Time-Lapse Image Fusion

Francisco J. Estrada

University of Toronto at Scarborough,
1265 Military Trail,
M1C 1A4, Toronto, On., Canada
http://www.cs.utoronto.ca/~strider

Abstract. Exposure fusion is a well known technique for blending multiple, differently-exposed images to create a single frame with wider dynamic range. In this paper, we propose a method that applies and extends exposure fusion to blend visual elements from time sequences while preserving interesting structure. We introduce a time-dependent decay into the image blending process that determines the contribution of individual frames based on their relative position in the sequence, and show how this temporal component can be made dependent on visual appearance. Our time-lapse fusion method can simulate on video the kind visual effects that arise in long-exposure photography. It can also create very-long-exposure photographs impossible to capture with current digital sensor technologies.

1 Introduction

The problem of blending visual information from multiple source frames for the purpose of creating high dynamic range (HDR) images has been studied at great depth over the past two decades. The seminal work by Debevec and Malik [1] first proposed a sound framework for recovering a camera's response curve, for using this curve to generate a high-dynamic range radiance map from a set of differently exposed shots of a scene, and for rendering a single image from the radiance map that more faithfully approximates the perceived visual qualities of scenes with extreme illumination variations. Since then, a large volume of research on algorithms and techniques for HDR has been published. Thorough studies of existing techniques can be found in [2], [3], while [4] provides a sample of current research.

HDR techniques are typically complex as they are concerned with photometric accuracy and, in order to produce an image that can be displayed on a typical computer screen, require an additional and often computationally expensive step of tone mapping ([5], [6], [7], [8], [9]) during which the HDR information is appropriately compressed into the available dynamic range (typically that of an 8-bit RGB image).

Recently, exposure fusion ([10] and [11]) has been proposed as a faster, simpler alternative to HDR based on the argument that as long as we are interested only in the final blended image, neither radiance map estimation nor tone mapping

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 441–450, 2012.

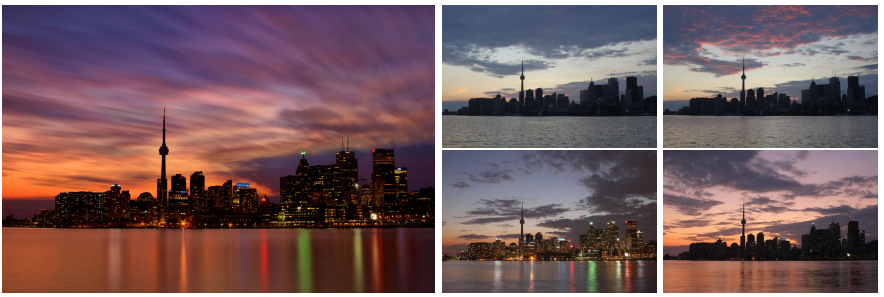

Fig. 1. Time-lapse fusion result on a sequence of 219 photographs taken over a period of 2 hours around sunset. Four of the source frames are also shown. The blended image contains visual elements and detail from all source frames, while softening the appearance of moving scene structure.

are required. Instead, exposure fusion produces an image by blending pixels from all source frames with suitable weighting for each pixel. The weight of each pixel depends on how valuable its visual information is for the final blend. Exposure fusion results in [11] show that it can yield images of comparable visual quality to those created from HDR techniques while greatly simplifying the image processing pipeline.

Until now, HDR and image fusion have been concerned with the blending of multiply-exposed shots to increase the dynamic range of the scene. Here we propose that image fusion can also be used to blend images taken over a (possibly very long) interval of time, blending visual information from a dynamically changing scene while preserving detail and interesting structure. We focus on two problems: Simulating long-exposure effects on motion video, which constrains the exposure time to be no longer than that allowed by the framerate; and the creation of very-long-exposure photographs such as that shown in Fig. 1 which are beyond what is possible to capture in practice with a camera. In what follows, we describe first the original exposure fusion formulation, then we introduce the temporal component that is the basis for time-lapse fusion, and finally show results of the algorithm on video and time-lapse photographic sequences.

2 Exposure Fusion

There is a wide array of existing algorithms that blend multiple images to enhance particular aspects of visual structure. Laplacian pyramid blending [12], for example, demonstrated early on that by properly mixing the coefficients of the Laplacian pyramids of two images, a single, blended frame could be created that preserved sharp detail from both source images. In the interval since the pyramid blending method was introduced, algorithms have been proposed to blend information from images taken at different apertures [13], for fusing multi-spectral data for image enhancement [14], [15], for seamlessly inserting components of one image into another [16], [17], and for removing undesired image content by

replacing it with suitable regions taken from images in a large database [18]. These are just a handful of representative examples. In this paper we explore a specific problem in image blending, namely, that of combining images taken at different points in time for photographic and video applications.

