Multispectral Bilateral Video Fusion

Eric P. Bennett, John L. Mason, and Leonard McMillan

***Abstract*—We present a technique for enhancing underexposed visible-spectrum video by fusing it with simultaneously captured video from sensors in nonvisible spectra, such as Short Wave IR or Near IR. Although IR sensors can accurately capture video in low-light and night-vision applications, they lack the color and relative luminances of visible-spectrum sensors. RGB sensors do capture color and correct relative luminances, but are underexposed, noisy, and lack fine features due to short video exposure times. Our enhanced fusion output is a reconstruction of the RGB input assisted by the IR data, not an incorporation of elements imaged only in IR. With a temporal noise reduction, we first remove shot noise and increase the color accuracy of the RGB footage. The IR video is then normalized to ensure cross-spectral compatibility with the visible-spectrum video using ratio images. To aid fusion, we decompose the video sources with edge-preserving filters. We introduce a multispectral version of the bilateral filter called the "dual bilateral" that robustly decomposes the RGB video. It utilizes the less-noisy IR for edge detection but also preserves strong visible-spectrum edges not in the IR. We fuse the RGB low frequencies, the IR texture details, and the dual bilateral edges into a noise-reduced video with sharp details, correct chrominances, and natural relative luminances.**

***Index Terms*—Bilateral filter, fusion, image decomposition, IR, multispectral, noise reduction, nonlinear filtering.**

I. Introduction

A SIGNIFICANT problem in night vision imaging is that, while IR imagery provides a bright and relatively low-noise view of a dark environment, it can be difficult to interpret due to inconsistencies with visible-spectrum imagery. Therefore, attempts have been made to correct for the differences between IR and the visible-spectrum. The first difference is that the relative responses in IR do not match the visible spectrum. This problem is due to differing material reflectivities, heat emissions, and sensor sensitivities in the IR and visible spectra. These differing relative responses between surfaces hinder the human visual system's ability to perceive and identify objects. The other difference is the IR spectrum's lack of natural color. Unfortunately, colorization (chromatic interpretation) of IR footage and correction of relative luminance responses are difficult because there exists no one-to-one mapping between IR intensities and corresponding visible-spectrum luminances and chrominances.

Alternately, visible-spectrum video is easy to interpret due to its natural relative luminances and chrominances, but visible-spectrum sensors typically fail in low-light and night-vision situations due to poor sensor sensitivity. To achieve sufficient responses, long exposure times must be used, making them impractical for video applications.

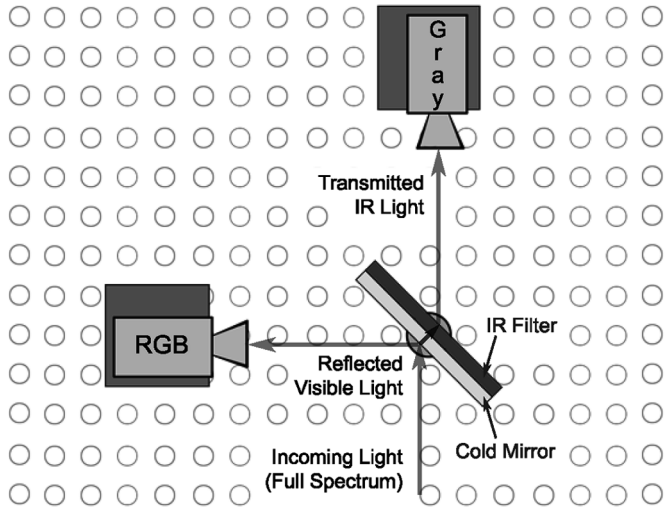


Fig. 1. Diagram of our prototype multispectral imaging system mounted on an optical bench. The incoming optical path is split with a cold mirror which provides an efficient separation of spectra.

Because RGB video has the perceptual characteristics we desire, we present a fusion technique that enhances visible-light video using information from a registered and synchronized IR video sensor (Fig. 1). Our goal is to create video that appears as if it was imaged only in the visible spectrum and under more ideal exposure conditions than actually existed. This differs from most multispectral fusion approaches that combine elements from all sensors, creating a mixed spectral representation [1]. It also differs from learning-based methods that rely on sparse priors of the visible-light spectrum to enhance IR [2] because we have an IR/RGB pair for every frame.

Our fusion decomposes the visible-spectrum and IR-spectrum videos into low frequencies, edges, and textures (detail features). Specifically, we consider the visible spectrum as 400–700 nm and IR as either Short Wave Infrared (SWIR, 900–1700 nm) or Near Infrared (NIR, 700–2500 nm). Our decompositions are enhanced and fused in a manner that corrects for their inherent spectral differences.

In this work, we present a series of contributions that enable our fusion method as follows:

- an extention to the bilateral filter (the "dual bilateral") that preserves edges detected in multiple spectra under differing noise levels;
- a per-pixel modulation (normalization) of the IR to transfer visual-spectrum relative luminance responses;
- a video decomposition model that specifically considers and processes edge components.

Manuscript received July 14, 2006; revised December 8, 2006. This work was supported by the DARPA-funded, AFRL-managed Agreement FA8650-04-2-6543. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Luca Lucchese.

The authors are with the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: bennett@cs.unc.edu; dantana@email.unc.edu; mcmillan@cs.unc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.894236

II. Related Work

Our fusion approach attempts to improve reconstruction of the visible-spectrum footage with the assistance of registered IR imagery, meaning that we do not include elements that appear only in IR. Traditional multispectral fusions attempt to combine elements from multiple spectra to communicate information from all sources. Two classic multispectral applications are remote sensing (aerial and satellite imagery) and night vision, which both fuse visible and nonvisble spectra.

To fuse amplified night-vision data with multiple IR bands, Fay *et al.* [3] introduce a neural network to create false-color (pseudo-color) images from a learned opponent-color importance model. Many other false-color fusion models are commonly used in the remote sensing community, such as intensity-hue saturation. A summary is provided in [1]. Another common fusion approach is combining pixel intensities from multiresolution Laplacian or wavelet pyramid decompositions [4] [5]. Also, physically based models that incorporate more than per-pixel image processing have been suggested [6].

Therrien *et al.* [7] introduce a method to decompose visible and IR sources into their respective high and low frequencies and processes them in a framework inspired by Peli and Lim [8]. A nonlinear mapping is applied to each set of spectral bands to fuse them into the result. Therrien *et al.* [7] also address normalizing relative luminance responses between spectra. However, our technique attempts to match the IR response to the relative luminances of the visible spectrum while [7] matches both spectra to a Sammon mapping [9].

