An Ensemble Prior of Image Structure for Cross-modal Inference*

S. Ravela A. Torralba W. T. Freeman
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139
ravela@mit.edu

Abstract

In cross-modal inference, we estimate complete fields from noisy and missing observations of one sensory modality using structure found in another sensory modality. This inference problem occurs in several areas including texture reconstruction and reconstruction of geophysical fields. We propose a method for cross-modal inference that simultaneously learns shape recipes between two modalities and estimates missing information by using a prior on image structure gleaned from the alternate modality. In the absence of a physical basis for representing image priors, we use a statistical one that represents correlations in differential features. This is done efficiently using a perturbation sampling scheme. Using just one example of the alternate modality, we produce a factorized ensemble representation of feature correlations that yields efficient solutions to large-sized spatial inference problems. We demonstrate the utility of this approach on cross-modal inference with depth and spectral data.

1. Introduction

Reconstruction of missing data in images arises as a key problem in several fields. In the geosciences, for example, one may have a digital elevation map (DEM) with missing depth values at several map locations. As shown in Figure 2(A), a large proportion of the DEM image lacks any data. In certain situations, this missing information can be reconstructed from an alternate modality. In this instance, Landsat data, which produces spectral measurements, can act as a surrogate for DEM in the sense that the visual structure in Landsat images are similar to the depth structure. The Landsat image corresponding to the DEM is shown in Figure 2(B). This image is spatially registered to the DEM data and has structures similar to what we see in the available DEM. Problems of this nature can be called as cross-modal inference problems.

We propose a solution to this spatial inference problem using state and parameter estimation. Parameter estimation refers to the fact that the relationship between the two modalities can be modeled by a function, with parameters that are estimated from available observations of each modality. State estimation refers to the fact that an unknown image, the state, is constructed by combining information from both sources by respecting their uncertainties. We combine these two classical estimation methods into a single framework. We do this because a model is necessary to relate the two modalities, and yet it is difficult to build the perfect model. By combining the two estimation schemes, we are able to compensate for model imperfections. This methodology is related strongly to inverse problems studied in the physical and computational sciences. In a Bayesian sense, our method can be viewed as a MAP estimate of state and parameters, whose uncertainties we model as Gaussian distributions.

The only source of fully-observed information comes from an alternate modality, often from a single image. A key step in this process is to capture the structural correlations in the alternate modality to produce the state estimates. We show that cross-correlation of differential features is a useful way to capture correlation between image structures. This feature cross-correlation forms a useful prior for inference by functioning as an empirical error-covariance of the prior uncertainty. We then show that the covariance can be factorized into a square-root form and therefore need not be computed explicitly. Further we show that a well-ranked square-root can be produced from a single image of the alternate modality, by perturbing it in space (position) and scale-space (smoothness). We call the idea of representing feature responses as a square-root of a prior covariance as an ensemble prior of image structure. By doing so, we are able to solve the estimation problem that would otherwise have been computationally prohibitive. In contrast to other inference methods that simplify complexity of spatial interactions by restricting connectivity in space (e.g. a pair-

*Funding was provided by a grant from Shell Research and NSF Grant 0121182.

wise MRF or a tree), our square-root form is a reduced rank representation of the prior's error covariance, and one that preserves long-range correlations in images implicitly.

This framework is based on certain assumptions. First, we suppose that preserving long-range correlations is useful. This would be if we want to spread the adjustments that sparse observations of the primary modality prompt to other locations whose appearance correlates well with it. Second, we assume that correlations in differential features can be a good representation of the similarity we see in visual structures. This assumption is reasonable for differential features have had a long history of use in modeling visual appearance. Third, we assume that differential features are computed at a scale at which correlations are useful. Fourth, we assume that both modalities are spatially registered and if they aren't some algorithm has been applied to register them. This will be demonstrated in examples. Specifically, we demonstrate the cross-modal inference solution on for inferring DEM from sparse observations of DEM and complete observations of Landsat.

