Video Enhancement for Underwater Exploration Using Forward Looking Sonar

Kio Kim, Nicola Neretti, and Nathan Intrator

Institute for Brain and Neural Systems
Brown University, Providence RI 02906, USA
kio@brown.edu

Abstract. The advances in robotics and imaging technologies have brought various imaging devices to the field of the unmanned exploration of new environments. Forward looking sonar is one of the newly emerging imaging methods employed in the exploration of underwater environments. While the video sequences produced by forward looking sonar systems are characterized by low signal-to-noise ratio, low resolution and limited range of sight, it is expected that video enhancement techniques will facilitate the interpretation of the video sequences. Since the video enhancement techniques for forward looking sonar video sequences are applicable to most of the forward looking sonar sequences, the development of such techniques is more crucial than developing techniques for optical camera video enhancement, where only specially produced video sequences can benefit the techniques. In this paper, we introduce a procedure to enhance forward looking sonar video sequences via incorporating the knowledge of the target object obtained in previously observed frames. The proposed procedure includes inter-frame registration, linearization of image intensity, and maximum a posteriori fusion of images in the video sequence. The performance of this procedure is verified by enhancing video sequences of Dual-frequency Identification Sonar (DIDSON), the market leading forward looking sonar system.

1 Introduction

Advances in robotics and imaging technologies have expanded the boundary of human activity and perception to those areas that have been out of our reach for a long time. The exploration of underwater environments is an example of successful applications of novel imaging and robotic technologies.

In the study of underwater environments, the use of forward looking sonar (FLS) systems is increasing thanks to the high frame rate, relatively high resolution, low power consumption and portability [1,2,3]. Forward looking sonar is a type of sonar that produces a 2D image by stacking 1D images produced by a 1D transducer array. Unlike in conventional sonar, the beam forming of FLS is spontaneously achieved without additional computation, so it can produce relatively high resolution images at a frame rate comparable to that of optical video cameras. There are several high performance FLS systems that are commercially available, and the use of such sonar systems is increasing these days [4,5].

Despite the merits of FLS systems, it has shortcomings when compared to optical cameras [6]. First, the angular resolution is relatively low, typically less than 100 pixels.

J. Blanc-Talon et al. (Eds.): ACIVS 2006, LNCS 4179, pp. 554–563, 2006.

Second, the signal-to-noise ratio is still lower than that of optical cameras because of the nature of B-scan images.

In this paper, we present a procedure to enhance FLS video sequences by fusing information collected from different frames. This procedure can reduce noise and increase the spatial resolution of FLS video sequences.

2 Scope of Applicable Video Sequences

Video enhancement algorithms based on super-resolution techniques have been extensively studied [7,8,9]. Such algorithms are basically made up in three parts including i) the registration, ii) the transformation, and iii) the fusion of the images. For optical cameras, the scope of video sequences that can be processed by video enhancement algorithms is strictly limited by the requirements in each step of the procedure, while the FLS video sequences are free from such restrictions. Once an enhancement method for FLS is established, it can be applied to most of FLS video sequences. For this reason, the development of a good video enhancement method is more important in FLS than in conventional optical cameras.

For image registration, one needs to model the homographic relation between images based on the imaging geometry of the imaging device. For example, when a pinhole camera views a planar surface from different perspectives, a perspective transformation is sufficient to explain the homography of images. Most of enhancement methods for optical camera images use a perspective homography or similar homographies of lower hierarchy, such as affine homography. A projective homography requires, however, the target object to be a planar surface, or the camera undergoes only rotational motion without translational motion. For optical cameras, mostly those video sequences intentionally produced can satisfy this requirement. In contrast, FLS requires the target object to be on a planar surface from the image acquisition level—otherwise, the visibility of sonar is extremely narrowed, and the output images suffer significant vignetting. (See Fig. 1.) This property imposes a huge constraint to the variability of FLS images so that an affine transformation can explain most of FLS video sequences [6].

