GRADIENT BASED MULTIFOCUS VIDEO IMAGE FUSION

Stephen Frechette

Riverside Research Institute
Lexington, MA USA
sfrechet@comcast.net

V.K. Ingle

Northeastern University
Boston, MA USA
vingle@ece.neu.edu

ABSTRACT

Optics of lenses with a high degree of magnification suffer from the problem of a limited depth of field. The larger the focal length and magnification of the lens the smaller the depth of field becomes. As a result, fewer objects in the image are in focus. Multifocus digital image fusion attempts to increase the apparent depth of field through the fusion of object within several different fields of focus. In this work a novel multifocus video image fusion algorithm is proposed. Our unique approach requires orders of magnitude fewer calculations than all other known multifocus video image fusion algorithms.

1. INTRODUCTION

This paper examines a method specifically designed for the fusion of hundreds of frames of multifocus video. Our method requires that during video capture all the objects in the video, which are to be fused, are brought in and out of focus by moving the field of focus through the objects. Multiple video frames with different objects in focus are fused together to create one image. The goal of multifocus image fusion is to create a fused image with the maximum number of objects in focus.

This paper presents a unique multifocus video image fusion algorithm that achieves a high degree of parallelism as well as a scalability only limited by the memory size. Additionally, from a field of n frames of multifocus video our gradient based approach only requires $O(n)$ calculations for the selection of several video frames that contain the highest number of objects in focus.

Our unique method processes the information contained in the gradient of the pixels' intensity as the field of focus moves through an object, in addition to the information contained exclusively within a given image or video frame. This novel approach to multifocus video image fusion represents our contribution to the field.

2. PREVIOUS WORK

Many multifocus image fusion methods exist. These methods include: block-by-block, the fusion of image pyramids or discrete wavelet transforms, object-by-object image fusion using discrete wavelet coefficients, and pixel-by-pixel image fusion. All of these methods have been examined in previous literature. There is a consensus that the latter two methods perform better than methods that exclusively employ the image pyramids and discrete wavelet transforms [1], [2].

Multifocus image fusion can be performed in the domain of the image transform, the transformed images are then fused, finally an inverse transform results in a fused image. Additionally, block-by-block, and object-by-object image fusion may use an image transform as a metric.

An image pyramid is a collection of images of decreasing resolution in which the dimensions equal the horizontal cross section of a pyramid [3]. The base of the pyramid consists of the largest size image. The image resolution decreases as the levels are traversed upward.

The Laplacian pyramid is an image transform that effectively divides an image into subbands. The Laplacian image pyramid was first proposed in [4] and requires many fewer calculations compared to a bandpass filter bank structure. This reduction in the number of required calculations makes the Laplacian pyramid a popular transform, though object-by-object or pixel-by-pixel image fusion results in a fused image of higher quality.

A method described in [5] details the pixel-by-pixel fusion of two images through the examination of the Laplacian pyramid's nodes. This method uses the Laplacian pyramid as a transform and fuses the original images pixel-by-pixel rather than node-by-node. Pixel-by-pixel fusion uses an image transform as a metric. The first step of this approach is to transform the input images into Laplacian pyramids. Next, instead of creating a fused Laplacian pyramid, this algorithm selects pixels for the final fused image by examining the nodes of the pyramid that comprise the input image's Laplacian pyramid. The pixel that corresponds to the nodes that contain the most energy are

selected for the final image. This algorithm is detailed in [6] and is referred to as the maximal absolute value function.

Similar to the Laplacian pyramid, the 2-D fast wavelet transform is also a pyramidal algorithm. In addition to exhibiting the same computational efficiency as the Laplacian pyramid algorithm, the 2-D fast wavelet transform has an optimality of storage space and also represents orientation in the frequency domain [7].

In terms of the objective criterion of the quality of the final fused image, the best method for multifocus image fusion is the wavelet based object-by-object image fusion method [1]. The algorithm presented in [1], was employed to create a library of Matlab functions for object-by-object wavelet based image fusion [12].

