Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

Temporal image stacking for noise reduction and dynamic range improvement

Kalin Atanassov, James Nash, Sergio Goma, Vikas Ramachandra, Hasib Siddiqui
Qualcomm Inc.

ABSTRACT

The dynamic range of an imager is determined by the ratio of the pixel well capacity to the noise floor. As the scene dynamic range becomes larger than the imager dynamic range, the choices are to saturate some parts of the scene or "bury" others in noise. In this paper we propose an algorithm that produces high dynamic range images by "stacking" sequentially captured frames which reduces the noise and creates additional bits. The frame stacking is done by frame alignment subject to a projective transform and temporal anisotropic diffusion. The noise sources contributing to the noise floor are the sensor heat noise, the quantization noise, and the sensor fixed pattern noise. We demonstrate that by stacking images the quantization and heat noise are reduced and the decrease is limited only by the fixed pattern noise. As the noise is reduced, the resulting cleaner image enables the use of adaptive tone mapping algorithms which render HDR images in an 8-bit container without significant noise increase.

Keywords: noise reduction, video stabilization, high dynamic range imaging, anisotropic diffusion, image registration

1. INTRODUCTION

The dynamic range of an imager is determined by the ratio of the pixel well capacity to the noise floor. As the scene dynamic range becomes larger that the imager dynamic range, the choices are to saturate some parts of the scene or "bury" others in noise. Although it is possible to increase the sensor dynamic range by using sensors with a larger pixel well, it is not practical for cost and size limited systems. Recently, there have been several sensor designs proposed that utilize pixels level modifications that decrease pixel sensitivity at higher light level or split each pixel into low and high sensitivity parts [1]. Although these hardware solutions address the problem of increasing the pixel well capacity, they do not improve the performance at lower light levels. Software only solutions involving two or more frames with usually different exposures have been explored with satisfactory results; however, moving objects are problematic since they rely heavily on light level dependent weighting between the frames which results in ghosting artifacts [2].

In this paper we recognize the high dynamic range enhancement is essentially a noise reduction problem. By not allowing the frame to saturate one can ensure that the higher light levels are captured accurately. Noise reduction on the other hand uncovers the details buried in the noise. We propose an algorithm that produces high dynamic range images by "stacking" sequentially captured frames and thus reducing the overall noise and creating additional bits. Another advantage of stacking frames taken with relatively small exposures is avoidance of motion blur. In order for the frame to be added without any blur, possible misalignment between frames arising due to the hand motion should be estimated and compensated for. The frame alignment is done by detecting and matching points/regions in the image frame with a sharp autocorrelation function (typically corners [1-2]) and fitting the motion vectors into a low degrees of freedom transform. Projective and affine transforms are good approximations which however do not accommodate objects' disparity produced by possible translation between frames. Keypoints that lie on moving objects should be filtered so they do not affect the global image correction that is caused by the hand jitter.

Combining multiple identical frames makes it easier to handle moving objects do the capability to use different amount of samples (from different frames). Since perfect moving object alignment might not be possible, weighting heavily on the reference frame for that object provides clean (not blurred) albeit a noisy version of the object.

As alluded to above, not all pixels in the frame will benefit from full noise reduction, i.e., equal weighting from all captured frames. Since alignment for some pixels/regions might not be possible (edges on moving objects or parallax due to camera translation), a mechanism for guided averaging is needed. The temporal anisotropic diffusion [3] turns off the averaging on a per pixel basis if object motion has occurred or there is not sufficient correspondence for whatever reason.

Typical imager capabilities allow capturing high resolution still images at 30 or higher frames per second. In this scenario, it is possible to capture 30 images within a second. It is reasonable to expect a similar noise reduction and

SPIE USE: ____ DB Check, ____ Prod Check, Notes:

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

commensurately increase in the dynamic range. To achieve this one should consider the noise sources that contribute to the noise floor. The noise components are sensor heat noise, the quantization noise, and the sensor fixed pattern noise. It can be shown that by stacking images one can significantly reduce the quantization and heat noise and be limited by the fixed pattern noise only which may or may not be spatially correlated. At low signal levels, by design, the fixed pattern noise shows much lower that the heat noise and the quantization noise and is not a problem for single shot images. Once the temporally uncorrelated noise is removed via frame stacking, the fixed pattern noise becomes an issue. Further noise improvement comes not by adding more frames but by characterizing and removing the fixed pattern noise. Fixed pattern noise has an additive and multiplicative component that can be characterized independently as a function of the exposure time and signal level. The achieved noise reduction can be best demonstrated by applying an adaptive contrast enhancement. The allowed contrast increase is usually set as a function of the noise reduction. As a result, for typical scenarios, 10-20 times dynamic range increase (or noise reduction) can be realized.