2 Related Work

The inference problem presented here is related to several research threads. Techniques developed for various applications including texture synthesis, texture infilling, inpainting, image quilting [3] and image analogies [9] are related to the present work. In texture synthesis, Heeger and Bergen [8] used histograms of filter responses at multiple scales and orientations. This is useful for representing random textures, but less so for structured textures. De Bonet [2] uses a multiscale procedure that randomizes the fine scale texture patches in a way that preserves the conditional distribution of filter outputs over coarser scale outputs. Portilla and Simoncelli [10] algorithm for texture synthesis uses joint statistics of first and second order wavelet coefficients.

Example-based methods for synthesis, super-resolution and transfer are also related to the present work. We borrow the idea of using the best match as the first guess to the inference problem (see 4) from this work. Efros and Leung [4] synthesize textures one pixel at a time by finding closely matching neighborhoods and randomly choosing one. The evolving context constrains the new choices and provides sufficient "continuity" of the texture, but this algorithm is quite expensive/slow. In example-based super-resolution [6] a low-resolution image is used to find similar patches in a dictionary, whose high resolution counterparts are used for inference on a Markov network. Although this isn't directly related, example-based schemes need to solve a spatial inference problem. The Markov assumption in space, in particular, leads to a simplification that stands in contrast to what we do, rank reduction. We posit that long range correlations are easier to represent in our scheme.

Work in image-infilling is related, but only from an application point of view. In one PDE-based infilling framework, Bertalmio et al. [1], combine Efros and Leung's [4] texture synthesis scheme with an advection scheme. Advection is modulated by the image laplacian, which the isophotes and image brightness values are propagated. Although there is, in general, a relationship between PDE methods and variational formulations, what we propose is different.

From a methodological perspective, the proposed technique is close to the ensemble Kalman filter [5]. As noted in the introduction, our technique may be viewed as restricting rank instead of restricting connectivity in space or scale. Therefore, non-parametric inference on a Markov graph also bears some relation to the proposed method. Our technique is parametric because we assume the distributions are Gaussian. But the parameters of the Gaussians are obtained statistically. A low-rank factorization allows us to solve estimation quite quickly. Of course, if one were to disallow long-range correlations using a decaying kernel, then the proposed method can, in principle, be sped up using a multipole expansion [7]. The proposed method can also be implemented in a multiresolution manner, but these extensions are not a subject of this paper.

3 Cross-modal Fusion

One way to solve the cross-modal inference problem is via parameter estimation.

We can suppose that we have an observed image of modality B, called $Y^{(b)}$. This image is assumed to be incomplete and noisy, and therefore we may model the observation via the equation:

$$Y^{(b)} = H\ X^{(b)} + \eta \tag{1}$$

This equation states that the observations can be obtained from a true *state* (or image) of modality B, $X^{(b)}$, using additive noise η at few locations indicated by the binary incidence matrix H.

We also assume we have an observation of modality A called $Y^{(a)}$, and an associated state $X^{(a)}$. We model the relationship between the two modalities as a function of their states

$$X^{(b)} = f(X^{(a)}; r) \tag{2}$$

Where r is a vector of unknown parameters. This function is deterministic.

We can write the inference problem by characterizing the posterior density $P(X^{(b)}, X^{(a)}, r|Y^{(b)}, Y^{(a)})$. This can be written via Bayes rule as

$$P(X^{(b)}, X^{(a)}, r|Y^{(b)}, Y^{(a)}) \quad \propto \quad P(Y^{(b)}|X^{(b)})$$

$$P(X^{(b)}|X^{(a)}, r)$$
$$P(X^{(a)}|Y^{(a)})$$
$$P(r) \quad (3)$$

We assume that we observe modality A perfectly, that is $P(X^{(a)}|Y^{(a)}) = \delta(Y^{(a)})$. We also assumed that the model between modality A and B is deterministic, that is $P(X^{(b)}|X^{(a)}, r) = \delta(X^{(a)}, r)$ With these assumptions, it follows that the inference problem is