For the fusion of images, in order to combine multiple frames of optical camera video sequences, one needs a video sequence without any occlusion in it. Or, when an occluded area exists in a scene, one needs to add extra steps for segmenting and ex-

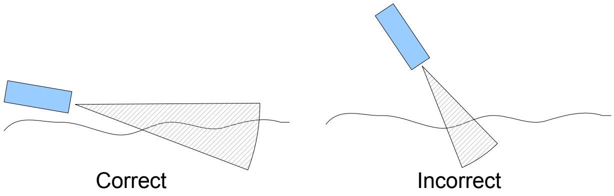

Fig. 1. Correct (left), and incorrect use (right) of a forward looking sonar system. When a FLS device is incorrectly used as describe above, the visible area in the produced image is significantly restricted.

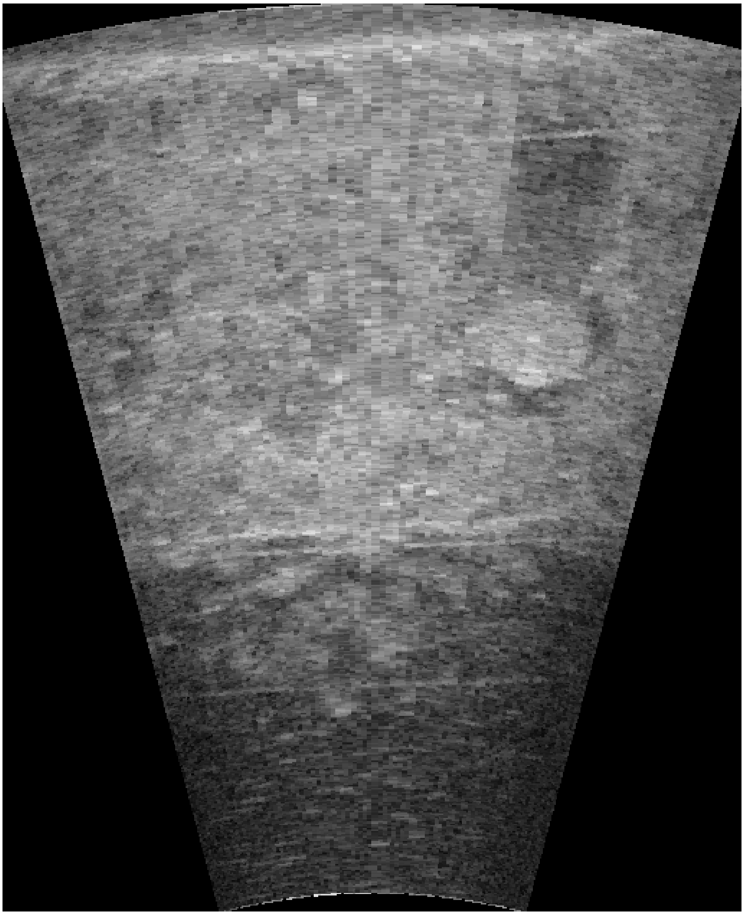

Fig. 2. Occlusion by a cylindrical object in a scene of a forward looking sonar video sequence. The occluded area appears darker than its usual appearance, instead of being replaced by the shape of the occluding object as in optical camera images.

cluding the occlusion. Another merit of FLS video sequences is that the fusion method can tolerate occlusions as far as the occlusion is static. In FLS images, an area turns dark when it is occluded, and recovers its original brightness as soon as it escapes the occlusion. This is simply a change of illumination, which is much easier to handle than an occlusion exclusion problem in optical camera video enhancement. (See Fig. 2.)

3 Methodology

The proposed procedure is largely made up of the following steps: i) separation of illumination profile, ii) inter-frame image registration, iii) linearization of brightness, and iv) maximum a posteriori (MAP) fusion of images.

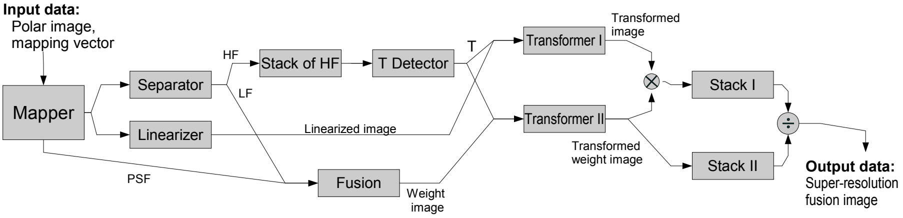

Fig. 3. Block diagram of the super-resolution image fusion process

The flow of data in the procedure is depicted in Fig. 3. The detailed description of each step in the procedure is presented in the following subsections.