The core of our fusion technique is the separation of detail features (textures) from the large-scale features (uniform regions) of an image. These features are then remixed between spectra. This decomposition and recomposition is akin to the high dynamic range (HDR) compression technique introduced by Durand and Dorsey [10]. In HDR, the dynamic range of the large-scale features is decreased, whereas the details are preserved within a single image source. This decomposition is accomplished via the edge-preserving bilateral filter [11], a nonlinear algorithm that filters an image into regions of uniform intensity while preserving edges. The bilateral filter is a specific instance of the SUSAN filter of Smith and Brady [12], which performs edge detection with both range (intensity) and domain (spatial proximity) metrics. Extending this idea, the trilateral filter, discussed by Garnett *et al.* [13], uses a rank order absolute difference metric to robustly detect and handle shot noise within a bilateral filter formulation. The identically named trilateral filter of Choudhury and Tumblin [14] is another extension of the bilateral filter that targets a piecewise-linear result as opposed to piecewise-constant by adaptively altering the kernel.

A variant of the bilateral filter that uses a second image as the edge identification source, called the "joint bilateral filter," was proposed by Petschnigg *et al.* [15] and by Eisemann and Durand [16] (who refered to it as "the cross bilateral filter"). Both of these papers consider the problem of combining the details of an image captured with the use of a flash with the "look" of an image captured under ambient illumination. These papers discuss flash shadows, which account for edge differences between images. The multispectral relative luminance differences we address are another source of edge differences seen at different wavelengths.

Image fusion and texture transfer have been explored in the gradient domain, using Poisson solvers to reintegrate processed gradient fields. Socolinsky [17] used a Poisson interpolation formulation to match the output dynamic range to the desired display range. Techniques such as Poisson image editing [18], for texture transfer, and day/night fusion [19] generate gradient fields that contain visual elements from all images. This differs from our approach that seeks to enhance visible images without introducing nonvisible elements.

IR colorization algorithms, such as [2] and [4], attempt to learn a mapping from IR to chrominance and then construct a plausible colorized output. For that reason, colorization can be considered a class of fusion operator that fuses a chrominance prior into IR footage. In our technique, we instead recover actual chrominance solely from registered, but noisy, visible-light footage.

Our IR normalization process parallels the ratio-image work of Liu *et al.* [21]. Their work addresses reconstructing faces under similar luminance conditions. Our technique transfers details between physical structures that appear different at varying wavelengths.

Our work is also related to the topic of noise reduction in night-vision sensors. One approach to noise reduction is to use the bilateral spatial filters mentioned above [11], but this does not guarantee temporal coherence. Simple frame averaging for noise reduction is effective for static scenes, but it creates ghosting artifacts in dynamic scenes. We suggested an algorithm to reduce sensor noise without ghosting by adaptively filtering temporally adjacent samples [22], but it is forced to recover features in motion areas using only spatial filtering. The NL-means noise reduction used in [23] uses similar neighborhoods, which may not be spatially or temporally aligned, to attenuate noise. Although we employ noise reduction in the visible-light video to improve the quality of large-scale feature fusion, we acquire detail features from the less-noisy IR video.

Our capture rig, which consists of two registered cameras sharing a common optical path, is influenced by recent work in multisensor matting [24]. Their system was configured using similar cameras at differing focuses, while our rig uses cameras with varying spectral sensitivities.

III. Fusion Overview

Our video fusion can be broken down into four distinct stages:

1) noise reduction of the RGB video;
2) IR video normalization using ratio images;
3) decomposition of input videos into RGB lumininace low frequencies, edges, and IR detail features;
4) fusion of multispectral components into RGB output.

We reduce the visible spectrum's noise using temporal-edge-preserving bilateral filters (Section IV, "Prefilter" in Fig. 2). This noise reduction improves the accuracy of the decompositions, particularly in static image areas. It also filters chrominance, which is provided by the RGB and is processed in a separate pipeline (Fig. 5).

Many visible-spectrum textures are corrupted by video noise and must instead be acquired from the IR video. However, the

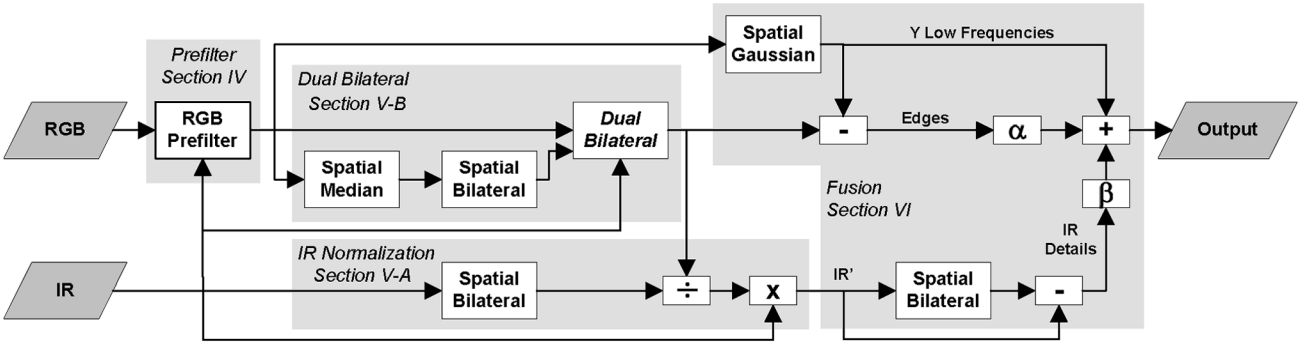


Fig. 2. Illustration of the luminance processing of our fusion technique. The RGB luminance signal (Y) provides the low frequencies. Assisted by the IR signal, the edges are extracted as well. The IR signal is normalized by a ratio of bilateral filters (large-scale features) then its detail features (textures) are isolated. The right side of the diagram demonstrates our linear combination of image components via α and β.

IR textures cannot be transferred directly due to relative luminance differences. Thus, we normalize the IR video to exhibit similar relative luminances to the RGB image (Section V-A, "IR normalization" in Fig. 2).

In order to extract sharp RGB edges we introduce a novel filter called the "dual bilateral" (Section V-B, "Dual Bilateral" in Fig. 2). This filter uses shared edge-detection information from both spectra simultaneously while considering sensor noise tolerances. It also enables more robust IR normalization.

Finally, we fuse the extracted components into a single video stream that contains reduced noise, sharp edges, natural colors, and visible-spectrum-like luminances (Section VI, "Fusion" in Fig. 2).

IV. RGB Video Noise Reduction

We first filter the visible spectrum video to improve the signal-to-noise ratio (SNR) of static objects and to provide improved color reproduction. This allows for more accurate decomposition and fusion later in the pipeline (Fig. 2).

We assume a noise model similar to that of [25]. At a high level, an image (I) can be decomposed into signal (E), fixed pattern noise (N_f), and temporal Poisson noise which we approximate with zero-mean Gaussian distributions. Thermal sensor noise (N_c) is modeled with constant variance while shot noise (N_s) is modeled with a variance dependent on exposure time and intensity

$$I = E + N_s + N_c + N_f. \tag{1}$$

We calculate a total noise variance, σ_i, for each sensor and label the sum of temporal noise as N_t

$$I = E + N_t + N_f. \tag{2}$$

In the case of a fixed camera, static objects may be reconstructed via temporal filtering and fixed pattern subtraction. The fixed pattern image, N_f, can be obtained by averaging many images taken with the lens cap on. Temporal filtering is achieved by averaging multiple static frames, reducing the contribution of the zero-mean noise $N_{t(m)}$ from each frame

$$\lim_{M\to\infty} \frac{1}{M}\left[\sum_{m=0}^{M}\left[E + N_{t(m)} + N_f\right] - N_f\right] = E. \tag{3}$$

In our fusion pipeline, noise is decreased in static areas using a temporal filter based on the bilateral filter [11] and the visible-spectrum temporal filtering of [22]. In the following sections, we describe our filter's design.