A current generation multifocus video image fusion system [13], was released in 2001. This hardware offers six main image processing applications, one of which is image fusion. A Laplacian pyramid is used in the image fusion algorithm [15]. This board is limited to fusing two 720×480 pixel video images with a throughput of 30 frames/sec and a latency of two frames [14].

3. OVERVIEW

Our novel multifocus video image fusion algorithm is implemented in two phases. Phase I results in a fused image of reasonable quality, given the small number of calculations necessary to produce the image. Phase II results in an improvement of the final image through the use of wavelet transforms, but requires several orders of magnitude more calculations as compared to the exclusive use of Phase I.

3.1. Out of Focus Model

The blurring of an object that is out of the field of focus can be modeled by a Gaussian low pass filter. A Gaussian smoothing function that simulates the effects of blurring is given in 1.

$$f(x,y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{1}$$

where σ is the standard deviation of a Gaussian function.

A more accurate transfer function that estimates the effects of blurring is given in [8], but our algorithms do not require this degree of accuracy.

3.2. Focal length requirement

Our video image fusion algorithm requires that the objects to be fused are brought in and out of focus by varying the field of focus. Therefore the focal length of the video camera must increase incrementally in order to record the objects to be fused in various degrees of focus. Objects within the field of focus are considered to be in focus.

As the field of focus moves outward away from the camera, objects will first appear out of focus, then in focus, then out of focus again. To capture objects in various degrees of focus, the field of focus is moved toward the object, then through the object, then away from it.

To examine the effectiveness of each phase of our multifocus video image fusion method, we demonstrate the fusion of four distinct objects. Due to the requirements that the focal length must increase incrementally during video image capture, experimental data is used to illustrate the effectiveness of our method. The experiment consists of five seconds of video recorded at 30 $\frac{frames}{sec}$ with a frame size of 720×480 pixels. In the first frame of the video all objects in the frame are out of focus, then in sequence ordered by the object's distance from the camera, the objects move in and then out of the field of focus as the focal length of the camera increases.

The video capture of the experimental data is detailed as follows. In Fig. 2, the book on the right is approximately one third of a meter from the camera, the book on the left is 1M from the camera, the book in the center is 2M from the camera, and the paper in the top left corner is on the wall 4M from the camera. Although, with the correct aperture, it would be possible to record an image with all the books in Fig. 2 in focus, if an image is captured from a large distance that requires the use of a telephoto lens and the books were placed further apart, this would not be true. Our experiment uses restrictive camera settings that allow, at most, only one book to be in focus in any given frame of the video.

4. PHASE I: PIXEL-BY-PIXEL IMAGE FUSION

The input to Phase I is a video clip in which all objects are initially out of focus, then the field of focus moves sequentially through each object until all objects are out of focus again. The algorithm first places each pixel of each frame of video into N one-dimensional arrays represented by , the pixel intensity, where N is the number of pixels per frame in the video. The gradient of the array $\hat{x}$ is represented by (2) and calculated in (3).

$$\bigtriangledown\hat{x} \equiv \frac{d\hat{x}}{dx} \tag{2}$$

$$\frac{d\hat{x}}{dx} = \hat{x}(x+1) - \hat{x}(x) \tag{3}$$

The peak value of that gradient is determined, and the frame of video that corresponds to the peak value of that pixel's gradient, is estimated to correspond to an object that is within the field of focus at that frame. Therefore,

it is estimated that the pixel is within an object that is in focus, or the pixel in that frame is more in focus than the other frames. The initial Phase I calculations result in the pixel-by-pixel fusion of various video frames, with different books in focus. The Phase I algorithms follows in Fig. 1.

Phase I

INPUT: $video(x, y)(frame)$

OUTPUT: $f(x, y)$ **AND**

$peak\_gradient\_frame\_index(x, y)$

1 N = |video frames|

2 $\hat{x}(x, y) = video(x, y)(1 : N)$

3 $peak\_gradient\_frame\_index(x, y) = Max\{\bigtriangledown \hat{x}(x, y)\}$

4 $f(x, y) = video(x, y)(peak\_gradient\_frame\_index(x, y))$

Fig. 1. Phase I algorithm.

