Image Super-Resolution for Improved Automatic Target Recognition

Raymond S. Wagner[a] and Donald Waagen[b] and Mary Cassabaum[b]

[a]Rice University, Houston, TX
[b]Raytheon Missile Systems, Tucson, AZ

ABSTRACT

Infrared imagers used to acquire data for automatic target recognition are inherently limited by the physical properties of their components. Fortunately, image super-resolution techniques can be applied to overcome the limits of these imaging systems. This increase in resolution can have potentially dramatic consequences for improved automatic target recognition (ATR) on the resultant higher-resolution images. We will discuss super-resolution techniques in general and specifically review the details of one such algorithm from the literature suited to real-time application on forward-looking infrared (FLIR) images. Following this tutorial, a numerical analysis of the algorithm applied to synthetic IR data will be presented, and we will conclude by discussing the implications of the analysis for improved ATR accuracy.

Keywords: Super-Resolution, Automatic Target Recognition, FLIR

1. INTRODUCTION

The fidelity of data gathered by forward-looking infrared (FLIR) imagers is limited by the quality of the optical and electronic components of the system. Such images suffer, of course, from blurring effects and noise from both thermal and electronic sources, and dealing with such phenomena is the domain of classic image restoration. On a more basic level, though, the images gathered by FLIR detectors are inherently limited in their resolution by the detection array used to capture the images. Specifically, the density of cells in the detection array fundamentally restricts the resolution of the images gathered by such systems. Details in the image plain smaller than the size of a cell are averaged out during the image capture process and are unavailable to further image processing operations such as automatic target recognition (ATR). Fortunately, image super-resolution techniques can be applied to overcome the limits of these imaging systems. This increase in resolution can have potentially dramatic consequences for improved ATR on the resultant higher-resolution images.

Image super-resolution entails fusing a set of low-resolution (LR) images, related to each other primarily through random translations and rotations in the image plane, in order to create a single, high-resolution (HR) image of the original scene. Such techniques can, for example, be applied to a sequence of video frames to generate a still image of higher resolution than any single frame in the feed. To generate the HR image, the LR images must first be registered relative to a specific frame of reference. Following this registration, available LR pixels are used to sparsely populate an HR image grid, and non-uniform interpolation techniques are applied to the remaining gridpoints to generate an estimate of the HR image.

In this tutorial paper, we will discuss super-resolution techniques in general and specifically review the details of one such algorithm suited to real-time application on FLIR images. Section 2 will discuss in more detail the theory of image SR and briefly overview three families of SR methods.[1] Section 3 will review the mathematics behind the particular SR solution demonstrated here,[2,3] and Section 4 will provide a brief example of the technique as well as a quantitative analysis of the increase in fidelity which SR can provide. Section 5 will conclude by discussing the implications of the analysis for improved ATR accuracy.

RSW: rwagner@rice.edu, DW: donald_e_waagen@raytheon.com, MC: mlcassabaum@raytheon.com

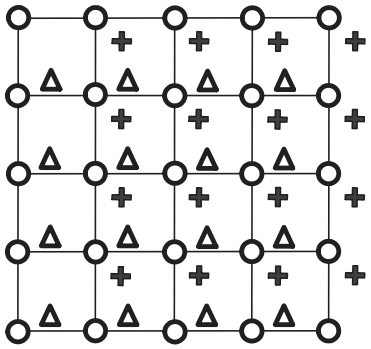


Figure 1. Example sub-pixel displacements.

2. IMAGE SUPER-RESOLUTION THEORY

The field of image super-resolution arose from the need to overcome the physical limitations of low-resolution (LR) imaging systems to generate higher-resolution images than would be otherwise possible with the available hardware. For example, in surveillance applications, single video frames are relatively low in image detail and not well-suited to tasks such as face recognition. Similarly, even the high-end optics on imaging satellites are not always sufficient to distinguish important scene features. Fortunately, when a moderate amount of scene motion exists between frames, the data in these low-resolution images can be fused to yield an image of higher resolution than any one of the frames. A variety of approaches can be found in the literature for exploiting this scenario, and a comprehensive overview of the state-of-the art in image SR can be found in,[1] to which we refer the interested reader. For the purpose of a tutorial for the ATR community, a brief overview of[1] is presented in this section.