A. Spatial and Temporal Bilateral Filtering

Edge-preserving noise-reduction filters (e.g., anisotropic diffusion, bilateral filtering, median filtering, and sigma filtering) are the ideal filters for reducing noise in our circumstance. Smoothing both spatially and temporally while preserving edges enhances sharpness, preserves motion, and improves the constancy of smooth regions.

The bilateral filter [11], shown in (4), is a noniterative edge-preserving filter defined over the domain of some kernel Ω. The bilateral filter combines each kernel's center pixel s with the neighboring pixels p in Ω that are similar to s. In the original bilateral formulation dissimilarity is determined by luminance difference, shown in (5).

In addition to noise reduction, the bilateral filter is used because it decomposes images into two components which have meaningful perceptual analogs [10] [15]. The bilateral's filtered image has large areas of low frequencies separated by sharp edges, called the "large-scale features" [16]. The complement image, found through subtraction, contains "detail features," which are the textures

$$J_s = \frac{\sum_{p\in\Omega} g(\|p-s\|,\sigma_h)g(D(p,s),\sigma_i)I_p}{\sum_{p\in\Omega} g(\|p-s\|,\sigma_h)g(D(p,s),\sigma_i)} \tag{4}$$

$$D(p,s) \equiv I_p - I_s \tag{5}$$

$$g(x,\sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}. \tag{6}$$

In our noise filter, we choose Ω to include temporally aligned pixels in adjacent frames. The resulting filter is a temporal bilateral filter, useful for removing noise from static objects without blurring the motion. Edge-preservation in the spatial domain translates to preserving "temporal edges," or motion, in time. In essence, detecting "temporal edges" is equivalent to motion detection. However, due to the low SNR, it is difficult in practice to choose a σ_i to differentiate noise from motion based solely on a single pixel-to-pixel comparison.

To solve the problem of separating noise from motion when temporally filtering, we use a local neighborhood comparison

to determine dissimilarity, reminiscent of [22]. Instead of just comparing the intensities of s and p, as in (5), we use a sum of squared differences (SSD) between small spatial neighborhoods Ψ (typically 3×3 or 5×5) around s and p, weighted to favor the kernel's center by the Gaussian $\omega(x, y)$. This reduces the ambiguity between noise and motion because the larger Ψ neighborhood reduces the impact of single-pixel temporal noise, instead requiring the simultaneous change of many pixel intensities indicative of motion

$$D(p, s) \equiv \sum_{x}^{\Psi} \sum_{y}^{\Psi} \omega(x, y)(I_p - I_s)^2. \tag{7}$$

B. Robust Temporal Joint Bilateral Filtering

To further improve our filtering we incorporate ideas from the joint bilateral filter introduced in [15] and [16]. Joint bilateral filters allow a second image to shape the kernel's weights. Thus, all dissimilarity comparisons are made in one image, but the filter weights are applied to another. In our temporal filtering, this causes Ψ-neighborhood SSD motion detection in the IR video to determine the visible image's filter support. This is accomplished by modifying (7) as follows:

$$D(p, s) \equiv \sum_{x}^{\Psi} \sum_{y}^{\Psi} \omega(x, y) \left(I_p^{\mathrm{IR}} - I_s^{\mathrm{IR}}\right)^2. \tag{8}$$

Our complete noise reduction technique is a temporal-only joint bilateral filter that uses SSD Ψ neighborhood dissimilarities in the IR video to filter the visible video. This de-noises the static regions of the RGB video and improves color reproduction.

In most cases, visible-spectrum motion can be detected in the IR video even in the presence of significant relative luminance differences between spectra. If the SSD Ψ neighborhood motion detection fails, the system can be made more robust by replacing (4) with (14) discussed in Section V-B.

V. Video Decomposition Techniques

In this section, we describe methods to decompose prefiltered visible and IR videos into separate components. These components will be assembled (in Section VI) into our final fusion result. First, we discuss a per-pixel scaling of the IR video that normalizes it to resemble the visible light video. This allows the "detail features" to be acquired from the IR and appear correct when fused with RGB components. However, this normalization mapping requires knowledge of the large-scale features from the visible imagery, which cannot be robustly extracted using existing bilateral filters because of the remaining confounding noise. Therefore, we present an extension to the bilateral filter (the "dual bilateral") to address this problem. Because of its robustness, this new filter is also used to extract the image components that provide sharp edges in the final fusion.

From this point on, we will use the term "Y" to refer to only the luminance channel of the visible-spectrum input. The chrominance channels, U and V, are separated from the RGB video in YUV color space after prefiltering (Section IV) and processed separately in Section VI.

A. Y and IR Video Normalization

Before decomposing the input videos for fusion, we adjust their characteristics to more closely resemble the desired system output. To prepare the dark and underexposed Y, its histogram stretched to the display's full range, often 0–255, or to an HDR range.

Since our goal is to combine IR detail features with visual elements from the visible image, the IR video, from which those detail features are extracted, is remapped to better resemble the stretched Y video. These sources differ in both absolute and relative luminances, so features transferred from IR to visible may not smoothly fuse. Therefore, we correct these luminance differences by modulating the IR per-pixel image statistics to resemble those of the Y video.

The concept of ratio images, discussed by Liu *et al.* [21], resembles our normalization. In their application, images were captured of two faces in neutral poses ($\mathcal{A}$ and $\mathcal{B}$). By assuming a Lambertian illumination model, given a new expression on the first person's face ($\mathcal{A}'$) a similar expression could be simulated on the face of the second person ($\mathcal{B}'$) at each pixel (u, v) with the following modulation:

$$\mathcal{B}'(u, v) = \mathcal{B}(u, v) \frac{\mathcal{A}'(u, v)}{\mathcal{A}(u, v)}. \tag{9}$$

In our normalization, we do not have access to a neutral pose image standard. Instead, to correct differing relative responses, our ratio is the surface-to-surface luminance ratio. Since relative response differences are characteristic of surface types, it follows that their ratios in uniform image regions are ideal for normalization. Uniform regions of the Y and IR videos can be approximated with the spatial bilateral result, the large-scale features (Y_{LS} and $\mathrm{IR}_{\mathrm{LS}}$).

Thus, the following formulation normalizes the IR video:

$$\mathrm{IR}'(u, v) = \mathrm{IR}(u, v) \frac{Y_{\mathrm{LS}}(u, v)}{\mathrm{IR}_{\mathrm{LS}}(u, v)}. \tag{10}$$

This normalization is also similar to the per-pixel log-space texture transfers in both [10] and [16] and to the linear-space modulation in [15]. However, our normalization is applied to the original images, not to just a single component (such as their detail features). Normalization is crucial because of the significant relative luminance differences between image sources. Normalizing the entire image before decomposition may substantially change the image structure, meaning that prenormalized large-scale features may become detail features after normalization, and vice versa.

