells. The axons of these ganglion cells make up the optic nerve, the route by which information leaves the eye. Rods are responsible for scotopic vision (i.e., at low light levels), while cones are responsible for photopic vision (i.e., at relatively high light levels). There are approximately 5 million cones and 100 million rods in each eye [86]. Even though rods are far more numerous than cones, visual acuity under dim light conditions is poor. This is due to the fact that signals from many rods are processed by a single neuron, which increases light sensitivity but decreases resolution. However, signals from each cone are processed by several neurons which explain their poor light sensitivity but high resolution and hence are mainly responsible for our ability to see fine details. There are also three types of cones that respond to different light wavelengths (short or blue, medium or green, and long or red) which form the basis for color perception.

The number of cones and rods vary greatly over the surface of the retina in function of eccentricity, i.e. distance from the center of the retina, known as the fovea. The highest density of cones, up to $300,000/mm^2$ [86], is concentrated in the fovea and significantly degrades towards the boundaries. This explains that the highest human visual acuity is attained in the fovea which covers approximately a region of 2 degrees of visual angle. The central fovea contains no rods at all. The highest rod densities (up to $200,000/mm^2$ [86]) are found along an elliptical ring near the eccentricity of retina. The blind spot around the retina, where the optic nerve exits the eye, is completely void of photoreceptors.
As shown in Fig. 3.2, there is a variety of different neurons in the retina that interconnect to encode the photoreceptor signals [85]. These neurons are identified into four major types; Horizontal, Bipolar, Amacrine, and Ganglion cells. Light activates the photoreceptors, which influences the activity of neurons along the visual pathway. The retinal area that influences the firing of a neuron is called the receptive field of the considered neuron. The Ganglion cells are the final layer of processing that relays the output signal through the optic nerve of the retina to other processing centers in the brain. The receptive fields of the retinal ganglion cells are concentric, consisting of a roughly circular central area and surrounding ring with antagonistic center-surround behavior as shown in Fig. 3.3 [87]. Hence the retinal ganglion cells have two basic types of receptive fields: on-center/off-surround and off-center/on-surround. Light falling directly on the center of a ganglion cell’s receptive field may either excite or inhibit the cell. In the surrounding region, light has the opposite effect. This excitatory and inhibition behavior causes the uniform light falling on the cell to cancel and it
amplifies differences in luminance such as edges and contours. Hence, the information supplied by the retina to the brain weights the visual scene differently by emphasizing features, such as boundaries and edges, that convey important information. This unique center-surround receptive field is also a property of the lateral geniculate nucleus (LGN) neurons.

Given the varying sampling grid of the photoreceptors in function of eccentricity, the optics of the eyeball can be modeled as a 2-D spatial impulse response function,
Figure 3.4: PSF of the human eyeball. The $x$ and $y$ axis are in degrees of visual angle and the foveated area corresponds to the high acuity of the fovea.

a spatially variant point spread function (PSF) [88]:

$$h(r) = 0.952e^{-2.59|r|^{1.36}} + 0.048e^{-2.43|r|^{1.74}}$$  \hspace{1cm} (3.1)

where $r$ represents the eccentricity from the center of the fovea in degrees of visual angle (Fig. 3.4). Also, when modeling the receptive field of the ganglion cells along with the LGN cells in the visual pathway, the center-surround mechanism can be modeled by a Difference-of-Gaussians (DoG) represented as [89]:

$$DOG(x,y) = a_1e^\frac{-(x^2+y^2)}{s_1^2} - a_2e^\frac{-(x^2+y^2)}{s_2^2}$$  \hspace{1cm} (3.2)

where $a_1$ and $a_2$ normalize the areas, and $s_1$ and $s_2$ are space constants in the ratio of 1 : 1.6 (Fig. 3.5). Their exact values will vary as a function of eccentricity. The LGN cells, that is the last station in the visual pathway, relay and control information from the retina to the visual cortex. Similar to the retinal ganglion cells, the LGN cells have a center-surround receptive field behavior and respond to different types of stimuli,
namely, motion and spatial details. The visual cortex is responsible for all higher-level processing of vision and a lot is still to be discovered about its functionality. The visual cortex has an enormous variety of cells classified into two major categories: simple and complex cells [83]. The simple cells have center-surround characteristics and are sensitive to spatial frequencies, orientations, and phase. These cells can be modeled as Gabor filters serving as oriented band-pass filters as shown in Fig. 3.6 (see also [89] for a more accurate model). Complex cells are orientation selective but do not exhibit the center-surround mechanism. Instead, they respond to a properly oriented stimuli in their receptive field. They are sensitive to corners, curvature and breaks in lines.

It is well known that the knowledge about the behavior of the visual cortex is rather limited and an enormous amount of ongoing research is being conducted to uncover the mysteries of the human mind, more specifically the human visual perception. The HVS perceives the outside world in a rather complicated way and its response
to visual stimuli is sometimes unpredictable. In the HVS, visual information is not perceived equally as some information may affect our visual perception more than others. This behavior can be due to many unpredictable factors and various limitations of the HVS, such as contrast discrimination, masking, and light adaptation. Exploiting these limitations and behavior of the human visual perception can be of great importance in many applications entailing digital image enhancement, quality assessment, and video compression. In the remaining part of this chapter, important perceptual vision concepts that can be exploited to achieve more efficient image processing and enhancement, are introduced.

3.2 Light Adaptation

Visual perception is continually adapting by controlling the sensitivity to match the properties of the environment. The HVS is capable to discriminate between an enormous range of light intensities, whether it is under scotopic or photopic levels. However, the visual system cannot operate over such a wide range simultaneously; rather it accomplishes this by a phenomenon known as light adaptation. Light adaptation, which influences luminance masking characteristics of the HVS, is the process of adjusting sensitivity to the average luminance in the scene. Thus, at any given adaptation

---

Figure 3.6: Receptive field profile of a simple cell in the primary visual cortex modeled by Gabor filters.
level, one can only distinguish between a small range of intensity levels. This arises from the fact that visual perception is sensitive to local luminance variations relative to the surroundings rather than the absolute luminance in the scene. Fig. 3.7 shows the different perception of a constant luminance depending on the surrounding adaptation levels. Two objects with identical intensity levels (inner square) over different surrounding intensities (outer square) are perceived as having different brightness.

The just-noticeable difference (JND) is the smallest intensity difference, $\Delta L$, that can be distinguished from a background of constant intensity, $L$. The Weber-Fechner law defines the Weber contrast as a measure of this relative variation of luminance to its surrounding and can be expressed as follows [39]:

$$C = \frac{\Delta L}{L}$$  \hspace{1cm} (3.3)

where $L$ represents the initial intensity of background, $\Delta L$ is the JND threshold between the foreground and background, and $C$ is the Weber constant or contrast threshold, defined as the minimum contrast necessary for an observer to detect a change in intensity $\Delta L$. $C$ remains nearly constant over a significant range of intensities $L$ due to the adaptation capabilities of the human visual system. Weber’s law states that the ratios of light levels are the main factors in determining the perceptual response.

The contrast sensitivity and luminance masking of our eye can be measured by determining the just noticeable difference luminance, $\Delta L$, required to produce a visual sensation over a certain background luminance level for stimuli at different frequencies. This can be achieved experimentally by recording the levels at which a subject can see a stimulus over a constant background level by increasing $\Delta L$, as shown in Fig. 3.8.

### 3.3 Contrast Sensitivity

Neurons in the HVS respond to stimuli above a certain contrast level. The necessary contrast needed over a uniform-intensity background to enable a response from the
neurons and thus to detect a stimulus in a scene is defined as the detection threshold. The inverse of this detection threshold is the contrast sensitivity [90]. Contrast sensitivity varies with spatial frequency, temporal frequency, and orientation. The contrast sensitivity can be modeled experimentally by a Contrast Sensitivity Function (CSF).

In measurements of the CSF, there are many possible definitions of contrast. The Michelson contrast usually applies in cases of periodic stimuli such as sine-wave gratings and is defined as follows [90]:

$$ C = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{max}} + L_{\text{min}}} $$  \hspace{1cm} (3.4)
where $L_{min}$ and $L_{max}$ are the luminance extrema of the pattern. The Michelson contrast is commonly used for patterns where both bright and dark features are equivalent and take up similar fractions of the area in a scene.

The root-mean-square (RMS) contrast, does not depend on the spatial frequency content or the spatial distribution of contrast in the image [36]. The RMS contrast is defined as the standard deviation of the pixel intensities:

$$C = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (I_{ij} - \bar{I})^2}$$  \hspace{1cm} (3.5)$$

where $I_{ij}$ is the pixel intensity at the location $(i, j)$ of the image of size $M \times N$ and $\bar{I}$ is the average intensity of the image.

The contrast sensitivity as a function of spatial frequency determines the Contrast Sensitivity Function (CSF) and can be modelled as follows [89]:

$$CSF(f) = 2.6 (0.0192 + 0.114f) \exp^{-0.114f^{1.1}}$$  \hspace{1cm} (3.6)$$

where $f$ is the spatial frequency in cycles/degree of visual angle. The CSF (Fig. 3.9) is typically bandpass, sensitive to spatial frequencies between 5 and 10 cycles per degree and less sensitive to very low and very high frequencies. This fact can be used to develop perceptual-based image processing algorithms.

### 3.4 Contrast Masking

The term masking refers to any effect among stimuli that have similar characteristics in space and time (frequency, orientation, color, etc.) [91]. The effects of masking may be a decrease in brightness, errors in recognition, or a failure to detect. The description here is concerned only with the effect of one stimulus on the detectability of another where the stimuli are coincident in space and simultaneous in time. Spatial contrast masking effects are usually quantified by measuring the detection threshold for a target stimulus when it is superimposed on a masker with varying contrast and frequency.
Figure 3.9: Contrast sensitivity function.

content [91]. Thus, it measures the variation of the detection threshold of a target signal as a function of the contrast and/or frequency of the masker. Spatial masking explains why similar artifacts can be disturbing in certain regions of an image and hardly noticeable elsewhere. Fig. 3.10 demonstrates how the same artifact of additive Gaussian noise can be perceived differently when added to different regions in an image due to masking effects. Compared with the original image on the right, the noise patch in the middle image added on the sky region is much more annoying than the noise patch masked by the grass texture. Note that many perceptual-based video coding techniques take advantage of the contrast masking effect by ignoring any perceptually undetected signal [92].

3.5 Human Visual Attention

The human visual perception is faced with limited resources when confronted with the vast information in a visual content. As stated earlier, the receptors in the retina are highly dense in the fovea and rapidly diminish with increasing eccentricity. The human visual system scans a visual scene through a small window, restricted by the foveal region, having high central resolution and degrading resolution towards the pe-
Figure 3.10: Demonstration of noise masking by adding a $30 \times 30$ pixels of a zero-mean Gaussian noise with variance 150, indicated by the arrow. Left image is the original image, Middle image with noise added on the sky, Right image with noise added on the grass.

ripheries. A large field of view is processed by a number of fixation points, attended to with high visual acuity, connected by fast eye movements referred to as saccades. The human visual attention or fixations are driven by saliency-specific bottom-up mechanisms, and by cognitive top-down mechanisms [93]. The bottom-up attention is driven by certain low-level features that can be salient and distinguishable in an image, such as contrast, oriented edges, luminance levels, motion, and colors. So bottom-up approaches are signal driven and are experimentally recorded using eye trackers under free viewing conditions. Top-down attention is driven by higher-level cognitive factors and external influences, such as experience, emotions, memory, personal preference, and viewing-task. The top-down fixations can be recorded using an eye tracker under task-driven conditions, such as finding a target in a scene. Visual perception is active only during fixations and suppressed during saccades [94]. Computational VA models have been the subject of extensive research aiming to automatically predict the gaze locations of human observers. The automatic detection of salient image regions is important for many applications involving adaptive region-of-interest-based image compression, video summarization, progressive image transmission, attention-driven image segmentation, image and video quality assessment, object recognition and track-
ing, and efficient perceptual image enhancement as will be proposed later in this work. Generally, eye tracking data under task-driven or free viewing experimental conditions can be recorded to assess the validity of these VA computational models.

In the following section, a general structure for computational VA systems is presented with a focus on bottom-up VA models, and a brief overview of some main computational VA models is provided to set a background for their application in imaging systems.

3.6 General Framework of Visual Attention Systems

Generally, bottom-up VA systems differ in details, but follow a similar structure as illustrated in Fig. 3.11. The basic idea is to compute several features in parallel and fuse these features in a linear or nonlinear approach creating a saliency map. A saliency map is simply a likelihood map in which regions with large values have higher probability
of being selected as a fixation region by the HVS as compared to regions with lower values. These maps can give a relative measure of the HVS attraction to spatial locations in a scene considering several competing low-level features. Extracted features can range from basic to complex stimuli that are of significance to human visual perception. Commonly used features are intensity, color, and orientation that are of major significance to the basic characteristics of the retinal receptors and neurons in the visual pathway and that can be mathematically modeled and computed. Other features that can be considered are, for example, spatial resolution, optical flow, or corners. Several VA models [36, 95–99] compute more complex features to extract image information. Examples for such features are entropy, Shannon’s self-information measure, ellipses, eccentricity, or symmetry. Considering more features usually results in more accurate and biologically plausible detection results, but it also reduces the processing speed. Hence, there is a trade-off between accuracy and speed in any VA model. In most of the existing VA systems, three to four features are used and seem to be a practical compromise.

Before each of the feature maps are combined into one saliency map, they are usually normalized and weighted. Normalization is done to remove the differences between a priori not comparable modalities with different extraction mechanisms. A widely used straightforward approach is to normalize all maps to a fixed range [34]. Usually, a weighting function is applied to each map before summing up the maps into one saliency map. This weighting function determines the significance of features, and it can take a linear or nonlinear form that represents the between-map interactions.

Top-down cues usually influence our attention according to the current situation. These cues include aspects like current tasks and prior knowledge about the target, the scene, or the objects that might occur in the environment, as well as emotions, desires, and motivations. Top-down features are still not clearly understood or
accurately modeled and constitute the subject of ongoing research in the field of computational VA. Whether these cues are computed in parallel and applied directly to the saliency map, or computed prior or after the saliency map and used in a feedback loop or a Bayesian framework is still a design specific option and not theoretically verified [100, 101].

The saliency map shows the saliency for each region of a scene. However, the output of the VA system is a trajectory of image regions, mimicking the sequence of human saccades and fixations, which starts with the highest saliency value. In Fig. 3.11, the max finder selects the image regions that are of local maximum significance in the saliency map. Max finder techniques can be implemented in various ways, one of which is the biologically plausible winner-take-all (WTA) neural network implemented by Koch et al. [93, 102]. A simpler and efficient alternative to the WTA is to simply determine the locations with the largest saliency values in the saliency map [34]. The focus of attention is usually not a single point but an attention region. The simplest approach is to determine a fixed-sized circular region corresponding to a foveal region of 1 – 2 degrees of visual angle around the most salient point [34]. More sophisticated approaches integrate segmentation approaches on feature or saliency maps to determine an irregularly shaped attention region [102]. In computational systems, inhibition of return (IOR) is implemented by inhibiting the surrounding region at the attended location of saliency map [34, 36]. It is observed in human vision that the speed and accuracy of attended regions is impaired after a target was attended, which prevents the region of attention from always staying at the most salient region. In the following, an overview of existing computational VA models is given and then several popular systems that are of relevance to the proposed resolution enhancement framework is focused upon.
3.7 Overview of Existing Computational Visual Attention Methods

The saliency model proposed by Itti et al. in [34] is one of the earliest and most popular computational saliency model. It is an expansion and implementation of the basic ideas introduced by Koch and Ullman in [93]. The model is inspired by the feature integration theory [103] and motivated by neurobiologically plausible modeling of the early visual pathway in the HVS. The model takes a hierarchical structure of an image as input, which is decomposed into three channels: intensity, color, and orientation. A center-surround operation, implemented by taking the difference of the filter responses from two scales of the hierarchy, yields a set of feature maps that mimic the center-surround behavior of the Ganglion cell receptive fields in the visual pathway. The feature maps for each channel are then normalized and combined across scales and orientations, creating conspicuity maps for each channel. The conspicuous regions of these maps are further enhanced by normalization, and the channels are linearly combined to form the overall saliency map. A Winner-Take-All (WTA) neural network is then implemented to detect the ordered fixations and the VA regions. This model has been shown to be successful in predicting human fixations and to be useful in object detection [34, 104].

In [95–97], a computational attention model is proposed based on the Information Maximization (AIM) model derived from efficient coding and information theory. The premise is that the saliency of visual content is equivalent to a measure of the local information present within a scene as defined by its surround, or more specifically, how unexpected the content in a local patch is based on its surrounding. AIM uses Shannon’s self-information measure to transform the image feature plane into a dimension that closely corresponds to visual saliency. The information conveyed by an image feature is inversely proportional to the likelihood of observing that feature. The features are coefficients of basis functions trained through Independent Component...
Analysis (ICA) [105] over a large set of random patches from a database of natural images. The linear combinations of the resulting basis functions can be used to describe each local neighborhood of any arbitrary image. A probability density function for each basis function corresponding to a particular local window is computed. These estimated likelihood measures, $p_c$, of each corresponding basis coefficients is translated into Shannon’s measure of Self-Information by $-\log(p_c)$. The resulting information map depicts the saliency attributed to each spatial location based on the aforementioned computation. VA information in the AIM approach is conveyed as a saliency map that maximizes the information measure of the scene patches.

Oliva et al. [100] also followed a similar statistical approach and defined the saliency of a location to be inversely proportional to its probability of occurrence in the image. Hence, the saliency of a location is large when the image features at that location are unexpected in the image. In [100], the probability of feature vectors at each location is approximated by fitting a Gaussian to the distribution of the local features in the image. A top-down mechanism is added in [100] based on contextual information for target detection. A contextual saliency map is estimated based on a Principal Component Analysis (PCA) training using the local features resulting in a trained set of basis functions. Then a Bayesian framework is used to detect the presence of an object at a certain location given the distribution of the contextual basis functions and the Gaussian feature fits.

In [101], Zhang et al. proposed a saliency approach using natural statistics (SUN) in a Bayesian framework that incorporates bottom-up saliency, as the self-information of visual features, and top-down information as the pointwise mutual information between the features and a specific target. The proposed method is target-specific achieved by estimating the probability of a target at every location given the visual features observed. As in [97], the features are learned using ICA from natural
image statistics and similar to [100] a Bayesian framework is incorporated for target-driven tasks.

In [106], Gao et al. defined saliency based on the idea that pixels are salient if they differ greatly from their surroundings. They use difference of Gaussians (DoG) filters and Gabor filters, measuring the saliency of a point as the Kullback-Leibler (KL) divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region. Seo et al. in [98, 99] computes a Saliency Detection method based on Self-Resemblance (SDSR) by measuring the similarity of the feature matrix at a pixel of interest relative to its neighboring feature matrices. Local steering kernels are used to capture local features and thus generate the feature matrices. A matrix cosine similarity, which is a generalization of cosine similarity, is employed to measure the likeness of each pixel to its surrounding. The saliency value at each pixel represents the statistical likelihood of its feature matrix given the feature matrices of its surrounding pixels.

In [36], a Gaze-Attentive Fixation Finding Engine (GAFFE) is proposed based on local low-level features at the point of gaze in a scene. Using a foveated framework, corresponding to the irregular sampling of the receptors in the retina, basic low-level features of luminance and contrast and their bandpass outputs are evaluated for their ability of attracting human fixations. An eccentricity based analysis is performed where varying diameters of circular patches around observer’s fixations, recorded by an eye tracker, are processed and the relevant image features are analyzed separately for each blur level (due to the different diameters of the foveated image). The VA model, proposed in [36], starts by foveating the center of the image and then calculates the low-level features on the foveated image. Then, the different features are combined into a saliency map by using a weighted sum approach. A max finder is used to detect the next fixation which will be foveated and the process repeats itself. The max finder
is simply a max operator on the foveated saliency map. Inhibition-of-Return (IOR) around each fixation point is achieved by weighting the resulting saliency map around each fixation point by an inverted Gaussian mask.

In [37], a computationally efficient Frequency-Tuned Salient (FTS) region detection algorithm is proposed that outputs well-defined boundaries of salient objects. This approach can be effectively used towards applications involving fast segmentation and target detection. The underlying hypothesis is that fixations are driven by local center-surround feature contrasts. This is achieved by using Difference of Gaussian (DoG) filters to compute a saliency map of local color and luminance features. This approach is implemented by simple operators, such as a $5 \times 5$ blurring kernel, a mean operator, and a Euclidean distance operator.

Most of the statistical-based approaches [95–98, 100, 101, 106] described earlier require training of a large set of specific natural images, which may not be effective for applications of efficient super-resolution techniques that can be widely used on any scene. The hierarchical and foveated saliency-based models, proposed in [34] and [36] respectively, are biologically plausible and utilize features that are essential to the HVS and that can be relevant for resolution enhancement processing, such as edge orientations, contrast, and luminance. Furthermore, the FTS method proposed in [37] is highly efficient and thus can be valuable to efficient image processing applications. Details regarding VA models presented in [34,36,37] are provided in the following section to set the path for the proposed selective efficient SR framework.

