Abstract—Super-resolution involves the use of signal processing techniques to estimate the high-resolution (HR) version of a scene from multiple low-resolution (LR) observations. It follows that the quality of the reconstructed HR image would depend on the quality of the LR observations. The latter depends on multiple factors like the image acquisition process, encoding or compression and transmission. However, not all images are equally affected by a given type of impairment. A proper choice of the LR observations for reconstruction, should yield a better estimate of the HR image, over a naive method using all images. We propose a simple, model-free approach to improve the performance of super-resolution systems based on the use of perceptual quality metrics. Our approach does not require, or even assume, a specific realization of the SR system. Instead, we select the image subset with high perceptual quality from the available set of LR images. Finally, we present the logical extension of our approach to select the perceptually significant regions in a given LR image, for use in SR reconstruction.

Keywords—super-resolution; perceptual quality metrics

I. INTRODUCTION

Super-resolution (SR) reconstruction [1] has been a very active research area ([2], [3], [4]) for several years. Even though SR techniques significantly outperform classical approaches in terms of perceived visual quality, it is clear that the image quality of the input LR sequence plays an important role in ensuring this improvement. Most SR algorithms assume that the low-resolution input is otherwise free of any artifacts - like compression or blur. This assumption may not always be valid for real-life sequences. A comprehensive evaluation of the performance of SR reconstruction techniques needs to consider test sequences that are captured in unconstrained scenarios, and with devices commonly at the disposal of the average user. For example, the mobile phone has become a ubiquitous device to generate and distribute multimedia content. However, videos recorded using mobile phones are often of significantly poor quality than those recorded using digital cameras or camcorders. As an example, we downloaded several videos from the mobile video streaming service Qik [5]. Most of this content is fairly low resolution. Some frames show significant blur, while others show strong block-edge impairments (see Figures 1 and 2 for some sample frames). Such content poses an interesting challenge to even the best SR systems. In this work, we present an approach to choose the high quality subset of LR frames for performing super-resolution. Perceptual quality metrics are used to define the optimality criteria. We are motivated by the need to enhance the quality of SR reconstruction algorithms without constraining the algorithm used.

This paper is organized as follows: in Section II we provide a very broad overview of perceptual quality metrics. Section III describes the different aspects of our method in detail. Section IV provides a discussion on our experiments and results. Finally, Section V concludes the paper with a look at possible future direction for this work.

II. PERCEPTUAL QUALITY METRICS

Perceptual Quality Metrics attempt to capture the quality of an image as perceived by a human observer [6]. They are classified into three types depending on the availability of the ground-truth reference signal

1) Full reference metrics require the original signal to compute a quality score for the test image.
2) Reduced reference metrics require some characteristics of the original signal, or a pilot to compute a quality score for the test image.
3) No reference metrics compute a quality score when no ground-truth or any knowledge of it is available.

In the context of SR reconstruction, we do not have access to the original (undegraded) LR observations. Indeed, the challenge is to process noisy LR observations in a robust manner to reconstruct a perceptually high quality HR version of the scene. Judicious use of no-reference metrics thus holds the best promise to overcome this challenge.

Because of the absence of ground-truth, no-reference quality metrics rely on measuring artifacts (like blur or block-edge impairment) in the image or video that affect perceived quality. Metrics that measure blur have been proposed by Ferzli et al [7] and Marziliano et al [8]. Similarly Wang et al [9] propose a metric to measure perceived quality of JPEG compressed images. Though many metrics have been proposed, their application in quality enhancement and monitoring continue to be areas of active research.
III. NO-REFERENCE PERCEPTUAL QUALITY METRICS FOR SUPER-RESOLUTION

Super resolution is the process of combining several low-resolution (LR) noisy images to produce a higher resolution image. The most common image acquisition model used is given by

\[ Y_k = DH_{\text{cam}}^k F_k X + V_k \quad k = 1, \ldots, N \]  

Here \( D \) represents the decimation operation, \( H_{\text{cam}}^k \) models the camera lens point spread function (PSF), \( F_k \) models the geometric motion between the HR frame \( X \) and the \( k \)th LR frame \( Y_k \), and \( V_k \) is the system noise.