The scene motion resulting from small camera displacements is typically referred to as sub-pixel displacement, and is illustrated in Figure 1, which shows a grid of pixels ($\circ$) composing a reference image frame. Pixels from two other images ($\triangle$ and $+$) with sub-pixel displacements are drawn relative to the reference frame. This picture provides a very intuitive understanding of the motivation for image SR. Since digital imaging systems discretely capture scene information as pixels, scene features falling between pixels in the reference image $\circ$ are lost (in reality, they are averaged among neighboring pixels, but the unique features are no longer available). Therefore, the images $\triangle$ and $+$ contain scene details not captured by $\circ$, information which can be fused to provide an overall higher-resolution look at the scene. Note that the $\triangle$ image contains strict sub-pixel displacement from the reference image $\circ$, while the image $+$ contains pixel-plus-subpixel displacement. Provided that the larger displacement of $+$ can be accurately estimated, both images are useful in super-resolving the scene. The only caveat is that $+$ will contribute fewer pixels to the enhancement of the left edge of the reference frame $\circ$, since the left-hand portion of $\circ$ lies more than a pixel outside of the scope of $+$. Similarly, $+$ contains pixels in its right-hand side which are outside the scope of $\circ$.

Before a high-resolution (HR) image (call it i) can be reconstructed, one must first specify a model by which each of K LR images g_k are generated from the unknown image i. Re-formatting images as vectors of pixel values ($\underline{i}$ and $\underline{g_k}$), one can form the general relation

$$\underline{g_k} = W_k \underline{i} + \underline{n_k},$$

where matrix W_k models the unique generation process for each low-resolution image, and $\underline{n_k}$ is a generalized noise vector which accounts for error in the specification of W_k. Assuming that i is discrete, non-aliased sampling of the continuous, infinite-resolution scene, it suffices to include in W_k motion relative to the reference frame (typically translations tx_k and ty_k and rotation θ_k), blurring, and downsampling (horizontally by a factor L_1 and vertically by a factor L_2) to the lower resolution. Figure 2 summarizes this process.

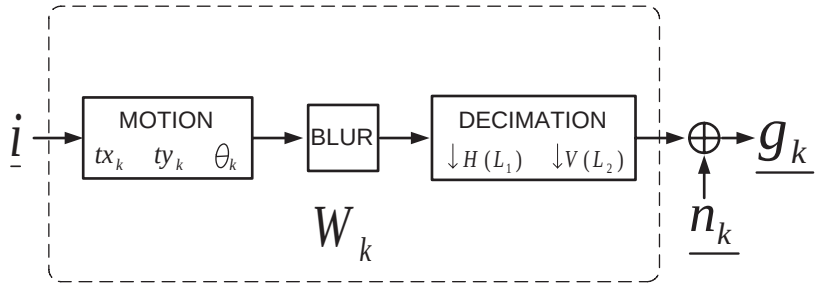


Figure 2. Model relating low-resolution to high-resolution images.

Once this model is established, the steps to super-resolving the K low-resolution images become clear. First, to account for scene motion, each of the $K-1$ non-reference LR images must be registered relative to the reference LR image. Once this is accomplished, the effects of downsampling are overcome by populating an HR image grid with the registered low-resolution pixels, as illustrated in the second frame of Figure 3 (again, the reference frame is denoted as $\circ$). As the $K-1$ non-reference LR images likely did not have perfect n/L_1 and m/L_2 horizontal and vertical displacements (n,m integers), the HR grid will end up with the majority of LR pixels falling between HR pixel locations. Thus, interpolation of the un-assigned HR pixels will be necessary. This is depicted in the third frame of Figure 3, where the remainder of the HR pixels are given interpolated values ($*$). Finally, as LR pixels were sampled from a blurred HR image, a de-blurring process is necessary to restore a clear estimate of the original, HR image.

A variety of approaches have been suggested for implementing the three steps (registration, interpolation, and de-blurring) of the SR problem, of which there are three principle families. The first, and most intuitive method, exactly proceeds as described above. First, LR images are registered relative to the reference, and the HR grid is estimated via non-uniform interpolation of the registered LR pixels followed by a noise-tolerant de-blurring scheme such as Wiener filtering. The example presented in this tutorial proceeds along these lines and is discussed in detail in Section 3.