### 3.8 Hierarchical VA Model

In [34], Itti et al. compute a hierarchically-inspired saliency map, $S_{IT}$, using center-surround differences of intensity, color, and orientation between different scales of a dyadic Gaussian hierarchical structure. However, color channels will not be considered
in this thesis. The center-surround differences of intensity and orientation features are representative of the behavior of the ganglion and LGN neurons in the visual pathway. The input image, $I$, is sub-sampled into a dyadic Gaussian pyramid of $\sigma$ levels obtained by progressively filtering and down-sampling each direction separately. The intensity information is the image intensity values at each pyramid level, $I_\sigma$. The orientation information is calculated by convolving the intensity pyramid with Gabor filters as follows [104]:

$$O(\sigma, \theta) = \sqrt{|I_\sigma * G_0(\theta)|^2 + |I_\sigma * G_{\pi/2}(\theta)|^2}$$

(3.7)

where $G_\psi(\theta)$ is a Gabor filter with phase $\psi = [0, \pi/2]$, and different orientations, $\theta = [0, \pi/4, \pi/2, 3\pi/4]$. The foveated visual perception and the antagonistic “center-surround” process are implemented as across-scale differences between fine levels, $c = \{2, 3, 4\}$ corresponding to center pixels, and coarse levels, $s = c + \delta$ with chosen $\delta = \{3, 4\}$ corresponding to surround pixels. The across-scale difference, $\ominus$, is calculated by interpolation to the finer scale followed by point-by-point subtraction. Feature maps which signify the sensitivity of the HVS to differences in intensity and orientation are calculated as follows:

$$F_I(c, s) = |I(c) \ominus I(s)|$$

$$F_O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|$$

(3.8)

At this point, 6 intensity feature maps and 24 orientation feature maps are created. Each group of feature maps is combined into two conspicuity maps through across-scale addition, $\oplus$, by down-sampling each map to the middle scale $m$ (e.g., scale 4 for 9 scales $\sigma = [0, 1, \ldots, 8]$) of the pyramid followed by point-by-point addition. A map normalization operator, $\mathcal{N}(.)$, is applied to scale the values of different ranges into a common fixed range, $[0 M]$. The conspicuity maps for intensity and orientation are
calculated as follows:

\[ C_I = \bigoplus_{c=2}^{c+4} \bigoplus_{s=c+3}^{c+4} \mathbb{K}(I(c,s)) \]

\[ C_O = \sum_{\theta} \mathbb{K}(\oplus_{c=2}^{c+4} \bigoplus_{s=c+3}^{c+4} \mathbb{K}(O(c,s,\theta))) \]  

(3.9)

A saliency map is then calculated by averaging over the two normalized conspicuities:

\[ S = \frac{1}{2}(\mathbb{K}(C_I) + \mathbb{K}(C_O)) \]  

(3.10)

At this point, the most salient locations corresponding to visual fixations and the focus of attention (FOA), can be selected at any given time by selecting the highest values of the saliency map [34]. The most active attention regions can be detected by identifying a circular patch of 1 degree of visual angle around a fixation at any point in time. The diameter of the circular patch can be between 1 – 1.6 degrees of visual angle corresponding to the area of high photoreceptors concentration around the center of the fovea. Moreover, a more neuronally plausible approach using a WTA neural network can be implemented to detect fixations but at the expense of significantly higher computational complexity. Fig. 3.12 shows the conspicuity maps generated by computing the intensity and orientation features. Fig. 3.13 shows the saliency mask and the attended regions generated by the hierarchical VA model using circular patches corresponding to 1 degree of visual angle which corresponds to 64 pixels in diameter. Visual fixations in the attended regions were generated using the WTA approach [34].

The operational overhead introduced by generating the foveated saliency, \( S_{IT} \), of [34] is analyzed in this fashion. Generating \( L \) levels of the intensity dyadic Gaussian pyramid involves convolution with a separable symmetric kernel of size, \( W \), followed by subsampling with a factor of 2. The convolution with a separable symmetric kernel requires \( 2(W-1) \) additions/pixel and \( 2\lceil W/2 \rceil \) multiplications/pixel. Subsampling by 2 involves selecting every other sample in both directions of each level, \( \sigma \), which
Figure 3.12: Calculated conspicuity maps of the hierarchical VA model.
Figure 3.13: Saliency map and attended regions of the hierarchical VA model.

does not require any computations. For a maximum of \( L \) levels, the first \( L - 1 \) levels of the dyadic pyramid contains a total of \( \sum_{\sigma=0}^{L-1} \left( N/2^{\sigma} \right) \) pixels, where \( N \) is the image total number of pixels, \( L \) is equal to \( \max(s) \), and \( s \) is the surround pixel level. The orientation pyramid generation, in (3.7), involves convolutions of each intensity level with Gabor kernels with different phases and orientations. Assuming that the number of phases and orientations are represented by \( \| \psi \|_o \) and \( \| \theta \|_o \), respectively, where \( \| \cdot \|_o \) is the \( l_o \)-norm, then a total of \( \| \psi \|_o \cdot \| \theta \|_o \) Gabor kernels, \( G_{\psi}(\theta) \), can be calculated offline and stored in a LUT. In [34], 2 orientations result in quadratic symmetric/anti-symmetric kernels (\( \theta = [0^\circ, 90^\circ] \)) and 2 result in symmetric/anti-symmetric kernels (\( \theta = [45^\circ, 135^\circ] \)). Then, the convolution with each quadratic symmetric/anti-symmetric Gabor kernel, of size \( [H \times H] \), requires \( H^2 - 1 \) additions/pixel and \( \lceil H/2 \rceil^2 \) multiplications/pixel, and convolution with each symmetric/anti-symmetric Gabor kernel requires \( H^2 - 1 \) additions/pixel and \( \lceil H^2/2 \rceil \) multiplications/pixel. Generating the orientation pyramids also involves computing the magnitudes of the Gabor filtering results. This requires \( 2 \| \theta \|_o \) of each of the additions and multiplications per pixel, and
\[ \|\theta\|_o \] square root operations per pixel. A total of \[ \sum_{r=\min(c)}^{\max(c)} (N/2^2)^r \] pixels need to be processed to generate the orientation features, where \( c \) and \( s \) are the center and surround pixel levels. Across-scale difference operators, used to generate the feature maps in (3.8), are calculated by nearest-neighbor interpolation to scale, \( c \), followed by point-by-point subtraction. Then, generating all the intensity feature maps requires \[ \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} (N/2^2)^r \] of each of the additions and absolute values operations, where \[ \|\delta\|_o \] is the number of \( \delta \) values used. Similarly, the calculation of the orientation maps requires \[ \|\delta\|_o \cdot \|\theta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} (N/2^2)^r \] of each of the addition and absolute value operations.

At this point, a total of \[ \|s\|_o \] intensity feature maps and \[ \|s\|_o \cdot \|\theta\|_o \] orientation feature maps are generated. Then, the conspicuity maps are generated by combining the intensity and orientation features into one intensity and one orientation conspicuity maps, using normalization operators and across-scale addition operators, as shown in (3.9). The normalization operator, \( \mathcal{N}(.) \), maps all the low-level feature maps into a common range, \([0,R]\), and is calculated as \[ [(I_\sigma - \min(I_\sigma))/\max(I_\sigma) - \min(I_\sigma)]/R. \] Thus, the total number of operations needed for all the normalizations used in generating the intensity conspicuity maps is, \[ \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} \max (N/2^2)^r + 1 \] additions, \[ \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} \max (N/2^2)^r + 1 \] multiplications, and \[ 2 \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} (N/2^2)^r - 2 \|\delta\|_o \] comparisons. The total number of operations used for normalizing the orientation conspicuity maps are \[ \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} \max (N/2^2)^r + 1 \] additions, \[ \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} \max (N/2^2)^r + \|\theta\|_o \cdot (N/2^2)^m + 1 \] additions, \[ 2 \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} \max (N/2^2)^r - 1 \] multiplications, and \[ 2 \|\delta\|_o \cdot \sum_{\sigma=\min(c)}^{\max(c)} \max (N/2^2)^r - 1 \] comparisons, where \( m \) is the level at which the saliency map is computed. Across-scale addition operators are calculated by interpolating (nearest neighbor) to scale, \( m \), followed by point-by-point additions. Then, the operations needed for the across-scale addition operators to generate the intensity and orientation conspicuity maps are \[ (\|s\|_o - 1) \cdot (N/2^2)^m \] and \[ (\|\theta\|_o \cdot \|s\|_o -
\(1 \cdot (N / 2^{2m})\) addition operations, respectively. Finally, generating the saliency map requires \(3 (N / 2^{2m}) + 2\) additions, \(2 (N / 2^{2m}) + 2\) multiplications, and \(4 (N / 2^{2m}) - 4\) comparisons.

Following the previous operation count analysis, the total operations required to generate the saliency map, \(S_{IT}\), for the hierarchical Visual Attention (VA) model is given by:

\[
O_{S_{IT}} (add) = 2 (W - 1) \sum_{\sigma=0}^{\max(s)-1} (N / 2^{2\sigma})
\]

\[
+ 4 (2H^2 - 1) \sum_{\sigma=\min(c)}^{\max(s)} (N / 2^{2\sigma})
\]

\[
+ \|\delta\|_o (\|\theta\|_o + 3) \sum_{\sigma=\min(c)}^{\max(c)} (N / 2^{2\sigma})
\]

\[
+ (\|s\|_o + \|\theta\|_o + \|\theta\|_o, \|s\|_o + 1) (N / 2^{2m}) + 2 \|\delta\|_o + \|\theta\|_o
\]

\[
O_{S_{IT}} (mult) = 2 \lceil W / 2 \rceil \sum_{\sigma=0}^{\max(s)-1} (N / 2^{2\sigma})
\]

\[
+ (4 \lceil H^2 / 2 \rceil + 4 \lceil H / 2 \rceil^2 + 8) \sum_{\sigma=\min(c)}^{\max(s)} (N / 2^{2\sigma})
\]

\[
+ 2 \|\delta\|_o \sum_{\sigma=\min(c)}^{\max(c)} (N / 2^{2\sigma}) + (\|\theta\|_o + 2) (N / 2^{2m})
\]

\[
+ 3 \|\delta\|_o + \|\theta\|_o + 2
\]

\[
O_{S_{IT}} (abs) = \|\delta\|_o (\|\theta\|_o + 1) \sum_{\sigma=\min(c)}^{\max(c)} (N / 2^{2\sigma})
\]

\[
O_{S_{IT}} (sqrt) = 4 \sum_{\sigma=\min(c)}^{\max(s)} (N / 2^{2\sigma})
\]

\[
O_{S_{IT}} (comp) = 4 \|\delta\|_o \sum_{\sigma=\min(c)}^{\max(c)} N / 2^{2\sigma} + 2 (\|\theta\|_o + 2) N / 2^{2m}
\]

\[
- 4 \|\delta\|_o - 2 \|\theta\|_o - 4
\]

55
where $N$ is the total number of image pixels, $\|c\|_o$ and $\|s\|_o$ are the center and surround number of levels respectively, $\|\delta\|_o$ is the number of possible values for $\delta = s - c$, $W$ is the size of the Gaussian kernel, $H$ is the size of the Gabor kernel with the number of phases $\|\psi\|_o$ and number of orientations $\|\theta\|_o$, and $m$ is the level at which the saliency map is generated. In our implementation, since $256 \times 256$ images/videos are used, $c = \{2, 3\}$ and $\delta = \{1, 2\}$, resulting in $\|\delta\|_o = 2$, $\|c\|_o = 2$, $\|s\|_o = 4$ with $\max(s) = 5$, $\min(c) = 2$ and $\max(c) = 3$. The Gaussian kernel size is $W = 6$, the Gabor kernel size is $H = 9$ with the number of phases $\|\psi\|_o = 2$ and number of orientations $\|\theta\|_o = 4$, and the map level is $m = 3$.

3.9 Foveated VA Model

The foveated visual attention model of [36] is based on the analysis of low-level features at the point of gaze in a scene. The saliency map, $S_{GAF}$, is generated by a foveated combination of low-level image features of mean luminance, contrast, and bandpass outputs of both luminance and contrast. The bandpass output of the luminance and contrast features is inspired by the concept that VA is drawn to regions that differ from their surrounding, which can be implemented by using Gabor filter responses to low-level features. For an image patch of size, $M \times M$, the mean luminance, $\bar{I}$, can be computed as follows:

$$
\bar{I} = \frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} I_i w_i \tag{3.16}
$$

where $I_i$ is the grayscale value of the pixel at patch location $i$, and $w_i$ is the raised cosine function given by:

$$
w_i = 0.5[\cos\left(\frac{\pi r_i}{R}\right) + 1] \tag{3.17}
$$

where $r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$ is the radial distance of pixel location $(x_i, y_i)$ from the center of the patch $(x_c, y_c)$ and $R$ is the patch radius. Then, the root-mean-squared
contrast, $C$, is computed as follows:

$$
C = \sqrt{\frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} w_i (I_i - \bar{I})^2 \over (\bar{I})^2}
$$

(3.18)

The Bandpass of Patch Luminance features are computed by using Gabor kernels designed for different eccentricity values. Eccentricity values, $e$, measured in degrees of visual angle, represent the distance needed to reach a particular patch from a given fixation point. Then the bandpass-luminance for each patch is computed as $G_{lum} = \max |G_{lum}(e) * I(e)|$, where $G_{lum}(e)$ is the designed luminance-bandpass Gabor filter at eccentricity, $e$. Similarly, the Bandpass of Patch Contrast is computed using $G_{grad} = \max |G_{grad}(e) * |\nabla(I(e))||$, where $G_{grad}(e)$ is the contrast-bandpass Gabor filter at eccentricity, $e$, that is designed using the image gradient, $|\nabla(I(e))|$. Please refer to [36] for details on the design of the Gabor filters. The saliency map is then calculated by weighting all the computed image features after being normalized to a fixed range, $[0, 1]$. To detect the ordered fixations at any point in time, the four low-level features are computed on a foveated version of the saliency map (initially, the input image is used in place of the saliency map) and a max operator is used to pick the next fixation point. A circular patch corresponding to 1.6 degrees of visual angle is used to detect the VA region around any fixation point. Each region is inhibited after detection with an inverted Gaussian kernel. Fig. 3.14 shows the feature maps generated by computing the low-level features of luminance, RMS contrast, bandpass luminance, and bandpass contrast. Fig. 3.15 shows the saliency mask and the attended regions for the first 5 fixation points that are generated by the foveated VA model [36].

The operational overhead to generate the foveated saliency map, $S_{GAF}$, of [36], is mainly due to calculating each of the low-level features of luminance, contrast, bandpass luminance, and bandpass contrast. The foveated processing of the low-level features is performed with patches of size $[M \times M]$. In [36], the patch sizes of $M = 96$
Figure 3.14: Generated feature maps of the foveated VA model.
are used corresponding to 1.6 degrees of visual angle. Foveation is applied to the input image before calculating the features. This involves multiplying the input image by another image consisting of a foveation function centered at the current fixation, which requires \( N \) multiplications. The mean luminance (3.16) can be implemented using a convolution operator with an \([M \times M]\) kernel size. Since \( M \) is large, the 2-D convolution is computed using Fast Fourier Transforms (\( FFT \)) requiring \((3N/8)\log_2 N\) complex multiplications and \( N\log_2 N\) complex additions for each \( FFT \) or \( FFT^{-1} \) operator [107], where each complex multiplication is equal to 4 multiplications and 2 additions, and each complex addition is equal to 2 additions. The kernels for several given patch sizes can be computed offline and their corresponding \( FFT \) coefficients can be stored in memory for retrieval. Then, convolution using \( FFT \) requires \( 3N\log_2 N + 4N \) multiplications and \((11N/2)\log_2 N + 2N \) additions. The RMS contrast (3.18) calculation is spatially variant for each patch so \( FFT \) convolutions cannot be used and a spatial convolution is used per patch. Each patch size is \( M \times M \) and RMS
contrasts are computed for $N$ patches centered at each pixel of the considered image of the saliency map. Then, each spatial convolution requires $[M/2]^2$ multiplications and $M^2 - 1$ additions per patch. To calculate the RMS contrast in (3.18) requires $N$ square root, $[M/2]^2N + 2N$ multiplications, and $M^2N$ additions for processing the total image pixels. Similar to the mean luminance calculation, the bandpass luminance and bandpass contrast can be calculated using FFT convolutions. The Gabor filters, $G_{lum}(e)$ and $G_{grad}(e)$, reflecting different eccentricities used for calculating the bandpass outputs of luminance and contrast, respectively, are designed offline and and their FFT coefficients can be pre-computed and stored in memory [36]. As in [36], five eccentricity values are used corresponding to values of 0.5, 1, 1.6, 2.5 and 3.0 degrees of visual angle for computing the feature maps. To calculate the bandpass luminance, one FFT-based convolution is used per patch. An extra $N(M^2 - 1)$ comparisons are needed for calculating the max() operator in addition to $N$ absolute value operations.

For calculating the bandpass contrast feature, an extra magnitude of the gradient is used which requires $3N$ additions, $2N$ multiplications, and $N$ square root operations. After the feature maps are generated, they are combined into a saliency map using weighted averaging. This requires $4N$ multiplications and $3N$ additions. Furthermore, as in [36], to reduce computations, the feature maps are computed on the input image downsampled by 2 in each direction, then $N$ in the following equations is equal to $(N_1 \times N_2)/4$, for an $N_1 \times N_2$ input image.

Following the previous operation count analysis, the total operations required to generate the saliency map, $S_{GAF}$, for the foveated VA model is given by:

$$O_{S_{GAF}}(add) = N \left( \frac{33}{2} \log_2 N + M^2 + 12 \right)$$

$$O_{S_{GAF}}(mult) = N \left( 9 \log_2 N + [M/2]^2 + 21 \right)$$

$$O_{S_{GAF}}(abs) = 2N$$
where $N$ is the $1/4^{th}$ of the total number of image pixels and $M$ is the patch size in each direction. In [36], the patch size is set to $M = 96$ pixels corresponding to 1.6 degrees of visual angle.

### 3.10 Frequency-Tuned VA Model

In [37], an efficient frequency-tuned saliency approach using a difference of Gaussians (DoG) of luminance intensity is proposed to generate the saliency map information, $S_{FT}$. The bandpass filters using DoG are designed by properly selecting the standard deviation of the Gaussian filters targeting saliency generation. In [37], the saliency map information ($S_{FT}$) is generated by taking the magnitude of the differences between a constant image with an intensity value equal to the mean of the input image and a
Gaussian blurred version of the input image using a $5 \times 5$ separable kernel as follows:

$$S_{FT} = |I_\mu - I_{w hc}|$$

(3.24)

where $|.|$ is the norm operator, $I_\mu$ is the arithmetic mean pixel value of the image, and $I_{w hc}$ is the Gaussian blurred version of the original image. This approach is targeted towards perceptual object segmentation applications. The VA regions can be estimated by segmenting the saliency map with an efficient fixed thresholding approach or with a complex adaptive thresholding approach using a modified K-means clustering technique [37, 108]. Fig. 3.16 shows the saliency mask and the attended regions generated by the Frequency-Tuned (FTS) VA model [37]. The attended regions are detected by using fixed thresholding on the saliency map to keep the highest 20% of the saliency information.

The operational overhead introduced by generating the foveated saliency, $S_{FT}$, of [37] is analyzed. To calculate the DC value of a $N = N_1 \times N_2$ image, $N - 1$ additions and 1 multiplication operations are needed. Generating the Gaussian blurred version using a $W \times W$ 2-D symmetric kernel requires $N(W^2 - 1)$ additions and $N\lceil W/2 \rceil^2$ multiplications. Then, the saliency map is generated by taking the magnitude of the difference vector, which requires $N$ differences and $N$ absolute value operations.

Following the previous operation count analysis, the total operations required to generate the saliency map, $S_{FT}$, for the frequency-tuned VA model is given by:

$$O_{S_{FT}}(\text{add}) = N(W^2 + 1) - 1$$

(3.25)

$$O_{S_{FT}}(\text{mult}) = N\lceil W/2 \rceil^2 + 1$$

(3.26)

$$O_{S_{FT}}(\text{abs}) = N$$

(3.27)

where $N$ is the total number of image pixels and $W = 5$ is the size of the Gaussian kernel.
Chapter 4

ATTENTIVE-SELECTIVE SUPER-RESOLUTION ESTIMATOR BASED ON PERCEPTUALLY SALIENT FEATURES

The high dimensionality of the SR image reconstruction problem demands high computational efficiency to be deemed of any practical value. As described in Section 2.2, iterative solutions were proposed to reduce the complexity and increase the stability of solving a very large system of linear equations. In [24], Hardie et al. used the gradient descent iterative procedure to solve a MAP-based SR problem. Other techniques for simplifying the problem of super-resolution, concentrate on reducing matrix operations in particular SR modeling scenarios [25, 60]. The Fusion-Restoration (FR) SR technique [25], reduces the matrix operations by eliminating the registration and decimation at each iteration. This was achieved through a median operator of the LR frames. Although these SR methods are theoretically justifiable and present reliable results in terms of image quality and robustness, still at each iteration all the pixels are processed on an HR grid inclusively and thus they still suffer from high computational complexity in solving the inverse problem. As a consequence, selective SR algorithms are introduced in which only a small set of significant pixels are super-resolved.