We run spatial bilateral filters on both the visible and IR videos to obtain Y_{LS} and $\mathrm{IR}_{\mathrm{LS}}$, respectively. For the well-exposed, relatively noise-free IR video, spatial bilateral filtering extracts the large-scale features as expected. However, directly recovering the large-scale features from the Y video using spatial bilateral filtering fails because it is confounded by the remaining noise. Recall, from Section IV, that many samples are required to significantly reduce noise and sufficient samples were unavailable in moving regions. To solve this problem, we use sensor readings from both video sources to accurately reconstruct the visible video large-scale features.

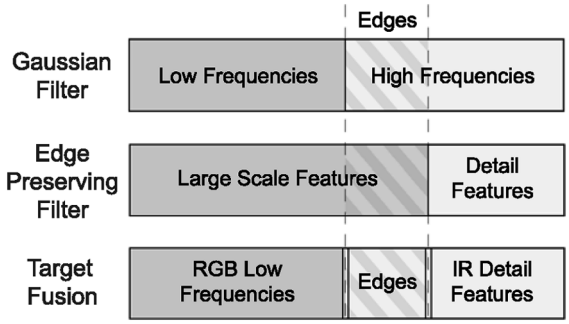


Fig. 3. Illustration of two common image decomposition methods and how those components are combined by our fusion method. Gaussian smoothing of an image extracts its low frequencies while the remainder of the image constitutes the high frequencies. Similarly, edge preserving filtering extracts large-scale features and details. We separate out the edges (the image components present in the high frequencies but not in the details) and use them in the output fusion.

B. Dual Bilateral Filtering

To filter the visible video while preserving edges in order to extract the large-scale features, we employ the registered IR video. We cannot, however, simply use the IR joint bilateral filter, discussed in Section IV-B, because of the inherent differences in spatial edges between the two sources (Fig. 10). As noted in Section I, features often appear in one spectra but not the other. We attempt to maintain all features present in the visible spectrum to avoid smoothing across edges. Therefore, we use multiple measurements to infer edges from our two non-ideal videos sources: the IR video, with its unnatural relative luminances, and the noisy Y video.

We use a bilateral filter which includes edge information from multiple sensors, each with its own estimated variance, to extract the Y large-scale features. Sensor noise variance estimates are determined through analysis of fixed-camera, static-scene videos. In the noisy visible video, edges must be significantly pronounced to be considered reliable. The less-noisy IR edges need not be as strong to be considered reliable.

This information is combined in the bilateral kernel as follows. The Gaussian distributions used by the bilateral filter's dissimilarity measures, shown in (5) and (8), can each be recast as the Gaussian probability of both samples p and s lying in the same uniform region given a difference in intensity, which we denote $U_{p,s}$

$$P(U_{p,s}\,|\,Y) = g\left(I_p^Y - I_s^Y, \sigma_Y\right) \tag{11}$$

$$P(U_{p,s}\,|\mathrm{IR}) = g\left(I_p^{\mathrm{IR}} - I_s^{\mathrm{IR}}, \sigma_{\mathrm{IR}}\right). \tag{12}$$

We wish to estimate the probability of samples p and s being in the same uniform region (i.e., no edge separating them) given samples from both sensors, $P(U_{p,s}\,|\,Y, IR)$. If we consider the noise sources in (11) and (12) to be independent, we can infer

$$P(U_{p,s}\,|\,Y, \mathrm{IR}) = P(U_{p,s}\,|\,Y)P(U_{p,s}\,|\,\mathrm{IR}). \tag{13}$$

From (13), it is clear that $P(U_{p,s}\,|\,Y, IR)$ will be low if either (or both) $P(U_{p,s}\,|\,Y)$ or $P(U_{p,s}\,|\,IR)$ are low due to detection of a large radiometric difference (an edge). We substitute (11), (12), and (13) into (4) to derive a "dual bilateral" filter which uses sensor measurements from both spectra to create a combined

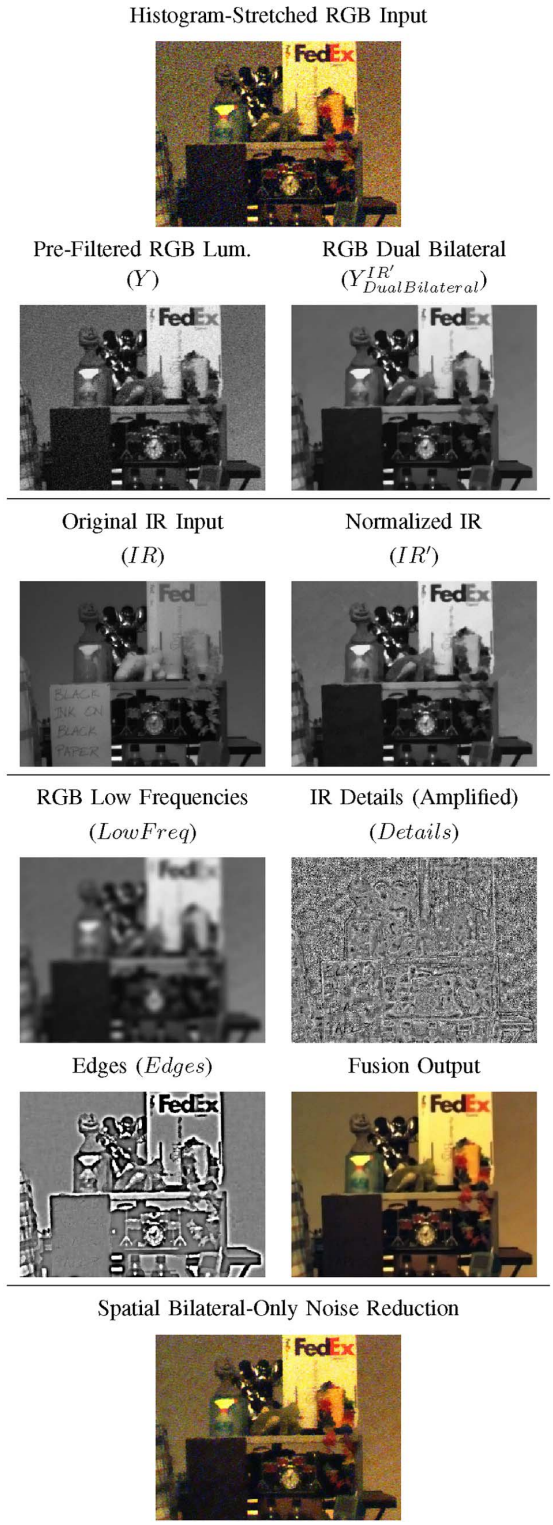


Fig. 4. Illustration of images at various stages of our processing pipeline associated with the variables used in Section VI. Specifically note the quality of the dual bilateral, the proper relative luminances of the normalized IR, and the image components which constitute the final fused output. For comparison, we show spatial bilateral-only noise reduction. Note that although at this size the normalized IR and dual bilateral Y images appear similar, the dual bilateral lacks texture details found in the IR′.