In the step i), an image is separated into the high frequency part and the low frequency part. In the step ii), the registration is performed both between two neighboring frames and between frames apart. The parameters of registration found therein are combined optimally to minimize the accumulation of registration errors. In the step iii), image intensity of the images is linearized for the maximum a posteriori fusion of the sequence in the following step. In the step iv), the previously observed frames are fused into one image to best render the frame displayed at the moment. The detailed procedures are described below.

3.1 Retinex Separator

Unlike in optical camera images, the illumination condition of FLS varies significantly within a sequence, and also within a frame, because the illumination depends merely on the ultrasound beam incident from the device itself. A slight change of the grazing angle of the FLS device and the curvature of the target surface can bring a variation of the illumination condition, which eventually makes the registration and fusion difficult. For this reason, one needs to separate the illumination profile from the reflectance profile of the target object.

Land has modeled an illumination process as homomorphic filtering [10], and the consequent researches disclosed algorithms to separate the illumination profile and the reflectance profile of an image in that regard [11,12]. For FLS images, a simple homomorphic filtering of the image is sufficient for the separation, say,

$$HF(\boldsymbol{x}) = I(\boldsymbol{x})/LF(\boldsymbol{x}), \tag{1}$$

where $I(\boldsymbol{x})$, $LF(\boldsymbol{x})$, and $HF(\boldsymbol{x})$ represent the intensity values at the position $\boldsymbol{x}$ in the original image, the low frequency part, and the high frequency part, respectively. The low frequency part is calculated by low-pass filtering the image. We consider that LF and HF are the illumination profile and the reflectance profile, respectively in this paper.

3.2 Inter-frame Registration

In previous work of the authors, it has been discussed that the cross-correlation based feature matching outperforms the conventional feature matching algorithms, particularly for detecting correspondence of images with low resolution [6]. With the outliers

```
(1)  anchor_frame = 1
(2)  for current_frame = 2:end_of_sequence
(3)    Register current_frame with anchor_frame.
(4)    if anchor_frame != current_frame-1,
(5)       Register current_frame with current_frame-1.
(6)    end if
(7)    if the registration above fails,
(8)       Optimize valid transformation parameters
            between anchor_frame and current_frame-1.
(9)       reset anchor_frame = current_frame
(10)   else if current_frame-anchor_frame == predefined_number,
(11)      Optimize transformation parameters until
            between anchor_frame and current_frame.
(12)      Reset anchor_frame = current_frame + 1.
(13)   end if
(14)end for
```

Fig. 4. The algorithm for inter-frame registration

removed via an appropriate algorithm such as RANSAC (Random Sampling Consensus) or LMedS (Least Median of Squares), these feature point pairs serve to register images with subpixel accuracy.

However, even when a subpixel accuracy registration between two frames is obtained, there still remains a concern about the accumulated registration error in registering multiple frames in a video sequence. Further more, the registration of FLS images is based on an affine approximation of the homography [6] instead of the exact geometrical model, the accumulation of registration error can lead to even more significant errors in the registration. A fine tuning of the registration parameters is performed in order to address this problem.

The ideal condition for the least registration error is when all the frames in the sequence are consistently registered with one another. In most of cases, however, the camera is in motion and the view of the targeted area can evolve during the image acquisition period, so most of the frames can be registered with only a few of their neighboring frames.

The proposed algorithm determines how many neighboring frames can be registered with one frame (called an 'anchor frame'). The first frame of the sequence of interest is set to be the anchor frame (Step `(1)` of Fig. 4), and the following frames are registered with the anchor frame (Step `(3)`), as well as their previous frame (Step `(5)`), until any of the registrations fails (Step `(7)`). When the size of the registrable section for the anchor frame is determined (Step `(7)`), one calculates the optimal set of transformation parameters that explains the homographies of all the frames in the section, with the minimal error (Step `(8)`). After this optimization step, it moves on to the remaining part of the sequence with the anchor frame reset to be the first first frame of the remaining part, until it reaches the end of the sequence (Step `(9)`). The structure of the algorithm is described in Fig. 4.