Early attempts on selective SR processing, as presented in [27, 28], used a gradient-based approach in order to detect active pixels that are significant for SR processing. Although the gradient-based (and other similar high-frequency detection) approach resulted in savings, it suffered from a significant drawback which consists of having to use a different threshold on the gradient for different images in order to be able to detect the pixels of interest and achieve good performance. This is not practical as it required manually tweaking the gradient threshold for each image differently. Furthermore, general high-frequency detection methods, such as gradient-based or entropy-based schemes, do not incorporate any perceptual weighting and cannot au-
automatically adapt to an image’s local content that is perceptually relevant to the Human Visual System (HVS). Thus, in the SELP SR method [29, 30], a set of perceptually significant pixels is determined adaptively using an automatic perceptual detection mechanism that proves to work over a broad set of images without any manual tuning.

The problem of devising automatic detection thresholds that can adapt to local image content perceptually is of major importance. Figs. 4.1 and 4.2 demonstrate the superior performance of perceptual automatic thresholds, such as the previously proposed SELP SR algorithm [29, 30], over the simple non-perceptual gradient-based and entropy-based SR approaches. To avoid the non-practical manual tweaking of the detection thresholds for each image differently and to avoid global thresholding that does not adapt to local image features and which was shown in Ivanovski et al. [28] to require manual tweaking to lead to good performance, the local non-perceptual detection thresholds for the non-perceptual gradient-based and entropy-based approaches, are computed as the mean value corresponding to the magnitude of the gradient or as the entropy of each block of $8 \times 8$ pixels. That is, the detection mechanism in [28] is replaced by locally thresholding the magnitude of the gradient (G-MAP), and locally thresholding the $9 \times 9$ block entropy (computed using entropyfilt.m in Matlab) (E-MAP).

Fig. 4.1 clearly demonstrates the superior perceptual quality of the SELP SR method (Fig 4.1(d)) as compared with the existing methods including the gradient-based (Fig. 4.1(e)) and the entropy-based (Fig. 4.1(f)) SR methods. Compared to the gradient-based (Fig. 4.1(e)) and the entropy-based (Fig. 4.1(f)) MAP SR methods, the SELP-MAP method (Fig. 4.1(d)) results in a better reconstruction of edges as it can be seen around regions such as the tripod of the camera, and the face of the cameraman. Furthermore, Fig. 4.2 illustrates the significant increase in SNR gains per iteration using the SELP-MAP method compared to the gradient-based and entropy-based MAP SR
Figure 4.1: Super-resolved $256 \times 256$ HR Cameraman image obtained using sixteen $64 \times 64$ low-resolution images with a $4 \times 4$ average blur and a zero mean Gaussian noise with standard deviation $\sigma_n = 10$. 
Figure 4.2: SNR comparison among the baseline MAP SR, the SELP-MAP SR, G-MAP SR with block mean thresholds, E-MAP SR with block mean thresholds, using sixteen $64 \times 64$ low-resolution images with noise standard deviation $\sigma_n = 10$ for the $256 \times 256$ Cameraman image.

methods. Figs. 4.1 and 4.2 also show the similar performance, in terms of visual quality and SNR gain, between the SELP-MAP and the baseline MAP SR algorithms.

To further enhance on our previously proposed SELP SR framework [30], our previous work in [31, 32] showed that not all the detail pixels detected by the SELP algorithm are needed to preserve the overall visual quality of an HR image. As described in Section 3.5, the human visual system scans a visual scene through a small window, restricted by the foveal region, having high central resolution and degrading resolution towards the peripheries. A large field of view is processed by a number of fixation points, attended to with high visual acuity, connected by fast eye movements referred to as saccades. The ordered selection of these regions of interest is predicted according to a visual attention model by studying the eye movement sensitivity to top-down mechanisms, such as image understanding, and bottom-up salient image features, such
as contrast of color intensity, edge orientations and object motion [109, 110]. Given the fact that the attended regions are processed at high visual acuity, artifacts present in these regions are better perceived by the HVS than artifacts present in non-attended areas. In consequence, the observer’s judgment of image quality is prejudiced by distortions present in salient regions as shown in our previous work in [35]. Following this logic, saliency maps generated by visual attention modeling, can play a fundamental role in reducing the number of processed pixels of the selective SR approaches. Hence, a selective SR approach that imitates the human visual perception and processing of visual content is proposed. Towards an effective SR solution, this chapter proposes a low complexity saliency detector designed for efficient attentive-selective SR estimators. Consequently, an improved ATtentive-SELective Perceptual (AT-SELP) SR framework is presented in order to reduce the computational complexity of iterative SR algorithms without any perceptible loss in the desired enhanced image/video quality. Moreover, different low-level features influenced by visual attention models presented in [34,36,37] are studied to illustrate the efficiency and quality of the proposed attentive SR framework.

4.1 ATtentive-SELective Perceptual Super-Resolution Framework

Previously proposed Bayesian MAP-based and regularized $l_p$-norm SR methods offer theoretically justifiable solutions with reliable results in terms of reconstruction quality and robustness. In [24, 25, 57–59, 62], iterative gradient steepest descent optimization methods are used to solve the SR problem. Consequently, in order to minimize the computational complexity, the selective SR framework [29–32] processes only a subset of active pixels at each iteration of the SR solution. In the selective SR framework, the gradient descent iterative solution of the SR problem in (2.10) and (2.11) is modified as follows:

$$\hat{z}_{n+1} = \hat{z}_n - \beta_n \cdot M_n \cdot \nabla f(\hat{z}) |_{\hat{z} = \hat{z}_n}$$

(4.1)
where $\beta_n$ is the step size in the direction of the gradient, $\nabla f(z)$, $n$ is the iteration number, and $M_n$ is a binary selective mask that signals the active pixel locations that need to be processed at each iteration. In this chapter, an efficient ATtentive-SELective Perceptual (AT-SELP) SR framework is proposed that integrates a contrast sensitivity threshold model and a new low-complexity perceptual-based saliency detector.

A block diagram of the proposed AT-SELP SR estimation framework is shown in Fig. 4.3. Initially, a rough estimate of the high resolution image, $z_0$, is obtained by either interpolating one of the LR images in the observed sequence or by fusing all the LR frames on one HR grid (also referred to as the shift-and-add technique [25]). Other techniques, such as learning-based approaches [64–66], can also be used to produce initial estimates. At each iteration and for each estimated HR image, active pixels are detected based on the human visual detection model described in Sections 4.2 and 4.3. Then, only these perceptually active pixels are updated by the SR estimation algorithm which is based generally on minimizing an SR cost function as in (2.10). As a result, at each iteration, only a subset of pixels (i.e., the active pixels) is selected for the SR processing phase.

The first phase of the proposed AT-SELP SR framework 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, $M_a$, is further iterated upon. The process of updating the HR estimates of the perceptual/attentive active pixels continues until a maximum number of iterations is reached or the system stabilizes, i.e. until $|M_a.(z_{n+1} - z_n)|/|M_a.z_n| < \epsilon$, in the attentive active region and $|M_p.(z_{n+1} - z_n)|/|M_p.z_n| < s.\epsilon$, 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. Only the selected perceptually active and salient attentive pixels need to
be included in computing the change in the estimated SR frames between iterations (since the other pixels do not change in value). It is necessary in the first phase of the algorithm to super-resolve the perceptually significant information of the non-attended regions to a certain acceptable quality level \((s.e)\) that will not attract and bias the HVS perception of the background quality. As a result, the salient regions are reconstructed with a higher visual acuity while maintaining a trade-off between smoothing the flat regions dominated by noise and sharpening the perceptually relevant edges. Furthermore, iterative SR estimation algorithms can be easily integrated within the proposed
AT-SELP SR framework. Hence, application of the proposed AT-SELP framework to an iterative MAP-based SR method [24], and a FR-based SR method [25] are described in Sections 4.4 and 4.5, respectively. The proposed AT-SELP SR scheme results in significant computational savings while maintaining the perceived SR image quality as compared to the iterative baseline SR schemes [24, 25].

The proposed AT-SELP framework is flexible in that any existing saliency-based VA model, such as [34, 36, 37], can be adopted to detect the attentive mask, $M_a$. Saliency maps, $S$, combine several low level image features that compete to attract the human attention, providing measures of the level of attention at every point in the visual scene. Existing saliency map generation techniques inspired by a hierarchical visual attention model (IT) [34], a foveated gaze attentive model (GAF) [36], and a frequency-tuned attention model (FT) [37] are adopted for further comparisons and analysis in the proposed AT-SELP SR framework. In [34], Itti et al. compute a saliency map, $S_{IT}$, using center-surround differences of intensity and orientation between different Gaussian scales. In [36], a saliency map, $S_{GAF}$, is computed using low-level features of patch luminance and contrast, and bandpass outputs of patch luminance and contrast. In [37], a Frequency Tuned Saliency approach using difference of Gaussians of luminance intensity is proposed to generate saliency information, $S_{FT}$. However, these existing saliency-based VA techniques result in high computational complexity [34, 36] which undermines the goal of efficient SR or result in a degraded reconstruction quality [37]. This necessitates the development of a new low-complexity saliency detector that is targeted towards SR applications. Details about the proposed low-complexity saliency detector that is used for the saliency map computation is described in the next section.
4.2 Perceptual Contrast Sensitivity Threshold Model

As described in Chapter 3, neurons in the primary visual cortex of the Human Visual System (HVS) are sensitive to various stimuli of low-level features in a scene such as color, orientation, contrast, luminance intensity, etc. [111]. The luminance sensitivity also referred to as light adaptation is the discrimination of luminance variations at every light level. Moreover, the contrast sensitivity is the response to local variations of luminance to the surrounding luminance [111]. Limits on the human visual sensitivity to low-level stimuli such as light and contrast are converted to masking thresholds that are used in perceptual modeling. Masking thresholds are levels above which a human can start distinguishing among several stimuli or distortions [89]. Thus, the human visual detection model discriminates between image components based on contrast sensitivity of local information to their surroundings. In [29, 30], the proposed SELP SR scheme attempts to exclude less significant information from SR processing by exploiting the masking properties of the human visual system through generating contrast sensitivity detection thresholds. 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.

Digital natural images can be represented without loss using linear weighted combinations of cosine functions using the Discrete Cosine Transform (DCT). This is exploited in the lossy JPEG standard and many image compression algorithms including lossless image representation and compression (such as lossless mode of JPEG and other). The models of [112] and [113] exploit this to derive contrast sensitivity thresholds for natural images in the DCT domain. The contrast sensitivity model considers the screen resolution, the viewing distance, the minimum display luminance, $L_{\text{min}}$, and the maximum display luminance, $L_{\text{max}}$. In our proposed scheme, the contrast sensitivity
thresholds are computed locally in the spatial domain using a sliding window of size $N_{blk} \times N_{blk}$. The obtained thresholds per block will be used to select the pixels to be super-resolved for each HR estimate. Our model involves first computing the contrast sensitivity threshold, $t_{128}$, for a uniform block having a mean grayscale value equals to 128, and then obtaining the thresholds for any block having arbitrary mean intensity using the approximation model presented in [113].

The contrast sensitivity threshold, $t_{128}$, of a block in the spatial domain is computed as:

$$t_{128} = \frac{TMg}{L_{max} - L_{min}}$$  \hspace{1cm} (4.2)

where $M_g$ is the total number of grayscale levels (i.e., $M_g = 255$ for 8-bit images), and $L_{min}$ and $L_{max}$ are the minimum and maximum display luminances, respectively. In (4.2), the threshold luminance, $T$, is evaluated based on the parametric model derived by Ahumada et al. [112] using a parabolic approximation where $T = \min(10^{g_0,1}, 10^{g_1,0})$ and $g_{0,1}$ and $g_{1,0}$ are given by:

$$g_{0,1} = \log_{10} T_{min} + K(\log_{10} \frac{1}{2N_{blk}w_y} - \log_{10} f_{min})^2$$  \hspace{1cm} (4.3)

$$g_{1,0} = \log_{10} T_{min} + K(\log_{10} \frac{1}{2N_{blk}w_x} - \log_{10} f_{min})^2$$  \hspace{1cm} (4.4)

In (4.3) and (4.4), $w_x$ and $w_y$ denote the horizontal width and vertical height of a pixel in degrees of visual angle, respectively. $T_{min}$ is the luminance threshold at the frequency, $f_{min}$, where the threshold is minimum, and $K$ determines the steepness of the parabola. The parameters $T_{min}$, $f_{min}$, and $K$ are the luminance-dependent parameters of the parabolic model and are experimentally modeled in [112] as follows:

$$T_{min} = \begin{cases} 
\left( \frac{L}{L_T} \right)^{\alpha_T} \frac{L_T}{S_0}, & L \leq L_T \\
\frac{L}{S_0}, & L > L_T
\end{cases}$$  \hspace{1cm} (4.5)
\[ f_{\text{min}} = \begin{cases} f_0 \left( \frac{L}{L_f} \right)^{\alpha_f}, & L \leq L_f \\ f_0, & L > L_f \end{cases} \]  
(4.6)

\[ K = \begin{cases} K_0 \left( \frac{L}{L_K} \right)^{\alpha_K}, & L \leq L_K \\ K_0, & L > L_K \end{cases} \]  
(4.7)

The values of the constants, in (4.5), (4.6) and (4.7), are given in [112] to be: \( L_T = 13.45 \) cd/m\(^2\), \( S_0 = 94.7 \), \( \alpha_T = 0.649 \), \( \alpha_f = 0.182 \), \( f_0 = 6.78 \) cycles/deg, \( L_f = 300 \) cd/m\(^2\), \( K_0 = 3.125 \), \( \alpha_K = 0.0706 \), and \( L_K = 300 \) cd/m\(^2\). For a background value of 128, the local background luminance is computed as:

\[ L = L_{\text{min}} + 128 \frac{L_{\max} - L_{\min}}{M_g} \]  
(4.8)

where \( L_{\text{min}} \) and \( L_{\max} \) denote, respectively, the minimum and maximum luminance of the display. Once the threshold at a grayscale value 128, \( t_{128} \), is calculated using (4.2), the Just Noticeable Difference (JND) thresholds for the other grayscale values are approximated using a power function [113] as follows:

\[ t_{\text{JND}} = t_{128} \left( \sum_{n_1=0}^{N_{\text{blk}}-1} \sum_{n_2=0}^{N_{\text{blk}}-1} I_{n_1,n_2} \right)^{a_T} \]  
(4.9)

where \( I_{n_1,n_2} \) is the intensity level at pixel location \((n_1,n_2)\) and \( a_T \) is a correction exponent that controls the degree to which luminance masking occurs and is set to \( a_T = 0.649 \), as given in [113]. Note that, if the block has a mean of 128, \( t_{\text{JND}} \) of (4.9) reduces to \( t_{128} \) as expected.

A DELL UltraSharp 1905 FP LCD monitor is used to display the images. For a screen resolution of 1280 \( \times \) 1024, and for a measured luminance of \( L_{\text{min}} = 0 \) cd/m\(^2\) and \( L_{\max} = 175 \) cd/m\(^2\), \( t_{128} \) is computed to be equal to 3.3092 for \( N_{\text{blk}} = 8 \).

### 4.3 Proposed ATtentive-SELEctive Perceptual Detection

The previously detailed saliency map detectors (Section 3.7) provide relatively good reconstruction quality that validate the effectiveness of the AT-SELP SR framework.
However, since the existing saliency map generators in [34,36] are highly complex, the previously proposed SR framework [31, 32] assumes that the attention information is already computed and stored offline, thus, ignoring any computational overhead introduced by the adopted complex visual attention model. Furthermore, the more efficient saliency detector presented in [37] does not detect image features relevant to SR reconstruction and may lead to minimal enhancement in quality or even divergence in the SR solution. To avoid these limitations, a low complexity saliency detector relevant to SR applications, is needed to deem this framework of any practical value. Thus, a low complexity saliency-based detector is proposed for attentive-selective SR processing, as described in this section.

In the proposed AT-SELP scheme, the locally computed JND thresholds, $t_{JND}$ described in (4.9) over a local block of size $8 \times 8$, are used in generating the perceptual mask, $M_p$, in addition to the attentive mask, $M_a$, in order to select the candidate pixels to be super-resolved for each HR estimate. As shown in Fig. 4.4, after the $t_{JND}$ is obtained, the perceptual mask ($M_p$) is generated by computing the difference between the center pixel of a $3 \times 3$ sliding window with its 4 cardinal neighbors. If any of the 4 absolute differences is greater than the local computed $t_{JND}$, then the corresponding center pixel location is flagged as a perceptually active pixel for SR processing. Then, the attentive mask, $M_a$, is generated by first creating an initial saliency map based on the already computed contrast sensitivity thresholds ($t_{JND}$), and then finding the most significantly attended regions. The initial saliency map, $S_{JND}$, is computed at each pixel by weighting the maximum of the locally computed $3 \times 3$ neighborhood differences with the corresponding locally computed $t_{JND}$. Then, a saliency detection rule is applied by finding the maximum threshold, $C$, where $C > 1$, above which the probability of saliency detection is greater or equal to $\tau\%$ as follows:

$$\Pr(S_{JND} > C) \geq \tau\%$$  \hspace{1cm} (4.10)
The salient active pixels, corresponding to pixels at which \( S_{JND} > C \), are then flagged as attentive locations for SR processing in the proposed AT-SELP framework. These detected pixels represent salient information with a relatively large intensity change in their local surrounding \((C \times t_{JND})\). Then, the main advantages of the proposed saliency mask are its ability to adapt to local perceptual high intensity changes such as edge information as well as noisy regions over smooth areas essential for SR processing without requiring high computational complexity. Note that the computational overhead for generating the perceptual and attentive mask is minimal compared to the complex and computationally demanding saliency-based VA methods (please refer to Chapter 3 for a detailed complexity analysis of the saliency-based VA methods).

For each estimated HR image, the \( t_{128} \) is computed only once according to (4.9) and stored in a lookup table (LUT) in memory. Also, for all 8-bit images, the other 255 \( t_{JND} \) values, corresponding to the 255 possible mean intensity values (other than 128), can be pre-computed once and stored in a LUT. Thus, for each image block, the total sum of the block intensities is computed and the corresponding \( t_{JND} \) is simply retrieved from the LUT. Pertaining to the computational overhead, the contrast sensitivity mask generation requires \( O_{M_p}(add) = 5 \) additions/pixel (four to compute the difference with the neighboring pixels and one addition per pixel contributing to the computation of the mean over \( 8 \times 8 \) block). Then, \( O_{M_p}(comp) = 4 \) comparisons/pixel.
and $O_{M}(abs) = 4$ absolute-values/pixel are needed for the detection process. To generate the saliency map, $S_{JND}$, one division/pixel is required for weighting the maximum differences with the computed thresholds, $I_{JND}$. Then, the saliency mask, $M_{a}$, is generated using (4.10). This is implemented by sorting the $S_{JND}$ pixels according to their values resulting in a sorted array of $N$ indexed elements, where $N$ is the total number of pixels in the considered image. Then finding $C$ consists of retrieving the element corresponding to the index of the smallest value among the highest $\tau\%$ values. This detection process requires $N\log_2 N$ comparisons for sorting [114]. Then $N$ comparisons are required to generate the binary mask $M_{a}$. Thus, following the previous analysis for generating the attentive mask, $M_{a}$, requires a total of $O_{M}(comp) = 5N + N\log_2 N$ comparisons, $O_{M}(mults) = N$ multiplications or divisions, $O_{M}(abs) = 4N$ absolute values, and $O_{M}(adds) = 5N$ additions for the entire image of size $N = N_1 \times N_2$ pixels.

Fig. 4.5 shows an example of the perceptual active pixels that are obtained by applying the perceptual mask and the attentive mask, respectively, to a blurred and noisy version of the $256 \times 256$ ‘Lena’ image. Comparing Figs. 4.5(a) & (b), one can easily notice the significantly lower number of active pixels that are selected for SR processing using the attentive masking phase as compared to only using the perceptual mask, $M_{p}$.

### 4.4 ATtentive-SELective Perceptual MAP-Based SR Estimation

Existing Bayesian MAP-based SR estimators present high quality estimation but suffer from high computational requirements [24,57–59]. In order to illustrate the significant reduction in computations for MAP SR techniques, the popular algorithm presented by Hardie et al. [24], which is based on a gradient-descent optimization, is integrated into the proposed AT-SELP SR framework. In [24], a uniform detector sensitivity is assumed over the span of the detector degradation model, then the point spread function (PSF), $H$, in the observation model (Section 2.3) is represented using an averaging
Figure 4.5: Illustration of the obtained set of active pixels that are obtained by generating and applying (a) the perceptual mask \( (M_p) \) and (b) the attentive mask \( (M_a) \) for the blurred and noisy version of the 256 × 256 ‘Lena’ image. ‘White’ intensities correspond to selected (active) pixels.

