GRADIENT DOMAIN CONTEXT ENHANCEMENT FOR FIXED CAMERAS

Adrian Ilie

Department of Computer Science,
University of North Carolina
at Chapel Hill, USA

Ramesh Raskar

Mitsubishi Electric
Research Laboratories,
Cambridge, USA

Jingyi Yu

Laboratory of Computer Science,
The Massachusetts Institute
of Technology, Cambridge, USA

Fig. 1. Enhancing a night time scene from an airport surveillance camera. A low quality nighttime image, and the final output of our algorithm.

ABSTRACT

We propose a class of enhancement techniques suitable for scenes captured by fixed cameras. The basic idea is to increase the information density in a set of low quality images by exploiting the context from a higher-quality image captured under different illumination from the same view. For example, a nighttime surveillance video can be enriched with information available in daytime images.

We propose a new image fusion approach to combine images with sufficiently different appearance into a seamless rendering. Our method ensures the fidelity of important features and robustly incorporates background contexts, while avoiding traditional problems such as aliasing, ghosting and haloing. We show results on indoor as well as outdoor scenes.

1. INTRODUCTION

In recent years, a number of techniques have emerged to extract useful information from multiple images taken from a fixed viewpoint. They include video summarization, generation of intrinsic images, multi-spectral image fusion and high dynamic range (HDR) compression.

In this paper we propose a different class of image and video enhancement techniques, which we call *context enhanced rendering* (CER, for simplicity and because there is no common term). The goal of CER is to extract context information of a scene from one image and important features from another image of the same scene. HDR imaging and image fusion are special cases of CER. We call the image that provides environmental context the *background image*, and the one that provides desirable features the *foreground image*.

A typical example of CER we consider in this paper is enhancing nighttime traffic or surveillance videos using daytime images taken from the same viewpoint. Usually the nighttime video is very difficult to understand because it lacks background context due to poor illumination. However, the elements of this background context, such as roads and buildings are critical to understanding the video as shown in Figures 1 and 3. While a trained traffic controller may easily recognize important features in nighttime videos, we think that our method would help non-specialists achieve the same performance. Moreover, even traffic controllers may benefit from our approach: they may confirm their suspicions by switching between normal imagery and our context-enhanced imagery.

1.1. Overview

Our approach is a new image fusion approach to combine snapshots of the same scene with sufficiently different appearance into a seamless rendering. For the rest of the paper, we limit ourselves to this type of CER. The method maintains the fidelity of important features and robustly incorporates background contexts, while avoiding problems encountered in traditional methods, such as aliasing, ghosting and haloing. We first encode the importance based on local variance in input snapshots or videos. Then, instead of a convex combination of pixel intensities, we combine the intensity gradients scaled by the importance. The reconstruction from the gradients achieves a smooth blend of the inputs, and at the same time preserves their important features.

Similar to compressing high dynamic range images, the result of CER should be “visually pleasing”, i.e., it should

have very few artifacts and it should exhibit a smooth transition from background to foreground. Our method accomplishes this by using the underlying properties of integration. We show how this can be used for synthetic as well as natural indoor and outdoor scenes.

A common artifact of gradient-based reconstruction is that it introduces observable color shifting. We will discuss in details the causes of these artifacts and show a color assignment scheme that can efficiently reduce them.

1.2. Contributions

Our main contribution is the idea of exploiting information available from fixed cameras to create context-rich, seamless results. Our technical contributions include a scheme for asymmetrically fusing two snapshots of the same scene while preserving useful features; and a method for context enhancement of videos in the context of unreliable frame differencing.

In addition, we modify the reconstruction from gradients method with a padding scheme to overcome integration artifacts and we employ a color assignment strategy to address color shifting problems.

1.3. Related Work

Methods to combine information from multiple images into a single result have been explored for various other applications. They range from image editing [1] to tone mapping for compression of variable-exposure high-dynamic range images [2],[3].