The use of only this preliminary gradient calculation, Phase I, results in Fig. 2.

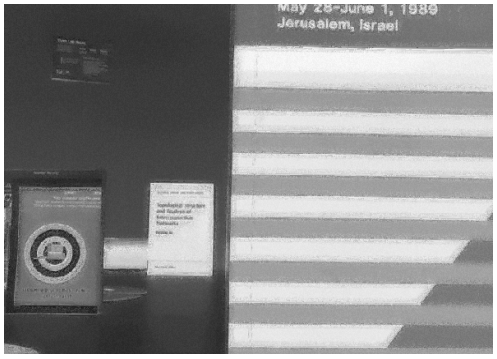

Fig. 2. Pixel-by-pixel image fusion using only the metric of divergence towards the peak of the one-dimensional gradient, as an object's pixels are brought in and out of focus because the field of focus moves away from the camera.

The image restoration of the Phase I image results in an improvement of the Fig. 2 image.

4.1. Restoration of PHASE I image

The image quality of Fig. 2 can be further improved through the use of an image restoration technique. A median filter is employed to improve the quality of Fig. 2. A 720×480

Restoration of Phase I

INPUT: $peak\_gradient\_frame\_index(x, y)$ **AND**

$video(x, y)(frame)$

OUTPUT: $\hat{f}(x, y)$

1 $median\_peak\_frame\_index(x, y) =$
$\underset{(s,t)\in S_{xy}}{\text{Median}} \left\{ peak\_gradient\_frame\_index(s, t) \right\}$

2 $\hat{f}(x, y) = video(x, y)(median\_peak\_frame\_index(x, y))$

Fig. 3. Phase I image restoration algorithm.

matrix of the frame indices corresponding to the selected in focus pixel is median filtered with a window size of 8×8 pixels. The median filter removes the isolated frame index elements that differ greatly from their neighbors. The restoration algorithm is given in Fig. 3, where S_{xy} represents the set of coordinates in a square subimage window of size 8×8 and centered at point (x, y).

The image resulting from the restoration of the initial Phase I image is illustrated in Fig. 4. The number of computations necessary to produce Fig. 4 is given by (4).

$$2nN + \frac{64 \times 65}{2} N \tag{4}$$

where n is the number of frames in the video and N is the number of pixels per frame. This median filter operation upon the frame indices results in a smoothing effect that improves the subjective image quality to the degree that the letters printed on the objects are clearly legible.

5. PHASE II: BLOCK-BY-BLOCK IMAGE FUSION

Further processing of the multifocus video results in a fused image of significantly better quality. This additional processing is referred to as Phase II and contains two steps. Phase II employs block-by-block image fusion, as appose to pixel-by-pixel image fusion employed in Phase I. The discrete wavelet transform, which requires several orders of magnitude more calculations than Phase I alone, is employed as a metric to determine the degree in which a given tile is in focus.

5.1. PHASE II Step 1: Creation of the focused frame array

In the first step, the matrix of frame indices is tiled at a tile size of 40×40 pixels. This results in 30 tiles which compose the 720×480 pixel image. The mode of each tile is determined by a histogram, with the number of bins equal to the number of the frames in the video. The mode of the

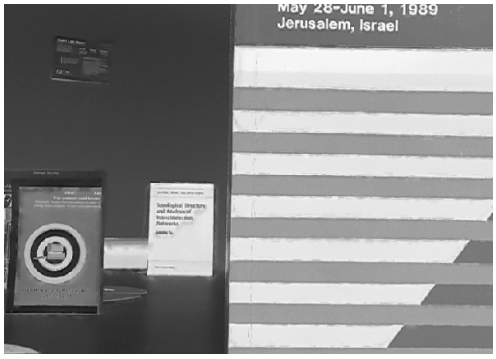

Fig. 4. A median filter of window size 8x8 pixels upon a matrix of the frame indices, the image is composed of pixels from various frames of the video in which the difference of the pixel intensity peaked as the field of focus moves toward, then through, then away from the objects.

