Multi-stream video fusion using local principal components analysis

Ravi K. Sharma

Department of Electrical and Computer Engineering
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000

Misha Pavel

Department of Electrical and Computer Engineering
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000

Todd K. Leen

Department of Computer Science and Engineering
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000

ABSTRACT

We present an approach for fusion of video streams produced by multiple imaging sensors such as visible-band and infrared sensors. Our approach is based on a model in which the the sensor images are noisy, locally affine functions of the true scene. This model explicitly incorporates reversals in local contrast, sensor-specific features and noise in the sensing process.

Given the parameters of the local affine transformations and the sensor images, a Bayesian framework provides a maximum a posteriori estimate of the true scene. This estimate constitutes the rule for fusing the sensor images.

We also give a maximum likelihood estimate for the parameters of the local affine transformations. Under Gaussian assumptions on the underlying distributions, estimation of the affine parameters is achieved by local principal component analysis. The sensor noise is estimated by analyzing the sequence of images in each video stream. The analysis of the video streams and the synthesis of the fused stream is performed in a multiresolution pyramid domain.

1. INTRODUCTION

Computational vision systems that provide visual guidance or are used in tasks such as detection and recognition need to be robust with respect to unpredictable environmental conditions. One way to approach this problem is to deploy multiple sensors, each specialized for a different environmental condition.[1–3] This approach is becoming increasingly popular due to the advances in sensing devices as well as the increase in available computing power. In addition, using multiple sensors can increase reliability with respect to single sensor systems.[4,1]

The application that we consider is the Autonomous Landing Guidance (ALG) system[5,6] in aviation. The term autonomous landing guidance refers to the use of synthetic or enhanced vision systems for landing aircraft autonomously in inclement weather. Another system that is used for autonomous landing of aircraft in bad weather

Other author information: (Send correspondence to R.K.S.)
R.K.S.: E-mail: ravi@ece.ogi.edu
M.P. : E-mail: pavel@ece.ogi.edu
T.K.L.: E-mail: tleen@cse.ogi.edu

Part of the SPIE Conference on Infrared Technology and Applications XXIV • San Diego, California • July 1998
SPIE Vol. 3436 • 0277-786X/98/$10.00

is the Instrument Landing System (ILS). The ILS is expensive to install as well as maintain, and is available at only a few big airports. Even with precision ILS, the pilot has no direct information about the positions of other aircraft or objects. Air traffic experiences delays because the air traffic control enforces greater separation between aircraft under such situations. At airports which do not possess ILS, operations cease completely when the visibility conditions drop below a specified minimum. The goal of ALG is to use multiple imaging sensors to provide visual guidance to pilots for landing the aircraft in low visibility conditions. The aim is to show the landing scene to the pilot on a suitable display in the cockpit. Such a system would support operation until visibility conditions drop below a much lower minimum, resulting in significant benefits to airlines and passengers. Since the equipment for ALG (sensors, processing modules) would be on the aircraft, it would increase the safety of operation at smaller airports. In addition, the ALG system would permit shorter separation between aircraft and at the same time enable the pilot to verify clear runway conditions.

Although, several different sensors have been studied for use in ALG, visible-band, infrared (IR) and millimeter wave (MMW) radar based imaging sensors are the most common. IR sensors can produce relatively high resolution images when imaging at night as well as through haze and some types of fog. MMW radar based imaging sensors can penetrate fog and have the least attenuation in rain. In the ALG application one needs to combine or *fuse* the sequences of video frames from the different sensors to generate one composite video stream. The fused video stream can then be displayed to the pilot, and would ideally provide all the necessary visual information for landing the aircraft safely.

If imagery from multiple sensors is similar then the most effective approach to fuse the images would be to employ some type of averaging. However, images from different sensors have different characteristics based on the underlying physical phenomena. The polarity of local contrast is often reversed between visible-band and IR images.[7–9] Infrared images may contain significant features that are different from visible-band images, say due to thermal differences caused by shadows. We call such sensor-specific features complementary features.[7] The noise characteristics of the sensors are also different. Fig. 1(a) and 1(b) are visual-band and IR images respectively, of a runway scene. Examples of local polarity reversal (marked rectangle) and complementary features (marked circle) are shown. These effects pose difficulties for fusion. In addition to these problems, image data from different sensors may have different geometric representations and may be mis-registered. In this paper we assume that images from the video sequences have been appropriately pre-processed such that they are in the projective representation[10] and are perfectly registered.[11]

2. BACKGROUND

In recent years, several image fusion techniques have been proposed. Direct methods such as pixel-wise averaging operate on pixels of sensor images to obtain the fused image. Averaging is optimal when images are noisy,[12] but causes cancellation of features when there are local polarity reversals, and reduced contrast when there are complementary features. Averaging is shown in Fig. 1(c).

Feature-based methods[13,8,14] use selection as the criterion for fusion and consist of three steps:

1. each sensor image is decomposed into features, such as a multiresolution pyramid (e.g. contrast pyramids, Laplacian pyramids, gradient pyramids),

2. a fused pyramid is constructed by selecting the most salient pyramid coefficient (based on, for example, maximum magnitude, local energy) at each pyramid location, and

3. the inverse pyramid transform is applied to synthesize the fused image.

Since features are selected rather than averaged, they are rendered at full contrast as shown in Fig. 1(d). However, techniques that are based on selection also have drawbacks. They do not explicitly model the sensor noise, and may thus heavily weight large noise spikes in the fused image. In addition, these techniques do not have the ability to adapt to changing sensor characteristics or environmental conditions. Finally, these techniques lack a formal mathematical basis for performing fusion.

More recently, a model-based approach[15] to fusion was proposed to overcome some of the drawbacks of the conventional fusion techniques that employ averaging and selection. This approach consisted of a framework that explicitly modeled the formation of the sequence of sensor images from the true scene, including the effects of sensor noise. This image formation model being probabilistic, could be inverted via Bayes' rule to arrive at the maximum a posteriori estimate of the true scene given the sensor images. This estimate of the true scene provided the rule for fusing the sensor images.

One limitation of this approach is that it requires a reference image for estimating the model parameters. In the absence of an actual reference (such as a rendered image from a terrain database), one of the sensors (e.g. visible-band sensor) was considered to be the primary sensor and used as a reference for computing the model parameters.

In this paper we present an extension to the probabilistic model-based approach. We review the image formation model and the Bayesian fusion in Section 3. In Section 4, we describe a method to estimate the model parameters using the sensor images. This method is based on local principal component analysis (PCA) of the sensor image data. We show that under some simplifying assumptions the fusion rule is closely related to local PCA of the sensor images. Our technique uses the Laplacian pyramid representation,[16] with step (2) described above replaced by our probabilistic fusion. Examples of fusion using this method are shown in Section 5.

3. PROBABILISTIC MODEL-BASED FUSION

Analysis of images of the same scene obtained from multiple sensors shows that the relationships between image features (caused by objects or patterns in the images) can be categorized into four main components[7] — common features (including polarity reversed features), complementary features, irrelevant features and noise. To maximize the benefits obtained from using multiple sensors, a fusion algorithm should enhance and display the common features, highlight the complementary features and suppress or eliminate noise and irrelevant features.

The image formation model provides a framework to incorporate knowledge about the formation of the sensor images and the relationships between the features.

3.1. The image formation model

The true scene, denoted s, gives rise to a sensor image through a non-linear and non-invertible mapping. For ALG, s would be an image of the landing scene under conditions of uniform lighting and unlimited visibility. We approximate the mapping between the true scene and the sensors by a local affine transformation defined at every hyperpixel of the Laplacian pyramid*.

$$a_i(\vec{l},t) = \beta_i(\vec{l},t)s(\vec{l},t) + \alpha_i(\vec{l},t) + \epsilon_i(\vec{l},t) \tag{1}$$

where,

$i = 1, \dots, q$ indexes the sensors,
$\vec{l} \equiv (x, y, k)$ is the hyperpixel location, with x, y the pixel coordinates and k the level of the pyramid,
t is the time,
a is the sensor image,
s is the true scene,
α is the non-random component (bias) that includes complementary features,
β is the sensor gain (which includes the effects of local polarity reversals), and
ϵ is the (zero-mean) sensor noise with diagonal covariance $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$.

In matrix form,

$$\boldsymbol{a} = \boldsymbol{\beta} s + \boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{a} = [a_1, a_2, \dots, a_q]^{\mathrm{T}}$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_q]^{\mathrm{T}}$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_q]^{\mathrm{T}}$, s is a scalar and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, ..., \epsilon_q]^{\mathrm{T}}$. The reference to location and time has been dropped for simplicity and will not be made explicit henceforth unless necessary.

Although this image formation model is only an approximation, it incorporates most of the components mentioned above. Since the image formation parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and the sensor noise covariance $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$ can vary from hyperpixel to

*The Laplacian pyramid transform decomposes an image into multiple levels such that each successive level is a band-passed subsampled and scaled version of the original image. The term hyperpixel refers to the value of the transform coefficient at a particular pixel of a particular level of the pyramid hyperpixel, the model can express local polarity reversals, complementary features and spatial variation of sensor gain. In addition, the model explicitly considers the noise characteristics of each sensor over time.

The model is defined for each hyperpixel of every video frame. We do assume, however, that the image formation parameters and sensor noise distribution vary *slowly* with location. Specifically, the parameters vary slowly on the spatio-temporal scales over which the true scene s may exhibit large variations. Hence, a particular set of parameters is considered to hold true over a spatial region of several square hyperpixels. We use this assumption implicitly when we estimate these parameters from data.

The model (2) fits the framework of the factor analysis model in statistics.[17,18] Here the hyperpixel values of the true scene s are the latent variables or common factors, $\boldsymbol{\beta}$ contains the factor loadings, and the sensor noise $\boldsymbol{\epsilon}$ values are the independent factors. Estimation of the true scene is equivalent to estimating the factors from the observations $\boldsymbol{a}$.

3.2. Bayesian estimation of true scene

Given the sensor intensities $\boldsymbol{a}$, we estimate the true scene s by appeal to a Bayesian framework. We assume that the probability density function of the latent variables s is a Gaussian with (locally varying) mean s_0 and (locally varying) variance σ_s^2 . The noise density is also assumed to be Gaussian with zero mean and a (locally varying) diagonal covariance $\boldsymbol{\Sigma_\epsilon} = \mathrm{diag}[\sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2, \ldots, \sigma_{\epsilon_q}^2]$. The density on the sensor images, conditioned on the true scene is,

$$\mathcal{P}(\boldsymbol{a}|s) = \frac{1}{(2\pi)^{\frac{q}{2}}\,|\boldsymbol{\Sigma_\epsilon}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\beta} s - \boldsymbol{\alpha})^{\mathrm{T}} \boldsymbol{\Sigma}_\epsilon^{-1} (\boldsymbol{a} - \boldsymbol{\beta} s - \boldsymbol{\alpha})\right] \tag{3}$$

The marginal density on $\boldsymbol{a}$ is

$$\begin{aligned} \mathcal{P}(\boldsymbol{a}) &= \int \mathcal{P}(\boldsymbol{a}|s)\mathcal{P}(s)ds \\ &= \frac{1}{(2\pi)^{\frac{q}{2}}\,|\mathbf{C}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\beta} s_0 - \boldsymbol{\alpha})^{\mathrm{T}} \mathbf{C}^{-1} (\boldsymbol{a} - \boldsymbol{\beta} s_0 - \boldsymbol{\alpha})\right] \end{aligned} \tag{4}$$

where $\mathbf{C}$ is the model covariance given by

$$\mathbf{C} = \boldsymbol{\Sigma_\epsilon} + \sigma_s^2 \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} \tag{5}$$

Finally, the posterior density on s, given the sensor data $\boldsymbol{a}$, is obtained by Bayes' rule:

$$\begin{aligned} \mathcal{P}(s|\boldsymbol{a}) &= \frac{\mathcal{P}(\boldsymbol{a}|s)\mathcal{P}(s)}{\mathcal{P}(\boldsymbol{a})} \\ &= \frac{1}{(2\pi\mathbf{M}^{-1})^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(s - \mathbf{M}^{-1}\left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{a} - \boldsymbol{\alpha}) + \frac{s_0}{\sigma_s^2}\right\}\right)^{\mathrm{T}} \mathbf{M}\left(s - \mathbf{M}^{-1}\left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{a} - \boldsymbol{\alpha}) + \frac{s_0}{\sigma_s^2}\right\}\right)\right] \end{aligned} \tag{6}$$

where the posterior covariance is,

$$\mathbf{M}^{-1} = \left[\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{\beta} + \frac{1}{\sigma_s^2}\right]^{-1} \tag{7}$$

The maximum a posteriori (MAP) estimate, $\hat{s}$, of the true scene is obtained by maximizing the posteriori density with respect to s, which for our Gaussian distributions is simply the posterior mean.

$$\hat{s} = \left[\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{\beta} + \frac{1}{\sigma_s^2}\right]^{-1} \left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{a} - \boldsymbol{\alpha}) + \frac{s_0}{\sigma_s^2}\right\} \tag{8}$$

This is our rule for fusion. For two sensors, it reads

$$\begin{aligned} \hat{s} &= \frac{\dfrac{\beta_1(a_1 - \alpha_1)}{\sigma_{\epsilon_1}^2} + \dfrac{\beta_2(a_2 - \alpha_2)}{\sigma_{\epsilon_2}^2} + \dfrac{s_0}{\sigma_s^2}}{\dfrac{\beta_1^2}{\sigma_{\epsilon_1}^2} + \dfrac{\beta_2^2}{\sigma_{\epsilon_2}^2} + \dfrac{1}{\sigma_s^2}} \\ &= \sum_{i=1}^{2} w_i(a_i - \alpha_i) + w_0 s_0 \end{aligned} \tag{9}$$

The fused image $\hat{s}$ is a weighted linear combination of the sensor images. The weights w_i change from hyperpixel to hyperpixel and through time as a result of the spatio-temporal variations in $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_\epsilon$ and σ_s^2.

As an example, in the case where the second sensor has a polarity reversal, β_2 is negative and the two sensor contributions are properly *subtracted.* On the other hand if the polarity is the same, then the two sensor contributions are added. Thus, in either case, averaging is performed using the correct polarity to add or subtract the sensor contributions. Now consider a case where a feature is missing from sensor 1. This corresponds to $\beta_1 = 0$. The model compensates by accentuating the contribution from sensor 2. In this case the result is same as selection of that sensor which contains the feature. Finally, consider the case where the sensors are noisy. If the first sensor has high noise (large $\sigma_{\epsilon_1}^2$), its contribution to the fused image is attenuated. At the same time, the contribution of the second sensor is increased.

The probabilistic framework provides an opportunity to use a terrain database to specify priors on the local distribution of s. This prior information can then be included in the MAP estimate $\hat{s}$ through the parameters s_0 and σ_s^2. In related work we demonstrated the use of such a terrain database to provide a reference image for estimating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.[15]

4. ESTIMATION OF MODEL PARAMETERS

4.1. Adaptive estimation of noise

We adaptively estimate the noise covariance $\boldsymbol{\Sigma}_\epsilon$ from successive video frames from each sensor, assumed to have been motion compensated. First the average value at each hyperpixel $(\overline{a_i}(t))$ is estimated by an exponential moving average of the hyperpixel intensity. The average square $(\overline{a_i^2}(t))$ is also estimated by an exponential moving average. The noise variance is then estimated by the difference $(\sigma_{\epsilon_i}^2(t) = \overline{a_i^2}(t) - \overline{a_i}^2(t))$. The details of this adaptive estimation of nosie variance have been described in our earlier work.[15]

4.2. Maximum likelihood parameter estimation using local principal component analysis

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we assume that these are nearly constant over small spatial regions (5 × 5 hyperpixels) surrounding the hyperpixel for which the parameters are to be estimated[†]. Second order statistics on the sensor data in these small regions provide estimates of our parameters.

To form a maximum likelihood estimate of $\boldsymbol{\beta}$, we write the log-likelihood of observing the data given the model, set its derivative with respect to $\boldsymbol{\beta}$ equal to zero, and recover

$$(\mathbf{C} - \boldsymbol{\Sigma}_a)\mathbf{C}^{-1}\boldsymbol{\beta} = 0 \tag{10}$$

where,

$$\boldsymbol{\Sigma}_a \equiv \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{a}_n - \boldsymbol{\mu}_a)(\boldsymbol{a}_n - \boldsymbol{\mu}_a)^{\mathrm{T}} \tag{11}$$

is the data covariance matrix, computed spatially over a (5 × 5 hyperpixel) local region with $N = 25$ points, and

$$\boldsymbol{\mu}_a \equiv \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{a}_n \tag{12}$$

is the data mean matrix computed over the same spatial region. The only non-trivial solution is

$$\boldsymbol{\beta}_{\mathrm{ML}} = \boldsymbol{\Sigma}_\epsilon^{\frac{1}{2}}\widetilde{\mathbf{U}}\frac{(\widetilde{\lambda} - 1)^{\frac{1}{2}}}{\sigma_s}r \tag{13}$$

[†] Essentially we are invoking a spatial analog of ergodicity, where ensemble averages are replaced by spatial averages, carried out locally over regions in which the statistics are approximately constant.

where $\widetilde{\mathbf{U}}$ an eigenvector and $\widetilde{\lambda}$ the corresponding eigenvalue, of the weighted data covariance matrix, $\widetilde{\boldsymbol{\Sigma}}_a \equiv \boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}\boldsymbol{\Sigma}_a\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}$, and $r = \pm 1$. The maximum likelihood occurs when $\widetilde{\mathbf{U}}$ is the principal eigenvector of $\widetilde{\boldsymbol{\Sigma}}_a$. The maximum likelihood estimate of the bias $\boldsymbol{\alpha}$ is,

$$\begin{aligned} \boldsymbol{\alpha}_{\text{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{a}_n - \boldsymbol{\beta} s_0) \\ &= \boldsymbol{\mu}_a - \boldsymbol{\beta} s_0 \ . \end{aligned} \tag{14}$$

An alternative to maximum likelihood estimation is the least squares approach.[17] Here the factor loadings $\boldsymbol{\beta}$ are obtained by minimizing

$$E = \operatorname{tr}(\boldsymbol{\Sigma}_a - \mathbf{C})^2 \ . \tag{15}$$

Differentiating E with respect to $\boldsymbol{\beta}$ and equating to zero yields,

$$(\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\boldsymbol{\beta} = \sigma_s^2 \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \tag{16}$$

The solution to (16) is,

$$\boldsymbol{\beta}_{\text{LS}} = \frac{\lambda^{\frac{1}{2}}}{\sigma_s}\mathbf{U}r \tag{17}$$

where $\mathbf{U}$ is the principal eigenvector, and λ is the principal eigenvalue of the noise-corrected covariance matrix $(\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and $r = \pm 1$.

The least squares solution is the same as the maximum likelihood solution when the model is exact $(\boldsymbol{\Sigma}_a = \mathbf{C})$[‡]. Under this condition,

$$\begin{gathered} \widetilde{\mathbf{U}} = (\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{U})^{-\frac{1}{2}}\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}\mathbf{U} \\ (\widetilde{\lambda} - 1) = \lambda(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{U}) \end{gathered} \tag{18}$$

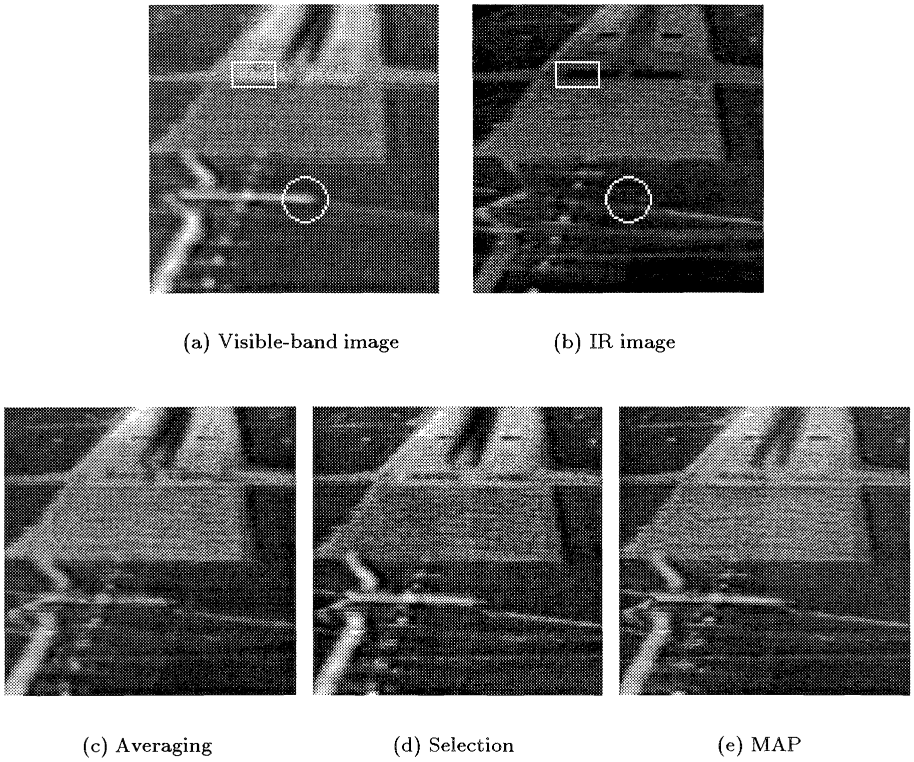
The sign of r is not specified. In a global model this poses no problem. However to properly piece together our local models we cannot allow arbitrary sign reversals between local regions. We therefore fix a convenient reference direction and choose the sign of r such that $\boldsymbol{\beta}$ lies on the same side of this reference for all the local regions. This ensures that the signs of $\boldsymbol{\beta}$ do not change in neighboring regions unless the underlying relationship between sensor images has changed. The choice of the reference direction also provides a way to constrain the polarity reversed features so that they appear in the polarity of a particular sensor. For example, the reference direction can be chosen such that in the fused image runway-markings appear white on a black runway as in a visible-band image of an asphalt runway.

Neither the maximum likelihood, nor the least squares approach provides an estimate of either σ_s^2 or s_0. As with the sign of r, this poses no problem for a global model, but we must impose a constraint in order to smoothly piece together our local models. We take $\sigma_s^2 = \lambda$. To see that this is reasoned, consider a small patch over which both σ_s^2 and the image formation parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ are constant. Any variation in sensor pixel intensities in this region must arise from variations in the true scene s, and the noise. The leading eigenvalue λ of the noise-corrected covariance matrix $\Sigma_a - \Sigma_\epsilon$ gives the scale of variations in $\boldsymbol{a}$ arising from variations in s. Thus, we should have $\lambda \propto \sigma_s^2$. To insure consistency, the proportionality constant should be the same in all local regions. From the least squares solution (17), this proportionality constant is just $\| \boldsymbol{\beta} \|^2$. Hence we take $\| \boldsymbol{\beta} \| = 1$ everywhere, or $\sigma_s^2 = \lambda$. The parameter s_0 causes a shift in $\boldsymbol{\alpha}$. In the absence of any prior information we take $s_0 = 0$.

The weighted data covariance matrix can be expressed as,

$$\widetilde{\boldsymbol{\Sigma}}_a = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\boldsymbol{a}_n - \boldsymbol{\mu}_a)(\boldsymbol{a}_n - \boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}} = \frac{1}{N}\sum_{n=1}^{N}(\widetilde{\boldsymbol{a}}_n - \widetilde{\boldsymbol{\mu}}_a)(\widetilde{\boldsymbol{a}}_n - \widetilde{\boldsymbol{\mu}}_a)^{\mathrm{T}} \tag{19}$$

[‡] Note that for the case of two sensors, the model is necessarily exact



(a) Visible-band image

(b) IR image

(c) Averaging

(d) Selection

(e) MAP

Figure 1. Fusion of visible-band and IR images. The rectangle highlights a region containing local polarity reversal. A region with complementary features is marked by the circle (Original data from SVTD[19]).

where $\widetilde{\boldsymbol{a}} = \boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}} \boldsymbol{a}$ is the weighted data and $\widetilde{\boldsymbol{\mu}}_a = \boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}} \boldsymbol{\mu}_a$ is the weighted bias.

Substituting the ML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from (13) and (14) in (8) and simplifying,

$$\hat{s} = \frac{\sigma_s}{\widetilde{\lambda}}(\widetilde{\lambda} - 1)^{\frac{1}{2}} \widetilde{\mathbf{U}}^{\mathrm{T}} (\widetilde{\boldsymbol{a}} - \widetilde{\boldsymbol{\mu}}_a) + s_0 \tag{20}$$

Thus, the MAP estimate of the true scene is a scaled principal component projection of the weighted data. When the noise variance is equal for all sensors ($\boldsymbol{\Sigma_\epsilon} = \sigma_\epsilon^2 \mathbf{I}$), then (20) simplifies to a scaled principal component projection of the data.

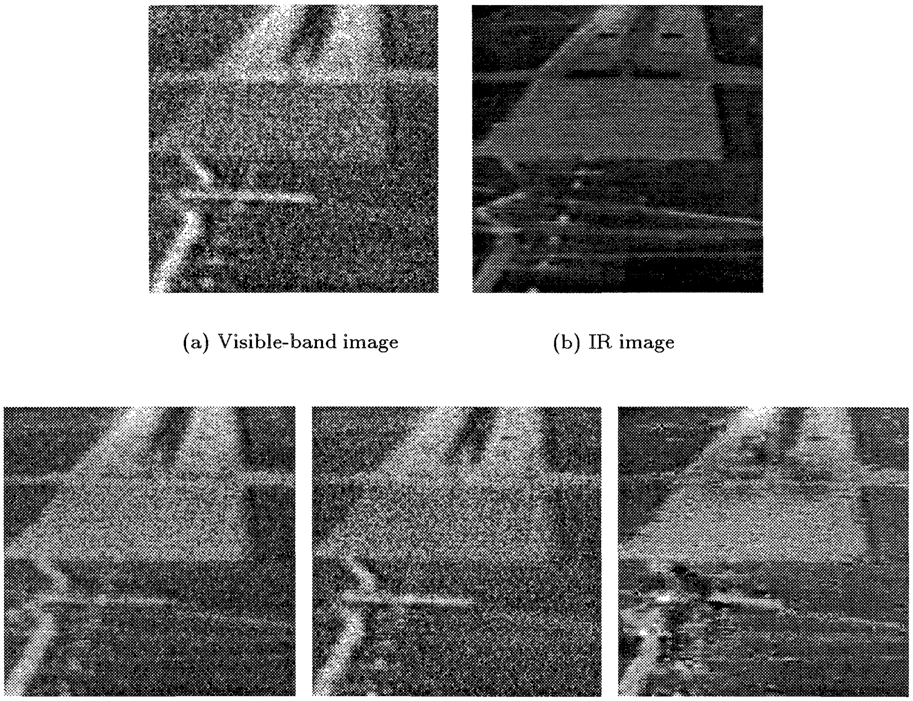
$$\hat{s} = \frac{\sigma_s}{\sigma_\epsilon^2 + \lambda} \lambda^{\frac{1}{2}} \mathbf{U}^{\mathrm{T}} (\boldsymbol{a} - \boldsymbol{\mu}_a) + s_0 \quad . \tag{21}$$

Since the computation of $\hat{s}$ is carried out separately for each hyperpixel using a local region around the hyperpixel, the solution for fusion is closely related local PCA.

5. EXAMPLES OF FUSION

We applied our fusion techniques to video streams from visible-band and IR sensors. The video frames from each of the sensors are decomposed into levels of a Laplacian pyramid. At each hyperpixel location, the hyperpixel of the fused pyramid is synthesized using (8) and the estimates of the parameters. The fused image is obtained by applying the inverse pyramid transform to the fused pyramid.

Figure 1 shows the fused image synthesized by our MAP technique using (21) with $\sigma_\epsilon^2 \to 0$ and $s_0 = 0$. In this case fusion is obtained by local PCA. The results of applying averaging and selection techniques are also shown for comparison. Averaging was performed directly on the pixels of the original images. Selection was performed using



(a) Visible-band image

(b) IR image

(c) Averaging

(d) Selection

(e) MAP

Figure 2. Fusion of visible-band and IR images with additive Gaussian noise (Original data from SVTD[19]).

the Laplacian pyramid. An area based salience measure (sum of squares of hyperpixels in a 5×5 area) was used to choose the sensor hyperpixel to be selected to form the fused pyramid. Comparing the results of MAP fusion with selection, it can be observed that the MAP-fused image is similar to the selection-fused image.

Figure 2 demonstrates our fusion technique in the presence of additive Gaussian noise. The noise in the visible-band sensor is considerably higher than in IR. The result of selection is noisy (Fig. 2(d)). The MAP-fused image is obtained by using (8). The noise is estimated as described in Section 4.1. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are estimated using (17) and (14) respectively. The result of MAP fusion is shown in Figure 2(e). The MAP-fused image does better than averaging and selection in the regions containing polarity reversals and complementary features. For example, the markings on the runway and the horizontal lines in the lower portion of the IR image, are better visible in the MAP-fused image. Although some artifacts can be observed in this image, it is less noisy compared to Figure 2(c) and 2(d).

6. DISCUSSION

We have presented a model-based probabilistic framework for fusion of video sequences from multiple sensors and applied the approach on visual-band and IR imaging sensors. The model is based on analysis of the relationships between image features in the video frames. The Bayesian solution for fusion, using maximum likelihood parameter estimation, is closely related to local PCA. The probabilistic framework provides the flexibiity to use a terrain database to specify priors on the local distribution for s. The benefits obtained by fusion can be further increased by using color mapping techniques that identify the different sensor contributions in the fused video stream.[7]

In the present model, the division into local regions is arbitrary, and better results might be obtained by segmenting the image according to its local covariance statistics. A soft-segmentation could be provided by a mixture model.[20] We are currently investigating the use of multiple video frames for the MAP estimation as well as for estimating the model parameters, to alleviate estimation difficulties caused by noise.

7. ACKNOWLEDGMENTS

This work was supported by NASA Ames Research Center grant NCC2-811. TKL was partially supported by NSF grant ECS-9704094.

REFERENCES

1. L. A. Klein, *Sensor and Data Fusion Concepts and Applications*, SPIE, 1993.
2. B. T. Sweet and C. Tiana, "Image processing and fusion for landing guidance," in *Enhanced and Synthetic Vision 1996*, *Proceedings of SPIE* **2736**, pp. 84–95, SPIE, 1996.
3. P. Pencikowski, "A low cost vehicle-mounted enhanced vision system comprised of a laser illuminator and range-gated camera," in *Enhanced and Synthetic Vision 1996*, *Proceedings of SPIE* **2736**, pp. 222–227, SPIE, 1996.
4. D. L. Hall, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood, MA, 1992.
5. J. R. Kerr, D. P. Pond, and S. Inman, "Infrared-optical multisensor for autonomous landing guidance," *Proceedings of SPIE* **2463**, pp. 38–45, 1995.
6. J. G. Verly, "Enhanced and synthetic vision," in *Enhanced and Synthetic Vision 1996*, *Proceedings of SPIE* **2736**, pp. ix–x, SPIE, 1996.
7. R. K. Sharma and M. Pavel, "Model-based sensor fusion for aviation," in *Proceedings of SPIE*, vol. 3088, pp. 169–176, SPIE, 1997.
8. P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *Fourth Int. Conf. on Computer Vision*, pp. 173–182, IEEE Comp. Soc., 1993.
9. H. Li and Y. Zhou, "Automatic visual/IR image registration," *Opt. Eng.* **35**(2), pp. 391–400, 1996.
10. M. Pavel and R. K. Sharma, "Fusion of radar images: rectification without the flat earth assumption," *Proceedings of SPIE* **2736**, pp. 108–118, 1996.
11. R. K. Sharma and M. Pavel, "Registration of video sequences from multiple sensors," in *Proceedings of the Image Registration Workshop*, pp. 361–366, NASA GSFC, 1997.
12. M. Pavel, J. Larimer, and A. Ahumada, "Sensor fusion for synthetic vision," in *Proceedings of the Society for Information Display*, pp. 475–478, SPIE, 1992.
13. P. Burt, "A gradient pyramid basis for pattern-selective image fusion," in *Proceedings of the Society for Information Display*, pp. 467–470, SPIE, 1992.
14. A. Toet, "Hierarchical image fusion," *Machine Vision and Applications* **3**, pp. 1–11, 1990.
15. R. Sharma and M. Pavel, "Adaptive and statistical image fusion," in *SID Digest*, pp. 969–972, Society for Information Display, 1996.
16. P. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications* **Com-31**, pp. 532–540, 1983.
17. A. Basilevsky, *Statistical Factor Analysis and Related Methods*, Wiley, 1994.
18. M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," tech. rep., NCRG/97/010, Neural Computing Research Group, Aston University, UK, 1997.
19. M. A. Burgess, T. Chang, D. E. Dunford, R. H. Hoh, W. F. Home, and R. F. Tucker, "Synthetic vision technology demonstration executive summary," Tech. Rep. DOT/FAA/RD-93/40,I, Research and Development Service, Washington, D.C., 1993.
20. M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," tech. rep., NCRG/97/003, Neural Computing Research Group, Aston University, UK, 1997.