The authors of [4] use multi-resolution splines for combining images into a seamless image mosaic. The source images are first decomposed into a set of band-pass filtered component images. Next, the component images in each spatial frequency band are assembled into a corresponding band-pass mosaic using a weighted average within a transition zone which is proportional in size to the wave lengths represented in the band. Finally, these band-pass mosaic images are summed to obtain the desired image mosaic.

In HDR imaging, a set of images taken under different levels of exposure are combined into a single image where details in all of the images are preserved while the overall contrast is reduced. However, in HDR, the pixel intensities increase monotonically. Usually cameras can only capture images at one fixed exposure time one at a time and hence can only capture part of the scene when the radiance range of the scene is high. The goal of HDR is to compensate context that is missing in one exposure setting from another. Two classes of approaches have been suggested: image space [5] and gradient space [2] methods. A recent approach [6] combines classic HDR techniques with motion estimation and other video methods to obtain HDR video.

On the other hand, our problem is quite different from combining high dynamic range images. For example, in day-night images we see intensity gradient reversals (such as objects that are darker than their surroundings during the day, but brighter than their surroundings during the night). An example of such a reversal is a building that is lit during the night so it becomes brighter than the nighttime sky, yet during the day it is darker than the daytime sky. These reversals do not appear in HDR, so special care needs to be taken for general CER methods.

Another example of CER is image fusion for multi-spec-tral imagery e.g. to merge satellite imagery captured at different wavelengths. Here, the images are relatively similar. Many ideas from multi-spectral image fusion can mutually benefit CER. Our approach is closest to the one proposed in [7]. They put forward a gradient space method by first forming a unified gradient image and then searching for an optimal image that satisfies gradient image.

A similar problem to enhancing images with context is removing or reducing undesirable context in images, such as shadows or fog. Authors of [8] remove shadows in an image by first computing its gradient, then distinguishing shadow edges, setting the gradient values at the shadow edges to zero and finally reintegrating the image. Nayar et al. use time-lapsing image sequences to model the effect of fog [9]. By setting appropriate parameters, they are able to efficiently enhance images and reduce undesirable weather artifacts.

Pèrez et al. [1] present a technique that uses integration of modified gradients from several images to produce one seamless result. However, since their goal is to provide a framework for seamless image editing, they rely heavily on user input to assign the areas from which the gradients are taken. The user designates which areas should come from which image, which is equivalent to a particular case of our method with simple or no blending and precise manual segmentation.

In this paper, we focus on enhancing poor-context nighttime snapshots or videos with context elements from high-quality daytime snapshots. Our proposed algorithm consists of two major steps: foreground extraction and background fusion. We believe robust foreground extraction in image space is difficult to achieve in practice, especially when dealing with low contrast and noisy snapshots and videos. Therefore we propose a gradient space algorithm that avoids a lot of undesirable artifacts like aliasing, ghosting and haloing that appear when using conventional methods.

We demonstrate our algorithm in different situations and show that our method is robust to poor foreground segmentation and generates day/night combinations with very few artifacts. We are inspired by many of the techniques mentioned here and aim to address some of their limitations.

2. BASIC TECHNIQUE

This section describes our basic fusion technique. We first present the basic algorithm, then our approach to ensure better reconstruction and color assignment.

2.1. Basic Algorithm

Our method combines information from two snapshots in a meaningful way, by picking high-quality background information from a daytime snapshot and using it to enhance the low-quality but important information from a nighttime snapshot. A straightforward approach is to use a linear combination of the input snapshots. We instead specify the desired local attributes of the final result and solve the inverse problem of obtaining a global solution that satisfies the local attributes. This leads to a non-linear combination, which means pixels with the same intensities map to different intensities in the final result. Our basic idea for determining the important areas of each snapshot relies on the widely accepted assumptions [10] that the human visual system is not very sensitive to absolute luminance reaching the retina, but rather responds to local intensity ratio changes. Hence, the local attribute is the local variance and we define an importance function for each input snapshot based on the spatial and temporal intensity gradients, which are a measure of the local spatial and temporal variance.