Phase II step 1

INPUT: $peak\_gradient\_frame\_index(x, y)$

OUTPUT: $focused\_frame\_array$

1 $focused\_frame\_array =$
$\underset{(i,j)\in S_{ij}}{\text{Mode}} \left\{ peak\_gradient\_frame\_index(i, j) \right\}$

Fig. 5. Phase II step 1 algorithm.

histogram is the largest number of pixels that contain the same value, i.e., the peak of the histogram. The histogram represents a probability mass function.

The mode of each tile is the frame index that is most common within the given tile. The mode of each of the 30 tiles, if unique, is noted and saved. Note that, this is the mode of a tile of frame indices, not the mode of pixel intensities, which is never calculated.

The result of Phase II step 1 is a set of frame indices, up to 30 elements, that represent the mode of one of the 30 tiles. This set of frame indices is referred to as the $focused\_frame\_array$.

The Phase II step 1 algorithm is given in Fig. 5, in which S_{ij} represents the set of coordinates in a square tile of size 40×40, the top left coordinate of the tile is point $(i \cdot 40, j \cdot 40)$. In this algorithm the mode for an entire tile is calculated once, the algorithm does not calculate the mode of a subimage window for each pixel.

5.2. PHASE II Step 2: Creation of the final fused image using the $focuse\_frame\_array$

In step two, the 2-D discrete Coiflet wavelet function is employed as a metric to determine the degree to which a tile in a given frame of the video is in focus. The $focuse\_frame\_array$ has an upper bound of 30 frame indices. The 2-D discrete Coiflet transform is applied to each frame in the $focuse\_frame\_array$. Next, video frame indexed in the $focuse\_frame\_array$ is tiled, i.e., the frame is partitioned into 30 tiles, or blocks, for the purpose of block-by-block image fusion. For each tile location, the 2-D discrete Coiflet wavelet transform is employed to determine, for the given tile location, which frame in the $focuse\_frame\_array$ is the most in focus. Each of the 30 tiles which is the most in focus is selected and, along with the other selected in focus tiles, create Fig. 7. This block-by-block method of image fusion is explained in more detail in the next subsection.

The next subsection describes the use of the Coiflet wavelet transform as a metric to aid in the determination of the degree in which a given tile is in focus.

5.3. Coiflet wavelet function

Through the use of the Coiflet wavelet transform the amount of energy in the high frequency region is measured and used to determine the degree to which a set of pixels are in focus. Different types of wavelet transforms contain different wavelet functions, these wavelet functions are dilated or compressed to form the wavelets. These differing wavelets also contain different frequencies. As a result, some wavelet classes better transform a signal into high frequency components. For image fusion applications, it is required to choose a wavelet function whose corresponding wavelet vectors as compared to other wavelet vectors, results in a difference equation that exhibits excellent transmission characteristic and above average frequency selectivity.

Through the use of a discrete wavelet function, the energy in high frequency components of the signal is represented by the detail coefficient and mapped in the spatial domain. This detail coefficient is the amplitude of the wavelet(s).

Several types of wavelet transforms can be used to represent images using fewer bits than the original image. The discrete wavelet transform (DWT) transforms an image into a sum of scaling functions with corresponding scaling coefficients, and wavelets with corresponding detail coefficients. Wavelet transforms out perform the Laplacian pyramid for the application of image compression [7].

The Coiflet wavelet function has above average transmission characteristics which leads to reasonably sharp frequency selectivity and an acceptable phase response [9]. We employ the Coiflet wavelet function of order 2. The

highest frequency Coiflet wavelet will only appear in objects that are in focus.

For certain types of wavelets the energy of each wavelet will be constant. This results in a constant Q factor of the frequency response of each wavelet. Also, the Coiflet wavelet transform may be implemented using a fast transform algorithm. As a result wavelets provide a fast transform that approximately represent the energy contained within each subband. These subbands are mapped to spatial locations and represent the strength of the signal at that frequency. The runtime of the Fast Wavelet Transform on a one-dimensional array is $O(N)$ [10], and can be implemented on a custom VLSI chip [11].