2. ALGORITHM DESCRIPTION

The algorithm relies on capturing and stacking multiple still images. To reduce computation requirements, it is necessary to capture the frames without any time gap between them. Furthermore, since frame alignment uses an approximate projective transformation, it is desirable that all frames are captured within one second. These limitations dictate that around 30 still images are taken at about 30 frames per second. A higher number of frames and/or longer exposure time are possible. The algorithm can be summarizes as follows:

A. Collect around 30 frames at 30 frames per second exposure time with no gap between them
B. Select a reference frame – that can be any fixed frame from the sequence or the frame with the highest frequency content (least amount of blur)
C. Detect key point (areas with sufficiently rich frequency content or steep auto-correlation function)
D. For each one of the remaining frames
    a. Match the key point from the reference frame and estimate motion vectors
    b. Fit motion vectors into an Affine transform
    c. Remove key points that belong to moving objects
    d. Fit remaining motion vectors into an Affine transform
    e. Align image to reference image by performing bilinear Affine transform
    f. For every pixel, compute each similarity with the corresponding pixel from the reference frame
    g. Weight the pixel based on the similarity and add to the reference

2.1 Keypoint detection

There is a well-established body of literature regarding the detection of salient points in an image [4]. The Harris corner detector [1] uses an approximation of the auto-correlation function to detect edges and corners. Several modifications are applied to the Harris corner detection rational in order to gain resolution independence and key point contrast resilience [2]. The following steps are applied:

- Resize reference image to a resolution of 640x480 to achieve consistency (same keypoint detected from image of different resolution of the same scene)
- Normalize the result of Harris corner detector as follows
    - Original Harris corner detector

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \det(M) = \lambda_1 \lambda_2, trace(M) = \lambda_1 + \lambda_2$$

with test for corners:

$$R = \det(M) - k\, trace(M)^2$$
$$R > 0$$

    - Modified Harris corner detector test

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

$$\frac{R}{max\left(1, \left(\frac{trace(M_L)}{C}\right)^2\right)} > threshold$$

The difference between $\boldsymbol{M}$ and $\boldsymbol{M_L}$ is that the latter has a larger support area. The values for C and *threshold* are chosen arbitrarily based on the trade-off between the number and the quality of the keypoints. The modified Harris corner detector is sensitive to the shape but not to the contrast as illustrated Figure 1.