Our approach is based on two heuristics: (a) we carry into the desired result the gradients from the nighttime snapshot that appear to be locally important and (b) we use gradients from the daytime snapshot to provide context to locally-important areas while maintaining intra-image coherence. Note that we do not improve the quality of the pixels themselves, but simply give sufficient context to improve human interpretation. Hence any operations such as contrast enhancement, histogram equalization, mixed Gaussian models for background estimation [11] are orthogonal to our approach and can be easily used alongside to improve the final result.

The regions of high spatial variance across each snapshot are computed by thresholding the intensity gradients, $G = (G^X, G^Y)$, for the horizontal and vertical directions using a simple forward difference. We then compute an *importance image* (a weighting function) W, by processing the gradient magnitudes $|G_D|$ and $|G_N|$ of the daytime snapshot D and the nighttime snapshot N, respectively. The weighted combination of the input gradients gives us the gradient of the desired output. The basic steps are as described in Algorithm 1.

As described in the following sections, the process of determining importance weights $W_{(x,y)}$, depends on the specific application.

Algorithm 1 Basic algorithm

Find gradient field of daytime snapshot $G_D = \nabla D$
Find gradient field of nighttime snapshot $G_N = \nabla N$
Compute importance image W from $|G_D|$ and $|G_N|$
for each pixel (x,y) **do**
  Compute mixed gradient field $G_{(x,y)} = G_{N(x,y)}W_{(x,y)} + G_{D(x,y)}(1 - W_{(x,y)})$
end for
Reconstruct result I' from gradient field G
Normalize pixel intensities in I' to closely match $N_{(x,y)}W_{(x,y)} + D_{(x,y)}(1 - W_{(x,y)})$

2.2. Image Reconstruction

Image reconstruction from gradients fields is an approximate invertibility problem, and still a very active research area. In 2D, a modified gradient vector field G may not be integrable. We use one of the direct methods recently proposed [2] to minimize the error $|\nabla I' - G|$. The estimate of the desired intensity function I', so that $G = \nabla I'$, can be obtained by solving the Poisson differential equation $\nabla^2 I' = divG$, involving a Laplace and a divergence operator. We use the full multigrid method [12] to solve the Laplace equation. We pad the input images to square images of size the nearest power of two before applying the integration, and then crop back the result to the original size.

One needs to specify boundary conditions for the solver (at the border of the image). A natural choice is Neumann condition $\nabla I' \cdot n = 0$ i.e. the derivative in the direction normal to the boundary is zero. This is clearly not true when high gradients are present near the image boundary, resulting in noticeable color bleeding and shifting artifacts. Padding the image to the nearest power of two for multigrid integration helps alleviate this problem.

Pseudo-integration of the gradient field involves a scale and shift ambiguity, $I''_{(x,y)} = c_1 I'_{(x,y)} + c_2$. To obtain the final image, I'', we compute the unknowns, c_1 and c_2, (in the least square sense) using a simple heuristic: the overall appearance of each part of the reconstructed image should be close to the corresponding part of the foreground and background images. Each pixel leads to a linear equation, $\sum W_{i(x,y)} I_{i(x,y)} = c_1 I'_{(x,y)} + c_2$. We do image reconstruction in all three color channels separately and compute the unknowns per channel.

3. ENHANCEMENT OF DYNAMIC SCENES

For dynamic scenes, our results are based on the observation that if the camera and most of the viewed geometry remain static, only illumination and minor parts of the scene change (e.g., moving objects like people, devices, vehicles). Thus, the intensity gradients corresponding to the stationary parts

in the nighttime snapshot can be replaced with better quality gradients from a daytime snapshot.

We use the notions of a static background and a dynamic foreground to provide context for an action or event. The static component can be captured at high resolution, under controlled illumination conditions. The dynamic component can be captured in multiple snapshots of lower quality. A good example is enhancing pictures of theme park visitors taken during a ride through a dark environment, when bright flashes cannot be used because they may harm the visitors' eyes. The static background can be inserted from a snapshot captured using brighter illumination, when there are no visitors in the scene. Also, using a higher resolution background image can increase the perceived resolution of the dynamic foreground.