Another family of SR solutions operates in the frequency domain of LR and HR images. The set of LR images are first registered as above, but rather than constructing the HR grid in the spatial domain via non-uniform interpolation, the frequency domain analog of the HR image is instead estimated. To realize this, a system equation is derived to generates discrete Fourier transform (DFT) coefficients of the LR images given samples of the continuous Fourier transform (CFT) of the unknown, HR image. This enables an inverse operation, by which the unknown CFT samples are recovered.

Finally, a family of techniques exists whereby the inversion to an unknown HR image is regularized to account for reconstruction in the presence of too few LR images and error-prone blur models. These techniques fall into two main camps: deterministic reconstruction via the method of constrained least squares (CLS) and stochastic reconstruction via maximum a posteriori (MAP) estimation. The CLS method uses Lagrange minimization of the error sum $\sum_{k=1}^{K} |\underline{g}_k - W_k \underline{i}|^2$ subject to a constraint such as smoothness of the reconstruction. This allows for a scalable tradeoff between relying on the potentially misleading LR data (due to registration error and noise) and forcing the result to satisfy the constraint. The stochastic version falls into the realm of Bayesian estimation, and requires specification of the conditional probability model by which LR images are generated given the HR image ($p(\underline{g}_1, ..., \underline{g}_K | \underline{i})$), as well as a prior model for the distribution of the HR image i

$(p(\underline{i}))$. The technique is based on maximizing the probability of i given the set $\{\underline{g}_1, ..., \underline{g}_K\}$, which decomposes as

$$\underline{i} = arg\ max\ p(\underline{i}|\underline{g}_1, ..., \underline{g}_K) = arg\ max\ \{\ln\ p(\underline{g}_1, ..., \underline{g}_K|\underline{i}) + \ln\ p(\underline{i})\}.$$

The solution to the estimation is obviously influenced by the choice of the prior for i. In fact, when a Gaussian prior is used for i, the MAP solution yields the same estimate as the CLS solution. Note that both regularized techniques assume registration for LR images, as with the non-uniform interpolation and frequency domain techniques. Stochastic techniques also exist to jointly estimate registrations and HR reconstructions.

3. GRADIENT-BASED REGISTRATION AND NON-UNIFORM INTERPOLATION

To illustrate the power of image SR, we implement a modified version of the algorithm proposed in Hardie et al..[2] This technique, which falls into the non-uniform interpolation category discussed above, is recommended for real-time resolution enhancement of infrared imaging sensor data. Speed of the algorithm is its key strength, and as such it is useful for consideration as a pre-processing step prior to ATR applications.

The registration portion of the algorithm is taken from Irani and Peleg,[3] and its application can be traced back to Keren et al.,[4] with origins in the 1981 work by Lucas and Kanade.[5] While Hardie[2] restricts LR image displacements to translations, Irani[3] incorporates both translations and rotations into its LR image generation model. As thus, this tutorial will adopt a modified version of the notation in Irani[3] to review the mathematics behind the registration, which is based on gradients of pixel intensities in the LR images.

Call the low-resolution reference image g_1, and label one of the remaining $K-1$ low-resolution images g_k. Pixel locations in g_k are related to those in the reference as

$$g_k(x,y) = g_1(x\cos\theta_k - y\sin\theta_k + tx_k, y\cos\theta_k + x\sin\theta_k + ty_k),$$

where tx_k and ty_k are the horizontal and vertical translations and θ_k is the rotation which takes g_1 to the frame of g_k. Replacing sin and cos above with their first-order Taylor series expansion gives

$$g_k(x,y) \approx g_1(x + tx_k - y\theta_k - x\theta_k^2/2, y + ty_k + x\theta_k - y\theta_k^2/2),$$

which can be further approximated via its own first-order Taylor series expansion as

$$g_k(x,y) \approx g_1(x,y) + (tx_k - y\theta_k - x\theta_k^2/2)\frac{\partial g_1}{\partial x} + (ty_k + x\theta_k - y\theta_k^2/2)\frac{\partial g_1}{\partial y}.$$