The maximum size of a registrable section has been limited in order to prevent the excessive dimensionality in optimization, to attain the desired latency of process under the allowed computational power, and also to meet the desired performance of the image enhancement (Step (10) of Fig. 4). The optimization of transformation parameters is done by minimizing the energy function defined as

$$\begin{aligned}
&E(\boldsymbol{p}_{1,2}, \boldsymbol{p}_{2,3}, \cdots \boldsymbol{p}_{n-1,n}) \\
&= \sum_{k=1}^{n_{1,2}} |FP^k_{1,2,2} - \mathrm{T}(\boldsymbol{p}_{1,2}) FP_{1,2,1}|^2 \\
&+ \sum_{j=3}^{n} \left\{ \sum_{k=1}^{n_{1,j}} |FP^k_{1,j,j} - \mathrm{T}(\boldsymbol{p}_{1,k}) FP^k_{1,j,1}|^2 \right. \\
&\left. + \sum_{k=1}^{n_{j-1,j}} |FP^k_{j-1,j,j} - \mathrm{T}(\boldsymbol{p}_{j-1,j}) FP^k_{j-1,j,j-1}|^2 \right\},
\end{aligned} \tag{2}$$

where n is the number of frames in the registrable section of the anchor frame, and $n_{p,q}$ is the number of inlier feature point pairs in the registration of the p-th and the q-th frames. $FP^k_{p,q,r}$ is the position vector of the k-th inlier feature point pair of the registration of the p-th and the q-th frames found in the r-th frame, and $\mathrm{T}(\boldsymbol{p}_{p,q})$ is the transformation operator defined by the registration parameter $\boldsymbol{p}_{p,q}$. For any two non-consecutive frame numbers i and j, $\boldsymbol{p}_{i,j}$ is obtained by combining all the transformation parameters of the consecutive frames between the i-th and the j-th frames, say, $\boldsymbol{p}_{i,i+1}, \cdots, \boldsymbol{p}_{j-2,j-1}, \boldsymbol{p}_{j-1,j}$.

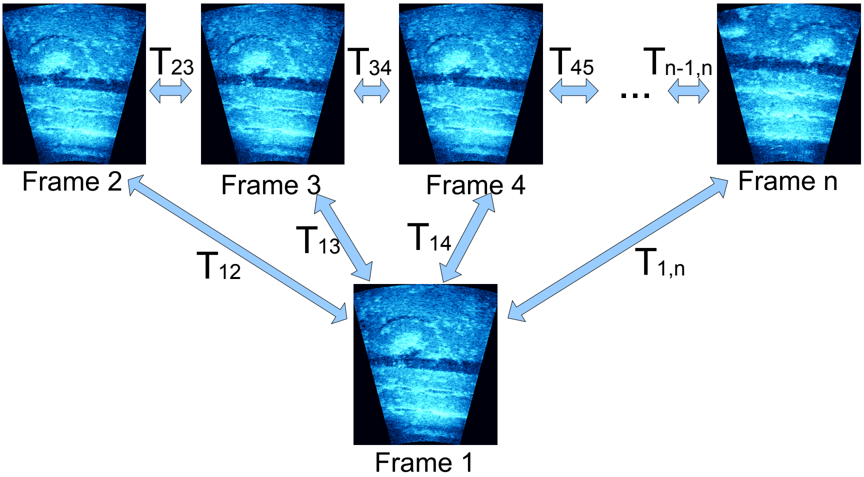

Fig. 5. Paring of the frames in an inter-frame registration of a section of frames. The first frame of a section, or the anchor frame, is registered with all other frames, and all the frames in this section are registered with its neighbors in the section as well.

3.3 Linearization of Image Intensity

When the strength of noise is comparable to the strength of signals as in ultrasound B-scan images, the noise structure is better explained by Rician statistics than Gaussian [13]. The Rician noise in general has non-zero mean, and the mean value of this additive noise is a function of the signal intensity, which in effect distorts the linearity of the image intensity. In addition to the distortion of linearity by the Rician noise, one has to consider the response property of the sensors in the imaging device, which might have been tuned to the precision that is sufficient only for visualization. Since the MAP fusion that will be described in the following subsection requires higher linearity of the sensor response, an additional tuning has to be performed.