Filter. Following the Bayesian MAP-based SR formulation in Section 2.4, the regularization smoothness constraint term, in the cost function (2.9), is represented according to [24] by \( \Gamma(z) = \frac{1}{2\lambda} z^T C_z^{-1} z \), where \( C_z \) is the covariance of the HR image prior model imposing a piecewise smoothness relationship between neighboring pixels in \( z \) as follows [24]:

\[
C_z^{-1} = \frac{1}{\lambda} \sum_{i=1}^{N} d_{i,k} \left( \sum_{j=1}^{N} d_{i,j} z_j \right) \tag{4.11}
\]
where $z_j$ is the pixel at the $j^{th}$ location of the lexicographically ordered vector $\mathbf{z}$, and $\lambda$ is a scaling factor controlling the effect of rapidly changing features in $\mathbf{z}$. The coefficients, $d_{i,j}$, express a priori assumptions about the relationship between neighboring pixels in $\mathbf{z}$ and are given by [24]:

$$d_{i,j} = \begin{cases} 
1, & \text{for } i = j \\
\frac{1}{4}, & \text{for } i \neq j: z_j \in \text{cardinal neighbors of } z_i 
\end{cases}$$

(4.12)

In the proposed AT-SELP-MAP SR scheme, the gradient descent minimization procedure is applied selectively and only to pixels that are determined to be perceptually significant by the proposed attentive-perceptual detector (Section 4.3). Following the steepest descent solution in the proposed AT-SELP framework and using (2.8) and (4.1), the estimated HR image can be computed as follows:

$$\mathbf{\hat{z}}_{n+1} = \mathbf{\hat{z}}_n - \beta_n \cdot \mathbf{M}_n \cdot \left\{ \frac{1}{\sigma_n^2} \sum_{k=1}^{K} \mathbf{W}_k^T (\mathbf{W}_k \mathbf{z}_n - \mathbf{y}_k) + \frac{1}{2} \mathbf{C}^{-1} \mathbf{z}_n \right\}$$

(4.13)

where $\mathbf{M}_n$ represents the attentive-perceptual mask for selecting active pixels at every iteration of the SR process, and $K$ is the total number of LR observations. Note that the initial HR image estimate is an interpolated version of one of the LR frames. The elements in $\mathbf{M}_n$ take binary values, where a value of ‘1’ indicates that the corresponding pixel is an active pixel that should be included in the SR update at the current iteration. Contrary, a mask value of ‘0’ indicates that the corresponding pixel is a non-active pixel that will not be processed in the current iteration.

Let $O_{MAP}(\text{add})$ and $O_{MAP}(\text{mult})$ denote the number of additions and multiplications operations, respectively, that are required to compute the gradient in (4.13) at any given pixel. The degradation matrix represented by $\mathbf{W}_k = \mathbf{DHF}_k$, forward projects the SR estimate to the LR grid. In [24], a uniform detector sensitivity is assumed over the span of the detector degradation model, and the point spread function (PSF), $H$, in the degradation matrix is represented using an averaging filter of size $L \times L$. \[78\]
where $L$ is the SR resizing factor. It follows that the forward projection, $W_k$, requires $L^2 - 1$ additions/pixel and 1 multiplication/pixel for the blur operator ($H$). The back projection matrix, $W_k^T$, brings back the estimation error to the HR grid. Then this requires 1 multiplication/pixel for the inverse blur (multiply the error by $1/L^2$). The convolution kernel used to obtain the image prior gradient in [24] is a $5 \times 5$ quadrant and diagonally symmetric filter, then calculating (4.11) requires 12 adds and 4 multiplies. Also, 1 addition/pixel for adding the error and prior terms, and 1 addition and 1 multiplication per pixel are required for the gradient descent step update at each iteration (4.13). Therefore, the total number of operations needed to calculate the gradient and the gradient-descent update in the MAP-based SR solution (4.13), consists of $O_{MAP}(add) = KL^2 + K + 13$ and $O_{MAP}(mult) = K + 6$ operations per pixel.

### 4.5 ATtentive-SELective Perceptual Fusion-Restoration SR Estimation

The proposed efficient AT-SELP framework is applied here to the Fusion-Restoration (FR) SR algorithms. Farsiu et al. [25] proposed a two-stage Fusion-Restoration (FR) super-resolution algorithm by using first a non-iterative data fusion step followed by an iterative gradient-descent deblurring-interpolation step. The algorithm in [25] models the relative motion between low resolution frames as translational and the point spread function (PSF) as an $L_1 \times L_2$ Gaussian lowpass filter with a standard deviation equals to 1. Following the regularized-norm SR formulation in Section 2.5, a fast Fusion-Restoration implementation of the minimization solution (2.12), can be achieved by solving for a blurred estimate, $z_b = Hz$, of the HR image followed by an interpolation and deblurring iterative step. In [25], an initial blurred version of the HR estimate, $z_b$, is estimated in the data fusion step by registration followed by a median operator of the LR frames on the HR grid, referred to as the “median shift and add” operator. As for the regularization term in (2.12), a bilateral total variation (TV) regularization that
preserves edges is adopted in [25] and is given by:

$$\Gamma_{BTV}(z) = \sum_{l=-R}^{R} \sum_{m=0}^{R} \alpha^{l+m+|l|} \|z - S_x^l S_y^m z\|_1$$ (4.14)

where $S_x^l$ and $S_y^m$ shift the HR image $z$ by $l$ and $m$ pixels in the horizontal and vertical directions, respectively, and $R \geq 1$ represents several scales of shifting values, such that $l + m \geq 0$. The weight $\alpha$ is applied as a decaying factor for convergence purposes, and is chosen between $0 \leq \alpha \leq 1$. The SR problem in (2.12) reduces to deblurring and interpolating for the missing pixels in the initial HR estimate, $z_b$, that is formulated as a regularized $l_1$-norm minimization problem [25]. In the proposed AT-SELP-FR SR scheme, the steepest descent minimization is applied selectively only to the active pixels detected by the proposed AT-SELP detection scheme that is presented as follows:

$$\hat{z}_{n+1} = \hat{z}_n - \beta_n M_n \left[ H^T A^T \text{sign}(AH\hat{z}_n - Az_b) + \lambda \sum_{l=-R}^{R} \sum_{m=0}^{R} \alpha^{l+m+|l|} \left[I - S_y^{-m} S_x^{-l}\right] \text{sign}(\hat{z}_n - S_x^l S_y^m \hat{z}_n)\right]$$ (4.15)

where $\beta_n$ is the step size in the direction of the gradient and $\lambda$ is a regularization weighting factor. Matrix $A$ is a $N \times N$ diagonal matrix with diagonal values equal to the square root of the number of measurements that contribute to make each element of $z_b$. Also, $S_x^{-l}$ and $S_y^{-m}$ define a shifting effect in the opposite directions of $S_x^l$ and $S_y^m$. The shifting values $l$ and $m$ should satisfy the condition $l + m \geq 0$. $M_n$ represents the perceptual-attentive masking that selects the active pixels that are processed at each iteration, thus reducing the computations required in [25].

The number of operations per pixel, that are required to compute the gradient at any given pixel of the FR-based SR method [25] using (4.15), are denoted by $O_{FR}(add)$, $O_{FR}(mult)$, and $O_{FR}(comp)$ representing the number of additions, multiplications, and comparison operations, respectively. A comparison operation is assumed to obtain the $\text{sign}$ operation. Also, a multiplication by a Gaussian blur matrix is the
same as convolving by a quadrant-symmetric Gaussian kernel of size $B \times B$ resulting in $\lceil B/2 \rceil^2$ multiplications and $B^2 - 1$ additions per pixel each time $H$ or $H^T$ occurs. Note that the operations needed to generate $A$ and $A^T$ are insignificant since these matrices are computed only once in the fusion stage and are retrieved from memory for later usage at each iteration. The shift matrices, $S^l_x$ and $S^m_y$ and their corresponding opposite direction shifts matrices $S^{-l}_x$ and $S^{-m}_y$, do not require any operations since the values of the pixels are not changed. In the regularization term of (4.15), $\alpha$ is raised to the power of $(a = |m| + |l|)$. Since the maximum value of $a$ is $2R$ (when $l = \pm R$ and $m = R$) and $R$ is typically set to a small value ($R = 2$ in [25]), one can compute all possible power values beforehand and store these in a Look-Up Table (LUT) in memory, thus saving computations. However, the double summation subject to $l + m \geq 0$ requires $\frac{3}{2}R^2 + \frac{5}{2}R + 1$ of the operations appearing inside the regularization term. Analyzing (4.15), the total number of multiplications, additions, and comparisons are given as $O_{FR}(add) = 2B^2 + 3R^2 + 5R + 4$, $O_{FR}(mult) = 2\lceil B/2 \rceil^2 + \frac{3}{2}R^2 + \frac{5}{2}R + 5$, and $O_{FR}(comp) = \frac{3}{2}R^2 + \frac{5}{2}R + 2$, respectively. These operation estimates do not include the computation of the blurred reconstructed image $\hat{z}_b$ since it is computed only once in the initial iteration and its contribution to the total number of operations is minimal.

Note that [25] assumes that $R = 2$ and $B = L$ ($L$ is the SR resizing factor), which will be reflected in the simulations that are presented in Section 4.6.

4.6 Simulation Results

In this section, the performance of the proposed AT-SELP SR framework is assessed using a set of simulated sequence of images where a sequence of LR images is generated from a single HR image. In this scenario, a single HR image is passed through the SR degradation model described in Section 2.3 to generate a sequence of blurred, shifted, and noisy LR images. Then, for a resolution factor of 4, the degradation process is applied by randomly shifting the reference $256 \times 256$ HR image in the horizontal and
vertical directions (Fig. 4.6), blurring the shifted images with a low-pass filter of size $4 \times 4$, and sub-sampling the result by a factor of 4 in each direction to generate sixteen $64 \times 64$ LR frames. The blur filters are modeled as an average filter and as a Gaussian filter with a standard deviation of one for the MAP-based [24] and the FR-based [25] SR observation models, respectively. Then, an additive Gaussian noise of variance $\sigma^2 = 16$, is added to the resulting LR sequence. Fig. 2.3 demonstrates the simulated sequence generation process for the ‘Monarch’ image cropped to size $256 \times 256$ from the LIVE database [115].

The images selected for this simulation are the $256 \times 256$ ‘Cameraman’, ‘Lena’ (resized from $512 \times 512$ to $256 \times 256$), and ‘Clock’ images from the USC image database [116], and the ‘Monarch’ image cropped to size $256 \times 256$ from the LIVE
Figure 4.7: Original images of size $256 \times 256$ used to test the proposed AT-SELP SR framework.
image database [115], as shown in Fig. 4.7. The selected images have different characteristics, for example, the ‘Clock’ and ‘Monarch’ images contain many smooth regions while the ‘Cameraman’ and ‘Lena’ images have more edges and texture variations. Moreover, the ‘Lena’ image can be used to demonstrate the application of SR to face recognition applications. The proposed ATtentive-SELective Perceptual MAP-based Super-resolution and ATtentive-SELective Perceptual Fusion-Restoration Super-resolution schemes referred to as AT-SELP-MAP and AT-SELP-FR, respectively, are compared with their existing non-selective counterparts MAP-SR [24] and FR-SR [25], as well as the previously proposed selective SR counterparts SELP-MAP and SELP-FR [29, 30]. The simulation parameters for the compared MAP-based SR methods are set to $\lambda = 100$, $\varepsilon = 0.0001$, $s = 100$, and a maximum of 20 iterations is performed, while the simulation parameters for the compared FR-based SR methods are set to $R = 2$, $\alpha = 0.6$, $\lambda = 0.08$, and $\beta = 8$, $\varepsilon = 0.0001$, $s = 120$ and a maximum of 30 iterations is performed. The parameter $\tau$ of the saliency map detection rule (4.10) is set to 20% to identify the attentive active pixels. Experimentally, a percentage value of 20% for $\tau$ was found to give good results in terms of the tradeoff between quality and computational efficiency.

For saliency mask detection comparisons, different saliency maps generated from existing VA models presented in [34, 36, 37] are also integrated and tested in the proposed AT-SELP SR framework. For comparison, Fig. 4.8 shows the attention mask, $M_a(JND)$, that is generated using the proposed JND-based saliency detection scheme compared with the attention masks, $M_a(IT)$, $M_a(GAF)$, $M_a(FT)$, that are generated from the computed saliency maps of $S_{IT}$ [34], $S_{GAF}$ [36], and $S_{FT}$ [37], respectively. As shown in Fig. 4.8, the attention mask generated by the proposed low-complexity JND-based detector, $M_a(JND)$, is better adapting to information essential to SR processing such as edges and perceived noise over smooth areas as compared to
Figure 4.8: Detected salient active pixels, $M_a$, for the $256 \times 256$ bicubicly interpolated Frame 9 of the Cameraman image with a $4 \times 4$ average blur, a Gaussian noise with $\sigma^2 = 16$, and $\tau = 20\%$.

the existing schemes. Furthermore, except for [37], the existing VA methods are more complex than the proposed JND-based saliency detector scheme, and are not suitable for real-time applications. The proposed JND-based detection method does not add high computational overhead since it reuses the JND thresholds computed in the first phase of the AT-SELP SR framework. Although, the VA method in [37] is efficient in
nature, it detects objects and flat areas that are not essential for SR processing and is more applicable towards segmentation applications as shown in Fig. 4.8 (c).

Fig. 4.9 gives a quantitative comparison, in terms of SNR gain (Fig. 4.9 (a)) and computational complexity signified by the number of processed pixels per iteration, or PPI (Fig. 4.9 (b)), among the baseline MAP [24], SELP-MAP [30], and proposed AT-SELP-MAP SR methods using different saliency detection techniques applied to super-resolve Frame 9 of the simulated Cameraman LR sequence. Similarly, Fig. 4.10 shows a quantitative comparison among the baseline FR [25], SELP-FR [30], and proposed AT-SELP-FR SR methods. It can be seen that, for the case of the baseline MAP and FR SR methods, all the pixels are processed at each iteration for all the images (i.e., for a 256 × 256 image, the total number of processed pixels is 65536 at each iteration). As shown in Figs. 4.9(b) and 4.10(b), for the SELP-MAP and SELP-FR SR methods, the number of processed pixels per iteration varies from one image to the other depending on the visual content. In Figs. 4.9(b) and 4.10(b), the visual attention processing takes effect around the 6th iteration, thus further reducing the detected active pixels processed. Due to the attentive selectivity, the proposed AT-SELP SR framework presents considerable savings in terms of the number of processed pixels leading to a significant reduction in computational complexity. Comparing the SNR measures in Figs. 4.9(a) and 4.10(a), it can be easily seen that the proposed $M_a(JND)$ mask integrated into the AT-SELP SR framework has the best error performance among the other implemented saliency mask detectors. Also, the relatively efficient and simple saliency mask generator, $M_a(FT)$, that is generated based on the saliency map of [37], does not enhance the overall quality of the SR estimate due to selecting objects and features in the image that are not relevant to SR processing applications. The overall quantitative assessment is also given in terms of PSNR measures in Tables 4.1 and 4.2.
Figure 4.9: Comparison between the baseline MAP, SELP-MAP, and the proposed AT-SELP-MAP SR estimators using sixteen 64 × 64 LR images, a resizing factor $L = 4$, and a noise variance $\sigma_\eta^2 = 16$ for Frame number 9 of the 256 × 256 Cameraman sequence.
Figure 4.10: Comparison between the baseline FR, SELP-FR, and the proposed AT-SELP-FR SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, and a noise variance $\sigma^2_\eta = 16$ for Frame number 9 of the $256 \times 256$ Cameraman sequence.
Table 4.1: PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of \( L = 4 \) and noise variance \( \sigma^2 = 16 \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>25.64</td>
<td>25.18</td>
<td>28.7</td>
<td>20.99</td>
<td>25.13</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>26.05</td>
<td>25.49</td>
<td>29.3</td>
<td>21.7</td>
<td>25.64</td>
</tr>
<tr>
<td>AT-SELP-MAP (JND)</td>
<td>25.82</td>
<td>25.34</td>
<td>29.09</td>
<td>21.59</td>
<td>25.46</td>
</tr>
<tr>
<td>AT-SELP-MAP (IT)</td>
<td>25.63</td>
<td>25.23</td>
<td>28.87</td>
<td>21.33</td>
<td>25.27</td>
</tr>
<tr>
<td>AT-SELP-MAP (GAF)</td>
<td>25.57</td>
<td>25.2</td>
<td>28.82</td>
<td>21.28</td>
<td>25.22</td>
</tr>
<tr>
<td>AT-SELP-MAP (FT)</td>
<td>25.19</td>
<td>24.95</td>
<td>28.59</td>
<td>20.86</td>
<td>24.90</td>
</tr>
</tbody>
</table>

Table 4.2: PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of \( L = 4 \) and noise variance \( \sigma^2 = 16 \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>26.36</td>
<td>25.53</td>
<td>29.17</td>
<td>22.38</td>
<td>25.86</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>26.36</td>
<td>25.56</td>
<td>29.12</td>
<td>22.4</td>
<td>25.86</td>
</tr>
<tr>
<td>AT-SELP-FR (JND)</td>
<td>26.3</td>
<td>25.59</td>
<td>29.07</td>
<td>22.34</td>
<td>25.83</td>
</tr>
<tr>
<td>AT-SELP-FR (IT)</td>
<td>26.2</td>
<td>25.5</td>
<td>29.01</td>
<td>22.19</td>
<td>25.73</td>
</tr>
<tr>
<td>AT-SELP-FR (GAF)</td>
<td>26.19</td>
<td>25.56</td>
<td>29.01</td>
<td>22.18</td>
<td>25.74</td>
</tr>
<tr>
<td>AT-SELP-FR (FT)</td>
<td>26.03</td>
<td>25.43</td>
<td>28.83</td>
<td>21.96</td>
<td>25.56</td>
</tr>
</tbody>
</table>

Note that the PSNR and SNR gain measures used to evaluate the proposed methods do not necessarily reflect the resulting visual quality [117]. Thus, for visual assessment, Figs. 4.11 - 4.12 show the obtained visual SR results for Frame 9 of the simulated Cameraman sequence, for the baseline and selective MAP-based and FR-based SR methods, respectively. From Figs. 4.11-4.12, it can be seen that, despite the fact that the proposed AT-SELP-SR(JND) scheme processes significantly less pixels per iteration, it results in a comparable visual quality to the existing selective and non-selective [24, 25, 30] SR schemes. Fig. 4.13 compares the visual quality of the zoomed in area around the building of the cameraman image. It can be easily noticed that the proposed AT-SELP SR method adopting the efficient JND-based saliency detector is perceptually adapting to the local image information by keeping a balance between sharpening edges and denoising smooth regions.
Figure 4.11: Super-resolved Frame 9 of the $256 \times 256$ HR Cameraman image obtained using MAP-based SR methods from sixteen $64 \times 64$ LR images with $\sigma^2_\eta = 16$ and $\tau = 20\%$. 

(a) Original.

(b) Bicubic interpolation, PSNR = 22.41 dB, Average PPI = 100%.

(c) MAP SR, PSNR = 25.64 dB, Average PPI = 100%.

(d) SELP-MAP SR, PSNR = 26.05 dB, Average PPI = 34.91%.