A simple choice for the weights $W_{(x,y)}$, used by the authors of [1], is to compute the desired gradient field as the local maximum of the input gradients, $G_{(x,y)} = max\ (G_{d(x,y)}, G_{n(x,y)})$. In this case importance weights are either 0 or 1. A better choice for our application is to give more importance to nighttime gradients in region of the nighttime snapshot where gradients or intensities are above a fixed threshold. This is to make sure that no information in the nighttime snapshot is lost in the final result.

To provide context to foreground illumination and geometry changes in the nighttime snapshot, we replace low-detail background areas using data from the daytime snapshot. This is where many of the traditional method using linear combination will fail to create seamless results. Let us consider the case where we want to provide context to nighttime snapshot N using information from another nighttime reference snapshot R and a daytime snapshot D. We create a mask image M, and set $M_{(x,y)} = |N_{(x,y)} - R_{(x,y)}|$ so that the importance is scaled by the difference between the two nighttime snapshots. Mask M is thresholded and normalized, then multiplied by the weights for snapshot N. (See Figure 2)

Although we use a very simple segmentation technique (pixel-wise difference in color space between snapshots N and R) to detect important changes at nighttime, our method is robust and does not need to rely on complicated segmentation techniques to obtain reasonable results. This is because we need to detect the difference between N and R only where gradients of N are sufficiently large. In a pair of snapshots, flat regions may have similar color but they naturally differ in regions of high gradient.

We allow for graceful degradation of the result when the underlying computer vision methods fail. More sophisticated segmentation techniques would bring marginal improvements to our results. Additionally, user input can help guide the algorithm by manually modifying the importance image.

Fig. 2. Enhancing a dynamic scene. (Top row) A low quality nighttime reference, and with a foreground person, a simple binary mask, the importance image obtained after processing. (Bottom row) A high quality daytime snapshot, the final output of our algorithm, compared with averaging and blending pixel intensities.

4. ENHANCEMENT OF VIDEOS

We also apply our technique to enhance low quality videos, such as the ones obtained from security and traffic surveillance cameras. In such videos, enhanced context can help answering questions such as: why is a person standing near a part of a building (they are looking at a poster), what is the person's hand hidden by (they are behind a dark object that is not illuminated), what are the reflections in the dark areas (car headlights reflecting from windows of dark buildings), what is a blinking light (traffic light clearly seen at daytime).

The static background, as in the previous section, comes from a single higher-quality daytime snapshot. The dynamic foreground is composed of regions of high variance, both spatial and temporal. Regions of high temporal variance between two video frames are computed by comparing the intensity gradients of corresponding pixels from the two frames.

Videos also present several additional challenges: (a) inter-frame coherence must also be maintained i.e. the weights in successive frames should change smoothly and (b) a pixel from a low quality frame may be important even if the local variance is small (e.g., the area between the headlights and the taillights of a moving car). Our solution is based on the simple observation that in a sequence of video frames, moving objects span approximately the same pixels from head to tail. For example, the front of a moving car covers all the pixels that will be covered by rest of the car in subsequent frames. Using temporal hysteresis, although the body of a car may not show enough intra-frame or inter-frame variance, we maintain the importance weight high in the interval between the head and the tail. The steps are as described in Algorithm 2.

The importance is based on the spatial and temporal

Algorithm 2 Video enhancement

Compute spatial gradients for daytime $G_D = \nabla D$
Smooth video using SUSAN
for each video frame F_i **do**
  Compute spatial gradients $G_{Ni} = \nabla F_i$
  Threshold temporal differences into binary masks M_i
  Create weights W_i using M_i, $|G_D|$ and $|G_{Ni}|$
end for
for each weight image W_i **do**
  Average into W'_i over 2c+1 time steps
end for
for each video frame F_i **do**
  for each pixel (x,y) **do**
    if $W'_{i(x,y)} > 0$ **then**
      Compute mixed gradient field as $G_{(x,y)} = G_{Ni(x,y)}W'_{i(x,y)} + G_{D(x,y)}(1 - W'_{i(x,y)})$
    else
      Set gradient field $G_{(x,y)}$ to the gradient with the greater magnitude between $G_{D(x,y)}$ and $G_{N(x,y)}$
    end if
  end for
  Reconstruct frame F'_i from gradient field G
  Normalize pixel intensities in F'_i to closely match $F_{i(x,y)}W'_{i(x,y)} + D_{(x,y)}(1 - W'_{i(x,y)})$
end for