Our method is based on the exposure fusion algorithm of Mertens et al. [11]. The original exposure fusion method provides an alternative to complicated and computationally expensive HDR processing. The exposure fusion method takes as input a set of images $\{I_1, I_2, \ldots, I_K\}$, and generates a single output image by computing a weighted sum of corresponding pixel values over the input image set. The weight of each pixel is based on three properties: Contrast, saturation, and well-exposedness, and it is assumed that the images are aligned.

Contrast computation is based on local edge energy provided by a Laplacian pyramid decomposition of the source images. Saturation is defined as the standard deviation of the RGB colour components of each pixel, and well-exposedness gives higher weights to pixels away from brightness extremes. Given these components, pixel weights are computed as:

$$W_{i,j,k} = C_{i,j,k}^{\alpha_c} + S_{i,j,k}^{\alpha_s} + E_{i,j,k}^{\alpha_e}, \tag{1}$$

where the indices i, j, k refer to pixel (i, j) in image k, $C_{i,j,k}$, $S_{i,j,k}$, and $E_{i,j,k}$ are the pixel's contrast, saturation, and well-exposedness values respectively, and the α exponents control the influence of each of these terms. Pixel weights are normalized to that the weight of pixels at (i, j) across all frames sums to 1: $\hat{W}_{i,j,k} = \frac{W_{i,j,k}}{\sum_{k'=1}^{K} W_{i,j,k'}}$. Given the pixel weights for all pixels in all source frames, the simplest exposure fusion algorithm would compute the colour of the final blended pixel $R(i, j)$ as:

$$\hat{R}_{i,j} = \sum_{k=1}^{K} \hat{W}_{i,j,k} I_{i,j,k}. \tag{2}$$

However, this can lead to artifacts around image edges and other fine structures due to high frequency variations in the weight maps themselves. To avoid this problem, Eq. 2 is actually implemented using pyramid blending similar to that described in [12]. Each input image is processed to obtain a Laplacian $L\{I_k\}$ pyramid with d levels. The corresponding weight map is processed to obtain a Gaussian pyramid $G\{\hat{W}_k\}$ with the same number of levels d. From these Laplacian and Gaussian pyramids, each level $l = \{d, d-1, \ldots, 1\}$ of the blended Laplacian pyramid for the result frame is computed as

$$L_l\{R_{i,j}\} = \sum_{k=1}^{K} G_l\{\hat{W}_{i,j,k}\} L_l\{I_{i,j,k}\}. \tag{3}$$

The result frame R is finally obtained by pyramid reconstruction from $L\{R\}$.

Exposure fusion was shown to produce blended images of a visual quality similar to those produced by more complicated HDR/tone-mapping algorithms. Given its relative simplicity, it has been adopted in applications such as panoramic imaging, through an open source project called *enfuse*. It should be

noted that the exposure fusion algorithm in no way constrains the source images to be taken under different exposure settings. In fact, if the parameters of the method are set to ignore the well-exposedness term, standard exposure fusion can be used to blend any set of aligned photographs of the same scene.

However, for video applications such as video processing we would like to be able to control the contribution of each pixel not only through its photometric properties, but also through its relative position within the sequence of frames. To this end, we expand the original formulation of exposure fusion to include a temporal decay term. This time decay can be applied in a per-frame manner to achieve visual effects similar to those in long-exposure photography, or it can be made structure dependent so that the specific image structure is highlighted in the blended frames.

3 Time-Lapse Fusion

Exposure fusion provides a way to extend the dynamic range of still photographs. We are interested in extending the temporal range of the events that can be recorded in photographs and video. There are two factors that limit exposure length: First, the need for short exposure times under bright illumination conditions. This applies to both still photography and video, and though it can be controlled to a certain degree through the use of appropriate filters and small apertures, it is still technically not feasible to obtain exposure times in the order of tens of minutes or even hours under daylight. Secondly, for motion video, the longest exposure achievable using a video camera is limited by the frame rate. This is an unavoidable limitation of video processing and places a hard upper-bound on the maximum exposure time for each frame.

Here we introduce time-lapse fusion as a means for artificially extending the exposure time of both still photography and motion video. The process takes as input a sequence of frames $I_1, \ldots, L_K$ either from a video, or from a time-lapse photographic sequence in which photographs are taken at regular intervals over an arbitrarily long period. We assume the sequence is taken with the camera mounted on a sturdy tripod so that static structure in the scene remains aligned.

Time-lapse fusion produces an output sequence $R_1 \ldots R_K$ in which image R_t is produced by blending source frames $I_t, I_{t-1}, \ldots, I_{t-\tau}$ effectively incorporating visual components from τ previous frames as well as the current one. The value of τ is set by the user, and controls the length of the virtual exposure time for each result frame. For motion video the virtual exposure time (VET) of each frame is given by $VET = \tau/\text{framerate}$. For time-lapse sequences, the virtual exposure time is dependent on the interval between shots. At first sight, it may seem reasonable to simply average frames $I_t, I_{t-1}, \ldots, I_{t-\tau}$ into the result frame; however, averaging over time tends to weaken and blur fast-moving structure, losing detail and giving a large weight to uniform backgrounds as seen in Fig. 2. Careful blending is required to preserve detail and produce the desired long-exposure effects.

We modify the pixel weight equation 1 to include a time-decay factor that modulates the overall contribution of pixels to the final blended image R_t:

$$W_{i,j,k} = (C_{i,j,k}^{\alpha_c} + S_{i,j,k}^{\alpha_s} + E_{i,j,k}^{\alpha_e})T(k,t). \tag{4}$$

Here, $T(k,t)$ is a Gaussian-shaped envelope with $\sigma = \tau/3$

$$T(k,t) = e^{-\frac{(t-k)^2}{2*\sigma^2}}, \tag{5}$$

with t the index of the current frame. The motivation for a Gaussian envelope is that it provides smooth transitions between consecutive output frames. A box window would be simpler, but would result in visible changes between frames especially for time-lapse sequences with large intervals between shots.

Figure 2 shows the output of time-lapse fusion for two frames from a fireworks sequence. The output frames clearly show the effects of having a long virtual exposure. Structure and fine detail are well preserved, and unlike frame averaging, time-lapse fusion does not create a blurry output that blends heavily with the black background. Figure 3 shows another example of time-lapse fusion with different virtual exposure times.

Fig. 2. Left: Original input frames from a movie sequence of fireworks. Middle: Results of averaging each frame with the previous 25 in the sequence. Right: Time-lapse fusion results with $\tau = 25$ for a virtual exposure time of .83 seconds. Averaging generates blur and blends structure with the background. Time-lapse fusion produces a pleasing long-exposure effect and preserves sharp detail.

For very-long-exposure photography, we follow the same process with the exception that only the last output frame R_K is computed, and we set τ to be very large so that effectively the entire sequence contributes to the final frame. The blended image in Figure 1 was produced in this way for a virtual exposure time of about 2 hours. Very long exposures are thus possible even under bright daylight conditions. During the process, the photographer retains full control

over focus, aperture, and hence depth of field. With careful implementation, it is possible to process pixel weights frame-by-frame, making arbitrarily long exposures possible. Figure 4, for example, shows long exposure results from long time-lapse sequences, including a portrait produced by blending photos of the same person taken over a period of 1 year.

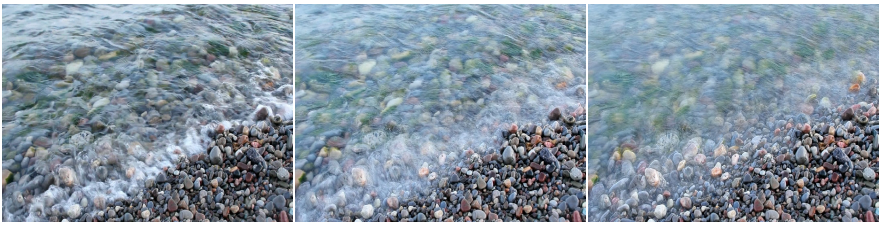

Fig. 3. Left: Original input frame from a movie sequence of waves. Middle: Time-lapse fusion results for $\tau = 7$ (virtual exposure time .23 seconds). Left: Time-lapse fusion with $\tau = 25$ (virtual exposure time .83 seconds).

While the use of Gaussian envelope is appropriate for simulating long exposure effects, the time decay factor can be an arbitrary function of relative frame position. In particular, structure-dependent envelopes that highlight specific image content are possible. In the following section we show how to make time-lapse fusion structure dependent, so that the decay time varies spatially over individual frames.

4 Stretch the Sunset

In the previous section we explored the use of a time decay factor that applies to entire frames in a sequence. However, the time decay factor can vary spatially over an image providing control at the pixel level of how long individual pixels will contribute toward the resulting blend. In this section we show that through the use of a structure-dependent time decay, computed independently for each pixel in each frame of the sequence, we can produce images that highlight specific image structure. The example we use is that of extending the duration of sunset colours over a time-lapse sequence taken at dusk.

Under this scheme, a time-decay map $D_{i,j,k}$ is computed for each frame k in the sequence. This map contains at each pixel location (i, j) a constant that controls how quickly the influence of the pixel will decay over time. To maintain consistency with the frame-wise time decay discussed in the previous section, we also model each pixel's temporal weight as a Gaussian envelope:

$$T(i,j,k,t) = e^{-\frac{(t-k)^2}{2*D^2_{i,j,k}}}, \tag{6}$$

Fig. 4. Very-long exposure photography. The portrait on the left is the result of blending images of the same person taken over a period of 1 year. The subject's face is carefully aligned in each source frame but the clothing, hair style, and background are otherwise unconstrained (source images from **clickflashwhirr.me**). Right side: Additional examples of very-long exposures from time-lapse sequences. Virtual exposure times range in the tens to hundreds of minutes.

where t is the current time index, k is the frame's time index, $D_{i,j,k} \in [\sigma_{min}, \sigma_{max}]$ is the decay constant for the pixel, and the parameters σ_{min} and σ_{max} determine the minimum and maximum intervals, respectively, over which pixels will contribute to the blended result.

In order to understand what Eq. 6 is doing, we need to specify how the time decay map $D_{i,j,k}$ is computed in the first place. This will depend on the input sequence and what the user wishes to accomplish through the blending. As a concrete example, here we use the same sequence of 219 images from which Figure 1 was generated. The sequence includes an interval covering about 2 hours around sunset.

We will generate an output sequence through time-lapse fusion so that the duration of the sunset colours is stretched. We also want some temporal smoothing over all other elements of the scene so that the different frames, which were taken about 30 seconds apart from each other, blend seamlessly. To detect sunset colours we rely on a simple measure of saturation. The time decay value for each pixel is given by: $D_{i,j,k} = \sigma_{min} + (S_{i,j,k} * (\sigma_{max} - \sigma_{min}))$, where $S_{i,j,k}$ is the saturation for the pixel in $[0, 1]$, and we set $\sigma_{min} = 7$ and $\sigma_{max} = 500$ which in effect means high-saturation pixels contribute all the way through the end of the sequence.

Figure 5 shows four frames from the sequence as well as their contribution over time as given by Eq. 6. Here t stands for the age of the frame, so the

Fig. 5. The leftmost column shows four frames of the input sequence. Each successive column then shows the temporal blending weights for pixels in the corresponding image depending on the age of the frame. For example, the third column shows the blending weights for pixels once the image is 100 frames old.

images show how strongly different pixels contribute to a blended image as the frame ages. By the time a frame is $3 * \sigma_{min}$ frames old, the contribution of non-saturated pixels has faded to almost zero. Saturated pixels have large temporal blending weights all through the sequence. Note that for each frame the regions that persist change in shape, and that indeed the regions that contain sunset-like colours fade the slowest.

Figure 6 shows the results of structure-dependent time-lapse fusion compared against the standard time-lapse fusion described in the previous section (i.e. using a single time decay value per frame, with a fixed $\tau = 25$). Results are similar for the earlier parts of the sequence; however, as the sequence moves along the standard time-lapse fusion method loses the sunset colours in the clouds. The structure-dependent version, on the other hand, preserves the red glow in the clouds and the colours reflected in the water for much longer. The resulting sequence does indeed produce the illusion of a longer sunset full of reds and oranges.

A final example is shown in Fig. 7 which shows selective smoothing of waterfall structure. In this case standard time-lapse fusion causes blurring of people walking through the sequence which is not desirable. A simple threshold on distance from pure white was used to automatically obtain a coarse segmentation of the waterfall structure in each frame. Waterfall regions are given $\sigma_{max} = 11$ for a virtual exposure time of 1.16 seconds. The rest of the scene is assigned $\sigma_{min} = .33$ so that non-waterfall pixels contribute only while the frame is the current frame. This effectively renders the moving water smooth while leaving the rest of the scene untouched. In the above examples no user interaction was required.

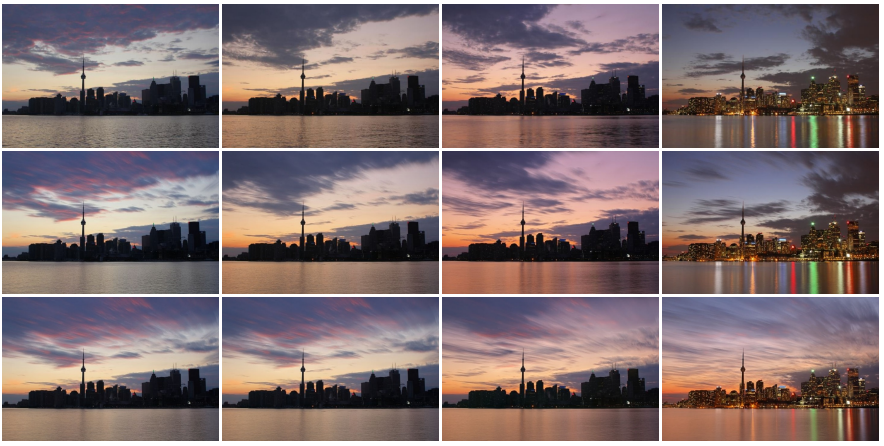

Fig. 6. Stretched sunset sequence. The top row shows four of the original sequence frames. The middle row shows the result of applying standard time-lapse fusion (as in Section 3). The bottom row shows blended results using the structure dependent decay map described in this Section.

Fig. 7. Selective smoothing of water. The left column shows two frames from motion video containing waterfalls. The middle column shows standard time-lapse fusion results which blur everything in the scene. The right column shows structure-dependent fusion which selectively smooths over the waterfall regions leaving the rest of the scene unchanged.

5 Conclusion

In this paper we proposed a time-lapse fusion algorithm that extends the exposure fusion framework to blend images taken at different points in time. We showed that a simple temporal decay factor with a Gaussian profile can be used to simulate long exposure effects on video, and that arbitrarily long exposures not physically realizable with a real camera are possible while allowing the photographer to maintain control of focus, aperture, and depth of field. We then extended the method to allow for pixel-level control of time-decay so that specific image

content can be highlighted in the final blend. At the pixel level, time-lapse fusion allows for the creation of video sequences that change the temporal profile of events in the scene to produce visually striking results. Time-lapse fusion can, in this way, provide photographers with a tool to expand their ability to create depictions of the world that are beyond physical or practical limitations.

References

1. Debevec, P., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: SIGGRAPH, pp. 369–378 (1997)
2. Bloch, C.: The HDRI Handbook: High Dynamic Range Imaging for Photographers and CG Artists. Rocky Nook (2007)
3. Hoefflinger, B. (ed.): High-Dynamic-Range (HDR) Vision. Springer (2005)
4. Hasinoff, S., Durand, F., Freeman, W.: Noise-optimal capture for high dynamic range photography. In: CVPR, pp. 553–560 (2010)
5. Reinhard, E., Devlin, K.: Dynamic range reduction inspired by photoreceptors. IEEE Transactions on Visualization and Computer Graphics 11, 13–24 (2005)
6. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. ACM Transactions on Graphics 21, 257–266 (2002)
7. Fattal, R., Lischinski, D., Werman, M.: Gradient domain high dynamic range compression. ACM Transactions on Graphics 21, 249–256 (2002)
8. Mantiuk, R., Myszkowski, K., Seidel, H.: A perceptual framework for contrast processing of high dynamic range images. ACM Transactions on Applied Perception 3, 286–308 (2006)
9. Kim, M., Kautz, J.: Characterization for high dynamic range imaging. Eurographics 27, 691–697 (2008)
10. Goshtasby, A.: Fusion of multi-exposure images. Image and Vision Computing 23, 611–618 (2005)
11. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: Pacific Conference on Computer Graphics and Applications, pp. 382–390 (2007)
12. Burt, P., Adelson, E.: A multi-resolution spline with application to image mosaics. ACM Transactions on Graphics 2, 217–236 (1983)
13. Hassinof, S., Kutulakos, K.: A layer-based restoration framework for variable apperture photography. In: ICCV, pp. 1–8 (2007)
14. Schaul, L., Fredembach, C., Süsstrunk, S.: Color image dehazing using the near-infrared. In: ICIP, pp. 1629–1632 (2009)
15. Bennet, E., Mason, J., McMillan, L.: Multispectral bilateral video fusion. IEEE Transactions on Image Processing 16, 1185–1194 (2007)
16. Kwatra, V., Schödl, A., Essa, I., Turk, G., Bobick, A.: Graphcut textures: Image and video synthesis using graph cuts. In: SIGGRAPH, pp. 277–286 (2003)
17. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: SIGGRAPH, pp. 313–318 (2003)
18. Hays, J., Efros, A.: Scene completion using millions of photographs. In: SIGGRAPH (2007)