While this model has proven to be fairly effective for videos in controlled environments, it is too simplistic to account for many real-life scenarios. For example, Equation (1) does not model the motion-blur resulting from a hand-held camera. It also does not account for compression of the recorded (low-resolution) sequence, which is typical in almost all mobile phones. Theoretically, it is possible to model these impairments - for example, by assuming a time varying blur function \( H_k \), which is the composite of the (time-invariant) lens PSF \( H_{\text{cam}}^k \) and the (time-varying) motion blur \( H_{\text{motion}}^k \). However, this additional complexity makes the problem analytically intractable. Therefore, measuring the impairments in the LR sequence and effectively managing them to arrive at the best possible estimate of the HR image is an important research area.

A. Related Work

In previous work [10], [11], researchers have proposed selecting pixels that are perceptually significant and updating only these pixels in the iterative reconstruction algorithm. They propose a model based on contrast sensitivity threshold based on the perceived visibility of edges. We adopt a different approach: we attempt to discard the frames that impair the SR reconstruction rather than focus on pixels or regions based on the Visual Attention model. By pruning out the frames that have degraded quality relative to the other frames in the sequence, we improve performance over ‘blind’ SR systems.

Our method has three advantages:

1) The frame selection is based on the human-perceived notion of quality, rather than the statistical notion of outliers.
2) The quality metrics we use are quite efficient to compute and comprise little overhead as part of the overall system.
3) Finally, our method is a model-free approach and does not assume any particular implementation of the super-resolution algorithm used to generate the SR outcome.

B. Proposed Method

In this work we use no reference perceptual quality metrics to select a high quality set of frames for SR. (For brevity, henceforth we use just quality metrics to imply no-reference perceptual quality metrics.) Instead of explicitly modeling all possible impairments in the LR frame, we define quality metrics corresponding to such impairments, and use only the frames having a quality score \( Q \) above a certain threshold. We thus modify Equation (1) as follows:

\[ Y_k = DH_{\text{cam}}^k F_k X + V_k \quad \{Q_k > \theta\} \quad k = 1, \ldots, N \]  

where \( Q_k \) is the quality score for the \( k \)th low resolution frame \( Y_k \), and \( \theta \) is a quality threshold. We further extend our method to select regions in the \( k \)th frame that will yield high quality SR reconstruction.

We consider blur and block-edge artifacts as specific examples of impairments in the LR frames, and use appropriate quality metrics to measure them. While it may be argued that blur and block-edge artifacts are not the only impairments that may affect the input video sequence, these are, in practice, the most widely occurring ones. For measuring image blur, we used the image sharpness metric proposed by Ferzli et al [7]. They propose a blur metric based on the notion of Just Noticeable Blur. The sharpness metric is then defined as the reciprocal of the blur metric. For measuring compression artifacts, we used the metric of Wang et al [9]. For both the metrics, we use parameter values suggested by the authors in their original work, with the (possible) exception of the window-size over which these metrics are computed. We compute the metrics over a \( W_1 \times W_2 \) window, with non-overlapping windows covering...
the entire LR image. To compute the overall quality score for a frame corresponding to particular impairment, we simply average the metric scores over all the windows in the frame. This gives us the quality scores \( Q_{t,k} \in \{ Q_{\text{blur}}, Q_{\text{block}} \} \). Note that high values of \( Q_{\text{blur}} \) or \( Q_{\text{block}} \) for a frame imply higher perceived quality for the frame in question.

**C. Combining different quality scores**

It is possible for a video sequence to have more than one source of impairment - that is blur and block-edge artifacts. Individual scores corresponding to each can be combined using a weighted average scheme to yield an overall quality score for an image. However, this requires that the quality scores corresponding to different metrics are normalized in some way (e.g., they are bounded by the same limits). In practice, however, this will not be the case. To normalize the individual scores, we compute each quality score \( Q_t \) (\( Q_{\text{blur}} \) and \( Q_{\text{block}} \) in our experiments) over the \( N \) LR frames. We then compute the bounds \( (Q_t^{\min}, Q_t^{\max}) \) over this set and use them to normalize \( Q_t \).