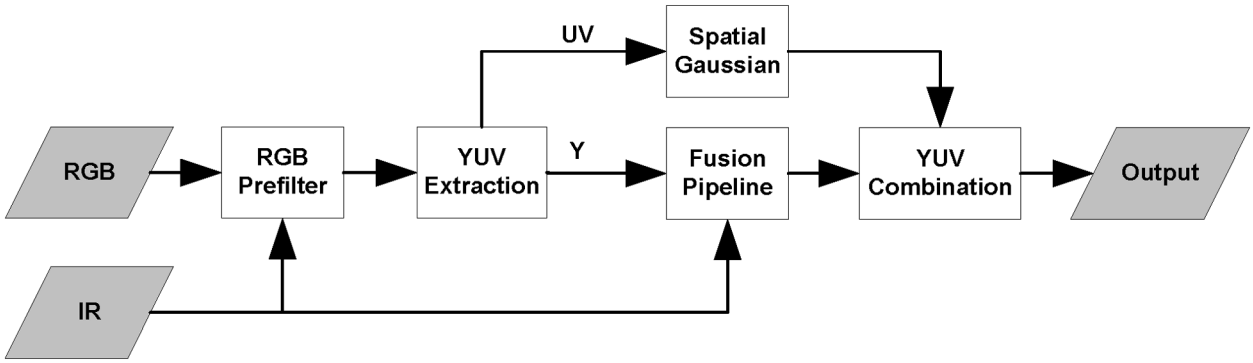


Fig. 5. Diagram of chrominance processing in our pipeline. After the RGB temporal noise prefiltering, the signal is converted to YUV. The Y component goes through the pipeline in Fig. 2, while the U and V channels are Gaussian smoothed to remove any noise where it was not removed by the prefiltering (i.e., in areas of motion). Note, prefiltering is shown in both figures to illustrate when in the overall pipeline the luminance and chrominance signals are split, but prefiltering is performed only once.

dissimilarity metric

$$J_s = \frac{\sum_{p\in\Omega} g(\|p-s\|,\sigma)P(U_{p,s}\,|\,Y)P(U_{p,s}\,|\,IR)I_p}{\sum_{p\in\Omega} g(\|p-s\|,\sigma)P(U_{p,s}\,|\,Y)P(U_{p,s}\,|\,IR)}. \tag{14}$$

This dual bilateral is now used to extract the large-scale features from the visible-spectrum video. The advantages of this approach beyond joint bilateral filtering are illustrated in Fig. 10.

In the presence of appreciable single-pixel shot noise, the $P(U_{p,s}\,|\,Y)$ measure can be confounded, resulting in edges being detected where none exist. We, therefore, assume that no single-pixel detail in the noisy Y video should be considered an edge. To incorporate this notion, we calculate the $P(U_{p,s}\,|\,Y)$ term in (14) using a median-filtered Y video that eliminates this shot noise (the Y video filtered by the dual bilateral is unaffected). If desired, any remaining Gaussian temporal noise in the Y edge-detection source can be further attenuated via bilateral filtering. This additional filtering is depicted prior to the dual bilateral in Fig. 2.

Our framework supports additional sensors by multiplications of $P(U_{p,s}\,|\,\mathrm{Sensor})$ in both the numerator and denominator. Because of the bilateral form, any associated scalars will cancel out.

VI. Multispectral Bilateral Video Fusion

The final step is to gather the necessary image components and fuse them together into our result. However, first we will discuss the optimal fusion for creating enhanced RGB visible-spectrum images. To reiterate, our goal is to reconstruct the RGB source in an enhanced manner with the assistance of the IR imager only as needed.

Fig. 3 shows two methods for decomposing images: Gaussian decomposition into low and high frequencies and edge-preserving decomposition into large-scale features and detail features. The image's sharp edges lie in the area indicated by the dashed lines. To construct the fusion, we combine RGB luminance low frequencies, IR detail features, edges, and chrominance. We now summarize our rationale for our filtering and fusion choices.

Even in the presence of noise, the RGB luminance video (Y) contains low frequencies of sufficient quality. These provide correct relative luminances for large, uniform image regions. We extract the low frequencies $(LowFreq)$ by Gaussian smoothing the prefiltered RGB luminance from Section IV-B (Y_{Gaussian}).

Because the Y details are most corrupted by visible-spectrum sensor noise, we seek evidence for them in the normalized IR footage (IR'). Detail features $(Details)$ are obtained by subtracting the IR spatial bilateral's large-scale features $(IR'_{\mathrm{Bilateral}})$ from its unfiltered image (IR') (Fig. 3). We use IR' detail features for the entire output image, including static regions because we know from [26] that the minimum signal recoverable from a video source is the mean of the dark current noise at any pixel. Therefore, there are textures in dark areas of the visible-spectrum video that luminance averaging cannot reconstruct. In our case, the better-exposed IR' footage provides those unrecoverable details.

Obtaining accurate edges $(Edges)$ is crucial to the sharpness of our fusion output image, but the visible-spectrum edges were corrupted by noise during capture. Alternately, not all the edges are present in the IR footage, preventing a direct IR edge transfer. However, the dual bilateral filter in Section V-B can extract enhanced visible-spectrum large-scale features with additional IR measurements $(Y^{\mathrm{IR}'}_{\mathrm{DualBilateral}})$. The edge components are isolated by subtracting a Gaussian with matching support (Y_{Gaussian}). Considering our image deconstruction model (Fig. 3), the edges complete the fusion along with the RGB luminance low frequencies and the IR detail features.

The equations below detail the entire luminance fusion process. This pipeline is also shown in Fig. 2 and depicted with step-by-step images in Fig. 4

$$\begin{aligned}
\mathrm{LowFreq} &\equiv Y_{\mathrm{Gaussian}} \\
\mathrm{Edges} &\equiv Y^{\mathrm{IR}'}_{\mathrm{DualBilateral}} - Y_{\mathrm{Gaussian}} \\
\mathrm{Details} &\equiv IR' - IR'_{\mathrm{Bilateral}} \qquad (15)\\
Y' &= \mathrm{LowFreq} + \alpha(\mathrm{Edges}) + \beta(\mathrm{Details}). \qquad (16)
\end{aligned}$$

A linear combination of the image components determines the final reconstruction. For our examples, α was set at 1.0 and β was varied between 1.0 and 1.2 depending on the texture content. Values of α greater than 1.0 result in sharper edges but would lead to ringing artifacts. When $\alpha = 1.0$, it is unnecessary to decompose LowFreq and Edges, as $Y^{\mathrm{IR}'}_{\mathrm{DualBilateral}}$

Fig. 6. Result 1–From top to bottom: A frame from an RGB video of a person walking, the same frame from the IR video, the RGB frame histogram stretched to show noise and detail, and our fusion result. Notice the IR video captures neither the vertical stripes on the shirt, the crocodile's luminance, nor the plush dog's luminance. Furthermore, note the IR-only writing on the sign. These problem areas are all properly handled by our fusion.