(e) Proposed AT-SELP-MAP(JND) SR, PSNR = 25.82 dB, Average PPI = 67.56%.
Figure 4.12: Super-resolved Frame 9 of the $256 \times 256$ HR Cameraman image obtained using FR-based SR methods from sixteen $64 \times 64$ LR images with $\sigma_\eta^2 = 16$ and $\tau = 20\%$. 
Figure 4.13: Zoomed in super-resolved Frame 9 of the 256 × 256 HR Cameraman image using MAP-based SR methods.
Tables 4.3 and 4.4 show similar pixel savings achieved by the different VA methods when integrated in the AT-SELP SR framework. However, those VA methods are more computationally complex than the proposed JND-based saliency detector scheme as detailed in the operational complexity analysis in Chapter 3. Tables 4.5 and 4.6 show the total number of operation savings for all the test images. The operations are classified into additions (adds), multiplications (mults), absolute values (abs), comparisons (comps), and square roots (sqrts) operations. The total number of operations, $O_{SR(tot)}$, for the selective SR methods are calculated as follows:

$$O_{SR(tot)} = \sum_{n=1}^{T} \|M_n\|_o \times O_{SR}/\text{pixel}$$  \hspace{1cm} (4.16)

where $\|M_n\|_o$ is the $l_o$-norm of the mask $M_n$, which is the total number of non-zero elements in $M_n$ and $T$ is the total number of iterations. Then the percentage of total operations savings, $S_{SR}$, is calculated as follows,

$$S_{SR(tot)} = \frac{O_{\text{base}(tot)} - O_{SR(tot)}}{O_{\text{base}(tot)}} \times 100\%$$  \hspace{1cm} (4.17)

where $O_{\text{base}(tot)}$ is the total operations used by the baseline SR methods (MAP and FR), and $O_{SR(tot)}$ is the total operations used by the selective SR schemes (SELP, AT-SELP, and AT discussed in Chapter 5). It can be noticed that the complex saliency detection methods adopted from [34, 36] are highly inefficient and cannot be used in real-time applications. 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. It is shown in Table 4.5 that up to 57% in total operation savings can be achieved by the proposed AT-SELP-MAP SR over the baseline MAP SR, while around 30% in operations savings is achieved over the efficient SELP-MAP SR. Table 4.6 shows that the proposed AT-SELP-FR framework saves around 43% of total operation savings over the non-selective baseline-FR and around 13% in total.
Table 4.3: Percentage of Pixel Savings for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>34.91%</td>
<td>9.81%</td>
<td>52.83%</td>
<td>28.51%</td>
<td>31.52%</td>
</tr>
<tr>
<td>AT-SELP-MAP (JND)</td>
<td>67.56%</td>
<td>59.29%</td>
<td>71.20%</td>
<td>68.34%</td>
<td>66.60%</td>
</tr>
<tr>
<td>AT-SELP-MAP (IT)</td>
<td>67.55%</td>
<td>59.28%</td>
<td>71.20%</td>
<td>68.34%</td>
<td>66.59%</td>
</tr>
<tr>
<td>AT-SELP-MAP (GAF)</td>
<td>67.55%</td>
<td>59.28%</td>
<td>71.20%</td>
<td>68.34%</td>
<td>66.59%</td>
</tr>
<tr>
<td>AT-SELP-MAP (FT)</td>
<td>67.61%</td>
<td>59.34%</td>
<td>71.32%</td>
<td>68.61%</td>
<td>66.72%</td>
</tr>
</tbody>
</table>

Table 4.4: Percentage of Pixel Savings for FR-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>43.06%</td>
<td>35.46%</td>
<td>61.66%</td>
<td>33.58%</td>
<td>43.44%</td>
</tr>
<tr>
<td>AT-SELP-FR (JND)</td>
<td>71.32%</td>
<td>70.03%</td>
<td>74.28%</td>
<td>69.40%</td>
<td>71.26%</td>
</tr>
<tr>
<td>AT-SELP-FR (IT)</td>
<td>71.32%</td>
<td>70.00%</td>
<td>74.28%</td>
<td>69.40%</td>
<td>71.25%</td>
</tr>
<tr>
<td>AT-SELP-FR (GAF)</td>
<td>71.32%</td>
<td>70.00%</td>
<td>74.28%</td>
<td>69.40%</td>
<td>71.25%</td>
</tr>
<tr>
<td>AT-SELP-FR (FT)</td>
<td>71.42%</td>
<td>70.37%</td>
<td>74.33%</td>
<td>69.59%</td>
<td>71.43%</td>
</tr>
</tbody>
</table>

operation savings over the efficient SELP-FR SR while resulting in the same visual quality.

The proposed AT-SELP SR algorithms are tested on a real video sequence. The used sequence is the ‘Alpaca’ sequence from the UCSC super-resolution testing database [118]. Similar to [25], the method described in [119] is used to compute the motion vectors. The first 16 frames of the sequence are processed to reconstruct Frame 9 with a magnification factor of 4 in each direction. The simulation parameters of the compared MAP-based SR methods are set to $\lambda = 100$, $\varepsilon = 0.0001$, $s = 100$ and a maximum of 20 iterations is performed. The simulation parameters of the compared FR-based SR methods are set to $R = 2$, $\alpha = 0.6$, $\lambda = 0.08$, and $\beta = 8$, $\varepsilon = 0.0001$, $s = 120$ and a maximum of 30 iterations is performed. The parameter $\tau$ of the saliency map detection rule (4.10) is set to 20% to identify the attentive active pixels. Figs. 4.14 and 4.15 show the obtained visual results. From Fig. 4.14, it can be seen that the
Table 4.5: Percentage of Operations Savings in computations for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>29.93%</td>
<td>4.16%</td>
<td>47.85%</td>
<td>24.27%</td>
<td>26.55%</td>
</tr>
<tr>
<td>AT-SELP-MAP (JND)</td>
<td>57.69%</td>
<td>48.10%</td>
<td>61.34%</td>
<td>59.95%</td>
<td>56.77%</td>
</tr>
<tr>
<td>AT-SELP-MAP (IT)</td>
<td>32.40%</td>
<td>19.43%</td>
<td>36.04%</td>
<td>38.46%</td>
<td>31.58%</td>
</tr>
<tr>
<td>AT-SELP-MAP (GAF)</td>
<td>-2119%</td>
<td>-2419%</td>
<td>-2115%</td>
<td>-1790%</td>
<td>-2110%</td>
</tr>
<tr>
<td>AT-SELP-MAP (FT)</td>
<td>57.16%</td>
<td>47.51%</td>
<td>60.87%</td>
<td>59.73%</td>
<td>56.32%</td>
</tr>
</tbody>
</table>

Table 4.6: Percentage of Operations Savings in computations for FR-based SR methods for all test sequences for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>29.39%</td>
<td>21.77%</td>
<td>47.98%</td>
<td>19.91%</td>
<td>29.76%</td>
</tr>
<tr>
<td>AT-SELP-FR (JND)</td>
<td>42.75%</td>
<td>41.46%</td>
<td>45.72%</td>
<td>40.83%</td>
<td>42.69%</td>
</tr>
<tr>
<td>AT-SELP-FR (IT)</td>
<td>-34.42%</td>
<td>-35.74%</td>
<td>-31.47%</td>
<td>-36.35%</td>
<td>-34.50%</td>
</tr>
<tr>
<td>AT-SELP-FR (GAF)</td>
<td>-6600%</td>
<td>-6602%</td>
<td>-6597%</td>
<td>-6602%</td>
<td>-6600%</td>
</tr>
<tr>
<td>AT-SELP-FR (FT)</td>
<td>41.10%</td>
<td>40.05%</td>
<td>44.01%</td>
<td>39.26%</td>
<td>41.11%</td>
</tr>
</tbody>
</table>

visual quality is comparable for the baseline-MAP [24], SELP-MAP [29, 30], and proposed AT-SELP-MAP SR methods. Similar observations can be made for the baseline-FR [25], SELP-FR [29, 30], and proposed AT-SELP-FR SR methods in Fig. 4.15. Furthermore, the total number of operations is reduced by 66% and 45%, respectively, for the proposed AT-SELP-MAP and AT-SELP-FR methods compared to their non-selective baseline-MAP and FR SR counterparts (Tables 4.7 and 4.8). Also, as shown in Tables 4.7 and 4.8, the total number of operations is reduced by around 15% and 3%, respectively, for the proposed AT-SELP-MAP and AT-SELP-FR methods compared to their selective SELP-MAP and SELP-FR SR counterparts.

To validate the obtained perceived visual quality, subjective tests were conducted using the processed versions of the images of Fig. 4.7 and the processed Frame 9 of the ‘Alpaca’ video sequence shown in Figs. 4.14 and 4.15. The SR images obtained by the baseline MAP-SR [24] and the proposed AT-SELP-MAP(JND) schemes
Figure 4.14: Super-resolved Frame 9 of the 96 × 128 Alpaca real video sequence using the first sixteen frames with magnification factor $L = 4$, $\sigma^2_\eta = 0$ and $\tau = 20\%$. (a) Bicubicly interpolated Frame; (b) Baseline MAP SR; (c) SELP-MAP SR; (d) proposed AT-SELP-MAP SR.

are displayed side by side for comparison. This is also done for the FR-SR [25] and the proposed AT-SELP-FR(JND) schemes. Each case is rated from 1-5 corresponding, respectively, to the reconstructed image produced by the proposed AT-SELP-SR framework is ‘worse’, ‘slightly worse’, ‘same’, ‘slightly better’, and ‘better’ than the non-selective SR methods [24] and [25]. The images are randomly displayed and each case is randomly repeated 4 times with the left and right images swapped to obtain a better subjective response statistics [120]. Eleven subjects took the test with normal
and corrected to normal vision and the MOS is calculated by averaging the responses of all the subjects for each different pair of images. Fig. 4.16 shows a snapshot of the subjective test interface. Experiments are conducted using a 19” DELL LCD monitor having a resolution of $1024 \times 1280$. The MOS results are shown in Table 4.9 for MAP-SR versus the proposed AT-SELP-MAP(JND) method and FR-SR versus the proposed AT-SELP-FR(JND) method for $\sigma_\eta^2 = 16$. The listed MOS values suggest that the two compared methods are achieving comparable visual results with an average MOS of
Table 4.7: Savings in computations when applying the proposed AT-SELP-MAP SR method on Frame 9 of the 96 × 128 Alpaca real video sequence for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 0$.

<table>
<thead>
<tr>
<th></th>
<th>Alpaca</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>50.97%</td>
</tr>
<tr>
<td>AT-SELP-MAP (JND)</td>
<td>65.49%</td>
</tr>
</tbody>
</table>

Table 4.8: Savings in computations when applying the proposed AT-SELP-FR SR method on Frame 9 of the 96 × 128 Alpaca real video sequence for a magnification factor of $L = 4$ and noise variance $\sigma^2_\eta = 0$.

<table>
<thead>
<tr>
<th></th>
<th>Alpaca</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>41.69%</td>
</tr>
<tr>
<td>AT-SELP-FR (JND)</td>
<td>44.81%</td>
</tr>
</tbody>
</table>

Table 4.9: Mean Opinion Scores (MOS): scores 1,2,3,4,&5 correspond to AT-SELP-SR(JND) is “Worse”, “Slightly Worse”, “Same”, “Slightly Better”, & “Better”, respectively, than the baseline SR scheme.

<table>
<thead>
<tr>
<th></th>
<th>Cam.</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Alpaca</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP vs. AT-SELP-MAP</td>
<td>2.955</td>
<td>3.273</td>
<td>2.909</td>
<td>3.273</td>
<td>2.909</td>
<td>3.064</td>
</tr>
<tr>
<td>FR vs. AT-SELP-FR</td>
<td>2.977</td>
<td>3.364</td>
<td>3.045</td>
<td>3.000</td>
<td>3.045</td>
<td>3.086</td>
</tr>
</tbody>
</table>

scores of around 3. This further verify that the proposed AT-SELP SR framework utilizing the proposed low-complexity saliency detector will maintain the perceptual visual quality while significantly reducing the computational complexity.

4.7 Conclusion

An ATtentive SELective Perceptual-based (AT-SELP) super-resolution framework is presented with an improved computational efficiency while achieving the same visual quality as compared to existing popular selective (SELP) and non-selective SR schemes. The proposed scheme is based on a low-complexity saliency detector designed for efficient attentive selective SR estimators in which the SR estimator is applied to just a subset of perceptual and attentively significant pixels. These pixels are
Figure 4.16: Snapshot of performed subjective test comparing the baseline MAP-SR with the proposed AT-SELP-MAP(JND) method, and the baseline FR-SR with the proposed AT-SELP-FR(JND) method.

adaptively selected based on a perceptual JND-based model that exploits the local image luminance masking characteristics and a saliency map based on salient low-level image features. The total number of processed pixels is significantly reduced resulting in a considerable decrease in the required number of operations. The proposed framework also proves to be perceptually adaptive to local image features by keeping a balance between sharpening and denoising of local image content.
Information prioritization for digital processing has been the subject of ongoing research for the past two decades. Limitations on bandwidth, power, and processing capabilities in many of the wireless networks implemented today, necessitates new avenues for innovation when faced with the vast amounts of digital multimedia content. Given current restrictions in mechanisms of processing multimedia information in many devices and network infrastructures such as cellular phones and wireless sensor networks, and the increasing need to process and transmit large amounts of digital information, there is a growing need for smarter and more efficient techniques to transmit and process visual information. Image and video compression technologies have devised bitstream scalability techniques and region-of-interest (ROI) coding capabilities to overcome network limitations while prioritizing important information for transmission and processing. Bitstream scalability in coding prioritize information as needed by the restrictions on the wireless environment. JPEG2000 compression algorithms achieve this goal, at the bitplane coding layer, by shifting the bitplanes of the user-selected ROI pixels above the rest of the pixels in an image [10]. Thus, these ROI regions are coded with higher bit-rates and are transmitted early in the bitstream. Scalable video coding techniques provide quality, temporal, and spatial scalability to meet hard limitations on bandwidth and computational complexity [121]. Spatial and temporal scalability describe cases in which subsets of the bitstream represent the source content with a reduced picture size or frame rate, respectively. With quality scalability, the substream provides the same spatio-temporal resolution as the complete bitstream, but with a lower quality. Spatial scalability can be useful in transmission to different devices with varying screen resolution. Temporal scalability can target applications of
low power and low processing constraints by reducing the frame rate of the transmitted video. Quality scalability is effective in real-time processing applications and progressive multimedia exchange. Wireless sensor networks implement smart protocols to communicate in an efficient manner to preserve battery power. These wireless nodes are awake only when needed and utilize parallel processing mechanisms to manage processing large amounts of data in a collective approach.

Automatic detection of the relevant information for processing is application specific. For example, surveillance applications need to detect and identify objects or track specific moving targets such as humans, cars, or military vehicles. In image enhancement type of applications, the information relevant for processing should have direct impact on the perceived quality. Computational Visual Attention models, as described earlier, imitate the HVS behavior in detecting fixations and regions of attention. The main premise of visual attention is that the HVS cannot handle a vast amount of data at once and devises automatic ways of prioritizing information to perceive visual content. Regions that attract our attention are processed with high visual acuity compared to non-attended regions and thus can directly affect our overall perceptual quality of digital media. Hence, these computational models can serve as an automatic detection mechanism for extracting regions important for SR processing and perceived image quality. Content-aware detection for SR processing can be effective in many applications of real-time HD media transmission to ubiquitous end-user devices. Real-time multimedia processing requires high processing capabilities that are often scarce in ubiquitous environments. Content-aware SR processing can also be effective under limited computational capabilities by allocating the computational resources for processing only part of the information while maintaining the target visual quality.

In this chapter, we propose an Attentive Super-Resolution (AT-SR) framework that enhances on the computational efficiency of previously proposed selective SR
methods while maintaining the perceived quality of the enhanced image. Computational VA models are used to detect relevant information for unequal enhancement of digital media inspired by the HVS attraction and prioritization of perceptual visual processing. Simulation results show enhancement on computational efficiency with minimal perceptible loss in SR quality.

5.1 ATtentive Super-Resolution Framework

A new highly efficient SR framework that is driven by attention information is illustrated in Fig. 5.1. The SR algorithm is initialized by a crude HR estimate, $z_0$, generated by interpolating one of the LR observations or fusion of all the LR observations as described previously. Attention regions, detected by the computational attention models described in Sections 3.8-3.10 or the proposed low-complexity attention detector described in Section 4.3, are processed at each iteration of the SR algorithm. Hence, only pixels falling in attended regions are updated in the SR estimation process. These attended regions, also referred to as attentive pixels, constitute a small subset of pixels selected for enhancement at each iteration to save on computations without degrading the image quality perceived by the user in the attentive regions. Attentive regions mimic the highly-salient regions that the HVS fixates upon and processes with high visual acuity to make up a visual scene.

In our previously proposed framework (Section 4.1), a perceptually selective SR framework is proposed based on utilizing low-level image features and luminance masking properties of the HVS to reduce the computational requirements of iterative SR solutions. In this enhanced framework, we will further enhance on the computational savings of iterative SR methods by only processing the visually attended regions detected by computational VA models. Fig. 5.1 shows a block diagram of the proposed ATtentive SR (AT-SR) scheme. As shown in Fig. 5.1, the active pixels that are signaled by the visually attentive mask, $M_a$, will be iterated upon by the SR algorithm
until a maximum number of iterations is reached or the system stabilizes, i.e., until

\[ \| \mathbf{M}_a \cdot (\hat{z}_{n+1} - \hat{z}_n) \| < \varepsilon, \]

in the attentive active region, where \( \varepsilon \) is a predetermined threshold which represents the desired accuracy of the SR algorithm.

The highly efficient AT-SR framework is also flexible in that several SR estimation algorithms can be integrated within this framework. To illustrate the performance of the proposed framework, the iterative MAP-based SR method [24] and the FR-based SR method [25] are integrated within the proposed AT-SR framework and are denoted by AT-MAP and AT-FR, respectively. The proposed AT-MAP and AT-FR schemes are
implemented following the formulation in Sections 4.4 and 4.5, respectively, where the mask, $M_n$, at each iteration $n$, is updated using an attentive mask, $M_a$, computed at each iteration using a VA detector. Enhancing the attended regions with high accuracy and roughly interpolating the background (non-attended) regions will significantly reduce the computational complexity with minimal or no loss in the perceived quality of the reconstructed SR image under time-constrained viewing conditions.

5.2 Visual Attention Mask Detection

The saliency map shows the saliency at each pixel of a scene by combining low-level features that compete on different modalities. The VA detectors are signified by a Max-finder technique (Fig. 3.11) that detects attentive regions mimicking the sequence of eye movements and fixations [122]. Max-finder techniques can be implemented in various ways and at different complexity levels. A simple and popular approach entails finding the maximum saliency at any given point in time and extracting a circular patch around it spanning a diameter of $1 - 2$ degrees of visual angle corresponding to the foveated vision property of the HVS [36]. In [34], Itti et al. detected the fixation points by implementing a biologically inspired Winner-Take-All (WTA) neural network, then the attention regions are extracted by simply extracting the corresponding circular patches or by more elaborate spreading of the activation in the neural network to a proto-object concept as introduced in [104]. The neural network approach with the proto-object type of region detection is complex and needs to be implemented on dedicated hardware for real-time VA detection applications. Another approach introduced in [37] is steered towards segmentation and target detection applications. In [37, 108], a segmentation algorithm is applied on the saliency information to detect well-defined salient objects in an image. Segmentation algorithms can be complex and may detect flat regions that may not be relevant for resolution enhancement.
In the proposed AT-SR framework, the attention mask, $M_a$, is detected using the low-complexity detection rule formulated in (4.10) to select the most prominent saliency locations that are essential for SR processing. This approach proves to be efficient and effective when targeting applications of highly efficient resolution enhancement. In the following section, simulation results are provided to show the enhanced efficiency of the proposed AT-SR framework and the preserved quality of reconstruction when integrated in the MAP-based and FR-based SR approaches. Furthermore, for comparison, existing VA models, namely the hierarchical VA model (IT) [34], the foveated gaze attentive model (GAF) [36], and the frequency-tuned attention model (FT) [37] are adopted to generate the attentive mask in the proposed AT-SR framework.

5.3 Simulation Results

The performance of the proposed AT-SR framework is assessed using a set of simulated sequence of images where a sequence of LR images is generated from a single HR image, as described in Section 4.6. The same set of images shown in Fig. 4.7 and their corresponding degraded and noisy LR observations are used in the following simulation. The proposed ATtentive MAP-based SR and ATtentive Fusion-Restoration SR schemes referred to as AT-MAP and AT-FR, respectively, are compared with their existing non-selective counterparts MAP-SR [24] and FR-SR [25]. The simulation parameters for the compared MAP-based SR methods are set to $\lambda = 100$, $\varepsilon = 0.0001$, and a maximum of 20 iterations is performed. The simulation parameters for the compared FR-based SR methods are set to $R = 2$, $\alpha = 0.6$, $\lambda = 0.08$, and $\beta = 8$, $\varepsilon = 0.0001$, and a maximum of 30 iterations is performed. The parameter $\tau$ of the VA detection rule (4.10) is set to $20 - 40\%$ to identify the attentive active pixels.

Figs. 5.2 and 5.3 show detected VA regions of Frame 9 of the Monarch image (Fig. 4.7(d)) generated by the proposed VA detector and by existing VA models, presented in [34, 36, 37], and integrated in the proposed AT-SR framework. Two different
values of $\tau = 20\%$ and $40\%$ are shown in Figs. 5.2 and 5.3, respectively. In cases of the hierarchical VA model (IT) [34] and the foveated VA model (GAF) [36], there are two different ways to apply the detection rule that generates the attentive mask, $M_n$, at each iteration $n$. One approach is by selecting the highest $\tau\%$ of the attended regions represented by foveated circular patches denoted by $M_a(IT_f)$ and $M_a(GAF_f)$, respectively. Note that, the subscript ‘$f$’ in IT$_f$ and GAF$_f$ corresponds to the ‘foveated’ version of the IT and GAF models, respectively. A circular patch with a diameter of 60 pixels corresponding to 1 degree of visual angle is selected for the IT model [34] and for the GAF model [36], shown in Figs 5.2-5.3(a) & (b), respectively. Another approach is by selecting the highest $\tau\%$ of the saliency mask represented by $M_a(IT)$ and $M_a(GAF)$, as shown in Figs 5.2-5.3(c) & (d), respectively. In cases of the Frequency Tuned (FT) [37] and the proposed JND-based (JND) VA models, the detection rule is applied by selecting the highest $\tau\%$ of the saliency mask represented by $M_a(JND)$ and $M_a(FT)$, respectively, and shown in Figs 5.2-5.3(e) & (f). The more complex adaptive thresholding approach, following a modified K-means clustering technique described in [37, 108], may detect flat regions and objects that are not essential for SR processing and add more computational complexity. As shown in Figs. 5.2 and 5.3, the attention mask generated by the proposed low-complexity JND-based detector, $M_a(JND)$, is better adapting to information essential to SR processing such as edges and perceived noise in attentive areas as compared to the existing schemes. Furthermore, the proposed JND-based detection is simple to implement and highly efficient in nature. One can easily notice the increased area of detected VA regions when increasing $\tau$ from 20% to 40% in Figs. 5.2 and 5.3, respectively.