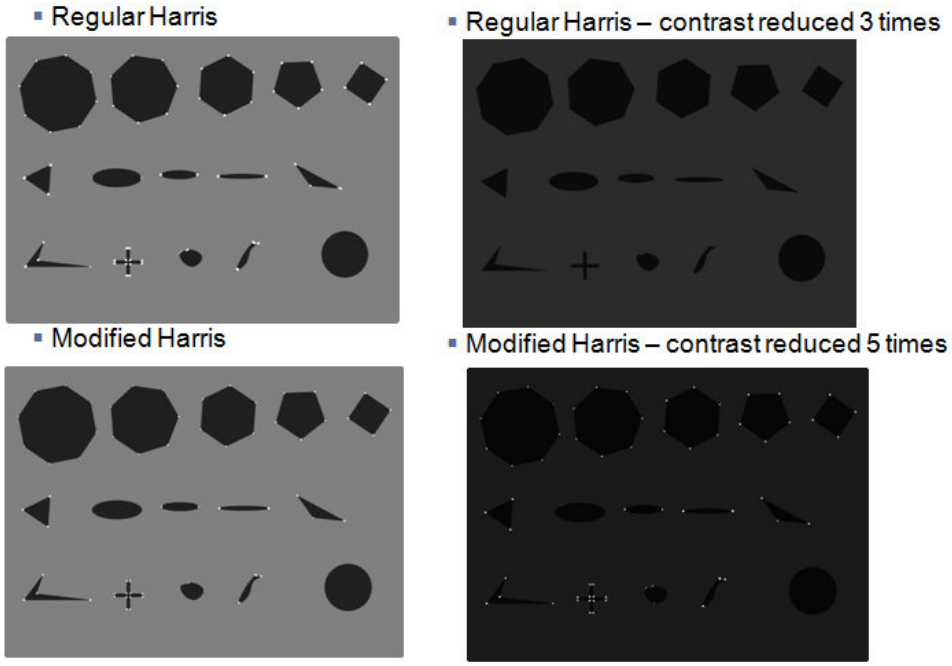


Figure 1. Modified Harris corner detector comparison

2.2 Keypoint matching

Once the keypoints in the reference image are detected, each keypoint is matched to its counterparts in the rest of the captured images. The matching is done based on the minimal mean square error and matches with minimum square error higher that an acceptable threshold are discarded. The matching is done using the luminance plane of the image and the search window is selected based on the time difference between the frames. To minimize the processing and improve the reliability, keypoints are matched sequentially as follow:

1. Select a keypoint from reference frame (n^{th} frame)
2. Using a search window of $K \times K$ (based on time difference between successive frames) around the position in the reference frame, find the corresponding match of the keypoint (in the reference frame) in the $(n+1)^{th}$ frame
3. Using a search window of $K \times K$ around the position in $(n+1)^{th}$ frame, find the corresponding match of the keypoint (in the reference frame) in $(n+2)^{th}$ frame

In this manner the keypoints in the reference frame are always used as the matching template.

2.3 Alignment transform selection

The following considerations apply when an alignment transform is selected:

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

- Movement of the camera (assuming it is handheld) while collecting the still image can be three dimensional rotation or translation
- If camera translation is very small (all images are taken within a second) compared to the closest object to the camera, then camera translation can be considered negligible
- Majority of phone cameras use CMOS technology in which pixel rows are exposed and read sequentially in time (rolling shutter)
- The rolling shutter causes skew of the imager columns

It is acknowledged that skew component can be added into a projective transform so the transformation between two frames may be characterized as a combination of roll, pitch, and yaw (three dimensional rotations) plus skew. We found experimentally that an affine transform containing skew and roll components works well while pitch and yaw are approximated with a simple shift. The transform is computed in via:

$$\begin{bmatrix} x1 & y1 & 1 \\ x2 & y2 & 1 \\ . & . & 1 \\ . & . & 1 \\ . & . & 1 \\ xn & yn & 1 \end{bmatrix}_{ref_{frame}} = \begin{bmatrix} x1 & y1 & 1 \\ x2 & y2 & 1 \\ . & . & 1 \\ . & . & 1 \\ . & . & 1 \\ xn & yn & 1 \end{bmatrix}_{reg_frame} x \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathrm{P_{ref_frame}} = \mathrm{P_{reg_frame}} \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix}$$

Where $\mathrm{P_{ref_frame}}$ is a matrix containing the keypoint positions in the reference frame and $\mathrm{P_{reg_frame}}$ is a matrix containing the same keypoints positions in the frame that is to be registered. The affine matrix that approximates the transition from registered frame to the reference frame is computed as

$$\mathrm{A} = pinv(\mathrm{P_{reg_frame}})\mathrm{P_{ref_frame}}$$

2.4 Moving object detection

Global alignment as described above depends on the movement of the camera while the still frames are taken. As an approximation it would not be affected by the actual depth of the keypoint. It will however be affected by moving objects which will cause an erroneous transform estimation and subsequent frame alignment. For this reason, keypoints that lie on moving objects should be detected and removed from the above affine matrix computation. Moving objects usually are detected with respect to the background. A typical assumption is that the moving object will occupy a much smaller area that the background. We propose a simple method for detecting key points that belong to the moving objects as follows:

- Use all key points to compute the affine transform matrix ***A***
- Using the computed matrix ***A*** project the keypoint positions of the registered image

$$\tilde{P}_{reg_frame} = \mathrm{P_{reg_frame}A}$$

- Compute differences between original keypoint from the reference frame and the corrected ones from the registered frame and compare with a threshold

$$\left|\tilde{P}_{reg_frame} - P_{ref_frame}\right| > threshold$$

- If the keypoints are not consistent with the affine fit , they most likely belong to moving objects and are removed from the computation of ***A*** (***A*** is recomputed without them)