For example, for DIDSON images, the following linearizing function significantly improves the quality of the fusion:

$$I'(\boldsymbol{x}) = \begin{matrix} \{I(\boldsymbol{x}) - \mu\}^2, & \text{if } I(\boldsymbol{x}) > \mu \\ 0, & \text{otherwise} \end{matrix} \tag{3}$$

where $I(\boldsymbol{x})$, μ, and $I'(\boldsymbol{x})$ represent the intensity at the position vector $\boldsymbol{x}$, the average background noise level, and the linearized intensity at $\boldsymbol{x}$, respectively.

3.4 Approximated MAP Image Fusion

Once the inter-frame registration and the linearization steps are complete, the frames are ready to undergo the final step of the procedure—the fusion step. Kim et al. have shown that the maximum a posteriori estimation of an image based on a set of low quality observation images of the image can be approximate by a weighted fusion of the low quality images [14]. This implies that one can perform the MAP image fusion without an iterative calculation that many of the super-resolution algorithms require. In addition, the method therein provides robustness under the inhomogeneous illumination condition which is occasional in FLS images.

The enhancement of a frame in a FLS video sequence is attained by fusing a predefined number of frames into the desired perspective using the MAP fusion as described in [14]; when N low resolution images $\beta_1, \cdots, \beta_N$ are fused,

$$\bar{\theta} \simeq \left(\sum_{i=1}^{N} \mathrm{W}_i + v_0 V_0^{-1} \right)^{-1} \left(\sum_{i=1}^{N} \mathrm{W}_i \mathrm{M}_i \beta_i \right), \tag{4}$$

where $\bar{\theta}$, W_i, M_i and $v_0 V_0^{-1}$ represent the calculated MAP fusion image, the i-th reliability matrix, the i-th up-sampling matrix, and the regularization factor, respectively. The up-sampling matrix M_i is a n_{HR}^2-by-n_{LR}^2 matrix, where n_{HR}^2 and n_{LR}^2 are the number of pixels in the high resolution image and in the low resolution image, respectively. The reliability matrix W_i is a n_{HR}^2-by-n_{HR}^2 matrix, which includes all the factors that affect the reliability of a pixel value, for example, illumination intensity, point spread function, etc. The regularization factor $v_0 V_0^{-1}$ is basically the inverse of the covariance matrix of pixel values normalized by v_0, the generic variance of pixel values. Ideally it includes non-zero off-diagonal terms, but for the simplicity of calculation, it is approximated by a diagonal matrix.

4 Results

An experiment was performed on a video sequenced used in a ship hull inspection [3]. In the inspection, a dual-frequency identification sonar (DIDSON) system mounted on a remotely controlled vehicle recorded the images of the surface of a ship hull while the vehicle was manipulated to move around on the ship hull surface.

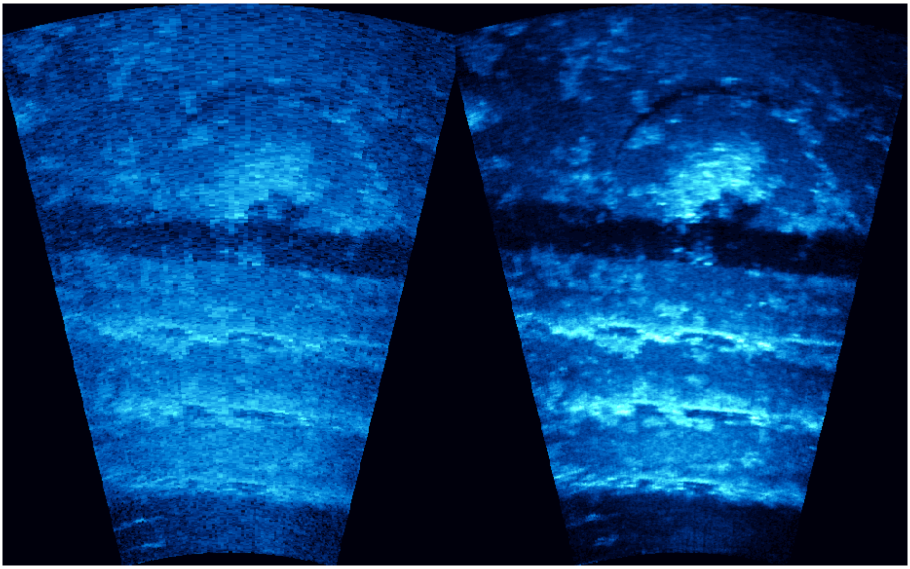

Fig. 6. Comparison of a frame in the original (left) and the enhanced (right) sequences. The frame in the enhanced sequence is a fusion of 7 consecutive frames portrayed in the same perspective.

The resolution of the initial polar coordinates images is 96x512, and the polar coordinate images are mapped to images of size 512x576 size in the Cartesian coordinate system. The size of the Cartesian coordinate image is approximately the size that the smallest pixel in the polar coordinates image can occupy at least one pixel in the Cartesian coordinate, and mostly more than one pixel. In this way, one pixel in the polar coordinates occupies from 1 pixel to up to 20 pixels in the Cartesian coordinates, due to the fixed field-of-view of a sensor and varying distance from the transducer to the observed area.

The suggested procedure has been applied to the video sequence. Figure 6 is the comparison of one of the frames in the original sequence, and the enhanced sequence, where up to 7 neighboring frames were fused to re-render a frame. In Fig. 6, one can verify that the fusion image (right) discloses crispier edges of the target object, than the original image (left). Also note that the surface texture that was difficult to identify in the original sequence can be easily identified in the enhanced sequence due to the reduced noise level and the increased solution.

5 Conclusion

In this paper, we presented a procedure to enhance a forward looking sonar video sequence. The procedure includes the separation of illumination profile, inter-frame registration, the linearization of the images, and non-iterative maximum a posteriori fusion of images for super-resolution.

Since the image acquisition method of FLS restricts the applicable target object to be on a planar surface, most of the FLS images can be registered using a simple affine homography. In addition, the occlusion problem, which often is an obstacle in processing optical video sequences, can be viewed simply as an illumination problem in FLS video sequences, which can be dealt with little trouble. This means there is no need to further consider the occlusion problem in FLS video sequences. For these reasons, video enhancement techniques for FLS in general are applicable to most of the FLS video sequences.

The proposed video enhancement procedure is largely made up of four steps including the separation of illumination profile, fine tuning of the registration parameters via inter-frame image registration, the linearization of brightness, and the maximum a posteriori (MAP) fusion of the images. All these steps are achievable with low computational power.

In future, further study for real time implementation of the proposed procedure is anticipated.

References

1. R. A. Moursund and T. J. Carlsonand R. D. Peters. A fisheries application of a dual-frequency identification sonar acoustic camera. *ICES Journal of Marine Science*, 60(3):678–683, 2003.
2. Yunbo Xie, Andrew P. Gray, and Fiona J. Martens. Differentiating fish targets from non-fish targets using an imaging sonar and a conventional sonar: Dual frequency identification sonar (didson) versus split-beam sonar. In *Journal of the Acoustical Society of America*, volume 117, pages 2368–2368, 2005.
3. J. Vaganay, M. L. Elkins, S. Willcox, F. S. Hover, R. S. Damus, S. Desset, J. P. Morash, and V. C. Polidoro. Ship hull inspection by hull-relative navigation and control. In *Proceedings of Oceans '05 MTS/IEEE*, pages –, 2005.
4. http://www.didson.com.
5. http://www.blueviewtech.com.
6. K. Kim, N. Neretti, and N. Intrator. Mosaicing of acoustic camera images. *IEE Proceedings—Radar, Sonar and Navigation*, 152(4):263–270, 2005.
7. M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, 1993.
8. S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. Technical report, 1998.
9. Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Journal*, 20(3):21–36, 2003.
10. E. H. Land and J. J. McCann. Lightness and the retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.
11. E. H. Land. An alternative technique for the computation of the designator in the retinex theory of color vision. *PNAS*, 83:3078–3080, 1986.

12. D. J. Jobson, Z. Rahman, and G. A. Woodell. Properties and performance of the center/surround retinex. *IEEE Transactions on Image Processing*, 6:451–462, 1997.
13. R. F. Wagner, S. W. Smith, J. M. Sandrik, and H. Lopez. Statistics of speckle in ultrasound b-scans. *IEEE Transactions on Sonics and Ultrasonics*, 30(3):156–163, 1983.
14. K. Kim, N. Neretti, and N. Intrator. Maximum a posteriori fusion method for super-resolution of images with varying reliability. pages –, 2006.