Figs. 5.4 and 5.5 give a quantitative comparison among the baseline MAP [24] and the proposed AT-MAP SR methods using different VA detection techniques applied to super-resolve Frame 9 of the simulated Monarch LR sequence in terms of SNR.
Figure 5.2: Detected VA regions, $M_d$, for the $256 \times 256$ bicubicly interpolated Frame 9 of the Monarch image with a $4 \times 4$ average blur, a Gaussian noise with $\sigma_\eta^2 = 16$, and $\tau = 20\%$. 
Figure 5.3: Detected VA regions, $M_a$, for the $256 \times 256$ bicubicly interpolated Frame 9 of the Monarch image with a $4 \times 4$ average blur, a Gaussian noise with $\sigma_n^2 = 16$, and $\tau = 40\%$. 
Figure 5.4: Comparison between the baseline MAP and the proposed AT-MAP SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2 = 16$, and $\tau = 20\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.
Figure 5.5: Comparison between the baseline MAP and the proposed AT-MAP SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2_{\eta} = 16$, and $\tau = 40\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.
Figure 5.6: Comparison between the baseline FR and the proposed AT-FR SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2_{\eta} = 16$, and $\tau = 20\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.
Figure 5.7: Comparison between the baseline FR and the proposed AT-FR SR estimators using sixteen $64 \times 64$ LR images, a resizing factor $L = 4$, a noise variance $\sigma^2_\eta = 16$, and $\tau = 40\%$ for Frame number 9 of the $256 \times 256$ Monarch sequence.
Table 5.1: PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of \( L = 4 \), noise variance \( \sigma_{\eta}^2 = 16 \), and \( \tau = 20\% \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>25.64</td>
<td>25.18</td>
<td>28.70</td>
<td>20.99</td>
<td>25.13</td>
</tr>
<tr>
<td>AT-MAP (JND)</td>
<td>25.50</td>
<td>24.72</td>
<td>28.15</td>
<td>21.16</td>
<td>24.88</td>
</tr>
<tr>
<td>AT-MAP (IT(_f))</td>
<td>23.93</td>
<td>23.13</td>
<td>25.87</td>
<td>19.67</td>
<td>23.15</td>
</tr>
<tr>
<td>AT-MAP (GAF(_f))</td>
<td>23.31</td>
<td>22.90</td>
<td>25.84</td>
<td>19.45</td>
<td>22.88</td>
</tr>
<tr>
<td>AT-MAP (IT)</td>
<td>23.37</td>
<td>23.08</td>
<td>26.15</td>
<td>19.16</td>
<td>22.94</td>
</tr>
<tr>
<td>AT-MAP (GAF)</td>
<td>23.83</td>
<td>23.41</td>
<td>26.36</td>
<td>19.82</td>
<td>23.36</td>
</tr>
<tr>
<td>AT-MAP (FT)</td>
<td>22.20</td>
<td>22.15</td>
<td>25.23</td>
<td>17.97</td>
<td>21.89</td>
</tr>
</tbody>
</table>

Table 5.2: PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of \( L = 4 \), noise variance \( \sigma_{\eta}^2 = 16 \), and \( \tau = 20\% \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>26.36</td>
<td>25.54</td>
<td>29.21</td>
<td>22.37</td>
<td>25.87</td>
</tr>
<tr>
<td>AT-FR (JND)</td>
<td>26.32</td>
<td>25.65</td>
<td>28.96</td>
<td>22.44</td>
<td>25.84</td>
</tr>
<tr>
<td>AT-FR (IT(_f))</td>
<td>25.63</td>
<td>25.27</td>
<td>28.43</td>
<td>21.81</td>
<td>25.29</td>
</tr>
<tr>
<td>AT-FR (GAF(_f))</td>
<td>25.56</td>
<td>25.31</td>
<td>28.08</td>
<td>21.75</td>
<td>25.18</td>
</tr>
<tr>
<td>AT-FR (IT)</td>
<td>25.68</td>
<td>25.21</td>
<td>28.15</td>
<td>21.65</td>
<td>25.17</td>
</tr>
<tr>
<td>AT-FR (GAF)</td>
<td>25.82</td>
<td>25.33</td>
<td>28.37</td>
<td>21.88</td>
<td>25.35</td>
</tr>
<tr>
<td>AT-FR (FT)</td>
<td>24.97</td>
<td>24.53</td>
<td>27.72</td>
<td>21.14</td>
<td>24.59</td>
</tr>
</tbody>
</table>

gain (Figs. 5.4 and 5.5(a)) and computational complexity signified by the number of processed pixels per iteration, or PPI (Figs. 5.4 and 5.5(b)). Similarly, Figs. 5.6 and 5.7 show a quantitative comparison among the baseline FR [25] and the proposed AT-FR SR methods. Figs. 5.4 and 5.6 show quantitative measures for \( \tau = 20\% \) and Figs. 5.5 and 5.7 show quantitative measures for higher values of \( \tau = 40\% \). Note that the subscript ‘\( f \)’ corresponds to the AT-SR methods applying the ‘foveated’ attention masks \( M_a(IT_f) \) and \( M_a(GAF_f) \) of the IT and GAF models, respectively. It can be seen that, for the case of baseline MAP and baseline FR SR methods, all the pixels are processed at each iteration for all the images. As shown in Figs. 5.4(b) and 5.6(b), for the AT-MAP and AT-FR SR methods, the number of processed pixels per iteration are 20% of the total number of pixels corresponding to the most salient locations in the image. Simi-
larly, in Figs. 5.5(b) and 5.7(b), the number of processed pixels per iteration are 40%. Thus, due to the attentive selectivity, the proposed AT-SR framework presents considerable savings in terms of the number of processed pixels leading to a significant reduction in computational complexity. Comparing the SNR gain measures for $\tau = 20\%$ in Figs. 5.4(a) and 5.6(a), it can be concluded that the proposed $M_a(JND)$ mask integrated into the AT-SR framework has the best SNR gain performance among the other implemented VA region detectors. From Figs. 5.5(a) and 5.7(a), one can observe the increasing performance in terms of SNR gain per iteration as $\tau$ increases to 40% of the detected VA regions. In almost all the cases shown in Figs 5.4(a)-5.7 (a), the proposed highly efficient AT-SR(JND) performs similar to or even better than the non-selective SR methods due to pooling the error criteria locally over pixels that are significant for SR enhancement rather than the whole image, as described in Section 5.1. Also, the relatively efficient and simple VA region detector, $M_a(FT)$, that is generated based on the VA region detector of [37], does not significantly enhance the overall quality of the SR estimate due to selecting objects and features in the image that are not relevant to SR processing applications. The overall quantitative assessment in terms of PSNR measures for $\tau = 20\%$ is given in Tables 5.1 and 5.2. Similar assessment for $\tau = 40\%$ is given in Tables 5.3 and 5.4. These PSNR measures also verify the similar performance between the AT-SR schemes and their non-selective counterparts, and the superior performance of the AT-SR based on the proposed $M_a(JND)$ as compared to adopting the existing VA models ($M_a(IT_f)$ [34], $M_a(GAF_f)$ [36], $M_a(IT)$ [34], $M_a(GAF)$ [36], and $M_a(FT)$ [37]). The FR-based methods adopt a median shift and add initial SR estimate that was proved in [25] to be an approximate solution of the $l_1$ norm SR estimation. This initial phase gave a good quality estimate to be iterated upon in the restoration phase that further deblurs the image. Thus, in these FR-based methods, the gain in SNR is not large due to the good initial estimate of the first fusion phase.
Table 5.3: PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma^2_\eta = 16$, and $\tau = 40\%$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>25.64</td>
<td>25.18</td>
<td>28.70</td>
<td>20.99</td>
<td>25.13</td>
</tr>
<tr>
<td>AT-MAP (JND)</td>
<td>25.88</td>
<td>25.38</td>
<td>29.32</td>
<td>21.77</td>
<td>25.59</td>
</tr>
<tr>
<td>AT-MAP (IT$_f$)</td>
<td>25.22</td>
<td>24.14</td>
<td>28.32</td>
<td>21.29</td>
<td>24.74</td>
</tr>
<tr>
<td>AT-MAP (GAF$_f$)</td>
<td>24.61</td>
<td>23.85</td>
<td>27.28</td>
<td>20.31</td>
<td>24.01</td>
</tr>
<tr>
<td>AT-MAP (IT)</td>
<td>24.94</td>
<td>23.94</td>
<td>27.95</td>
<td>21.00</td>
<td>24.46</td>
</tr>
<tr>
<td>AT-MAP (GAF)</td>
<td>24.76</td>
<td>24.23</td>
<td>27.42</td>
<td>21.21</td>
<td>24.41</td>
</tr>
<tr>
<td>AT-MAP (FT)</td>
<td>22.46</td>
<td>22.80</td>
<td>25.42</td>
<td>18.40</td>
<td>22.27</td>
</tr>
</tbody>
</table>

Table 5.4: PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma^2_\eta = 16$, and $\tau = 40\%$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>26.36</td>
<td>25.54</td>
<td>29.21</td>
<td>22.37</td>
<td>25.87</td>
</tr>
<tr>
<td>AT-FR (JND)</td>
<td>26.34</td>
<td>25.63</td>
<td>29.11</td>
<td>22.44</td>
<td>25.88</td>
</tr>
<tr>
<td>AT-FR (IT$_f$)</td>
<td>26.13</td>
<td>25.42</td>
<td>28.88</td>
<td>22.23</td>
<td>25.67</td>
</tr>
<tr>
<td>AT-FR (GAF$_f$)</td>
<td>25.93</td>
<td>25.33</td>
<td>28.67</td>
<td>22.00</td>
<td>25.48</td>
</tr>
<tr>
<td>AT-FR (IT)</td>
<td>26.09</td>
<td>25.33</td>
<td>28.70</td>
<td>22.16</td>
<td>25.57</td>
</tr>
<tr>
<td>AT-FR (GAF)</td>
<td>26.03</td>
<td>25.45</td>
<td>28.72</td>
<td>22.24</td>
<td>25.61</td>
</tr>
<tr>
<td>AT-FR (FT)</td>
<td>25.03</td>
<td>24.63</td>
<td>27.84</td>
<td>21.32</td>
<td>24.71</td>
</tr>
</tbody>
</table>

For visual assessment, Figs. 5.8-5.11 show the obtained visual SR results for Frame 9 of the simulated Monarch sequence, for the baseline and attentive MAP-based and FR-based SR methods, respectively. From Figs. 5.8-5.11, it can be seen that, despite the fact that the proposed AT-SR(JND) scheme processes significantly less pixels per iteration ($\approx 80\%$ less), it results in a comparable visual quality to the non-selective MAP-based [24] and FR-based [25] SR schemes. It can also be noticed that the proposed AT-SR framework adopting the efficient JND-based region detector is performing the best in terms of visual quality among the existing VA region detectors [34, 36, 37]. At very low computational resources, $\tau = 20\%$, the proposed AT-SR(JND) is better adapting to the local image features, such as edges, texture, and noise that are essential for SR enhancement and the overall visual quality. When $\tau$ is
increased to 40%, the visual quality as seen in Figs. 5.12-5.15 is significantly enhanced at the expense of doubling the computational requirements and a good visual performance is achieved by the adopted and proposed VA detectors. The visual quality of the AT-MAP(FT) approach, seen in Figs. 5.8(e) and 5.13(e), is poor due to the detection of regions that do not contribute to the overall image enhancement. However, the visual quality of the AT-FR based methods is considerably good in all cases due to the good initial SR estimate of the FR-based [25] method in the LR fusion phase, that does not leave a large window of enhancement.

The percentage of pixels detected at each iteration of the AT-SR framework is similar for all the adopted VA detectors and the proposed low-complexity JND-based detector. However, the computational complexity of the hierarchical [34] and foveated [36] VA detectors, as presented in previous simulations and according to the computational analysis in Chapter 3, is high and may not be practical for applications of real-time processing. Furthermore, the frequency-tuned VA detector proposed by [37] is efficient in nature but suffers from low reconstruction quality since it detects pixels insignificant for SR processing. Tables 5.5 and 5.6 show the total number of operation savings for all the test images using a detection threshold of $\tau = 20\%$. Also, the same set of measures for a detection threshold of $\tau = 40\%$ are shown in Tables 5.7 and 5.8. It is shown from Tables 5.5 and 5.8 that adopting the foveated VA approach, AT-SR(GAF$_f$) and AT-SR(GAF), for both the MAP-based and FR-based SR schemes, are inhibitive in computational savings due to having to calculate the VA features multiple times (at each fixation) during each iteration. Furthermore, the proposed AT-SR(JND) framework proved to be highly efficient when compared to the non-selective counterparts, while keeping a comparable visual quality. For $\tau = 20\%$ in Table 5.5, it is shown that around an average of 65% in total operation savings can be achieved by the proposed AT-MAP(JND) SR over the baseline MAP SR, and Table 5.6 shows that the
Figure 5.8: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma^2 = 16$ and $\tau = 20\%$. 

(a) Original. 

(b) MAP SR, PSNR = 20.99 dB, Average PPI = 100\%. 

(c) AT-MAP(IT), PSNR = 19.67 dB, Average PPI = 78.55\%. 

(d) AT-MAP(GAF), PSNR = 19.45 dB, Average PPI = 79.17\%. 

(e) AT-MAP(FT), PSNR = 17.97 dB, Average PPI = 80.45\%. 

(f) AT-MAP(JND), PSNR = 21.16 dB, Average PPI = 80.01\%. 

Figure 5.8: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma^2 = 16$ and $\tau = 20\%$. 

117
Figure 5.9: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma_{\eta}^2 = 16$ and $\tau = 20\%$. 

(a) Original.
(b) MAP SR, PSNR = 20.99 dB, Average PPI = 100\%.
(c) AT-MAP(IT), PSNR = 19.16 dB, Average PPI = 80\%.
(d) AT-MAP(GAF), PSNR = 19.82 dB, Average PPI = 80\%.
(e) AT-MAP(FT), PSNR = 17.97 dB, Average PPI = 80.45\%.
(f) AT-MAP(JND), PSNR = 21.16 dB, Average PPI = 80.01\%.
Figure 5.10: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using FR-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma^2_\eta = 16$ and $\tau = 20\%$. 

(a) Original. 

(b) FR SR, PSNR = 22.37 dB, Average PPI = 100\%.

(c) AT-FR(IT), PSNR = 21.81 dB, Average PPI = 79.73\%.

(d) AT-FR(GAF), PSNR = 21.75 dB, Average PPI = 78.65\%.

(e) AT-FR(FT), PSNR = 21.14 dB, Average PPI = 80.23\%.

(f) AT-FR(JND), PSNR = 22.44 dB, Average PPI = 80.01\%. 
Figure 5.11: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using FR-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma^2_\eta = 16$ and $\tau = 20\%$. 
Figure 5.12: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with \( \sigma^2_\eta = 16 \) and \( \tau = 40\% \).
Figure 5.13: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using MAP-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma^2_{\eta} = 16$ and $\tau = 40\%$. 
Figure 5.14: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using FR-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma_{\eta}^2 = 16$ and $\tau = 40\%$. 

(a) Original.  
(b) FR SR, PSNR = 22.37 dB, Average PPI = 100\%.  
(c) AT-FR(IT), PSNR = 22.23 dB, Average PPI = 59.12\%.  
(d) AT-FR(GAF), PSNR = 22.00 dB, Average PPI = 58.27\%.  
(e) AT-FR(FT), PSNR = 21.32 dB, Average PPI = 60.82\%.  
(f) AT-FR(JND), PSNR = 22.44 dB, Average PPI = 60.01\%. 

123
Figure 5.15: Super-resolved Frame number 9 of 256 × 256 HR Monarch image obtained using FR-based SR methods from sixteen 64 × 64 low-resolution images with $\sigma_\eta^2 = 16$ and $\tau = 40\%$. 
Table 5.5: Percentage of Operations Savings in computations for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma^2_\eta = 16$, and $\tau = 20\%$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>AT-MAP (JND)</td>
<td>65.01%</td>
<td>60.31%</td>
<td>64.98%</td>
<td>70.23%</td>
<td>65.13%</td>
</tr>
<tr>
<td>AT-MAP (IT$_f$)</td>
<td>36.13%</td>
<td>26.69%</td>
<td>35.37%</td>
<td>45.16%</td>
<td>35.84%</td>
</tr>
<tr>
<td>AT-MAP (GAF$_f$)</td>
<td>-11553%</td>
<td>-13107%</td>
<td>-11554%</td>
<td>-9805%</td>
<td>-11505%</td>
</tr>
<tr>
<td>AT-MAP (IT)</td>
<td>31.27%</td>
<td>22.10%</td>
<td>31.27%</td>
<td>41.58%</td>
<td>31.56%</td>
</tr>
<tr>
<td>AT-MAP (GAF)</td>
<td>-2837%</td>
<td>-3229%</td>
<td>-2837%</td>
<td>-2397%</td>
<td>-2825%</td>
</tr>
<tr>
<td>AT-MAP (FT)</td>
<td>64.27%</td>
<td>59.77%</td>
<td>64.36%</td>
<td>70.03%</td>
<td>64.61%</td>
</tr>
</tbody>
</table>

Table 5.6: Percentage of Operations Savings in computations for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma^2_\eta = 16$, and $\tau = 20\%$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>AT-FR (JND)</td>
<td>48.47%</td>
<td>48.50%</td>
<td>48.46%</td>
<td>48.46%</td>
<td>48.47%</td>
</tr>
<tr>
<td>AT-FR (IT$_f$)</td>
<td>-28.23%</td>
<td>-27.69%</td>
<td>-28.36%</td>
<td>-27.69%</td>
<td>-27.99%</td>
</tr>
<tr>
<td>AT-FR (GAF$_f$)</td>
<td>-31864%</td>
<td>-31864%</td>
<td>-31864%</td>
<td>-31864%</td>
<td>-31864%</td>
</tr>
<tr>
<td>AT-FR (IT)</td>
<td>-44.13%</td>
<td>-44.13%</td>
<td>-44.14%</td>
<td>-44.14%</td>
<td>-44.14%</td>
</tr>
<tr>
<td>AT-FR (GAF)</td>
<td>-7923%</td>
<td>-7923%</td>
<td>-7923%</td>
<td>-7923%</td>
<td>-7923%</td>
</tr>
<tr>
<td>AT-FR (FT)</td>
<td>46.47%</td>
<td>46.80%</td>
<td>46.42%</td>
<td>46.58%</td>
<td>46.57%</td>
</tr>
</tbody>
</table>

proposed AT-FR(JND) SR scheme saves around an average of 49% of total operation savings over the non-selective baseline-FR. For $\tau = 40\%$ in Table 5.7, it is shown that around an average of 42% in total operation savings can be achieved by the proposed AT-MAP(JND) SR over the baseline MAP SR, and Table 5.8 shows that the proposed AT-FR(JND) SR scheme saves around an average of 29% of total operation savings over the non-selective baseline-FR.