$$P(r|Y^{(a)}, Y^{(b)}) \propto P(Y^{(b)}|Y^{(a)}, r)P(r) \quad (4)$$

This is a parameter estimation problem and can be solved by regression. If we model η as an i.i.d. random variable, that is $\eta \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and assume we have a uniform prior on r, then the solution $\hat{r}$ is one that minimizes the norm $J_o(r) = \frac{1}{\sigma^2}||Y^{(b)} - Hf(Y^{(a)}; r)||$.

We suppose that the model f is linear. In particular, we model it as a convolution, and write $X^{(b)}[p] = (X^{(a)} \star a)[p] + b$ following Torralba and Freeman [11]. By writing the vector $r = [a \quad b]^T$ this model can be written as $f(X^{(a)}) = \xi r$, where ξ is constructed from $X^{(a)}$ and r is a vector of parameters. We can also construct a convolution matrix from a and write the linear model as $f(X^{(a)}) = AX^{(a)} + B$, where B is a vector of replicated b values.

If models were perfect, then parameter estimation would suffice, but this is often not the case. For example, Figure 2(C) shows the result of applying a linear model to a Landsat image. As can be seen in an overlay of DEM data on the reconstructed output, in Figure 2(D), the reconstruction is smooth and captures some variability, but there is a significant difference that is not just white noise. We posit that by combining state estimation with parameter estimation we can address model imperfections.

3.1 State and Parameter Estimation

The motivation for the proposed approach is that by making adjustments to the state $X^{(a)}$ in some consistent way, we may be able to compensate for model imperfections. This immediately suggests that the conditional prior $P(X^{(a)}|Y^{(a)})$ is no longer degenerate. We can expand equation 3 now to be:

$$P(X^{(a)}, r|Y^{(a)}, Y^{(b)}) \propto P(Y^{(b)}|X^{(a)}, r)P(X^{(a)}|Y^{(a)}) \quad (5)$$

We have once again assumed that the model f is deterministic and we have a uniform prior on r. If we assume the conditional prior is Gaussian, we can derive a quadratic objective:

$$J(X^{(a)}, r) =$$
$$(Y^{(b)} - H(AX^{(a)} + B))^T C_{bb}^{-1}(Y^{(b)} - H(AX^{(a)} + B))$$
$$+ \quad (X^{(a)} - Y^{(a)})^T C_{aa}^{-1}(X^{(a)} - Y^{(a)}) \quad (6)$$

Here $C_{bb} = \sigma^2\mathbf{I}$ because we assume the noise to be iid. To solve this objective we compute the Euler-Lagrange equations, which lead to the following coupled equations. The first is parameter estimation:

$$r_i = (H\xi_i)^{-1}Y^{(b)} \quad (7)$$

The second is state estimation:

$$\begin{aligned} X_{i+1}^{(a)} &= X_i^{(a)} \\ &+ C_{aa}H^T(HC_{aa}H^T + C_{bb})^{-1} \\ &\quad (Y^{(b)} - H(A_iX_i^{(a)} + B_i))) \end{aligned} \quad (8)$$

The subscript i is used to denote the fact that these coupled equations must be iterated. Starting at $i = 0$, we set $X_0^{(a)} = Y^{(a)}$, solve equation shape to compute r_0. Then we rewrite it as A_0, B_0 and estimate $X_1^{(a)}$. Then we iterate until the objective does not improve. The final solution is $f(X_n^{(a)}; r_n)$ at some iteration n

The matrix C_{bb} is the covariance of the noise model and is conveniently assumed to be diagonal and much smaller in energy than C_{aa}, indicating high confidence in modality B data where it is observed. The matrix C_{aa} can be viewed as a conditional Gaussian prior on modality A. But just what is this prior supposed to be?