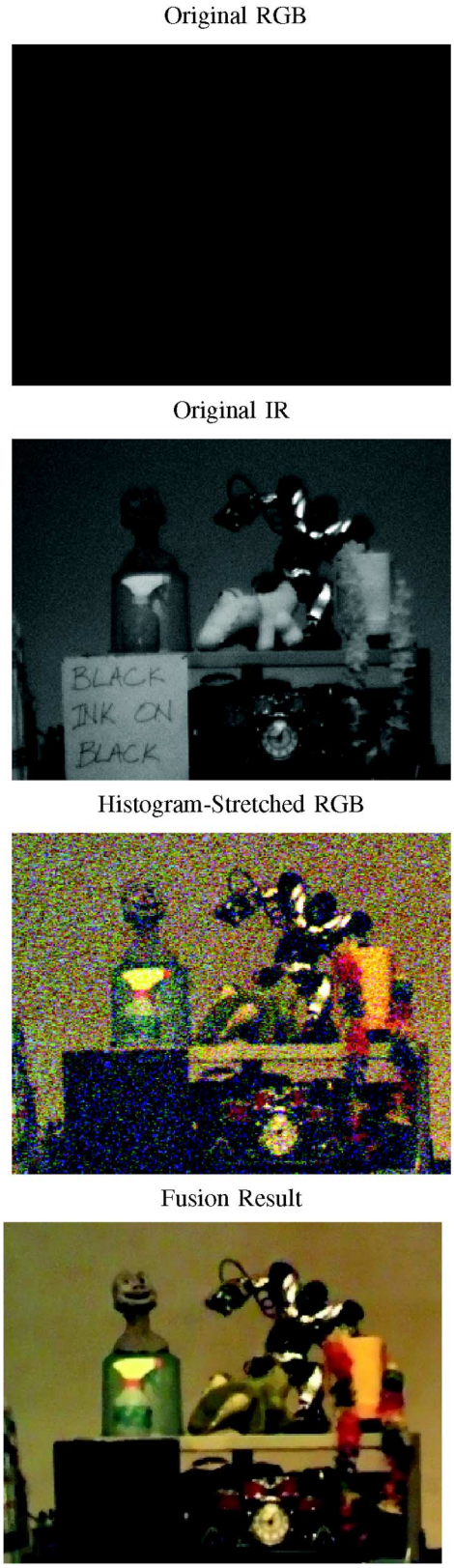


Fig. 7. Result 2–From top to bottom: A frame from an RGB video of a moving robot, the same frame from the IR video, the RGB frame histogram stretched to show noise and detail, and our fusion result.

contains both. Subsequently, (16) becomes

$$Y' = Y^{\mathrm{IR}'}_{\mathrm{DualBilateral}} + \beta(\mathrm{Details}). \tag{17}$$

The UV chrominance is obtained from the prefiltered RGB from Section IV-B. Gaussian smoothing is used to remove chrominance noise (especially in the nonstatic areas not significantly improved by prefiltering). The full chrominance pipeline is shown in Fig. 5. Although it is possible to filter the UV in the

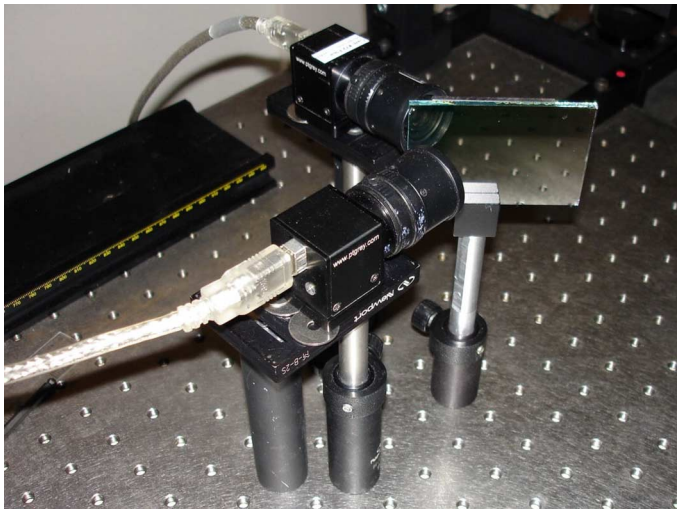


Fig. 8. Photograph of our capture setup with a Point Grey Color Flea capturing the visible-spectrum RGB and a filtered Point Grey Grayscale Flea capturing the nonvisible IR spectrum.

same manner as the luminance (i.e., using the detected edges to limit filtering across edges) doing so limits each pixel's support size compared to Gaussian smoothing. Insufficient support leads to noise artifacts and local "blotchiness." We trade off sharpness for lower chrominance noise and, thus, rely on the low spatial chrominance sensitivity of the human visual system to limit blurring artifacts.

VII. Results

Our RGB and IR videos are captured using two synchronized (genlocked) video cameras sharing a common optical path. Two PointGrey Flea cameras (one grayscale and one RGB) are used with the grayscale camera covered by a longpass filter passing only IR light (780 nm 50% cutoff, Edmund Optics #NT32-767). The two cameras are arranged as shown in Figs. 1 and 8. A cold mirror (reflects $\sim$90% of the visible spectrum, transmits $\sim$80% of the IR spectrum, Edmund Optics #NT43-961) is used as a beamsplitter because the spectral sensitivities of our sensors are mutually exclusive. Thus, we increase the number of photons reaching the appropriate CCD over traditional beamsplitting. Since each camera has its own lens and sensor alignment, their optical paths may differ slightly. Therefore, a least-squares feature-based homography transform is used to register the RGB video to the IR video prior to processing. The RGB sensor has a higher resolution, so some downsampling occurs during the homography registration. A benefit of this two sensor setup is that in well-lit environments, this same capture rig can also capture visible RGB footage.

Because our IR sensor is an unmodified off-the-shelf imager, it is significantly less sensitive to IR than specialized IR sensors, such as InGaAs SWIR sensors. Such high-end sensors would be ideal for our fusion algorithm. Yet even under our circumstances, our IR sensor is sensitive up to roughly 1000 nm and provides sufficiently bright imagery for fusion. Also, we benefit from having similar Flea camera bodies, allowing for accurate alignment between imagers.

The noise reduction filters in Sections IV-A and V-B rely upon σ values (6) derived from sensor noise characteristics in

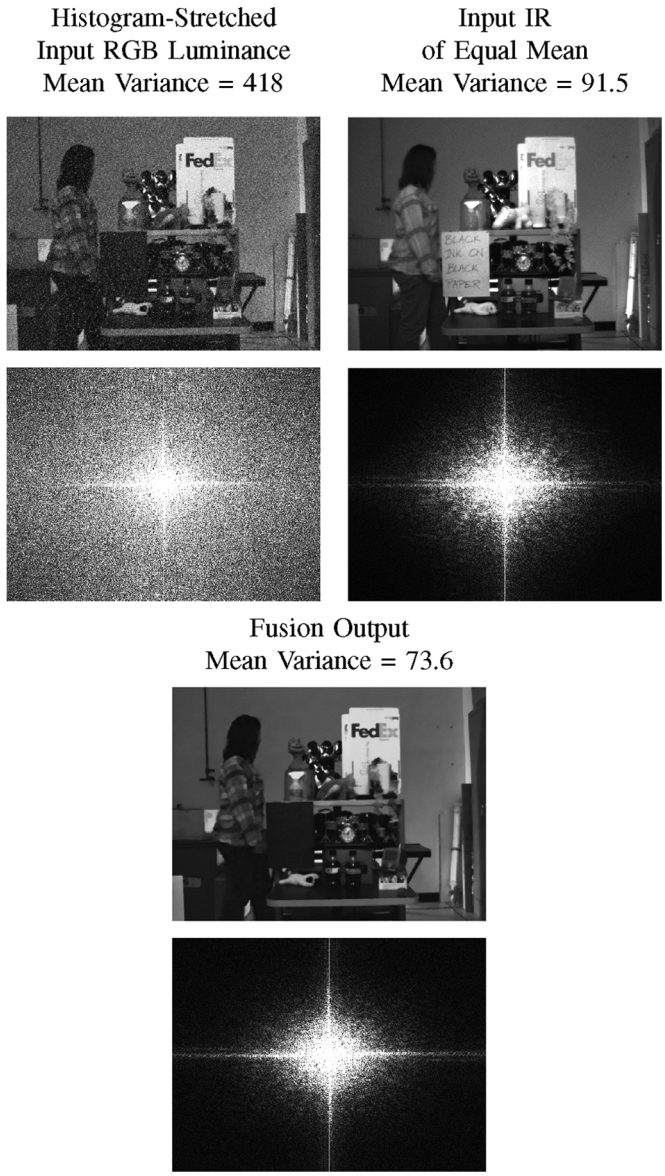


Fig. 9. Comparison of the mean spatial variance within a 3×3 window and power spectrum of each of our input images and the fused output. (Upper left) The original noisy RGB luminance input is shown with its mean variance and spectral noise. As in our fusion, it is histogram stretched to use more of the display's range. (Upper right) The less-noisy IR input exhibits less high-frequency noise and a lower mean variance than the visible spectrum sensor. For a fair comparison, the histogram of the IR was also stretched to match the visible-spectrum mean, a step not part of our fusion. (Bottom) Our fusion result is significantly improved with reduced noise and mean variance while still preserving high-frequency content. These statistics are similar to the IR video, yet we achieve them with a visible-spectrum-like response.

static environments. Experimentally, we found an average σ_Y of 8.78 for the RGB sensor and σ_{IR} of 2.25 for the IR sensor. However, we chose values of $\sigma_Y = 7.5$ and $\sigma_{\mathrm{IR}} = 2.5$ to account for subsequent median and bilateral processing.

Our first example, shown in Fig. 6, shows a frame from a video sequence processed using our method. In the video, a person walks across the camera's view. Note that the plaid shirt, the plush crocodile, the flowers, and the writing on the paper in the IR video do not match the RGB footage (Fig. 10). With our approach, we preserve details and also show noise reduction in all image regions. Fig. 9 shows the improvement in signal

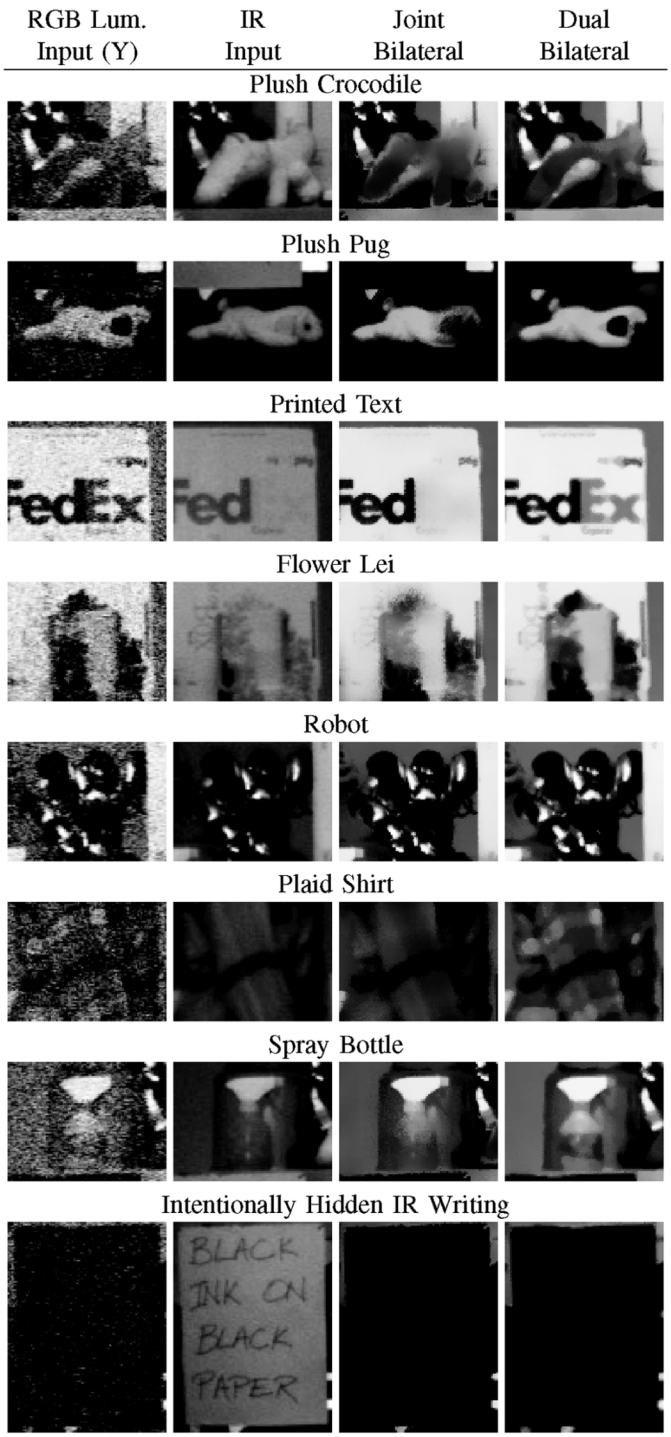


Fig. 10. Illustration of the difference in quality between the joint bilateral filter [15], [16] and our dual bilateral filter, each configured for the best output image quality. The desired output is a enhanced version of the RGB luminance (Y) that preserves all edges. Because the joint bilateral filter relies on IR edges to filter the Y, it cannot correctly handle edges absent in the IR due to relative luminance response differences. This results in blurring across the nondetected edges in the result. However, our dual bilateral filter detects edges in both inputs (weighted by sensor noise measurements) and is, thus, better at preserving edges only seen in the Y video. Again, note that our desired filter output should resemble the visible spectrum, meaning objects visible only in IR should not be included.

quality (mean variance) without loss of sharpness for a frame of this video. Our second example, shown in Fig. 7, shows the reconstruction of a moving robot video. This video poses similar problems to the previous example in addition to proper handling of specular highlights.