To further validate the obtained perceived visual quality, subjective tests were conducted using the super-resolved versions of the images of Fig. 4.7. The SR images obtained by the baseline MAP-SR [24] and the proposed AT-MAP schemes are each displayed in random sequence for a no-reference based comparison. This is also done
Table 5.7: Percentage of Operations Savings in computations for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma_\eta^2 = 16$, and $\tau = 40\%$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>AT-MAP (JND)</td>
<td>41.61%</td>
<td>33.65%</td>
<td>41.47%</td>
<td>50.25%</td>
<td>41.75%</td>
</tr>
<tr>
<td>AT-MAP (IT$_f$)</td>
<td>12.69%</td>
<td>0.56%</td>
<td>13.17%</td>
<td>26.11%</td>
<td>13.13%</td>
</tr>
<tr>
<td>AT-MAP (GAF$_f$)</td>
<td>-29022%</td>
<td>-32904%</td>
<td>-29021%</td>
<td>-22181%</td>
<td>-28282%</td>
</tr>
<tr>
<td>AT-MAP (IT)</td>
<td>7.74%</td>
<td>-4.56%</td>
<td>7.74%</td>
<td>21.58%</td>
<td>8.13%</td>
</tr>
<tr>
<td>AT-MAP (GAF)</td>
<td>-2861%</td>
<td>-3256%</td>
<td>-2860.74%</td>
<td>-2417%</td>
<td>-2848%</td>
</tr>
<tr>
<td>AT-MAP (FT)</td>
<td>41.86%</td>
<td>33.17%</td>
<td>45.38%</td>
<td>50.28%</td>
<td>42.67%</td>
</tr>
</tbody>
</table>

Table 5.8: Percentage of Operations Savings in computations for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, noise variance $\sigma_\eta^2 = 16$, and $\tau = 40\%$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>AT-FR (JND)</td>
<td>28.49%</td>
<td>28.47%</td>
<td>28.54%</td>
<td>28.47%</td>
<td>28.49%</td>
</tr>
<tr>
<td>AT-FR (IT$_f$)</td>
<td>-48.15%</td>
<td>-47.80%</td>
<td>-47.72%</td>
<td>-48.27%</td>
<td>-47.99%</td>
</tr>
<tr>
<td>AT-FR (GAF$_f$)</td>
<td>-71809%</td>
<td>-79797%</td>
<td>-87784%</td>
<td>-71810%</td>
<td>-77800%</td>
</tr>
<tr>
<td>AT-FR (IT)</td>
<td>-64.14%</td>
<td>-64.14%</td>
<td>-64.14%</td>
<td>-64.14%</td>
<td>-64.14%</td>
</tr>
<tr>
<td>AT-FR (GAF)</td>
<td>-7943%</td>
<td>-7943%</td>
<td>-7943%</td>
<td>-7943%</td>
<td>-7943%</td>
</tr>
<tr>
<td>AT-FR (FT)</td>
<td>26.84%</td>
<td>26.42%</td>
<td>32.03%</td>
<td>27.17%</td>
<td>28.12%</td>
</tr>
</tbody>
</table>

for the FR-SR [25] and the proposed AT-FR schemes. Each image is displayed for a 3 second interval followed by a gray image during which the observers enter their ratings. Limiting the viewing time of each image to 3 seconds simulates the behavior of the HVS in attending to salient regions in the image. Each case is rated from 1-5 corresponding, respectively, to the reconstructed image quality is ‘poor’, ‘fair’, ‘average’, ‘good’, and ‘excellent’. The images are randomly displayed and each case is randomly repeated 2 times to obtain a better subjective response statistics [120]. Ten subjects took the test with normal and corrected to normal vision and the MOS is calculated by averaging the responses of all the subjects for each different case. Fig. 5.16 shows a snapshot of the subjective test interface. Experiments are conducted using a
Table 5.9: Mean Opinion Scores (MOS) for $\tau = 20\%$: scores 1,2,3,4,&5 correspond to SR subjective quality is “Poor","Fair","Average","Good", & “Excellent”, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Monarch</th>
<th>Lena</th>
<th>Clock</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>1.05</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.16</td>
</tr>
<tr>
<td>MAP</td>
<td>3.10</td>
<td>3.45</td>
<td>3.35</td>
<td>2.90</td>
<td>3.20</td>
</tr>
<tr>
<td>AT-MAP(JND)</td>
<td>3.15</td>
<td>2.80</td>
<td>2.55</td>
<td>2.60</td>
<td>2.78</td>
</tr>
<tr>
<td>AT-MAP(IT)</td>
<td>2.40</td>
<td>2.05</td>
<td>1.85</td>
<td>2.15</td>
<td>2.11</td>
</tr>
<tr>
<td>AT-MAP(GAF)</td>
<td>1.65</td>
<td>1.55</td>
<td>1.25</td>
<td>1.50</td>
<td>1.49</td>
</tr>
<tr>
<td>AT-MAP(IT)</td>
<td>1.40</td>
<td>1.50</td>
<td>1.35</td>
<td>1.60</td>
<td>1.46</td>
</tr>
<tr>
<td>AT-MAP(GAF)</td>
<td>2.05</td>
<td>2.10</td>
<td>1.60</td>
<td>2.00</td>
<td>1.94</td>
</tr>
<tr>
<td>AT-MAP(FT)</td>
<td>1.00</td>
<td>1.15</td>
<td>1.00</td>
<td>1.65</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Table 5.10: Mean Opinion Scores (MOS) for $\tau = 20\%$: scores 1,2,3,4,&5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, & “Excellent”, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Monarch</th>
<th>Lena</th>
<th>Clock</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>1.05</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.16</td>
</tr>
<tr>
<td>FR</td>
<td>3.85</td>
<td>4.35</td>
<td>3.75</td>
<td>3.65</td>
<td>3.90</td>
</tr>
<tr>
<td>AT-FR(JND)</td>
<td>3.90</td>
<td>4.20</td>
<td>3.95</td>
<td>3.75</td>
<td>3.95</td>
</tr>
<tr>
<td>AT-FR(IT)</td>
<td>3.85</td>
<td>4.05</td>
<td>3.75</td>
<td>3.45</td>
<td>3.78</td>
</tr>
<tr>
<td>AT-FR(GAF)</td>
<td>3.90</td>
<td>3.95</td>
<td>4.00</td>
<td>3.35</td>
<td>3.80</td>
</tr>
<tr>
<td>AT-FR(IT)</td>
<td>3.50</td>
<td>3.75</td>
<td>3.65</td>
<td>2.90</td>
<td>3.45</td>
</tr>
<tr>
<td>AT-FR(GAF)</td>
<td>3.85</td>
<td>3.95</td>
<td>4.15</td>
<td>3.50</td>
<td>3.86</td>
</tr>
<tr>
<td>AT-FR(FT)</td>
<td>3.35</td>
<td>4.00</td>
<td>3.50</td>
<td>3.30</td>
<td>3.54</td>
</tr>
</tbody>
</table>

19” DELI LCD monitor having a resolution of 1024 $\times$ 1280. The MOS results are shown in Tables 5.9 and 5.10 for MAP-based and FR-based SR methods for $\sigma^2_{\eta} = 16$ and $\tau = 20\%$. The listed MOS values suggest that the proposed AT-SR framework adopting the JND-based detector is achieving better subjective scores among the other adopted VA detectors. Also, a very similar performance in MOS scores is shown when comparing the highly efficient proposed AT-SR(JND) approach with the non-selective baseline-SR schemes. This further verify that the proposed AT-SR framework utilizing the proposed low-complexity saliency detector maintains the perceptual visual quality while significantly reducing the computational complexity. Similarly, the MOS results
Table 5.11: Mean Opinion Scores (MOS) for $\tau = 40\%$: scores 1,2,3,4,&5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, & “Excellent”, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Monarch</th>
<th>Lena</th>
<th>Clock</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>1.05</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.16</td>
</tr>
<tr>
<td>MAP</td>
<td>3.10</td>
<td>3.45</td>
<td>3.35</td>
<td>2.90</td>
<td>3.20</td>
</tr>
<tr>
<td>AT-MAP(JND)</td>
<td>3.15</td>
<td>3.25</td>
<td>3.30</td>
<td>2.85</td>
<td>3.14</td>
</tr>
<tr>
<td>AT-MAP(IT$_f$)</td>
<td>2.25</td>
<td>2.70</td>
<td>2.35</td>
<td>2.60</td>
<td>2.48</td>
</tr>
<tr>
<td>AT-MAP(GAF$_f$)</td>
<td>2.35</td>
<td>1.95</td>
<td>1.60</td>
<td>1.95</td>
<td>1.96</td>
</tr>
<tr>
<td>AT-MAP(IT)</td>
<td>2.50</td>
<td>2.25</td>
<td>1.90</td>
<td>2.55</td>
<td>2.30</td>
</tr>
<tr>
<td>AT-MAP(GAF)</td>
<td>2.60</td>
<td>2.80</td>
<td>2.30</td>
<td>2.30</td>
<td>2.50</td>
</tr>
<tr>
<td>AT-MAP(FT)</td>
<td>1.15</td>
<td>1.20</td>
<td>1.30</td>
<td>1.35</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Table 5.12: Mean Opinion Scores (MOS) for $\tau = 40\%$: scores 1,2,3,4,&5 correspond to SR subjective quality is “Poor”, “Fair”, “Average”, “Good”, & “Excellent”, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Monarch</th>
<th>Lena</th>
<th>Clock</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>1.05</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.16</td>
</tr>
<tr>
<td>FR</td>
<td>3.85</td>
<td>4.35</td>
<td>3.75</td>
<td>3.65</td>
<td>3.90</td>
</tr>
<tr>
<td>AT-FR(JND)</td>
<td>4.30</td>
<td>4.45</td>
<td>4.30</td>
<td>3.65</td>
<td>4.18</td>
</tr>
<tr>
<td>AT-FR(IT$_f$)</td>
<td>4.10</td>
<td>4.55</td>
<td>4.05</td>
<td>3.80</td>
<td>4.13</td>
</tr>
<tr>
<td>AT-FR(GAF$_f$)</td>
<td>4.10</td>
<td>4.00</td>
<td>3.60</td>
<td>3.60</td>
<td>3.83</td>
</tr>
<tr>
<td>AT-FR(IT)</td>
<td>4.10</td>
<td>4.30</td>
<td>3.80</td>
<td>3.30</td>
<td>3.88</td>
</tr>
<tr>
<td>AT-FR(GAF)</td>
<td>3.95</td>
<td>4.35</td>
<td>4.15</td>
<td>3.80</td>
<td>4.06</td>
</tr>
<tr>
<td>AT-FR(FT)</td>
<td>3.35</td>
<td>4.05</td>
<td>3.70</td>
<td>3.40</td>
<td>3.63</td>
</tr>
</tbody>
</table>

for $\tau = 40\%$ are shown in Tables 5.11 and 5.12 that prove the enhanced subjective quality with increasing $\tau$ values due to more pixels processed in the AT-SR framework.

5.4 Conclusion

A highly efficient ATtentive Super-Resolution (AT-SR) framework is presented that improves on the computational efficiency of selective and non-selective SR schemes with minimal or no reduction in the perceived quality of the reconstructed image. The proposed scheme is completely driven by a low-complexity JND-based saliency detector that selects the most prominent salient locations, corresponding to attention regions in a scene, for selective SR processing. The proposed AT-SR framework proved to be flexi-
In adopting several iterative SR methods and different visual attention models. The iterative MAP-based SR [24] and the FR-based SR [25] with several saliency based VA models, namely the proposed JND-based VA detector (JND), a hierarchical VA detector (IT [34]), a foveated VA detector (GAF [36]), and a frequency tuned VA detector (FT [37]), are integrated in the AT-SR framework for comparisons and analysis. On average, quantitative measures of PSNR and SNR gains showed the superior performance of the proposed AT-SR(JND) scheme as compared to the other adopted VA detectors.
in the proposed framework. Subjective tests are conducted for perceived quality analysis and MOS showed a similar performance among the attentive and non-attentive FR-based SR schemes due to the good initial estimation quality of the fusion phase. Moreover, the proposed AT-MAP(JND) presented a slightly lower MOS compared to the non-attentive MAP SR scheme at the gain of 65% reduction in computational complexity and also presented the highest MOS among the efficient attentive SR schemes.
Chapter 6

PERCEPTUALLY WEIGHTED SUPER-RESOLUTION

In this chapter, a different approach to the SR reconstruction solution is considered. Perceptual weighting parameters are used in minimizing the cost function of the SR problem to locally enhance the perceptually relevant image features. Consequently, we propose a Perceptually Weighted (PW) SR technique that enhances on the reconstruction quality of iterative SR techniques with a faster convergence rate.

6.1 Motivation

The HVS is an extremely complicated system and a lot of research have been conducted to mathematically model its different stages and layers. However, when it comes to decoding the signals or mimicking the behavior of the visual cortex of the brain, current vision research is still facing uncharted waters. On the bright side, photoreceptors and neurons of different layers of the visual pathway (Fig. 3.1) are better understood and modeled and perceptual computational models are being exploited in many applications of digital image processing. It is well known that the HVS does not perceive visual information equally, as some information may affect our visual perception more than others. This behavior can be due to various limitations on the resources in the eye and visual pathways and many factors, such as contrast discrimination, masking, and light adaptation.

From Chapter 3, as a result of the center-surround mechanism of the HVS, the information supplied by the retina to the brain weights the visual scene differently by emphasizing features, such as boundaries and edges. This center-surround mechanism is also present in the lateral geniculate neurons (LGN) in the visual pathway leading to the visual cortex of the brain. The visual cortex has an enormous variety of neurons that are specifically sensitive to different types of stimuli or low level features in an image, such as, color, luminance levels, contrast, spatial frequencies, and edge orientations.
Subsequently, different perceptual characteristics of the HVS arise. Light adaptation arises from the fact that visual perception is sensitive to local luminance variations relative to the surroundings rather than the absolute luminance in the scene (Fig. 3.7). Consequently, the necessary contrast needed to enable a response from the neurons and thus to detect a stimulus in a scene is defined as the detection threshold. Contrast sensitivity varies with spatial frequency, temporal frequency, and orientation. Taking these perceptual concepts into consideration, we proposed in Chapters 4 and 5 SR frameworks that exploit the contrast sensitivity properties of the HVS by adopting a contrast sensitivity threshold model that can detect visible signals over a uniform background. Then, these perceptual thresholds, $t_{\text{JND}}$, are used to create a saliency map indicating the levels of perceptual relevance of pixel locations in a scene.

In the previously proposed AT-SELP SR (Chapter 4) and AT-SR (Chapter 5) frameworks, the perceptual selectivity is employed to reduce the computational complexity of iterative SR schemes. Only the attentive-perceptually significant pixels are selected for processing to reduce the computational requirements without degrading the perceived or quantitative reconstruction quality of the image. In this chapter, we follow a different approach inspired by the HVS unequal treatment of visual information by weighting the visual information in a scene differently thus emphasizing features, such as boundaries and edges, that convey the most important information. Thus, a perceptually weighted (PW) SR estimator is proposed that treats regions in the image according to their perceptual significance to SR processing resulting in a superior reconstruction quality in terms of SNR gain at a faster convergence rate. In the following section, the proposed PW-SR approach is detailed and adopted to the baseline MAP [24] and the baseline FR [25] SR schemes.
6.2 Proposed Perceptually Weighted SR Estimator

In the SR observation model (2.1), estimating the HR image \( z \) given a sequence of LR observations \( y_k, k = 1, 2, \ldots K \), is commonly formulated as an optimization problem minimizing an error criteria, \( E(y_k, W_k z) \), and a regularization term, \( \Gamma(z) \), as shown in (2.11), where \( E(.) \) is in function of the LR images, \( y_k \), the observation model \( W_k \), and the HR image estimate, \( z \). In Section 2.5, a general SR cost function is derived for the MAP-based SR methods [24,57] and the robust FR-based SR methods [25] following a regularized \( l_p \)-norm optimization problem as described in (2.12). The error criteria is represented by an \( l_p \)-norm with \( p \) set to 2 for MAP-based SR and \( p \) set to 1 for FR-based SR. The regularization term, \( \Gamma(z) \), takes different forms that statistically model a wide range of images while taking into consideration features that need to be preserved in the estimation process. In the MAP-based SR approach of [24], a piece-wise smooth regularization term is considered that imposes smoothness constraints on the image prior. In the FR-based SR approach of [25], a regularization approach based on the bilateral total variation theory is proposed that preserves edges and fills in missing pixels.

The SR cost function in (2.12) is a balance between two types of functions. The error term is minimized when the estimate, \( z \), projected through the observation model matches the observations, \( y_k \). Minimizing this term alone in the presence of noise can lead to excessive noise magnification due to the ill-posed nature of the inverse problem. The second term, which is a smoothness model of the target HR image, serves as a regularization term that imposes smoothness on the SR solution. This term is minimized when the HR estimate, \( z \), is generally smooth. The weight of each of these competing criteria in the cost function (2.12) is controlled by constant factors \( \kappa = 1/\gamma \) and \( \lambda \). For example, if the fidelity of the observed data is high (i.e., the noise
is small), the error dominates the cost function. If the observed data is very noisy, the cost function will emphasize the regularization term. This will generally lead to smoother image estimates. Then the constant factors are used to globally control the balance between sharpening and denoising the SR solution. However, these constant weighting parameters follow the global statistics of a set of images, thus treating each pixel in the estimation process equally without considering the local perceptual features of different regions in the image. For example, noise over different image regions can be theoretically justified by a constant variance (tuning parameter). However, due to masking and contrast sensitivity, perceptually this noise can be masked in regions of high texture and over edges while it can appear to be prominently annoying over flat areas in the image. In the proposed Perceptually Weighted (PW) SR approach, we assign a different weighting parameter for each pixel in the error term of the cost function that is determined according to the perceptual criteria that measures the significance of that location to the human perception of visual information. Then the PW-SR cost function can be represented as follows:

$$f(z) = \frac{\Psi}{\gamma} \sum_{k=1}^{K} \|y_k - W_k z\|_p^p + \lambda \Gamma(z)$$  \hspace{1cm} (6.1)

where $\Psi$ is the perceptual weighting function, $W_k$ is the degradation matrix for frame $k$, $z$ is the HR frame in lexicographical vector form, and $y_k$ are the observed LR frames also in vector form. The $\| . \|_p^p$ operator is the $l_p$-norm raised to the power $p = \{1, 2\}$. $K$ is the total number of LR observations, $y_k$. The weighting factor $\gamma$ is set to $\{1, \sigma^2\}$ for $p = \{1, 2\}$, respectively, and $\sigma^2$ is the noise variance. $\lambda$ is a regularization weighting factor, and $\Gamma(z)$ is a smoothness regularization term. In the following, the application of the proposed Perceptually Weighted SR approach is applied to the MAP-based [24] and FR-based [25] SR solutions.
In the proposed PW-MAP SR technique, the estimated HR image, \( \hat{z} \), described in [24] and using (6.1) can be represented as follows:

\[
\hat{z}_{n+1} = \hat{z}_n - \beta_n \cdot \left\{ \Psi_n \sum_{k=1}^{K} W_k^T (W_k z_n - y_k) + \frac{1}{2} C_z^{-1} z_n \right\}
\]

(6.2)

where \( \Psi_n \) represents the perceptual weights at every iteration, \( n \), of the SR process, \( \beta_n \) is the step size in the direction of the gradient, \( C_z \) is the covariance of the HR image prior model imposing a piecewise smoothness relationship between neighboring pixels in \( z \) as described in (4.11), and \( \lambda \) is a scaling factor controlling the effect of rapidly changing features in \( z_n \). Similarly, in the proposed PW-FR SR scheme, using the gradient descent solution of (6.1), the estimated HR image described in [25] and can be represented as follows:

\[
\hat{z}_{n+1} = \hat{z}_n - \beta_n \cdot \left[ \Psi_n H^T A^T \text{sign}(A \hat{z}_n - A z_b) + \lambda \sum_{l=-R}^{R} \sum_{m=0}^{R} \alpha^{|m|+|l|} \left[ I - S_y^{-m} S_x^{-l} \right] \text{sign}(z_n - S_y^m S_x^l z_n) \right]
\]

(6.3)

where \( \beta_n \) is the step size in the direction of the gradient and \( \lambda \) is a regularization weighting factor. Matrix \( A \) is a \( N \times N \) diagonal matrix with diagonal values equal to the square root of the number of measurements that contribute to make each element of \( z_b \), and \( H \) is a \( N \times N \) blur matrix. \( \Psi_n \) represents the different perceptual weights per iteration assigned to the errors at each location. \( S_y^l \) and \( S_x^m \) shift the HR image \( z \) by \( l \) and \( m \) pixels in the horizontal and vertical directions, respectively, and \( R \geq 1 \) represents several scales of shifting values. The weight \( \alpha \) is applied as a decaying factor for convergence purposes, and is chosen between \( 0 \leq \alpha \leq 1 \). \( I \) is an identity matrix. In the next section, the generation of the perceptual weights, \( \Psi \), is described.

### 6.3 Perceptual Weights Generation

The perceptual weights represent the perceptual significance of each location in an image. Since texture and edges mask noise and need to be more sharpened than smoothed,
these locations should be weighted with higher values than noisy locations over flat regions that can be smoothed out by the regularization term. Thus the perceptual weighting matrix, $\Psi$, aims at controlling the balance between sharpening and smoothing the local information in an image. The proposed PW-SR approach utilizes the JND-based saliency map detailed in Sections 4.2 and 4.3 to detect perceptual edges and texture. The proposed map, $S_{JND}$, given in Section 4.3, proved to be adapting to the image local features at each iteration of the SR solution. The proposed perceptual weighting matrix is formulated as follows:

$$\Psi = \frac{c + S_{JND}}{\max (S_{JND})}$$

where $S_{JND}$ is the saliency map generated at each iteration of the SR iterative solution and $c$ is a constant tuning parameter that biases the weights depending on the SR cost function. For the PW-MAP SR approach, $c$ was experimentally tuned to be 2. For the PW-FR SR approach, $c$ was set to 1.2. These values for $c$ gave consistently good results as will be shown in the simulations in Section 6.4. The JND-based saliency map is computed at each pixel by weighting the maximum of the locally computed sliding window differences (4.4) with the corresponding locally computed $t_{JND}$, as described in Section 4.3. Note that the detection rule (4.10) is not needed to calculate the perceptual weights.

Pertaining to the computational overhead, the calculation of the proposed perceptual weights follows the same analysis as the JND-based saliency map generation without the detection rule operational overhead. Thus, as discussed in Chapter 4, the contrast sensitivity mask generation requires 5 additions/pixel (four to compute the difference with the neighboring pixels and one addition per pixel contributing to the computation of the mean over $8 \times 8$ block). Then, 3 comparisons/pixel, 4 absolute-values/pixel, and one division/pixel are needed for computing the saliency map, $S_{JND}$, that is generated by weighting the maximum differences with the computed thresh-
Figure 6.1: Perceptual weights, $\Psi$, used by the proposed PW-MAP SR method for all test sequences for a magnification factor of $L = 4$, an average blur of size $4 \times 4$, and noise variance $\sigma_\eta^2 = 16$.

olds, $t_{JND}$. Then, one extra addition, comparison, and division per pixel is required to generate the proposed perceptual weights, $\Psi$. Therefore, following the previous analysis, generating the perceptual weights, $\Psi$, requires a total of $O_{\Psi}(comp) = 4N$ comparisons, $O_{\Psi}(mults) = 2N$ multiplications or divisions, $O_{\Psi}(abs) = 4N$ absolute values, and $O_{\Psi}(adds) = 6N$ additions for the total image of size $N = N_1 \times N_2$ pixels.