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

2.5 Pixel combining

Once the relationship between the coordinates of the reference image and the registration is computed, i.e., the affine transform with matrix ***A***, the registered image is re-projected with an affine bilinear transformation as follows:

- For every pixel with coordinate ***i*** and ***j*** in the corrected registration image, produce its coordinates from the original registration image

$$[i \quad j \quad 1]A^{-1} = [\tilde{i} \quad \tilde{j} \quad 1]$$

- To produce a pixel with coordinate ***i*** and ***j*** in the corrected registered image, one needs to synthesize a pixel with coordinates $\tilde{i}$ and $\tilde{j}$ (not necessarily integer numbers)
- Using a bilinear interpolation as shown below the pixel and hence the entire registration image can be corrected to match the reference image

$$\mathrm{i}_m = \lfloor \tilde{i} \rfloor \; and \; \Delta_i = \tilde{i} - \mathrm{i}_m$$

$$\mathrm{j}_m = \lfloor \tilde{j} \rfloor \; and \; \Delta_j = \tilde{j} - \mathrm{j}_m$$

$$k = \begin{bmatrix} 1-\Delta_i \\ \Delta_i \end{bmatrix} [1-\Delta_j \quad \Delta_j]$$

$$S(\tilde{i},\tilde{j}) = [S(i_m,j_m) \quad S(i_m+1,j_m) \quad S(i_m,j_m+1) \quad S(i_m+1,j_{m+1})] \begin{bmatrix} k_{11} \\ k_{21} \\ k_{12} \\ k_{22} \end{bmatrix}$$

After the registration frame is aligned to the reference frame, it needs to be selectively averaged with it. The averaging is done in the following manner:

- Compute pixel similarities by weighting a window around the squared differences between the reference image and the corrected registration image

$$D = w \otimes \left(S_{ref} - \tilde{S}_{reg}\right)^2$$

- The pixel similarity is used as the argument of a diffusion function ***C(.)*** [3] with the following properties
  - ***C(0)=1***
  - ***C(∞) =0***
  - ***C(.)*** is monotonically decreasing
- The final averaging is done as follows (for all registration images, i.e., all the images except the reference)

$$\mathrm{S}_{\mathrm{sum}} = 0$$

$$S_{sum} = S_{sum} + S_{ref}\left(1 - C(D)\right) + \tilde{S}^i_{reg} C(D)$$

where ***i*** is the index of the registration image. The sum image then needs to be divided by the number of all images in the set, say 30 , which reduces the noise and adds an additional 5 bits.

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

3. RESULTS

Noise reduction not only is a desirable image operation in it of itself but also increase the image dynamic range by reducing the noise floor and adding additional bits. To characterize the performance of the algorithm one needs to consider first and foremost its noise reduction and detail preservation capabilities. As of now, there is no reliable methodology that can be used to evaluate the combination of noise reduction and detail preservation algorithms due to the adaptive nature of the latter. To demonstrate the performance of the algorithm, we provide a comparison between short exposure, long exposure, and multiple stacked short exposures images. We demonstrate that the stacked image visually yields equivalent sharpness and details preservation as the short exposure image while exhibiting equal noise level as the long exposure image.



Figure 2. Original scene

In order to produce image quality that is consistent with images that are passed through a typical image processing pipeline, we have applied basic operations such as white balance, gamma correction, and edge enhancement. Shown bellow are some cropped image areas that illustrate the noise reduction capabilities of the algorithm.