The energy contained in the high frequency subband of the Coiflet wavelet transform is calculated for each tile of the video indexed in the focused_frame_array. The out of focus objects will not contain high frequencies. For the purposes of multifocus image fusion, it is desirable for the wavelet transform to contain a large percentage of the total energy in the high frequency wavelets. The absolute value of the wavelet coefficient of a DWT was employed as a metric for multifocus image fusion in [16].

The detail coefficients of the highest frequency subband of the Coiflet wavelet transform are summed. The tile that contains the most energy in this subband is selected for the final fused image shown in Fig. 7. The discrete Coiflet wavelet transform upon image I is given:

$$[D, A] = DWT(I) \tag{5}$$

The detail coefficients of the discrete Coiflet wavelet transform is denoted by $D_{(HH)}$ where (HH) represent highest frequency subband. The approximation coefficients of the wavelet transform, A , are not processed by our algorithms.

The Phase II step 2 algorithm is given in Fig. 6, in which S_{ij} represents a tile of size 40×40 pixels within a frame of the video indexed in the $focus\_frame\_array$.

The blocking effect, due to the use of tiling, can be reduced if the Phase II step 2 algorithm sums the wavelet coefficients in a window based pixel-by-pixel method, or within each object for region based object-by-object image fusion.

6. FUTURE WORK

Through the use of a 30 $\frac{frames}{sec}$ video camera this method of video image fusion fuses a sequence of video frames that contain objects which move in then out of the field of focus. The result is one fused image, i.e., one frame. The current limitation is the rate at which the focal length of the camera can oscillate. If a final result of 30 $\frac{frames}{sec}$ video is desired, then a means to vary a high speed camera's focal length at a high rate will have to be developed.

Phase II step 2

INPUT: $peak\_gradient\_frame\_index(x, y)$ **AND**

$video(x, y)(frame)$

OUTPUT: $f_{final}(x, y)$

1 FOREACH Element in $focused\_frame\_array$ is $focused\_frame$

2 $D_{HH}(x, y)focused\_frame$ =
Upsampled $DWT\{video(focused\_frame)\}$

3 $focused\_frame\_score(i, j)(focused\_frame)$ =
$\left| \sum_{(i,j)\in S_{ij}} D_{(HH)}(i, j)(focused\_frame) \right|$

4 END

5 $f_{final}(x, y) = video(i, j)\left(\underset{(i,j)\in S_{xy}}{\text{Max}} \left\{ focused\_frame\_score(i, j)(focused\_frame\_array) \right\} \right)$

Fig. 6. Phase II step 2 algorithm.

Future work could include the investigation of the effectiveness of various image restoration techniques, such as an adaptive median filter upon the image produced by the gradient method, e.g., Fig. 2.

7. CONCLUSION

Our method is best suited for applications in video surveillance that require the use of a telephoto lens, i.e., video surveillance at a distance that requires magnification. Potential uses of multifocus video image fusion include applications in video surveillance such as event detection, facial recognition, and target recognition.

For the special case of video image fusion, where the field of focus is varied and objects are brought in focus then out of focus, our method requires only subtraction, comparison, and a median operation to produce Phase I's image, Fig. 2. The image that results from Phase I is of low quality, but this may be acceptable in applications where the image quality may be compromised for computational efficiency.

For the fusion of hundreds of frames, in which the field of focus moves through various objects, our Phase I method outperforms all others in terms of the number of computations required, because our method does not require the transform of a frame, e.g., a Laplacian pyramid or discrete wavelet transform. Our pixel-by-pixel image fusion is performed without any transformations.

Although Phase I results in some blur in the final image, important features of the sample image, e.g. letters and edges, contain less blur in the final fused image than other

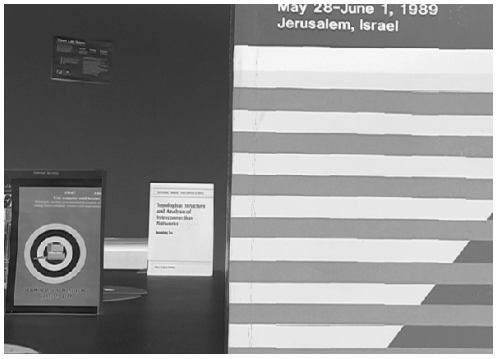

Fig. 7. Final results of the tile-based image fusion function.

regions that are composed of mostly lower frequencies.