6.4 Simulation Results

The performance of the proposed PW-SR framework is assessed using a set of simulated sequence of images where a sequence of LR images is generated from a single HR image.
Figure 6.2: Perceptual weights, $\Psi$, used by the proposed PW-FR SR method for all test sequences for a magnification factor of $L = 4$, a Gaussian blur of size $4 \times 4$, and noise variance $\sigma^2 = 16$.

image, as described in Section 4.6. The same set of images shown in Fig. 4.7 and their corresponding degraded and noisy LR observations are used in the following simulation. The proposed Perceptually Weighted MAP-based SR and Perceptually Weighted Fusion-Restoration SR schemes referred to as PW-MAP and PW-FR, respectively, are compared with their existing baseline counterparts, the MAP-SR [24] and FR-SR [25], as well as our previously proposed efficient SELective Perceptual SR schemes [30], the SELP-MAP SR and SELP-FR SR schemes. The simulation parameters for the compared MAP-based SR methods are set to $\lambda = 100$, $\varepsilon = 0.0001$, and a maximum of 20
iterations are performed. The simulation parameters for the compared FR-based SR methods are set to \( R = 2, \; \alpha = 0.6, \; \lambda = 0.08, \; \text{and} \; \beta = 8, \; \epsilon = 0.001, \) and a maximum of 30 iterations are performed.

Figs. 6.1 and 6.2 show the perceptual weights used by the proposed PW-MAP and PW-FR SR schemes, respectively, for all the test images. The perceptual JND-based weighting function assigns higher weights to edges and texture regions essential to SR processing and lower weights to flat areas that do not require extra sharpening.
and can be smoothed by the regularization term. Also, edge and texture regions are perceptually known to mask noise and thus need to be more sharpened by the error term and less smoothed by the regularization term of the SR functional. Weighting pixels in the error term differently according to a perceptual weighting criteria, mimics the HVS unequal perception of local image information and results in a sharper image at a faster convergence rate. As shown in Figs. 6.1 and 6.2, the perceptual weights generated at each iteration, $\Psi$, are adapting to information essential to SR processing such as edges.
Table 6.1: PSNR values in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>25.64</td>
<td>25.18</td>
<td>28.7</td>
<td>20.99</td>
<td>25.13</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>26.05</td>
<td>25.49</td>
<td>29.3</td>
<td>21.7</td>
<td>25.64</td>
</tr>
<tr>
<td>PW-MAP</td>
<td>26.38</td>
<td>25.75</td>
<td>29.44</td>
<td>22.28</td>
<td>25.96</td>
</tr>
</tbody>
</table>

Table 6.2: PSNR values in dBs for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>26.23</td>
<td>25.54</td>
<td>29.09</td>
<td>22.09</td>
<td>25.74</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>26.21</td>
<td>25.54</td>
<td>29.02</td>
<td>22.07</td>
<td>25.71</td>
</tr>
<tr>
<td>PW-FR</td>
<td>26.42</td>
<td>25.56</td>
<td>29.11</td>
<td>22.5</td>
<td>25.90</td>
</tr>
</tbody>
</table>

In all the test cases. In the proposed PW-MAP SR scheme (Fig. 6.1), the perceived edges and textures are assigned higher weights than the other flat areas or imperceptible contrast variations in the enhanced image at each iteration. However, noise regions are also detected in the background which may be sharpened in the enhanced image. This is due to the fact that perceived noise can result in a high $S_{JND}$ since it is perceived by the HVS. An improved scheme that can differentiate between edges/textures and noise needs to be devised and integrated into the perceptual weighting as part of future work. Similar observations can be made for the proposed PW-FR SR scheme (Fig. 6.2) however, in this case, lower perceptual weights are assigned to noise in the background due to the good quality of the initial SR estimate following the median shift and add operator in the fusion phase [25].

Fig. 6.3 gives a comparison of the SNR gain per iteration among the baseline MAP [24], the efficient SELP-MAP [30], and the proposed PW-MAP SR methods applied to super-resolve Frame 9 of all the simulated test sequences. Similarly, Fig. 6.4 shows a quantitative comparison among the baseline FR [25], the efficient SELP-FR [30], and the proposed PW-FR SR methods. In all the cases, the proposed
Table 6.3: SNR gain measures in dBs for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_n = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>3.62</td>
<td>3.08</td>
<td>4.72</td>
<td>3.23</td>
<td>3.66</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>4.03</td>
<td>3.4</td>
<td>5.29</td>
<td>3.94</td>
<td>4.17</td>
</tr>
<tr>
<td>PW-MAP</td>
<td>4.35</td>
<td>3.66</td>
<td>5.47</td>
<td>4.52</td>
<td>4.50</td>
</tr>
</tbody>
</table>

Table 6.4: SNR gain measures in dBs for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_n = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>1.6</td>
<td>1.07</td>
<td>2.14</td>
<td>1.25</td>
<td>1.52</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>1.58</td>
<td>1.08</td>
<td>2.08</td>
<td>1.23</td>
<td>1.49</td>
</tr>
<tr>
<td>PW-FR</td>
<td>1.87</td>
<td>1.19</td>
<td>2.33</td>
<td>1.69</td>
<td>1.77</td>
</tr>
</tbody>
</table>

PW-SR methods are consistently achieving the best SNR gain performance for all the test images. In Figs. 6.3 and 6.4, the proposed PW-SR methods show a faster convergence rate when compared with the baseline and efficient SELP-SR counterparts. Thus, the proposed PW-SR schemes save on the computational requirements of the baseline SR methods due to their faster convergence rates. The overall quantitative error assessment is also shown in terms of PSNR measures and total SNR gains in Tables 6.1 - 6.2 and Tables 6.3 - 6.4, respectively. The average PSNR measures in Tables 6.1 and 6.2 further verify the consistent enhanced performance of the proposed PW-MAP and PW-FR SR schemes, respectively, over the other methods. The total SNR gains in Tables 6.3 and 6.4 also present a better error performance of the proposed PW-SR methods over the baseline and SELP SR methods. Note that, in the FR-based methods, the gain in SNR is not large due to the good initial estimate of the first fusion phase [25].

The visual SR results for Frame 9 of the simulated Lena and Cameraman sequences, for the MAP-based methods, are shown in Figs. 6.5 and 6.6, respectively. As shown in Figs. 6.5 (d) and 6.6 (d), the proposed PW-MAP SR method results in sharper details as compared to the other two methods. This can be clearly seen around the
face area of the Lena image and the camera area of the Cameraman image. Also, as shown in Fig. 6.5(d), the proposed method achieves a better recovery of high frequency information as it can be seen around the hat region of the Lena image. However, the drawback of the proposed PW-MAP SR technique is the sharpening of the noise in the background due to assigning relatively high weights to the noisy pixels, compared to the SELP-MAP SR technique. Similarly, visual SR results for Frame 9 of the simulated Lena and Cameraman sequences, for the FR-based methods, are shown in Figs. 6.7 and 6.8, respectively. The proposed PW-FR SR scheme reaches a comparable visual quality as the baseline FR SR [25] method with less number of iterations thus leading to lower computational complexity. Furthermore, the visual results are comparable with the efficient SELP-FR SR scheme. Although the proposed PW-MAP scheme require more computations than the SELP-MAP scheme, it can achieve a faster convergence rate.

For computational complexity analysis, Tables 6.5 and 6.6 show the total number of operation savings when applying the MAP-based and FR-based SR methods, respectively, on all the simulated test sequences. The proposed PW-MAP SR and the proposed PW-FR SR methods proved to be more efficient than the baseline MAP [24] and the baseline FR [25], respectively, due to the faster convergence rate when weighting the error terms in the SR functional differently according to a perceptual criteria. However, the proposed PW-SR methods are still less efficient than the selective SR approaches proposed in our previous work [30]. In Table 6.5, it is shown that an average of 25% in total operation savings can be achieved by the proposed PW-MAP SR over the baseline MAP SR. Table 6.6 also shows that the proposed PW-FR SR scheme saves around 11% of total operations on average over the non-selective baseline-FR.

To validate the obtained perceived visual quality, subjective tests were conducted using the super-resolved set of images shown in Fig. 4.7. The SR images ob-
Figure 6.5: Super-resolved Frame 9 of 256 × 256 HR Lena image obtained using the baseline MAP SR, SELP-MAP SR, and the proposed PW-MAP SR from sixteen 64 × 64 low-resolution images with $\sigma^2_{\eta} = 16$. 
Figure 6.6: Super-resolved Frame 9 of $256 \times 256$ HR Cameraman image obtained using the baseline MAP SR, SELP-MAP SR, and the proposed PW-MAP SR from sixteen $64 \times 64$ low-resolution images with $\sigma^2_\eta = 16$. 

(a) Original.  

(b) MAP SR, PSNR = 25.64 dB.  

(c) SELP-MAP, PSNR = 26.05 dB.  

(d) PW-MAP, PSNR = 26.38 dB.
Figure 6.7: Super-resolved Frame 9 of 256 × 256 HR Lena image obtained using the baseline FR SR, SELP-FR SR, and the proposed PW-FR SR from sixteen 64 × 64 low-resolution images with $\sigma^2_\eta = 16$. 

(a) Original.  
(b) FR SR, PSNR = 25.54 dB.  
(c) SELP-FR, PSNR = 25.54 dB.  
(d) PW-FR, PSNR = 25.56 dB.
Figure 6.8: Super-resolved Frame 9 of 256 × 256 HR Cameraman image obtained using the baseline FR SR, SELP-FR SR, and the proposed PW-FR SR from sixteen 64 × 64 low-resolution images with $\sigma^2_\eta = 16$. 
Table 6.5: Percentage of operation savings in computations for MAP-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-MAP</td>
<td>29.93%</td>
<td>4.16%</td>
<td>47.85%</td>
<td>24.27%</td>
<td>26.55%</td>
</tr>
<tr>
<td>PW-MAP</td>
<td>26.81%</td>
<td>24.48%</td>
<td>26.19%</td>
<td>16.35%</td>
<td>23.46%</td>
</tr>
</tbody>
</table>

Table 6.6: Percentage of operation savings in computations for FR-based SR methods for all test sequences for a magnification factor of $L = 4$, and noise variance $\sigma^2_\eta = 16$.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>SELP-FR</td>
<td>45.37%</td>
<td>43.54%</td>
<td>55.66%</td>
<td>43.06%</td>
<td>46.91%</td>
</tr>
<tr>
<td>PW-FR</td>
<td>15.19%</td>
<td>11.34%</td>
<td>3.38%</td>
<td>12.80%</td>
<td>10.68%</td>
</tr>
</tbody>
</table>

tained by the baseline MAP-SR [24] and the proposed PW-MAP scheme are displayed side by side for comparison. This is also done for the FR-SR [25] and the proposed PW-FR scheme. Similarly, the selective SELP-SR schemes [30] are compared side by side with the proposed PW-SR counterparts. Each case is rated from 1-5 corresponding, respectively, to the reconstructed image produced by the proposed PW-SR framework is ‘worse’, ‘slightly worse’, ‘same’, ‘slightly better’, and ‘better’ than the non-selective SR methods [24] and [25] or the selective SELP-SR [30] methods. Each case is randomly repeated 4 times with the left and right images swapped to obtain a better subjective response statistics [120]. Ten subjects took the test with normal and corrected to normal vision and the MOS is calculated by averaging the responses of all the subjects for each different pair of images. The images are randomly displayed and Fig. 4.16 shows a snapshot of the subjective test interface. Experiments are conducted using a 19” DELL LCD monitor having a resolution of $1024 \times 1280$. The MOS results are shown in Tables 6.7 and 6.8 for the baseline SR methods versus the proposed PW-SR methods, and for the SELP-based SR methods versus the proposed PW-SR methods, respectively, for $\sigma^2_\eta = 16$. The listed MOS values suggest that, in the
Table 6.7: Mean Opinion Scores (MOS): scores 1,2,3,4,&5 correspond to PW-SR is “Worse”, “Slightly Worse”, “Same”, “Slightly Better”, & “Better”, respectively, than the baseline SR scheme.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP vs. PW-MAP</td>
<td>3.80</td>
<td>3.60</td>
<td>3.75</td>
<td>3.83</td>
<td>3.74</td>
</tr>
<tr>
<td>FR vs. PW-FR</td>
<td>2.78</td>
<td>2.35</td>
<td>3.08</td>
<td>2.65</td>
<td>2.71</td>
</tr>
</tbody>
</table>

Table 6.8: Mean Opinion Scores (MOS): scores 1,2,3,4,&5 correspond to PW-SR is “Worse”, “Slightly Worse”, “Same”, “Slightly Better”, & “Better”, respectively, than the SELP-SR scheme.

<table>
<thead>
<tr>
<th></th>
<th>Cameraman</th>
<th>Lena</th>
<th>Clock</th>
<th>Monarch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELP vs. PW-MAP</td>
<td>3.55</td>
<td>2.93</td>
<td>3.53</td>
<td>2.78</td>
<td>3.19</td>
</tr>
<tr>
<td>SELP vs. PW-FR</td>
<td>2.75</td>
<td>2.35</td>
<td>3.23</td>
<td>3.05</td>
<td>2.84</td>
</tr>
</tbody>
</table>

case of the MAP-based SR, the proposed PW-MAP is performing better than the MAP and SELP-MAP SR schemes with less computational requirements than the MAP SR scheme. However, the proposed PW-FR scheme is performing slightly worse than the FR and SELP-FR schemes. Furthermore, the proposed PW-FR method results in a higher computational efficiency compared to the baseline FR-SR method.

6.5 Conclusion

A Perceptually Weighted Super-Resolution (PW-SR) technique that improves on the convergence rate, and in some cases, the reconstruction quality of iterative SR schemes, is presented. Inspired by the concept of unequal processing of the HVS to stimuli in an image, perceptual weighting is integrated into the cost function of the SR minimization problem. The proposed PW-MAP SR and the PW-FR SR techniques showed enhancement in the reconstructed image quality at a faster convergence rate than the baseline MAP [24] and baseline FR [25] SR schemes.
Chapter 7

CONCLUSION AND FUTURE RESEARCH

This thesis investigates efficient selective multi-frame Super-Resolution (SR) techniques targeting two main categories of SR problem formulations: MAP-based and FR-based SR methods. The MAP-based SR methods that are based on Bayesian formulations of the SR problem strongly depend on prior HR information. The MAP SR approaches were shown to have a good reconstruction quality but with high computational requirements. The FR-based SR methods that are efficient implementations of the regularized-norm minimization methods by reducing on matrix operations, proved to be more robust to errors with higher computational efficiency as compared to the MAP-based SR methods. The objective of this work is to reduce the computational complexity of these popular SR solutions, while maintaining the perceptual visual quality of the reconstructed image. This chapter summarizes the major contributions of this dissertation and proposes future research directions.

7.1 Contributions

The main contributions of this work are:

- An ATtentive-SELective Perceptual (AT-SELP) SR framework: this thesis proposes a new class of selective SR solutions that can reduce the computational complexity of iterative SR problems while maintaining the desired estimated HR media quality. Due to the human Visual Attention, not all the detail pixels are needed to preserve the overall visual quality of an HR image. Towards this goal, we propose an improved efficient selective SR framework jointly driven by the Human Visual System (HVS) contrast sensitivity and saliency-based Visual Attention (VA) models. Moreover, the proposed AT-SELP SR framework is shown to be easily integrated into a MAP-based SR algorithm as well as a FR-based SR
estimator. Simulation results show significant reduction on average in computational complexity with comparable visual quality.

- A highly efficient ATtentive (AT) SR framework: this thesis proposes an enhanced selective framework that is completely driven by visual attention (VA) information. The proposed AT-SR framework proved to be very flexible by adopting several VA models. Existing VA detectors proved to be computationally complex or, in many cases, failed to detect regions essential for efficient SR processing. Simulation results of the proposed AT-SR framework showed that significant enhancement in computational efficiency and minimal reduction in visual quality can be obtained by the proposed attentive SR scheme. Subjective tests with limited viewing time are conducted to further verify the comparable visual quality of the proposed framework as compared to the baseline non-selective schemes and the enhanced quality over the existing integrated VA models.

- A low-complexity JND-based saliency detector: the general high-frequency detection methods, such as gradient-based or entropy-based, do not incorporate any perceptual weighting and cannot automatically adapt to an image’s local high frequency content that is perceptually relevant to the Human Visual System (HVS). Thus, the problem of devising automatic detection thresholds that can adapt to local image content perceptually is not an easy problem. As a result, we propose a low-complexity saliency detector that is based on notions of Just Noticeable Differences and targeted towards efficient selective SR. By adopting the proposed low-complexity saliency detector, the proposed AT-SEL and AT SR frameworks are shown to be able to adapt locally to image content by enhancing the edges and denoising the flat areas differently according to human perception and visual attention.
• A Perceptually Weighted SR approach: the perceptual weights determined by means of the proposed JND-based saliency map are incorporated in the error function of the SR minimization problem. Different perceptual weights are assigned to pixel locations mimicking the unequal treatment of the human visual perception to local image stimuli. Using these perceptual weights resulted in a higher quantitative quality of SR reconstruction in terms of SNR gain and PSNR at a faster convergence rate. Subjective tests also showed that the proposed PW-SR approach achieves a higher visual quality among the MAP SR methods, and a slightly lower perceived visual quality among the FR SR methods.

7.2 Future Research

The contributions of this thesis provide a new efficient SR framework and new insights in relation to the use of Visual Attention (VA) models for image resolution enhancement and reconstruction problems. The extensions and possibilities for future research include:

• Efficient Perceptual Color Super-Resolution: in this work, only luminance channels are considered when calculating the JND thresholds based on blurred horizontal and vertical edges. It is interesting to investigate the effect of color channels on the JND thresholds and the VA mask detection algorithms. Based on the color JND thresholds and color VA saliency models, an efficient color SR approach can be investigated, where significant pixels are detected on the chrominance channels as well.

• Perceptual 3D Super-Resolution: 3D multimedia processing has been an active area of research in the recent years. High-resolution 3D multimedia applications require low complexity signal processing algorithms due to the high amount of data involved. Feature points need to be detected for registering the multiple
views in the stereo sequence to compute the disparity maps. Thus, the perceptual
detection algorithms can be employed to select pixels that are significant for the
improved generation of disparity maps. Furthermore, the proposed efficient se-
lective SR approaches can be extended to deliver HD 3D content under limited
computational requirements.

- **Perceptual edge detection:** the contrast sensitivity threshold detection model can
detect perceived contrast levels that can correspond mainly to significant edges.
The same concept can be applied to select a perceptual threshold that can be used
in edge detection schemes. Perceptual edge detection can be incorporated into
segmentation algorithms essential for target detection and manufacturing defects
detection.

- **Perceptual model-based SR techniques:** model-based SR techniques have pro-
liferated recently due to the low cost of data storage and data processing. Fea-
tures of the visual content can be modeled through a parametric model or learnt
through training from a large dataset of images. Perceptual modeling can play
an important role in devising better approaches that take the HVS into consid-
eration. Thus, perceptual model-based SR approaches can emerge to provide
enhanced results that are perceptually relevant to the user.

- **Wavelet-based adaptive kernel-based SR schemes:** The non-iterative Fusion-
Interpolation (FI) SR approaches are inherently less computationally intensive
but suffer from limited reconstruction quality depending on the accuracy of the
assumed statistical model or training set. As shown in Section 1.3, the non-
iterative FI-SR approach [26] requires around 64% and 82% of total multipli-
cation and addition operations less than the iterative MAP-based SR [24] and
FR-based SR [25], respectively, but suffers from a limited reconstruction quality
in areas of high variations in pixel intensities, such as strong edges as shown in Fig. 1.2. A wavelet-based Adaptive Wiener Filter (WAWF) SR approach can be proposed, wherein the statistical model parameters for designing the adaptive filter are estimated in the wavelet domain from correlated information in each subband independently. These SR schemes can be improved upon by integrating perceptual or edge directed techniques. Currently, a circularly symmetric autocorrelation parametric model is utilized in the wavelet domain to solve for the adaptive kernel coefficients. This parametric model adapts to the estimated SR image local statistics and the motion vectors between the observed LR frames. Thus, future work should define an autocorrelation model that can also adapt to local perceptually significant features, such as edge orientations and contrast levels. One approach towards this goal can be achieved by weighting the autocorrelation model with a perceptual detection model and steering the autocorrelation towards edge oriented processing.

- Adaptive perceptually weighted SR scheme: The proposed Perceptually Weighted (PW) SR scheme can be improved to better adapt to edges and noise thus maintaining a balance between denoising flat regions and sharpening edges and textures in the SR solution. An enhanced perceptual weighting function can be proposed that can differentiate between edges and noise by incorporating contrast masking in addition to contrast sensitivity and luminance masking in the improved detection threshold model.
REFERENCES