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

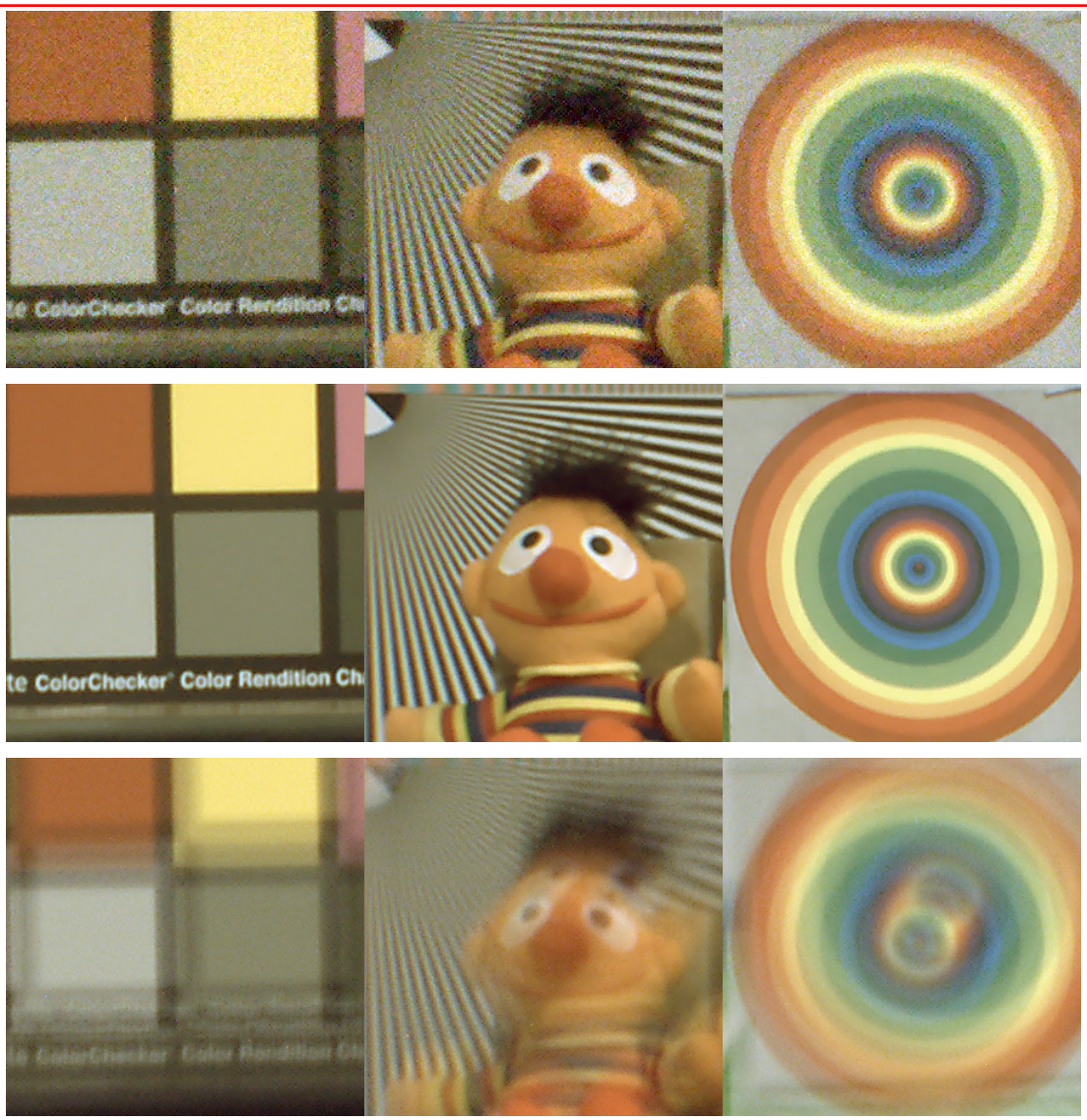


Figure 3. Short exposure (top), short exposure stacked (middle), long exposure(bottom)

To illustrate that frame stacking offers a unique approach to reducing the noise, we compare it with a commercial state-of-the-art spatial noise reduction algorithm known as BM3D[7].