In Phase II the output of the gradient method is further processed to produce the $focused\_frame\_array$. Through the pruning of all the frames of video down to one frame per tile, i.e. 30, the number of frames which are further processed is significantly reduced to only the frames that the gradient method indicates contain objects in focus.

For applications in video surveillance that require a telephoto lens, an increase in the apparent depth of field could be highly desirable. Our unique approach to increasing the apparent depth of field requires several orders of magnitude fewer calculations than any other multifocus video image fusion method.

8. REFERENCES

[1] Z. Zhong, "Investigations of Image Fusion," overview of Ph.D. dissertation, Electrical Engineering and Computer Science Department, Lehigh University, Bethlehem, PA, Sept. 2002. [Online]. Available: http://www.ece.lehigh.edu/SPCRL/IF/image_fusion.htm

[2] R. Blum, "Image Fusion with Some Emphasis on CWD," U.S Army Research Office Electrical Engineering and Computer Science Department Lehigh University, Bethlehem, PA., Tech. Rep. DAAD19-00-1-0431

[3] R.C. Gonzalez and R.E. Woods, *Digital Image Processing.*, 3rd. ed. Upper Saddle River, NJ.: Prentice Hall, pp. 1257, 2002.

[4] P. Burt and E. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. on Communications*, vol. COM-31, no. 4, Apr. 1983.

[5] S. Chang, "Multi-focused Image Fusion," August 2001.

[6] L. Bogoni and M. Hansen, "Pattern-Selective Color Image Fusion," *10th Intl Conf. on Image Analysis and Processing*, 1999.

[7] L. Prasad and S.S. Iyengar, *Wavelet Analysis with Applications to Image Processing.*, New York, NY: CRC Press, pp. 224-231, 1997.

[8] Y. Xiong and A. Shafer, (1993 Mar.). "Depth from Focusing and Defocusing," The Robotics Institute Carnegie Mellon University, PA, Tech. Rep. CMU-RI-TR-93-07 [Online]. Available: http://www.ri.cmu.edu/pubs/pub_298.html

[9] S. Fu, B. Muralikrishnan, and J. Raja, "Engineering Surface Analysis With Different Wavelet Bases," *J. of Manufacturing Science and Engineering*, vol. 125, no. 4, pp. 844-852, Nov 2003.

[10] L. Ward, "Introduction to Wavelets and their Applications," Lecture, Department of Mathematics Harvey Mudd College, 2000. [Online]. Available: http://www.math.hmc.edu/faculty/ward/wavelets/

[11] A. Abbate, C.M. DeCusatis and P. Das, *Wavelets and Subbands Fundamentals and Applications.*, Boston: Birkhauser, pp. 101-503, 2002.

[12] S. Frechette, "A Multifocus Image Fusion Algorithm for Parallel Image Processing," M.S. thesis, Dept. Elect. and Comp. Eng., Northeastern Univ., Boston, MA, 2003.

[13] "Acadia I PCI Vision Accelerator Single-Chip Video Processing for Consumer and Professional Electronics," Data Sheet Pyramid Vision Technologies a Sarnoff Vision Company, Princeton, NJ, April 2001.

[14] "Software Kits Add Video Surveillance, Enhancement, Targeting, Mapping to Acadia I Vision Card For PCs," Data Sheet, Sarnoff Corporation, Princeton, NJ, April 2001.

[15] "Pyramid-Based Video Processing," Sarnoff Corporation, Princeton, NJ, Feb 2005. [Online]. Available: http://www.sarnoff.com/products_services/government_solutions/vision_technology/core_technology/pyramid_video_processing.asp

[16] S. Li, J. Kwok, and Y. Wang, "Combination of Images with Diverse Focuses Using the Spatial Frequency," *J. of Information Fusion*, vol. 2, no.

3, pp. 169-176, Nov 2001. [Online]. Available: www.cs.ust.hk/faculty/jamesk/papers/if01.pdf