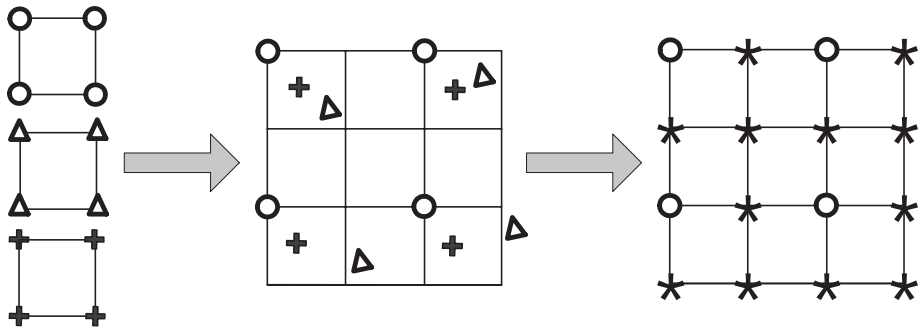


Figure 3. Illustration of registration and interpolation processes.

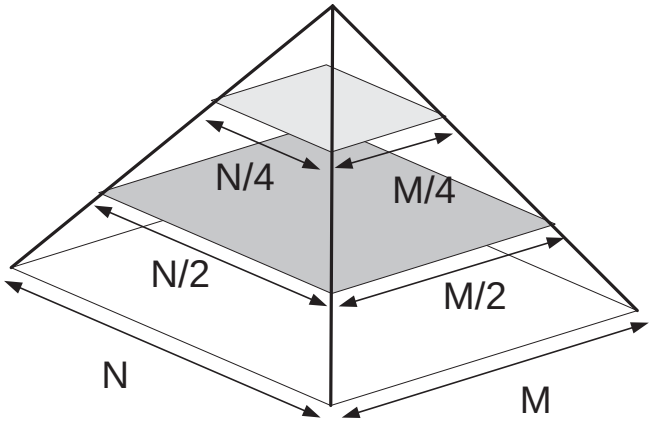


Figure 4. Gaussian pyramid multi-resolution structure.

This allows for an error expression between g_k and the transformed g_1 parametrized by the registration terms tx_k, ty_k, and θ_k:

$$E(tx_k, ty_k, \theta_k)$$
$$= \sum \left[g_1(m,n) + (tx_k - n\theta_k - m\theta_k^2/2)\frac{\partial g_1}{\partial m} + (ty_k + m\theta_k - n\theta_k^2/2)\frac{\partial g_1}{\partial n} - g_k(m,n) \right]^2,$$

where the sum is taken over the overlapping portions of g_1 and g_k, and variables of summation have been changed to m and n to represent the discrete horizontal and vertical pixel locations.

$E(tx_k, ty_k, \theta_k)$ can be minimized with respect to each of tx_k, ty_k, and θ_k , yielding the system

$$\begin{bmatrix} \sum g_m^2 & \sum g_m g_n & \sum A g_m \\ \sum g_m g_n & \sum g_n^2 & \sum A g_n \\ \sum A g_m & \sum A g_n & \sum A^2 \end{bmatrix} \begin{bmatrix} tx_k \\ ty_k \\ \theta_k \end{bmatrix} = \begin{bmatrix} \sum g_m g_t \\ \sum g_n g_t \\ \sum A g_t \end{bmatrix},$$

where $g_m = \frac{\partial g_1(m,n)}{\partial m}$, $g_n = \frac{\partial g_1(m,n)}{\partial n}$, $g_t = g_k(m,n) - g_1(m,n)$, and $A = m g_n - n g_m$, and the sum is again taken over overlapping pixels of g_1 and g_k. Note that $\frac{\partial g_1(m,n)}{\partial m}$ corresponds to the horizontal gradient and $\frac{\partial g_1(m,n)}{\partial n}$ the vertical gradient of the reference image.

Since the approximations above are based on Taylor expansions about the estimated parameters, tx_k, ty_k, and θ_k must be relatively small for the estimation to be accurate. To allow for larger registration parameters, a multi-resolution, iterative technique is employed. First, both images g_1 and g_k are artificially reduced in resolution to a minimum size via a structure known as a gaussian pyramid[6] (see Figure 4). Where the base of the pyramid is the original-resolution image, each upper tier is derived from the lower one by convolving the image with a 2-D Gaussian kernel and downsampling by a factor of 2 in each direction. Performing the registration at lower resolutions allows large translations to reduce to small values which can be accurately estimated. Once registration has converged at a specific resolution level, the algorithm then moves up one level and proceeds to refine the registration parameters.