variation as well as the hysteresis computed at a pixel. A binary mask M_j for each frame F_i is calculated by thresholding the difference with the previous frame, $|F_i$ - $F_{i-1}|$. To maintain temporal coherence, we compute the importance image W_j by averaging the processed binary masks M_k , for frames in the interval k=i-c..i+c. We chose the extent of influence c, to be 5 frames in each direction. Thus, weight due to temporal variation W_i is a mask with values in [0,1] that vary smoothly in space and time. Then for each pixel of each frame, if $W_{i(x,y)}$ is non-zero, we use the method of context enhancement of dynamic scenes i.e., blend the gradients of the nighttime frame and daytime snapshot scaled by $W_{i(x,y)}$ and $(1 - W_{i(x,y)})$. If $W_{i(x,y)}$ is zero, we revert to a special case of the method of enhancement for static scenes i.e., choose the gradient with the larger magnitude. Finally, each frame is individually reconstructed from the mixed gradient field for that frame (See Figure 3).

The input video is noise reduced by using feature-preserving bilateral filtering in three dimensions (space and time). This eliminates false-positives when frame-differences are computed. For a practical implementation we repeatedly applied a 3D SUSAN filter [13] (3x3x5 neighborhood, sigma = 15 and t = 20). The high-quality daytime snapshot used for filling in the context is obtained by median filtering a daytime video clip (about 15 seconds).

Just as in the case of dynamic scenes, a good quality video segmentation or optical flow technique will marginally improve our results. User input can also elasily be incorporated in the process. Since the camera position is static, the user can either designate areas to be filled from the daytime image for all frames, or for each frame separately.

Fig. 3. Enhancing traffic video. A high quality daytime and a low quality nighttime video frame, the importance image obtained after processing, and the final output of our algorithm.

5. DISCUSSION

We introduced a practical method for improving a low-quality nighttime image by combining it with a high-quality daytime scene. This idea appears to be very simple in retrospect. However, despite our search efforts, the idea appears to have been unexplored in image enhancement.

A naïve approach to automatically combining a daytime and nighttime snapshot would be to use a pure pixel substitution method based on some importance measure. This works well only when the inputs are almost identical (e.g. two snapshots of the same scene with different focus [14]). Similarly, blending strategies such as $max_i(I_{i(x,y)})$ or $average_i(I_{i(x,y)})$ also create problems. For example, when combining day-night snapshots, one needs to deal with high variance in daytime snapshots and with mostly low contrast and patches of high contrast in nighttime snapshots. Taking the average simply overwhelms the subtle details in the nighttime snapshot, and presents "ghosting" artifacts around areas that are bright at nighttime. Furthermore, juxtaposing or blending pixels usually leads to visible artifacts (e.g. sudden jumps from dark night pixels to bright day pixels) that distract from the subtle information conveyed in the night snapshots. Figure 2 shows a comparison of our method with averaging pixel values and blending pixel values using an importance function.

We have shown that our algorithm avoids most of the visual artifacts as ghosting, aliasing and haloing. However our method may cause observable color shifts in the results. This phenomenon unfortunately has been a common problem of gradient-based approaches and can be observed in most previous works [8], [2]. There are two major reasons that cause the color shifting. First of all, a valid vector field is not guaranteed to be maintained when modifying it with non-linear operators. The gradient field of the result com-

puted by our method is only an approximation of the desirable one. Secondly, in some cases, it is difficult to maintain the perception of high contrast in the result because the daytime and nighttime snapshots are taken at significantly different exposure times.

A possible extension to our work will be to maintain a valid vector field when computing the gradients of the result. This requires using analytical operators to approximate our non-linear mask and blending function. Separating intrinsic [15] and color images, then applying our algorithm on intrinsic images and fusing them back with the color images could be another possible solution.

6. RESULTS

Our data for video enhancement is from the Washington State Dept. of Transportation website (used by permission). The data for enhancement using the basic algorithm was captured with a Canon PowerShot G3^{TM} camera, placed on a fixed tripod. We show an example of a dynamic outdoor scene combined from a day and a night snapshot (see Figure 1). Notice the dark regions of the nighttime snapshot are filled in by daytime snapshot pixels but with a smooth transition. We also show enhanced videos of traffic cameras (see Figure 3). The camera resolution is 320x240 pixels and it is very difficult to get an idea of the context, especially at nighttime. In our experience, even on a well-organized website, where cameras are labelled and placed on a map, it is still hard to correctly evaluate the traffic situation because architectural features, which are essential for location recognition, cannot be readily discerned.

Processing was done offline as proof of concept and took approximately one second per frame after noise removal. We are working on a faster version of our method that can be applied to enhance traffic camera videos in real time.

7. CONCLUSION

We have presented techniques to extract useful information from multiple snapshots taken using fixed cameras. By providing context to dark or low-quality snapshots or videos, we can create more useful images and easier to interpret surveillance videos. Our methods are suitable for processing low-contrast and noisy inputs while avoiding artifacts present in conventional combining methods such as aliasing, ghosting or haloing.

8. REFERENCES

[1] Gangnet M. Pèrez P. and Blake A., "Poisson Image Editing," in *Proceedings of SIGGRAPH 2003*, 2003, pp. 313–318.

[2] R. Fattal, D. Lischinski, and M. Werman, "Gradient Domain High Dynamic Range Compression," in *Proceedings of SIGGRAPH 2002*. ACM SIGGRAPH, 2002, pp. 249–256.

[3] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic Tone Reproduction for Images," in *Proceedings of SIGGRAPH 2002*. ACM SIGGRAPH, 2002, pp. 267–276.

[4] P. J. Burt and Adelson E. H., "A Multiresolution Spline With Application to Image Mosaics," .

[5] F. Durand and J. Dorsey, "Fast Bilateral Filtering for High-Dynamic-Range Images," in *Proceedings of SIGGRAPH 2002*. ACM SIGGRAPH, 2002, pp. 257–266.

[6] Winder S. Kang S. B., Uyttendaele M. and Szelinski R., "High Dynamic Range Video," in *Proceedings of SIGGRAPH 2003*, 2003, pp. 319–325.

[7] D. Socolinsky and L. Wolff, "A New Visualization Paradigm for Multispectral Imagery and Data Fusion," in *Proceedings of IEEE CVPR*, 1999, pp. 319–324.

[8] G.D. Finlayson, S.D. Hordley, and M.S. Drew, "Removing Shadows from Images," in *Proceedings of ECCV*, 2002, vol. 4, pp. 823–836.

[9] Nayar S.K. and Narasimhan S.G., "Vision in Bad Weather," in *Proceedings of ICCV*, 1999, pp. 820–827.

[10] J.M. DiCarlo and B.A. Wandell, "Rendering High Dynamic Range Images," in *Proceedings of SPIE: Image Sensors*, 2000, vol. 3965, pp. 392–401.

[11] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and Practice of Background Maintenance," in *ICCV*, 1999, pp. 255–261.

[12] W. H. Press, S.A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* , Pearson Education, 1992.

[13] S.M. Smith and J.M. Brady, "SUSAN - a new approach to low level image processing," *Int. Journal of Computer Vision*, vol. 23, no. 1, pp. 45–78, 1997.

[14] P. Haeberli, "A Multifocus Method for Controlling Depth of Field," Availavle at: http://www.sgi.com/grafica/depth/index.html, 1994.

[15] Y. Weiss, "Deriving intrinsic images from image sequences," in *Proceedings of ICCV*, 2001, vol. 2, pp. 68–75.