FAST DISAMBIGUATION OF SUPERIMPOSED IMAGES FOR INCREASED FIELD OF VIEW

Roummel F. Marcia, Changsoon Kim, Jungsang Kim, David J. Brady, and Rebecca M. Willett

Department of Electrical and Computer Engineering,
Duke University, Durham, NC 27708, USA

ABSTRACT

Many infrared optical systems in wide-ranging applications such as surveillance and security frequently require large fields of view. Often this necessitates a focal plane array (FPA) with a large number of pixels, which, in general, is very expensive. In this paper, we propose a method for increasing the field of view without increasing the pixel resolution of the FPA by superimposing the multiple subimages within a scene and disambiguating the observed data to reconstruct the original scene. This technique, in effect, allows each subimage of the scene to share a single FPA, thereby increasing the field of view without compromising resolution. To disambiguate the subimages, we develop wavelet regularized reconstruction methods which encourage sparsity in the solution. We present results from numerical experiments that demonstrate the effectiveness of this approach.

Index Terms— Image reconstruction, Image sampling, Video cameras

1. INTRODUCTION

The performance of a typical imaging system is characterized by the resolution (smallest feature the system can resolve) and the field of view (FoV: the maximum angular extent that is observed at a given instance). In most video imaging systems today, the detector element is a focal plane array (FPA) typically made out of semiconductor photodetectors. While low-cost, large pixel count charge-coupled device (CCD) and complementary metal-oxide semiconductor (CMOS) sensors are widely available for imaging in visible wavelengths, the FPAs in the mid- and long-wave infrared (3-20 μm in wavelength) remain very expensive. Many thermal-imaging surveillance systems utilize this wavelength range, and technologies that enable a wide FoV using a small pixel count FPA will have an important impact in this application. It has been long known that human vision can effectively differentiate two superimposed images moving relative to each other [1]. In this paper, we propose and demonstrate a computational imaging technique where the overall FoV of an imaging system is broken into smaller scenes and superimposed onto a single FPA, effectively increasing the FoV of the imaging system. The superimposed image is disambiguated using an efficient video processing algorithm using small relative motion among the scenes, and restores the complete scene corresponding to the overall FoV.

The proposed technique has two key advantages over possible alternative methods of generating high FoV, high-resolution image data in the infrared domain: (1) unlike shutter-based systems which could measure different regions of a scene sequentially, the proposed system is mechanically stable, robust, and easy to assemble, and (2) the proposed physical system does not require complicated calibration or tuning.

In the proposed approach, a video is produced in which the different subimages are moved relative to one another, resulting in a collection of frames with different composite images. The video data thus generated is considered as a linear mixture, which is then separated by an optimization technique based on sparse representation bases. This technique, in effect, allows each subimage to share a single FPA, thereby increasing the FoV without compromising resolution. Our numerical experiments, where superimposed videos are generated numerically from digital images, show that our optimization technique can reconstruct the constituent images with small mean square errors.

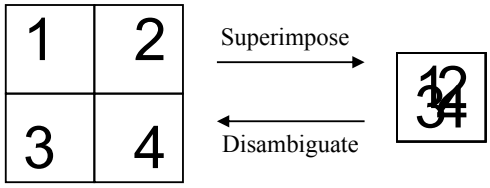


Fig. 1. Superimposition and disambiguation.

In this paper, we propose a simple infrared camera architecture for collecting composite images of the type described above and an associated technique for disambiguating the superimposed subimages. The paper is organized as follows: In Sec. 2, we discuss the concept of our technique, followed by a more detailed description of the proposed architecture and the basic mathematical formulation in Sec. 3. Sec. 4 shows how the video disambiguation problem can be solved using optimization techniques based on sparse representation algorithms. In Sec. 5, we describe the numerical experiments.

2. PROBLEM FORMULATION

Fig. 1 schematically shows the basic concept of superimposition and disambiguation. In the superimposition process, multiple subimages are merged to form a composite image (shown on the right side of Fig. 1) in a straightforward manner; the intensity of each pixel in the composite image is the simple summation of the intensities of the corresponding pixels in the individual images. However, the inverse process – the disambiguation of the individual subimages from this composite image – is more challenging. For this, we must determine how the intensity of each pixel in the composite image is distributed over the corresponding pixels in the individual subimages so that the

The authors were partially supported by DARPA Contract No. HR0011-04-C-0111, ONR Grant No. N00014-06-1-0610, and DARPA Contract No. HR0011-06-C-0109.

resulting reconstruction accurately approximates the original scene. Our technique achieves this task by measuring a composite video sequence, where the position of each subimage is slightly altered at each frame. It is the movement of these individual subimages that allows disambiguation to succeed.

The disambiguation problem can be modeled mathematically at the t^{th} frame as

$$\mathbf{z}_t = \mathbf{A}_t\mathbf{x} + \mathbf{n}_t, \tag{1}$$

where $\mathbf{z}_t \in \Re^{m\times 1}$ is the observed composite image, $\mathbf{x} \in \Re^{n\times 1}$ is the scene, $\mathbf{A}_t \in \Re^{m\times n}$ is the mixing matrix, and $\mathbf{n}_t$ is noise at frame t. We assume in this paper that $\mathbf{n}_t$ is zero-mean white Gaussian noise. In this setting, $n > m$, which makes (1) underdetermined. There are several techniques for approaching this ill-posed statistical inverse problem, many of which exploit the sparsity of $\mathbf{x}$ in one or more bases (cf. [2, 3, 4]).

We formulate the reconstruction problem as a sequence of nonlinear optimization problems, minimizing the norm of the error $\|\mathbf{z}_t - \mathbf{A}_t\mathbf{x}\|$ at each time frame and using the computed minimum as the initial value for the following frame. Since the underlying inverse problem is underdetermined, we include a regularization term $\tau\|\mathbf{x}\|$, where τ is a tuning parameter, in the objective function to make the disambiguation problem well-posed. This formulation of the reconstruction problem is similar to the $\ell^2 - \ell^1$ formulation of the compressed sensing problem [5, 6, 7] for suitably chosen norms: using the Euclidean norm for the error term gives the least-squares error while using the one norm for the regularization term induces sparsity in the solution. When the number of observation frames is large, solving for $\mathbf{x}$ using all the data $\{\mathbf{z}_t\}_{t=1}^{T}$ simultaneously can be overdetermined yet computationally prohibitive. This problem can be circumvented by solving (1) for each successive t, and using the t^{th}-frame solution to initialize the $(t+1)^{\text{th}}$-frame optimization.

3. PROPOSED CAMERA ARCHITECTURE AND SYSTEM MODEL

Superimposed images which are shifted relative to one another at different frames can easily be recorded using a simple camera architecture, depicted for two subimages in Fig. 2. Constructed using beamsplitters and movable mirrors, the proposed assembly merges the subimages into a single image and temporally varies the relative position of the two subimages as they hit the detector. The optical field from the left half of the scene (denoted S) propagates directly through the beamsplitter (denoted BS) and hits the FPA in the camera at the same relative position for every frame. The optical field from the right half of the scene, however, is reflected by a movable mirror (denoted M) followed by the beamsplitter before hitting the focal plane array. When the mirror, mounted on a linear stage, is translated, the right half of the scene is translated by a proportional amount on the FPA. The image recorded by the FPA is then the sum of the stationary left subimage and translated right subimage for each frame, resulting in a superimposed video sequence. Using this setup, building a superposition imaging system is straightforward, making the methods described in this paper readily applicable to practical, real-world settings.

For ease of notation, we will assume that we are only superimposing two subimages, but the approach we describe can easily be extended to more general cases. (In our numerical experiments, we superimpose and disambiguate up to four subimages; see Sec. 5.) Without loss of generality, we can also assume that one subimage is stationary relative to the other. If $\mathbf{x} = [\mathbf{x}^{(1)}; \mathbf{x}^{(2)}]$ are the parameters corresponding to the two images, then $\mathbf{A}_t$ is the underdetermined matrix $[\mathbf{I} \;\; \mathbf{S}_t]$, where $\mathbf{I}$ is the identity matrix and $\mathbf{S}_t$ describes the

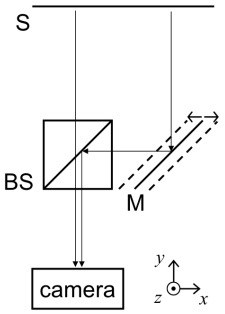


Fig. 2. Proposed camera architecture for superimposing two subimages. The scene (marked with the S) is split into two halves. The optical field from the left half propagates directly through the beamsplitter (marked with the BS) to hit the FPA in the camera. The optical field from the right half hits a movable mirror (marked with the M) before propagating to the beamsplitter and being reflected to the FPA in the camera.

movement of the second subimage in relation to the first. Here, we assume that $\mathbf{x}^{(1)}$ corresponds to the stationary subimage while $\mathbf{x}^{(2)}$ corresponds to the moving subimage. Then the above system can be modeled as

$$\mathbf{z}_t = [\mathbf{I} \;\; \mathbf{S}_t] \begin{bmatrix}\mathbf{x}^{(1)} \\ \mathbf{x}^{(2)}\end{bmatrix} + \mathbf{n}_t.$$

4. SPARSE REPRESENTATION ALGORITHMS

4.1. Optimization problem formulation

The above camera architecture results in a sequence of frames where each frame is a superposition of several subimages; this disambiguation problem can thus be formulated as a sequence of underdetermined inverse problems as in (1). Let $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ denote the vectors of coefficients for the two subimages in some basis, *e.g.,* the wavelet basis, so that

$$\mathbf{x} = \begin{bmatrix}\mathbf{x}^{(1)} \\ \mathbf{x}^{(2)}\end{bmatrix} = \begin{bmatrix}\mathbf{W}\boldsymbol{\theta}^{(1)} \\ \mathbf{W}\boldsymbol{\theta}^{(2)}\end{bmatrix} \equiv \widetilde{\mathbf{W}}\boldsymbol{\theta},$$

where $\widetilde{\mathbf{W}} \equiv \begin{bmatrix}\mathbf{W} & 0 \\ 0 & \mathbf{W}\end{bmatrix}$, $\mathbf{W}$ is the matrix corresponding to the inverse wavelet transform and $\boldsymbol{\theta} = [\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(2)}]$ corresponds to the wavelet basis coefficients for the two subimages. We use the wavelet transform here because of its effectiveness with many natural images, but alternative bases could certainly be used depending on the setting.

To solve the problem of disambiguating two superimposed images, we formulate it as the nonlinear optimization problem

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\| \mathbf{z}_t - [\mathbf{I} \;\; \mathbf{S}_t]\widetilde{\mathbf{W}}\boldsymbol{\theta} \right\|_2^2 + \tau \left\|\boldsymbol{\theta}\right\|_1. \tag{2}$$

The first term in the objective function is the least-squares error between the observation and the image reconstruction while the second term regularizes the problem and drives the small wavelet coefficients in the solution to zero, thus ensuring a sparse solution. Sparse solutions in the wavelet domain provide good solutions since the wavelet transform typically retains the majority of natural images'

energy in a relatively small number of basis coefficients. The regularization parameter τ is calibrated to the noise level of the observations or to the normalized ℓ^2-norm of the observations for nearly noiseless cases.

Note that if we solve (2) for each frame independently, then the inverse problem is underdetermined and ill-posed, but the ℓ^1 regularization term can lead to reasonably accurate solutions, particularly when the true scene is *very* sparse in the wavelet basis. However, small subsets of subsequent frames of observations can be used simultaneously to achieve significantly better solutions. We explore this using the following three methods:

Method 1. For a scene that changes only slightly from frame to frame, the reconstruction from a previous frame is often a good approximation to the following frame. In Method 1, we use the solution $\widehat{\boldsymbol{\theta}}$ to (2) at the t^{th} frame to initialize the optimization problem for the $(t+1)^{\text{th}}$ frame.

Methods 2. We can improve upon the Method 1 approach by solving for multiple frames simultaneously. In Method 2 we solve for two frames at a time, and the optimization problem becomes

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\| \begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}_{t+1} \end{bmatrix} - \begin{bmatrix} \mathbf{I} & \mathbf{S}_t \\ \mathbf{I} & \mathbf{S}_{t+1} \end{bmatrix} \widetilde{\mathbf{W}}\boldsymbol{\theta} \right\|_2^2 + \tau \|\boldsymbol{\theta}\|_1 , \qquad (3)$$

where $\mathbf{z}_{t+1}$ and $\mathbf{S}_{t+1}$ are the observation and shifting operator in the $(t+1)^{\text{th}}$ frame.

Method 3. Method 3 is very similar to Method 2, but we solve for $\boldsymbol{\theta}$ using four successive frames instead of two, including $\mathbf{z}_{t+2}$ and $\mathbf{z}_{t+3}$ in the observation vector and $\mathbf{S}_{t+2}$ and $\mathbf{S}_{t+3}$ in the observation operator matrix. By requiring $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ to satisfy more equations, the resulting linear system becomes less underdetermined or overdetermined, making the problem less ill-posed and the solutions more accurate. The drawback, however, is that the corresponding linear systems to be solved are larger and require more computation time.

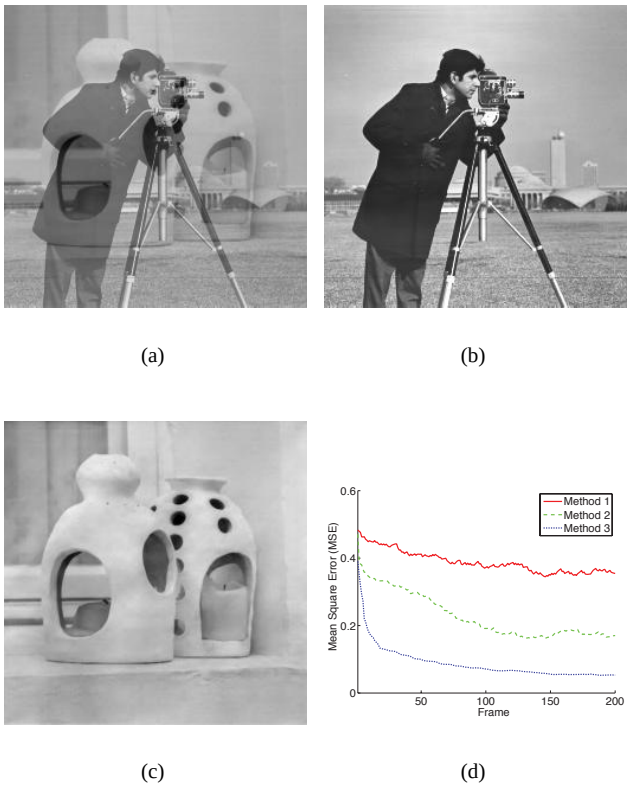
5. SIMULATION RESULTS

To verify that our optimization techniques are capable of disambiguating superimposed images and explore the tradeoffs associated with the above three methods, we perform two simulation studies, where superimposed videos were generated numerically from digital images. In this study, we use stationary scenes but note that the proposed approach will also be effective when motion within the scene is slow relative to the frame rate of the imaging system. More effective exploitation of inter-frame correlations for disambiguating moving scenes can yield additional improvements and is the subject of ongoing work.

In these experiments, we solve the optimization problems for the various proposed methods (e.g., (2) and (3)) using the Gradient Projection for Sparse Reconstruction (GPSR) algorithm of Figueiredo et al. [4]. GPSR is a gradient-based optimization method that is very fast, accurate, and efficient. In addition, GPSR has a *debiasing* phase, where upon solving the $\ell^2-\ell^1$ minimization problem, it fixes the non-zero pattern of the optimal $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ and minimizes the ℓ^2 term of the objective function, resulting in a minimal error in the reconstruction while keeping the number of non-zeros in the wavelet coefficients at a minimum. It has been shown to outperform many of the state-of-the-art codes for solving the $\ell^2-\ell^1$ minimization problem or its equivalent formulations. The computational bottleneck in GPSR is the multiplication by the mixing matrix $\mathbf{A}_t$ (and its transpose) in (1). In our setup, this can be performed very efficiently since the shifting operator $\mathbf{S}_t$ and the discrete wavelet transform are both $O(n)$, *i.e.*, the computational complexity is linear in the number of image pixels.

5.1. Simulation I: Two images superimposed

In the first experiment, the simplest case involving two distinct images is studied to quantitatively compare the performances of Methods 1-3 in terms of the execution time and the mean square error between the original and reconstructed images. Two images, "cameraman" and "lamp" (both 256×256 pixels in gray-scale), are shown in Fig. 3(a). In the composite video, the cameraman image is fixed while the lamp image is moved horizontally with some prescribed perturbations. We created a movie consisting of 200 frames, and added zero-mean white Gaussian noise. For each frame, we ran ten GPSR iterations and ten debiasing steps. The number of iterations was limited and the optimization algorithm was not allowed to run to convergence because of the real-time nature of the video applications for which we anticipate this approach would be most useful. As displayed in Fig. 3(a), the observed image is the sum of two different images, several details and other features are difficult to visualize, and it is not clear which features correspond to which image. Using the wavelet-based optimization method described above, however, very accurate reconstructions are possible, as displayed in Fig. 3(b) and (c); these reconstructions were computed using Method 3.



(a) (b) (c) (d)

Fig. 3. Numerical experiment I, with two superimposed images. (a) Observation $\mathbf{z}_0$. (b) Reconstructed cameraman image after 200 frames computed using Method 3; MSE = 0.031. (c) Reconstructed lamp image after 200 frames computed using Method 3; MSE = 0.023. (d) Plot of total MSE (cameraman + lamp) as a function of frame number.

Clearly, the separation of the two images is largely successful, and even the fine details of each image are preserved in the disambiguation. Subtle ghosting occurs in each image, resulting from the stark contrast in their intensities (the dark cameraman coat versus the bright lamp texture). Fig. 3(d) shows that the MSE values for each method decrease with frame number. They also show that Method 2 and 3 outperform Method 1, indicating that solving for more frames simultaneously per iteration leads to more accurate reconstruction. Furthermore, the MSE values for Method 3 decrease more rapidly initially than the other two methods, which is particularly important when the scene being recorded contains movement. Even though all three methods have the same initial MSE value, the MSE value for Method 3 by the 20th frame is already less than any of the MSE values of the other two methods. At the last frame, the MSE value for Method 3 was 0.053, a 69% improvement over that of Method 2. Finally, although the linear system in Method 3 is twice as large as the linear system in Method 2, Method 3 only took 21% more time: Method 1 took 404.5 sec to complete 200 frames, while Methods 2 and 3 took 467.3 sec and 566.6 sec, respectively. Thus, for the following simulation, we only used Method 3 for disambiguation.

5.2. Simulation II: Four quadrants of an image superimposed

The second numerical experiment consists of dividing a 512×512 image (Barbara) into four quadrants and superimposing them to form one 256×256 observation. To avoid parallel movements in this simulation, one image is held still, the second moves horizontally, the third vertically, and the last moves diagonally opposite the second and third images. These motion patterns were selected because they will be simple to implement in the proposed camera architecture described above. As in Simulation I, zero-mean white Gaussian noise was added to the observation.

The result of the second simulation shows that our technique can disambiguate the four superimposed quadrants to reconstruct the original Barbara image (Fig. 4(a)). Prominent features in the original image (*e.g.*, table, books, chair, and Barbara) were reconstructed without ambiguity. We note the low MSE value from about the 60^{th} frame on. We also note that the MSE is nearly monotonically decreasing in time and that the steep drop in MSE observed in the first simulation is present here as well. Like the first simulation, the disambiguation is not perfect. For example, the *average* pixel intensity of each quadrant cannot be distinguished using the proposed approach, producing artifacts at the boundaries of the four quadrants. This is particularly noticeable in the interface between the upper-left and lower-left quadrants. Also, while the striped fabric patterns exhibits some artifacts, many details are accurately reconstructed despite the large amount of fine-scale detail in this image.

6. CONCLUSIONS

In this paper, we propose a novel camera architecture for collecting high resolution, wide field-of-view videos in settings such as infrared imaging where large focal plane arrays are unavailable. This architecture is mechanically robust and easy to calibrate. Associated with this architecture is a fast and accurate technique for disambiguating the composite video image consisting of the superposition of multiple subimages. Simulation results demonstrate that our optimization approach can reconstruct the constituent images with small mean square errors, and that the errors decay rapidly as a function of frame number despite the very small number of optimization iterations allowed for each new frame. Ongoing work in this area includes disambiguating superimposed videos of slowly changing scenes by exploiting inter-frame correlations.

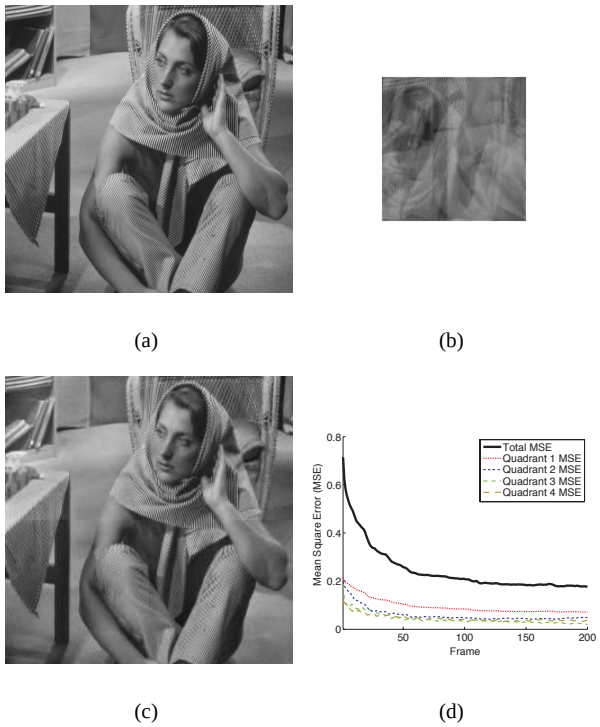


Fig. 4. Numerical experiment II, with four superimposed quadrants. (a) Original Barbara image. (b) Observation $\mathbf{x}_0$ consisting for four superimposed quadrants of the Barbara image. (c) Reconstructed image after 200 frames computed using Method 3; MSE = 0.177. (d) Plot of MSE of each quadrant and total image as a function of frame number.

7. REFERENCES

[1] W. R. Uttal, L. Spillmann, F. Stürzel, and A. B. Sekuler, "Motion and shape in common fate," *Vision Res*, vol. 40, no. 3, pp. 301–310, 2000.

[2] J. Bobin, J.-L. Starck, J. Fadili, and Y. Moudden, "Morphological diversity and source separation," *IEEE Transactions on Signal Processing*, vol. 13, no. 7, pp. 409–412, 2006.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61 (electronic), 1998.

[4] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, To appear.

[5] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," To be published in IEEE Transactions on Information Theory. http://www.acm.caltech.edu/ emmanuel/papers/OptimalRecovery.pdf, 2006.

[6] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ^1-norm minimization problems when the solution may be sparse," *Preprint*, 2006.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.