Finally, Fig. 4 illustrates the stages of our processing pipeline by showing images as they are filtered and fused through our system. The images were taken from a 20 frame video with no motion.

VIII. Future Work

The primary area for future work is in improved color reconstruction. The chrominance in our model relies entirely on the RGB video and does not consider any of the IR information. However, in areas of motion, our temporal-only filter cannot significantly improve the chrominance quality. Thus, a supplemental learned model might help reduce the blotchiness of the chrominance in those areas.

Second, the large filter kernels necessary to remove low-frequency image variations due to slight heat variations on the CCD or CMOS sensor cause our approach to be slow. Increasing the speed of these filters, possibly using the techniques of Durand and Dorsey [10], would be beneficial.

Finally, we have focused on bilateral functions to help classify edges, texture, and smooth areas while also providing de-noising. Wavelet decompositions might also provide similar functionality, possibly at reduced computational cost.

IX. Conclusion

We have shown how RGB and IR video streams captured using the same optical path can be fused into an enhanced version of the RGB video. This is accomplished by initially denoising the RGB video, normalizing the IR video, decomposing each, and then fusing select components back together. By using a variety of filters derived from the same root bilateral filter, we are able to reduce noise, preserve sharpness, and maintain luminances and chrominances consistent with visible-spectrum images.

Acknowledgment

The authors would like to thank J. Zhang, L. Zitnick, S. B. Kang, Z. Zhang, and the review committee for their feedback and comments.

References

[1] C. Pohl and J. V. Genderen, "Multisensor image fusion in remote sensing: Concepts, methods, and applications," *Int. J. Remote Sens.*, vol. 19, no. 5, pp. 823–854, 1998.

[2] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, 2002.

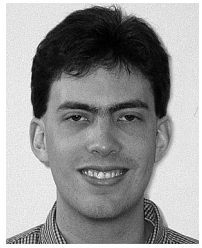
[3] D. Fay, A. Waxman, M. Aguilar, D. Ireland, J. Racamato, W. Ross, W. Streilein, and M. Braun, "Fusion of multi-sensor imagery for night vision: Color visualization, target learning and search," presented at the Int. Conf. Information Fusion, 2000.

[4] A. Toet, "Hierarchical image fusion," *Mach. Vis. Appl.*, vol. 3, no. 1, pp. 1–11, 1990.

[5] H. Li, B. Manjunath, and S. K. Mitra, "Multi-sensor image fusion using the wavelet transform," in *Proc. Int. Conf. Image Process.*, 1994, pp. 51–55.

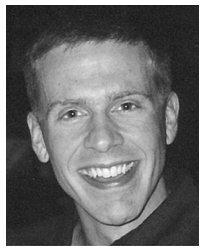
[6] N. Nandhakumar and J. Aggarwal, "Physics-based integration of multiple sensing modalities for scene interpretation," *Proc. IEEE*, vol. 85, no. 1, pp. 147–163, Jan. 1997.

[7] C. Therrien, J. Scrofani, and W. Krebs, "An adaptive technique for the enhanced fusion of low-light visible with uncooled thermal infrared imagery," in *Proc. Int. Conf. Image Process.*, 1997, pp. 405–408.
[8] T. Peli and J. S. Lim, "Adaptive filtering for image enhancement," *Opt. Eng.*, vol. 21, no. 1, pp. 108–112, 1982.
[9] C. Sammon, "A nonlinear mapping algorithm for data structure analysis," *IEEE Trans. Comput.*, vol. C-18, no. 5, pp. 401–409, May 1969.
[10] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, 2002.
[11] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. Int. Conf. Computer Vision*, 1998, pp. 836–846.
[12] S. M. Smith and J. M. Brady, "Susan-a new approcah to low level image processing," *Int. J. Comput. Vis.*, vol. 23, no. 1, pp. 45–78, 1997.
[13] R. Garnett, T. Huegerich, C. Chui, and W. He, "A universal noise removal algorithm with an impulse detector," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1747–1754, Nov. 2005.
[14] P. Choudhury and J. Tumblin, "The trilateral filter for high contrast images and meshes," in *Proc. Eurographics Symp. Rendering*, 2003, pp. 1–11.
[15] G. Petschnigg, M. Agrawala, H. Hoppe, R. Szeliski, M. F. Cohen, and K. Toyama, "Digital photography with flash and no-flash pairs," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 661–669, 2004.
[16] E. Eisemann and F. Durand, "Flash photography enhancement via intrinsic relighting," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 670–675, 2004.
[17] D. A. Socolinsky, "Dynamic range constraints in image fusion and visualization," in *Proc. Signal and Image Processing Conf.*, 2000, pp. 349–354.
[18] R. Perez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.
[19] R. Raskar, A. Ilie, and J. Yu, "Image fusion for context enhancement and video surrealism," in *Proc. Int. Symp. Non-Photorealistic Animation and Rendering*, 2004, pp. 85–94.
[20] A. Toet, "Colorizing single band intensified nightvision images," *Displays*, vol. 26, no. 1, pp. 15–26, 2005.
[21] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," *ACM Trans. Graph.*, vol. 20, no. 3, pp. 271–276, 2001.
[22] E. P. Bennett and L. McMillan, "Video enhancement using per-pixel virtual exposures," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 845–852, 2005.
[23] A. Buades, B. Coll, and J. M. Morel, "Denoising image sequences does not require motion estimation," in *Proc. IEEE Conf. Advanced Video and Signal Based Surveillance*, 2005, pp. 70–74.
[24] M. McGuire, W. Matusik, H. Pfister, J. Hughes, and F. Durand, "Defocus video matting," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 567–576, 2005.
[25] Y. Tsin, V. Ramesh, and T. Kanade, "Statistical calibration of ccd imaging process," in *Proc. IEEE Int. Conf. Computer Vision*, 2001, pp. 480–487.
[26] M. Grossberg and S. Nayar, "Modeling the space of camera response functions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1272–1282, Oct. 2004.



Eric P. Bennett received the M.S. degree from the Department of Computer Science, University of North Carolina, Chapel Hill, in 2004, where he is currently pursuing the Ph.D. degree in the same program.

He received the B.S. degree in computer engineering from Case Western Reserve University, Cleveland, OH, in 2002. His research explores new techniques for video processing including visualization, editing, noise reduction, and IR fusion.



John L. Mason received the B.S. degree in industrial design in 1997. He is currently pursuing the M.S. degree in computer science at the University of North Carolina, Chapel Hill.

He worked in the fields of computer-based training and multimedia from 1998 to 2004. His research interests span the fields of graphics and intelligent multimedia systems.

Leonard McMillan received the B.S. and M.S. degrees from the Georgia Institute of Technology, Atlanta, and the Ph.D. degree from the University of North Carolina, Chapel Hill.

He is an Associate Professor in the Department of Computer Science, University of North Carolina, Chapel Hill. His research interests include image-based approaches to computer graphics and applications of computer vision to multimedia.