The overall quality metric for the \( k \)th frame \( Q_k \) is then computed as

\[
Q_k = \sum_t w_{t,k} \cdot Q_t^{n}
\]  

where \( w_{t,k} \) is the weight for metric \( Q_t^n \) and the superscript on \( Q_t \) indicates that the normalized metric is used. The summation in the above equation is over the different types of metrics used. The weights \( w_{t,k} \) are computed based on the relative magnitudes of the normalized scores \( Q_t^n \). We presently use simple binary weights to combine the scores. That is, we assume that any given frame can have only one primary impairment. Thus, if the normalized metric score for a given type (that is, \( Q_{t,k} \)) is more than 0.5, we set its weight \( w_{t,k} \) to 1 and the other weights to 0.

**D. Rejecting poor quality frames**

The quality score for each frame \( Q_k \) is compared against a threshold \( \theta \) that is computed to yield the frames with highest relative quality among all the frames in the available set. Assuming that less than half the frames available for SR reconstruction are degraded, we can use the median as an estimate of the quality score of an undegraded frame in the sequence. The threshold \( \theta \) can then be computed as:

\[
\theta = Q^* - 1.1 \sigma_Q
\]  

where \( Q^* \) is the median of quality scores of all \( N \) frames and \( \sigma_Q^2 \) is the sample variance. We found this method of selecting frames to work fairly well across the different test sequences that we used.

Of course, the estimate of \( \theta \) will not be reliable if the fraction of significantly degraded frames is greater than 0.5. To estimate the fraction of significantly degraded frames we use the following approach: we maintain a history of the quality scores \( Q^h = \{ Q^{-K}, \ldots, Q^{-1} \} \) over the previous \( K \) sets of frames processed by the system. (A set is defined as the batch of \( N \) frames used for reconstruction.) We then define an indicator variable \( s_k \) such that

\[
s_k = \begin{cases} 
1 & \text{if } Q_k < \alpha Q_{\text{min}}^h \\
0 & \text{otherwise}
\end{cases}
\]

where \( Q_{\text{min}}^h \) is the minimum over the set \( Q^h \) and \( \alpha \in [0, 1] \) is a parameter. We found values of \( \alpha \) between 0.3 and 0.5 to work well in practice. Then the fraction of frames that is significantly degraded is computed as

\[
\frac{\sum s_k}{N}
\]

We detect such cases and use bilinear interpolation to generate the HR outcome for these frames.

**E. Rejecting poor quality regions within a frame**

We extend the idea of rejecting frames to rejecting poor quality regions in the frame. For this, we compute a mask to filter out blocks based on the quality score for the block. The threshold for the mask computation is computed in an adaptive fashion as follows: the blockwise quality scores for the LR image are inserted in a histogram. The threshold is then computed as a fixed percentile of the histogram. This allows us to specify the threshold simply using a single, intuitive parameter. The mask is used to blank out certain regions of the frame from being used in the SR reconstruction. The SR outcome for the masked areas is computed by simple bilinear interpolation.

**IV. EXPERIMENTS AND RESULTS**

We first examine our hypothesis that so-called robust SR reconstruction techniques cannot handle perceptual impairments. Such techniques aim for robustness to outliers in the data. However, outlier theory does not correlate well with the way the human visual system treats degradations in the image content. To test out our hypothesis, we use a set of \( N = 10 \) LR images and attempt to generate the SR image. For this we used the algorithm of Farsiu et al [3]. We chose this algorithm because of two reasons: first, it has shown excellent performance at a modest computational
cost on standard test sequences. Second, the method claims robustness to outliers and thus is an ideal candidate to test our hypothesis.

We use three test vectors as input to the algorithm:

1) TV1 - The original set of 10 LR images; Figure 3 (a) shows one of the LR images.
2) TV2 - We randomly choose one image in the test set and blur this image using a Gaussian blur model; see Figure 3 (b). The remaining images are left untouched.
3) TV3 - We leave out the blurred image using only the remaining 9 images.

The results obtained for these three test vectors are shown in Figure 4. The top image in the figure shows the result obtained with the original set of LR images (that is, using the test vector TV1). The bottom two images show the result obtained using the test vectors TV2 and TV3 respectively. It is clear that leaving out the blurred LR image has improved the visual quality of the SR reconstruction. Indeed, this is expected as the blurred image does not conform to the assumed model of uniform blur. However, the super resolution system does not have a mechanism to detect such frames to process them appropriately (or leave them out of the reconstruction). Such a mechanism would also prove to be beneficial to correct for any modeling constraints in the super-resolution algorithm.

A. Results on real-life sequences

We now present our results with real-life test sequences. We used several video sequences from the mobile streaming site Qik [5]. We first present results for the case of blur impairment: we used the image sharpness metric proposed by Ferzli et al [7]. Figure 5 shows the plot of the quality score $Q_{blur}$ for a set of frames from one of the test videos. Next we show the results of SR reconstruction (using the method of [3]) for this video. We used $N = 10$ LR frames to generate the SR outcome for a given frame. Four of the frames in the test sequence were significantly blurred (due to camera shake) as compared to the others. Figure 6 (a) shows the SR outcome using all the frames, and Figure 6 (b) shows the outcome using only the high quality frames. The difference in the quality obtained is clearly evident.

Next we present the results of masking out regions from the frames that are not optimal for SR reconstruction. Figure 7 shows the input frames and the corresponding masked frames for two of the frames in the test sequence. These results were generated using a block size of $16 \times 16$ to compute the local blurriness score. The blurriness scores are
inserted in a histogram and the threshold set at 35 percentile of the cumulative histogram to reject the blur values. Using a higher value allows regions with greater perceived blur to be included. Pixels from only the unmasked (perceptually high quality) regions are used in the SR reconstruction. The pixel values corresponding to the masked regions are filled in by bilinear interpolation of the pixels in the frame being super-resolved.

We now consider frames having compression artifacts. We show results of our method on a test sequence in which the frames were compressed using a different Q-factor for each frame. This resulted in some frames having heavy compression artifacts as compared to other frames. Figure 8 (a) shows the plot of the quality score $Q_{\text{block}}$ for a set of frames from our test sequence. Next we show the results of SR reconstruction using all the available frames, and compare it with the results obtained using only the high quality subset of frames. Figure 9 shows the SR outcomes for these two cases. Figure 9(a) is generated using all available frames, while Figure 9(b) is generated using only the high quality frames. Again, the advantage of filtering out degraded frames before doing SR reconstruction is quite evident.

1) Multiple impairments in video sequence: We now present results for the case of multiple impairments in a video stream. We assume that the impairment processes are mutually independent. We selected a frame sequence where we introduced blur and compression artifacts in randomly chosen frames. Specifically, 2 frames (Frames 1 and 9) were corrupted with motion blur and 2 frames (Frames 5 and 8) were compressed with a very low Q-factor. The quality scores $Q_{\text{blur}}$ and $Q_{\text{block}}$ were computed for the sequence. As noted earlier, these scores cannot be directly combined. We normalized these scores and compute the composite quality score for each frame as described in Section III-C. Figure 10 shows the (normalized) individual metric scores (bottom) and the composite metric score (top). We next present the SR outcome for this frame sequence using all the available frames (Figure 11 (a)) and using only the high quality frames (Figure 11 (b)). The difference in perceptual quality is immediately clear.

V. CONCLUSION

We presented a model-free approach based on using perceptual quality metrics to select high quality frames from the available set for doing SR reconstruction. We also extend this idea to selecting perceptually significant regions in a given frame. We have conducted informal tests on a small set of subjects to evaluate the subjective quality of our method. Our tests show that the subjective quality of the SR outcome using only optimal frames is indeed higher. This is encouraging, especially given the simplicity of our approach. We are in the process of conducting more comprehensive MOS tests using a larger set of both video sequences and subjects. Future research efforts may be directed toward exploring better quality metrics in context of super-resolution reconstruction.
Figure 10. Normalized metric scores for video sequence with multiple impairments. The composite quality score $Q_k$ is shown in (a) with the individual scores $Q_{blur}$ and $Q_{block}$ shown in (b) and (c) respectively.

Figure 11. SR reconstruction with all available frames (a) and only high quality frames (b).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers whose comments and suggestions helped improve the presentation of this paper.

REFERENCES