3.2 Ensemble Priors

We would like to construct C_{aa} so that it captures the correlation between structures in the image. This way, sparse measurements can be used to update unobserved state elements. If we suppose that visual structures can be represented by some features computed from the image, then we can represent C_{aa} using the feature correlations. One set of features that prove to be good representation of image structures is the differential structure of the image. As an example lets think of an image as a one-dimensional vector X, its differential structure is $\mathcal{J}_\mathcal{N} = [\frac{\partial X}{\partial p} \quad \frac{\partial^2 X}{\partial p^2} \cdots \frac{\partial^N X}{\partial p^N}]$ to some order N. Please note that here $\mathcal{J}_\mathcal{N}$ is a matrix of size $n \times N$ where each column of size n is the vector of spatial derivatives. In practice we restrict the derivatives to the first two orders, that is use $\mathcal{J}_2$.

So we can now think of $C_{aa} = \mathcal{J}_2\mathcal{J}_2^T$([1]). Computing C_{aa} explicitly in this manner is both space and hence time consuming. For example if the vector X is of size 360000, a 600x600 image, then C_{aa} is $360,000 \times 360,000$. It is impractical to construct this matrix, let alone inverting the

[1] We ignore the normalizing denominator in the rest of the paper. To be sure, For $\mathcal{J}_\mathcal{N}$ is N-1 and $\mathcal{J}_2$ is 1. It eventually cancels out in Equation 8

innovation covariance $HC_{aa}H^T + C_{bb}$ in equation 8, or using it to solve a linear system. We need a factorization that can be exploited usefully.

We will show that the innovation covariance can be factorized. $\mathcal{J}_2$ represents the *square root* form of the covariance, so one already has a nice factorization of C_{aa}. We can also factorize e C_{bb}. To do this construct an *ensemble* of observations arranged in a matrix $\mathcal{Y}^{(b)}$. Each column of this matrix is obtained by perturbing $Y^{(b)}$ using the noise model. We need as many columns as $\mathcal{J}_2$ has. Then $C_{bb} = \tilde{\mathcal{Y}}^{(b)}(\tilde{\mathcal{Y}}^{(b)})^T$ where $\tilde{\mathcal{Y}}^{(b)}$ is a matrix of observation deviations. If we assume that the observations are uncorrelated, we can write a square-root for the innovation as $C = H\mathcal{J}_2 + \tilde{\mathcal{Y}}^{(b)}$ and express the posterior covariance $HC_{aa}H^T + C_{bb} = CC^T$. The matrix C is small, and singular value decomposition $C = USV^T$ can be used to invert CC^T, without ever computing the latter. This appears to be a nice scheme. The key here was to use the square-root form of the observational noise using samples from the distribution that observations are drawn from. Since C_{bb} is diagonal it is easy to draw random samples, so we don't have a sampling problem per se.

But there is a problem! The number of columns in $\mathcal{J}_2$ is too small (2 for 1D or 5 for 2D); and so C is a 360000×5 matrix. Its rank is too low to be useful. We are faced with a dilemma. Computing the innovation covariance directly is nearly impossible, it is too big. The square-root form C may be too rank deficient unless we use "lots of features" but yet, intuitively we can argue that even the gradient correlations $\mathcal{J}_1\mathcal{J}_1^T$ must provide useful information about image structures to construct a prior covariance.

We propose a solution that represents C_{aa} using ensembles that are perturbations in some space (to be discussed) and its construction is similar to the discussion for C_{bb}. Let's start with the first derivative. Observe that $\mathcal{J}_1\mathcal{J}_1^T$ can be computed statistically. To see this, write the observed image of modality A as $Y^{(a)}(p)$ and consider a truncated Taylor expansion for a perturbation in position Δ. This perturbation is a normal random variable of two dimensions with mean 0 and some (user specified) standard deviation. All pixels are perturbed by the same amount. The expansion can be written as

$$Y^{(a)}(p+\Delta) - Y^{(a)}(p) \approx \Delta^T \frac{\partial Y^{(a)}}{\partial p} \tag{9}$$

The mean of this deviation is zero, because $E[\Delta^T \frac{\partial Y^{(a)}}{\partial p}] = 0$. We are now in a position to compute sampled representations that represent the gradient correlations of the image. We assemble the ensemble matrix $\mathcal{Y}_1^{(a)} = [Y^{(a)}(p+\Delta_0)\ \ Y^{(a)}(p+\Delta_1)\ldots Y^{(a)}(p+\Delta_N)]$. Its covariance is $\tilde{\mathcal{Y}}_1^{(a)}(\tilde{\mathcal{Y}}_1^{(a)})^T$. This covariance, by construction, contains the correlation of the first derivative responses. Similarly, by generating an ensemble matrix $\mathcal{Y}_2^{(a)}$ comprising of images that are blurred versions of each other, a covariance matrix that represents the correlations of the image laplacian is generated. The prior covariance $C_{aa} \propto (\tilde{\mathcal{Y}}_2^{(a)}(\tilde{\mathcal{Y}}_2^{(a)})^T + \tilde{\mathcal{Y}}_1^{(a)}(\tilde{\mathcal{Y}}_1^{(a)})^T))$. This can be generated from a square-root $\mathcal{Y}^{(a)} = [\mathcal{Y}_1^{(a)}\ \ \mathcal{Y}_2^{(a)}]$ as $C_{aa} = \tilde{\mathcal{Y}}^{(a)}(\tilde{\mathcal{Y}}^{(a)})^T$.

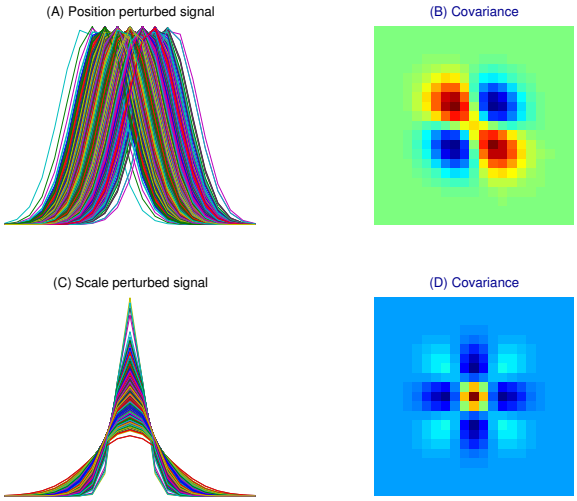

Figure 1. Ensemble generation with position and scale perturbations. The top-left image (A) shows a number of Gaussian signals that are displaced from one another. It is easy to see that their covariance, shown in image (B) captures the outer-product of the first derivative of the signal, or the correlation of the first derivative. Similarly, the second derivative correlations can be obtained by perturbing in scale (or smoothness) as shown in (C) and (D).

Figure 1 depicts an ensemble prior for a simple example. A one-dimensional Gaussian is perturbed in position, shown in image (A) and its covariance is a good representation of the outer-product of the derivative, shown in image (B). Image (C) shows a "scale" perturbation and its covariance, shown in image (D), represents the outer-product of the second derivative well. Thus an ensemble representation of feature correlations is generated.

As noted, the inversion of $HC_{aa}H^T + C_{bb}$ can be conducted from the square-root $C = H\ \tilde{\mathcal{Y}}^{(a)} + \tilde{\mathcal{Y}}^{(b)}$. If the prior square-root is generated using N samples each of displacement and scale-space (smoothness) perturbations, this matrix is of size $m \times 2N$, where there are m observed locations. The (pseudo) inverse of the innovation covariance CC^T then is calculated directly from the *square-root* C using SVD. That is $C = USV^T$ and $(CC^T)^{-1} = U_{\#}S_{\#}^2U_{\#}^T$ where $S^{\#}(i) = 1/S(i)$, $i = 1\ldots k$. $U_{\#} = U(:, 1:k)$. That is the k leading singular values and vectors are cho-

sen. Also note that $k \leq N << n$, and k is a tunable parameter. In experiments, we typically use 50 to 200 samples and have let k to be the same.

Using this ensemble prior the second Euler-Lagrange equation 8 can be evaluated. To do so, note that it can be written as

$$\begin{aligned} X_{i+1}^{(a)} &= X_i^{(a)} \\ &+ \tilde{\mathcal{Y}}^{(a)} \left(H(\tilde{\mathcal{Y}}^{(a)}) \right)^T (U_\# S_\#^2 U_\#^T) \\ &\quad (Y^{(b)} - H(A_i X_i^{(a)} + B_i))) \end{aligned} \tag{10}$$

A right to left multiplication solves this equation efficiently.

3.3 The Algorithm

We are now in a position to describe an algorithm for simultaneous state and parameter inference. The steps of this algorithm are

1. Inputs: $Y^{(a)}, Y^{(b)}$ and σ, the standard deviation of $Y^{(b)}$.
2. Assemble the ensemble matrix $\mathcal{Y}^{(a)}$. Assemble the ensemble matrix $\mathcal{Y}^{(b)}$ and therefore C. Compute the truncated singular values and vectors of the innovation square-root C.
3. Set i=0. Set $X_i^{(a)}$ to be $Y^{(a)}$.
4. Solve Equation 7 and compute the parameters r_i and equivalently A_i and B_i
5. Solve Equation 8 and compute an updated $X_{i+1}^{(a)}$, using Equation 10.
6. Set i = i+1. Repeat the last two steps until convergence.

4 Application of Methodology

We demonstrate cross-modal inference using an example illustrated in Figure 2. The top-left image (A) is DEM with lots of missing data. The Landsat image is shown in (B) and the two are spatially registered. They are each of size 600×600. A displacement of standard deviation 10 pixels and deviation in scale of 2 around a mean of 3 is used to generate a total of 200 samples. The parameter estimation or shape recipe output, without any state estimation is shown in (C). It is overlaid with the true DEM data in (D). This overlay replaces an estimated pixel with the observed DEM pixel, where available. This shows that the relation between DEM and Landsat is not completely captured.

Image (E) shows the output of the state and parameter estimation. An overlay of true DEM on the synthesized DEM (F) shows that the estimate is much better. This is also indicated by a reduced estimation error (not shown).

This example demonstrates that combining state and parameter estimation is useful.It also shows that correlation in differential structure can be useful for state estimation. It should be noted that it is important to capture the feature scales well. Because using differential features (whether by sampling or directly) includes some smoothing, we think this algorithm is useful for smoothly varying fields, such as the ones shown, and found in many geophysical problems. It should also be noted that good estimates of state depend on the balance between the prior uncertainty and the likelihood (noise). If we set the observation uncertainty to be too high, then the filter can depart from the observations and essentially "paste" the first guess. On the other-hand, if the observational uncertainty is set too low (in the limit, 0), then the filter will depart too, because the state will stick to the observations where they are available, but do little elsewhere.

5 Conclusions and Future Work

In this paper, a Bayesian formulation of the cross-modal inference problem is developed. The solution simultaneously learns a shape recipe model and estimates states. A statistical basis is used to represent the covariance of the Gaussian distribution representing the conditional prior. In particular, the prior is derived from a perturbation of approximation to the differential image structure. This allows us to produce a well-conditioned factorization of the prior's covariance that yields fast solutions. We are interested in developing multiscale versions of this inference problem as well as exploring other perturbation models appropriate for inference tasks in vision.

Acknowledgment

Data sets were provided by Shell including Landsat data from NOAA and the DEM data from SRTM/USGS. We thank particular efforts by Dr. Edward Bigert and Dr. Sandhya Devi for helping commission this work.

References

[1] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. In *Proc. CVPR*, 2003.

[2] J. S. D. Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In *SIGGRAPH*, 1997.

[3] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, 2001.

[4] A. A. Efros and T. Leung. Texture synthesis by nonparamertic sampling. In *Proc. ICCV*, 1999.

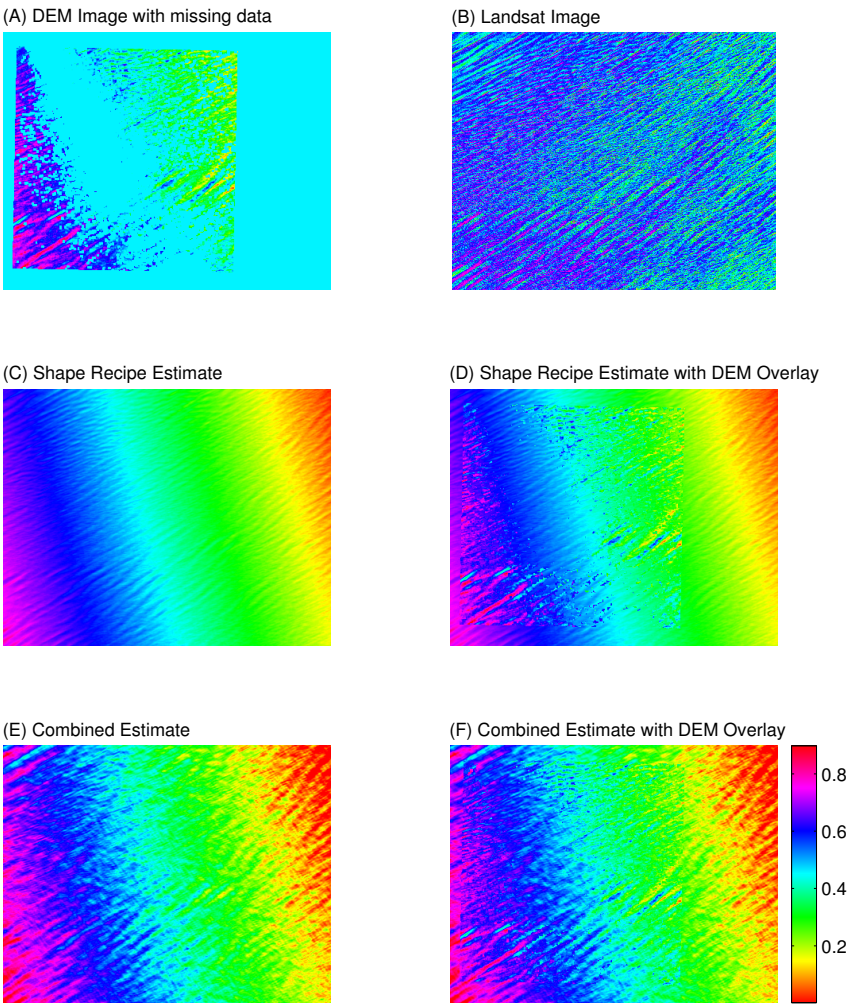

Figure 2. Cross-modal fusion on DEM and Landsat images. These color coded images illustrate the performance of the cross-modal fusion algorithm. Image (D) depicts an overlay of DEM data on shape recipe output. Every pixel where DEM data is available is replaces the recipe estimate. Similarly F shows the overlay of DEM on the cross-modal estimate.

[5] G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

[6] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002.

[7] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of computational physics*, 73(2):325–348, 1987.

[8] D. Heeger and J. Bergen. Pyramid-based texture analysis/synthesis. In *SIGGRAPH*, 1995.

[9] A. Hertzmann, C. E. Jacobs, N. Oliver, V. Curless, and D. Salesin. Image analogies. In *SIGGRAPH*, 2001.

[10] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000.

[11] A. Torralba and W. T. Freeman. Properties and applications of shape recipes. In *CVPR*, 2003.