At a given level, registration proceeds as follows. Motion estimates (zero at the first level, current registration estimate at all others) are applied to image g_k at the current resolution level. Note that this necessitates some sort of interpolation algorithm to generate the warped g_k. The estimation equations are then applied to the corrected g_k to obtain parameter refinements, which are aggregated with current estimates. The new

values are again used to warp g_k and refinements are again calculated. This proceeds until refinements become sufficiently small.

As it is $g_k(m,n)$ which is transformed at each iteration, the terms g_m, g_n, and A, all based on gradients of $g_1(m,n)$ need only be computed once at each iteration. This leads to efficient computation of the algorithm. Note also that registration values tx_k, ty_k, and θ_k are calculated to warp g_1 into the frame of g_k via a rotation, followed by a translation. Thus, to use these values to warp g_k into the reference frame, one must first translate g_k by $(-tx_k, -ty_k)$ and then rotate by $-\theta_k$ about the new origin.

A weighted nearest-neighbor interpolation scheme is suggested in[2] to generate the high-resolution grid after population by registered, LR image pixels. Rather than adopting this approach, we have chosen to use standard MATLAB non-uniform linear interpolation methods for their greater accuracy at the expense of an only slightly longer runtime.

For the purposes of this tutorial example, we assume an ideal de-blurring process. In practice, a restoration algorithm tolerant of interpolation and registration error would be required. This necessitates developing accurate estimates of signal power and noise power, as well as an accurate estimate of the blur operator convolved with the original, high-resolution image before downsampling.

4. SUPER-RESOLUTION EXAMPLE

As an example of the super-resolution technique, consider the case of super-resolving a reconstruction of a known high-resolution image from artificially-generated, lower-resolution images, a convenient exercise which allows for quantifying the fidelity of reconstruction. The target HR image is the 256x256 detail shown in Figure 5, which is a real FLIR image of an armored personnel carrier taken in a test at China Lake, CA. * Synthetic sets of low-resolution images are generated from the HR reference by applying random translations/rotations to the target image and then downsampling by a factor of L in each dimension. For the purposes of this example, two such sets are employed. A sample low-resolution image for the $L = 2$ case can be seen in Figure 6 (a), and one for the $L = 4$ case is found in Figure 7 (a). The original-resolution reconstruction using 8 $L = 2$ images is seen in Figure 6 (b), while Figure 7 (b) shows the HR estimate using 32 $L = 4$ images. Clearly, in both cases, resolution and image quality has been improved (substantially so in the $L = 4$ case) when HR estimates are compared to the LR images.

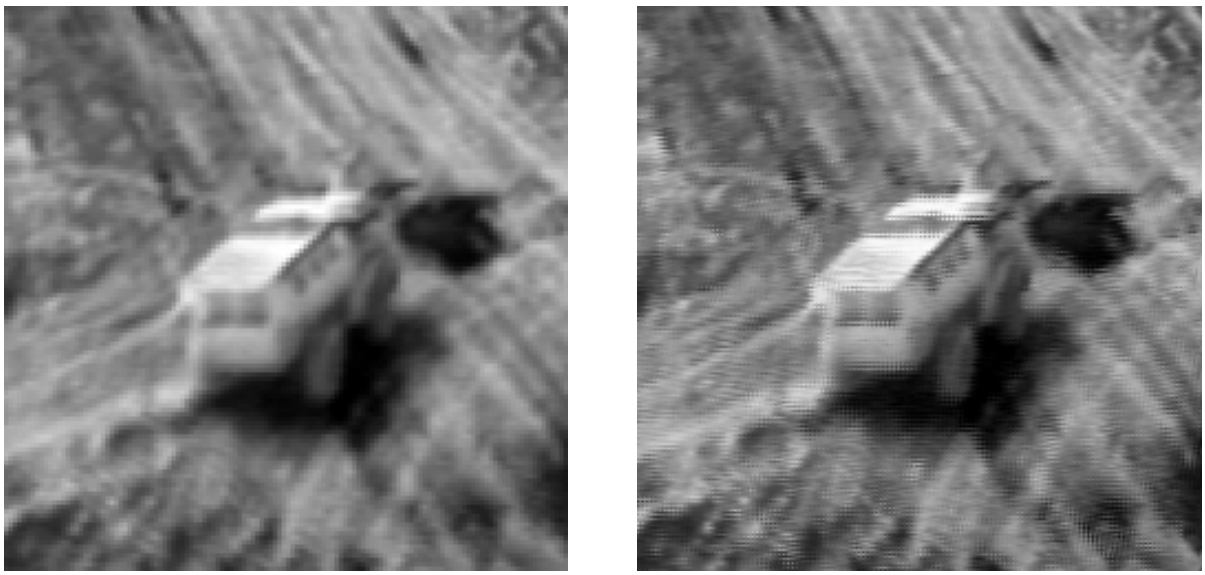
To quantify the results of super-resolution, distortion measures compared to the known image of Figure 5 are given in Table 1 for the $L = 2$ case and Table 2 for the $L = 4$ case. Distortion is measured in peak signal-to-noise ratio (PSNR). † In their final columns, both tables also give the distortion measure for the low-resolution images at each level (half resolution for Table 1, quarter resolution for Table 2).

Clearly, as measured in PSNR, the super-resolution process has substantially increased the distortion of the SR reconstructions compared to that of the original, LR image. Note that reconstruction quality scales with numbers of LR images used in the super-resolution. For the $L = 2$ case of Table 1, results are shown using $K = 4$, $K = 8$, $K = 16$, and $K = 32$ LR images. In the ideal case of perfect half-pixel displacements, one need only use 4 LR images to perfectly reconstruct the full-resolution HR estimate, but as displacements here are random (both translations and rotations), greater numbers of LR images yield a more densly populated HR grid prior to the non-uniform interpolation step, and hence more accurate interpolation. Additionally, increasing numbers of LR images are required as the resolution-enhancement factor L increases. For the $L = 4$ case of Table 2, $K = 16$ LR images corresponds to the number need for a perfect set of displacements. The $K = 32$ case merely represents a doubling of this "minimum." To achieve a similar performance trend as that seen in the $L = 2$ case (which uses up to 8 times the "minimum"), as many as 128 LR images could be required. Thus, there is an interplay between the number K of LR images available and the factor L to which

*Used with permission from `http://www.cis.jhu.edu/data.sets/nawc_flir/db_nawc_land.html`

†PSNR $= 10\log_{10}(M255^2/\text{MSE})$, where MSE is the mean squared error between the estimate and original and M is the number of data points in each image.

Figure 5. Known, target HR image



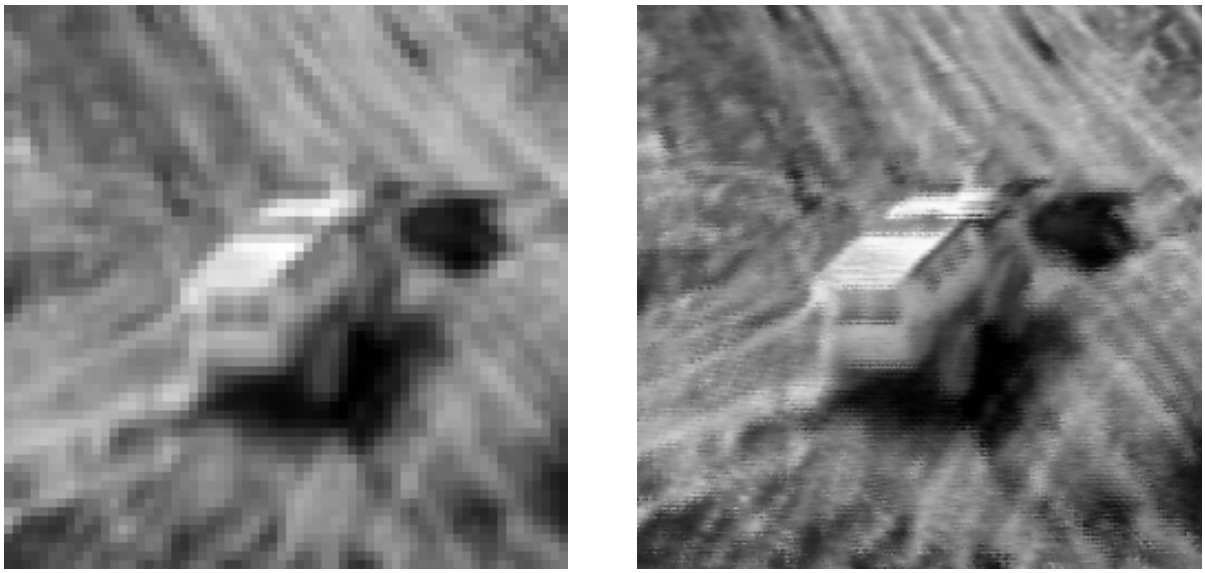
(a) (b)

Figure 6. (a) LR reference image (one-half resolution). (b) Super-resolved (factor of 2) HR estimate.

one can reliably attempt to super-resolve those images. Note, too, that as L increases, registration error of the LR images onto the HR grid will also increase, affecting overall reconstruction fidelity.

5. CONCLUSION

As the previous section illustrates, there is clearly a substantial amount of detail to be uncovered from fusing multiple images limited by the native resolution of an imaging system into a single, super-resolved image. The example from Section 4 demonstrates this improvement, both in image quality (Figures 6 and 7) and fidelity (Tables 1 and 2) as measured in PSNR. The implications for the ATR community are equally clear:



(a) (b)

Figure 7. (a) LR reference image (one-quarter resolution). (b) Super-resolved (factor of 4) HR estimate.

Table 1. Distortions (PSNR in dB) for factor-of-two SR estimates of known HR image using K= 8, 16, and 32 LR images. Compare against PSNR of single LR image.

$K = 32$	$K = 16$	$K = 8$	$K = 4$	LR (fact. of 2)
39.2635	36.3766	33.6324	31.3727	27.7190

improved detail of target images allows for clearer discrimination between targets - especially those which can be distinguished by details lost in the image acquisition process. Introducing image super-resolution as a pre-processing step to generate higher-resolution images from sequences such as IR video feeds has the potential to improve subsequent attempts at target identification.

In this tutorial, we have provided a brief introduction to image super-resolution and covered the details of one algorithm useful to super-resolving FLIR images, with the aim of introducing individuals interested in ATR to this powerful tool. We hope, in doing so, to help promote increased collaboration between the SR and ATR communities and encourage further investigation of the utility of image super-resolution in automatic target recognition.

Table 2. Distortions (PSNR in dB) for factor-of-two SR estimates of known HR image using K= 16, and 32 LR images. Compare against PSNR of single LR image.

$K = 32$	$K = 16$	LR (fact. of 4)
32.7507	30.5236	25.9024

ACKNOWLEDGMENTS

Thanks go to the Center for Imaging Science at John's Hopkins University for providing the test image used in Section 4. The primary author is also extremely grateful to Raytheon Missile Systems in Tucson, AZ for providing a friendly and stimulating work environment during his stay as a summer intern.

REFERENCES

1. S. Park, M. Park, and M. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Processing Magazine* **20**, pp. 21–36, May 2003.
2. M. Alam, J. Bognar, R. Hardie, and B. Yasuda, "Infrared image registration and high-resolution reconstruction using multiple translationally shifted aliased video frames," *IEEE Trans. on Instrumentation and Measurement* **49**, pp. 915–923, Oct. 2000.
3. M. Irani and S. Peleg, "Improving resolution by image registration," *Computer Vision Graphical Image Processing: Graphical Models and Image Processing* **53**, pp. 231–239, 1991.
4. D. Keren, S. Peleg, and R. Brada, "Image sequence enhancement using sub-pixel displacements," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 742–746, (Ann Arbor, MI), Jun. 1988.
5. B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artifical Intelligence*, pp. 674–679, (Vancouver, Canada), Aug. 1981.
6. A. Rosenfeld, ed., *Multiresolution Image Processing and Analysis*, Springer-Verlag, Berlin/New York, 1984.