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

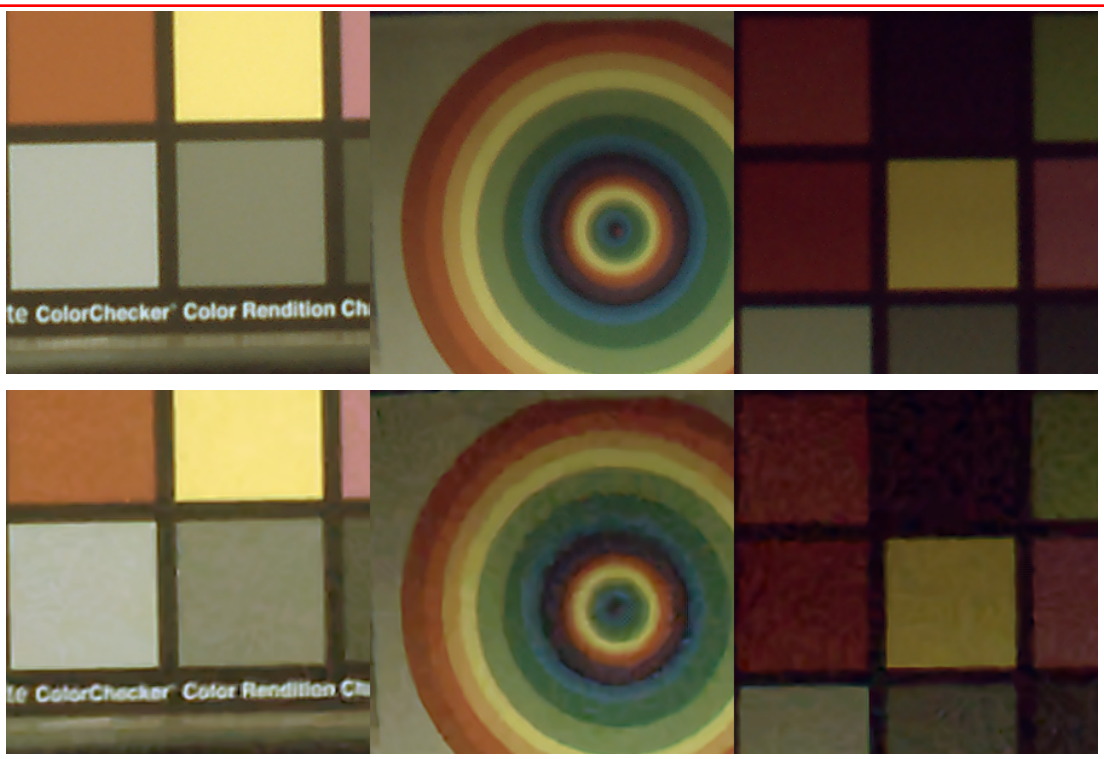


Figure 4. Frame stacking (top), spatial noise reduction (bottom)

To illustrate how noise reduction can translate to image dynamic range improvement, an HDR scene has been arranged as shown in Figure 5.



Figure 5. HDR scene

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

The original (short exposure) image and the stacked image have been processed with a local tone mapping function to illustrate what details have been "buried" in the noise. Further noise improvement is possible by characterizing and removing the fixed pattern noise.



Figure 6. Tone mapped adapted - short exposure (top) and stacked short exposures(bottom)

As demonstrated above, noise reduction at the lowest light levels leads to improved dynamic range.

4. CONCLUSIONS

Frame stacking provides noise reduction while preserving fine detail. In frame stacking an image is viewed not as an isolated snapshot but as a single realization of a noisy ensemble. With access only to a single frame, spatial averaging techniques such as BM3D are unable to reliably distinguish detail from noise and therefore trade image fidelity for noise performance. Temporal averaging over multiple aligned frames does not suffer from this limitation because regions of detail are superimposed while independent noise is averaged out. Frame stacking is especially advantageous in low light conditions where poor signal-to-noise ratio renders transform domain approaches ineffective. In HDR scenes frame stacking can be used to back down the exposure away from saturation in order to properly capture bright areas while relying on stacking to recover image content buried in low light. The combination of frame stacking with spatial averaging incorporates all available information and presumably would result in the best performance.

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/app/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

REFERENCES

[1] Darmont, A., Tessenderlo, M., "Methods to extend the dynamic range of snapshot active pixels sensors," SPIE Electronic Imaging 2008.
[2] Astola, A., Egiazarian, K., "Multiple exposures for high dynamic range imaging," SPIE Electronic Imaging 2011.
[3] Harris, C., & Stephens, M., "A combined corner and edge detector," Alvey vision conference, 1988.
[4] Lowe, D. "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110
[5] Atanassov, K., Ramachandra, V., Nash, J., and Goma, S., "Unassisted 3D camera calibration", SPIE Electronic Imaging 2012.
[6] Perona, P., & Malik, J, "Scale-space and edge detection using anisotropic diffusion," IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990.
[7] Dabov, K., Katkovnik, V., "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering," IEEE Trans. Im. Proc. 2007

SPIE USE: ____ DB Check, ____ Prod Check, Notes: