An EM-CI Based Approach to Fusion of IR and Visual Images

Siyue Chen and Henry Leung
Electrical and Computer Engineering
University of Calgary
Calgary, Alberta, Canada
chens, leungh@ucalgary.ca

Abstract – With the cost decline of Infrared (IR) cameras, it is envisaged that more IR camera will be deployed in vision surveillance system for around-clock day and night surveillance. Infrared camera senses radiation emitted by an object at a non-zero absolute temperature in the infrared spectrum which is not available in visual image, but lose information, such as texture, color and geometric, which is available in visual cameras. Fusion of IR and visual images can enhance features in both kinds of images. And more impressively, it can reveal some new features that might not be present either in IR images or in visual images. In this paper, a statistical signal processing approach based on expectation maximization (EM) is proposed for IR and visual image fusion. The sensor images are described as the true scene corrupted by additive Gaussian distortion. At each iteration of the EM, the fusion result is obtained by using covariance intersection (CI) in the E-step, while the model parameters are updated in the M-step. The simulation results using the real IR and visual images demonstrate the effectiveness of the proposed method.

Keywords: Image fusion, thermal images, Infrared (IR), expectation maximization, covariance intersection.

1 Introduction

Perceiving the real world with only one kind of sensors may result in ambiguities. The integration of the various signals from various sensors will eliminate these ambiguities and extract the maximum amount of information from the sensed environment. Image fusion can enhance the visual features of an image and improve reliability and capability of human or machine perception. In this study, we are particularly interested in fusion of Infrared (IR) and visual images. Compared to visual images, IR images have the advantages such that it is not sensitive to illumination changes, and it does not have the shading problem. However, IR images usually show little contrast due to noises of thermal camera. In addition, the resolution of IR camera is often lower than visual camera. Hence, fusion of IR and visual images is a good way to enhance features in both kinds of images. And more impressively it may reveal some new features that might not be present in either IR images or visual images.

The theory of image fusion has advanced rapidly in the past few years. Image fusion approaches ranging from extreme simplicity to considerable complexity have been proposed. For fusion of IR and visual images, several types of pyramid decomposition are developed, such as Laplacian Pyramid [1], Filter-subtract-decimate (FSD) Pyramid [2], Gradient Pyramid [3], Morphological Pyramid [4], Contrast Pyramid [5], and Ratio-of-low-pass (ROLP) Pyramid [6]. ROLP is a multi-resolution pyramidal technique that uses the maximum contrast information. Since a noisy image is typically of higher contrast than an image that is not, this method will select the noisier parts of the image to be retained in the composite. The gradient pyramid method came up with a match and saliency metric to determine which details are salient. The algorithm offers a potential for better noise reduction. It also allows the low contrast details to be preserved, if they are the salient features. The main disadvantage is that a template (weight matrix) is needed to decide which features are salient.

In this paper, we consider image fusion as a statistical estimation problem. More specifically, we assume that there is a true image from which the observed IR and visual images are generated. Hence, image fusion is posed as an estimation problem where the best fusion algorithm minimizes the mean square error between the fused image and the true scene. To do that, CI, which has been widely used in radar track fusion [7], is employed here. CI can produce consistent estimate for any degree of cross-correlation between the input sources through convex combination [8, 9]. In addition, it en-
forces estimation consistency by means of convex combination of the inverse of covariance matrices. When CI is applied to image fusion, it treats the true image as a linear combination of the IR and visual image, and it selects the fused coefficients through the convex combination of weight optimization.

To use the convex combination in CI, the covariance of the estimation error is required. But for many imaging applications, this information is usually not available. Here, we propose using expectation maximization (EM) to give a maximum likelihood estimate (MLE) of the covariance information. Formalized by Dempster et al. [10], the EM algorithm is an iterative procedure that estimates both the parameters and the missing or unobservable data during an iteration. The approach first computes the approximation to the expectation of the log-likelihood function of the complete data conditioned on the current parameter estimate. This is called the expectation step (E-step), and here, the current incomplete data estimate is calculated. Next, a new parameter estimate is computed by finding the value that maximizes the function found in the E-step. This is called the maximization step (M-step). In this study, the distortorless sensory image is treated as the missing data, and the covariances of them are treated as the parameters to be estimated by EM.

The remainder of this paper is organized as follows. In Section 2, the problem of fusing the IR and visual images is formulated. The proposed EM-CI method is presented in Section 3. Performance evaluation of EM-CI using real sensory images is reported in Section 4. Section 5 concludes this paper.

2 Problem Formulation

The observed IR and the observed visual images are denoted as \( \hat{x}_1 \) and \( \hat{x}_2 \), respectively, which are the vectors composed of pixel values. Due to sensor distortion and noise in the capture process of IR and visual images, the observed images can be modeled as “linear observation plus Gaussian noise” [11] [12]. That is,

\[
\hat{x}_i = H_i x_i + n_i, \quad i = 1, 2, \tag{1}
\]

where \( x_i \) denotes the vector of pixel values from an ideal sensory image, \( H_i \) represents the linear observation operator, and \( n_i \) is the white Gaussian noise. Typical examples of the linear observation operator \( H_i \) include optical blur, motion blur, tomographic projections, and etc. Since the study of \( H_i \) is not in the scope of this paper, we assume that it is known, and for simplicity, an identity matrix here.

From (1), it is noted that neither the ideal sensory image \( x_i \) nor its statistics are available. Assume a consistent estimate \( \hat{x}_i \) can be obtained. That is, if we define

\[
\hat{P}_{x_i x_i} = (\hat{x}_i - \bar{x}_i)(\bar{x}_i - \hat{x}_i)^T, \quad i = 1, 2, \tag{2}
\]

and

\[
P_{x_i x_i} = E \left[ (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \right], \quad i = 1, 2, \tag{3}
\]

we have [13]

\[
\hat{P}_{x_i x_i} \geq P_{x_i x_i}, \quad i = 1, 2. \tag{4}
\]

The inequality is in the sense of matrix positive definiteness, i.e., \( A > B \) if and only if \( A - B \) is positive definite. We further define the cross-correlation as

\[
P_{x_1 x_2} = E[(x_1 - \hat{x}_1)(x_2 - \hat{x}_2)^T]. \tag{5}
\]

Our objective is to construct a linear, unbiased estimate \( \hat{x} \) that combines \( x_1 \) and \( x_2 \), i.e.,

\[
\hat{x} = M_1 \hat{x}_1 + M_2 \hat{x}_2, \tag{6}
\]

so that

\[
E[\bar{x} - \hat{x}] = 0, \tag{7}
\]

provided that

\[
M_1 + M_2 = I. \tag{8}
\]

In addition, we want to determine a consistent estimate \( \hat{P}_{xx} \), for \( P_{xx} \) \( \hat{P}_{xx} = E[(\bar{x} - \hat{x})(\bar{x} - \hat{x})^T] \), and to find a pair of \( M_1 \) and \( M_2 \) such that the upper bound \( \hat{P}_{xx} \) is optimal in the sense of minimal trace.

According to the definition of \( P_{xx} \), we have

\[
P_{xx} = M_1 P_{x_1 x_1} M_1^T + M_1 P_{x_1 x_2} M_2^T + M_2 P_{x_2 x_1} M_1^T + M_2 P_{x_2 x_2} M_2^T. \tag{9}
\]

Similarly,

\[
P_{xx} = M_1 P_{x_1 x_1} M_1^T + M_1 P_{x_1 x_2} M_2^T + M_2 P_{x_2 x_1} M_1^T + M_2 P_{x_2 x_2} M_2^T. \tag{10}
\]

If \( P_{x_1 x_2} = 0 \), for any given \( M_1 \) and \( M_2 \), the estimate

\[
\hat{P}_{xx} = M_1 \hat{P}_{x_1 x_1} M_1^T + M_2 \hat{P}_{x_2 x_2} M_2^T \tag{11}
\]

will be consistent (\( \hat{P}_{xx} \geq P_{xx} \)) as a direct consequence of (4). The trace of the above \( \hat{P}_{xx} \) is then minimized by

\[
\hat{P}_{xx} = (\hat{P}_{x_1 x_1}^{-1} + \hat{P}_{x_2 x_2}^{-1})^{-1},
\]

\[
M_1 = \hat{P}_{xx} \hat{P}_{x_1 x_1}^{-1} = \hat{P}_{x_1 x_1} (\hat{P}_{x_1 x_1} + \hat{P}_{x_2 x_2})^{-1}, \tag{12}
\]

\[
M_2 = \hat{P}_{xx} \hat{P}_{x_2 x_2}^{-1} = \hat{P}_{x_2 x_2} (\hat{P}_{x_1 x_1} + \hat{P}_{x_2 x_2})^{-1}.
\]

Since the cross-correlation between \( x_1 \) and \( x_2 \) is generally non-zero and unknown, CI is employed here to deal with the situation when \( P_{x_1 x_2} \neq 0 \).

3 EM-CI for image fusion

First, we need to determine \( \hat{x}_i \) and \( \hat{P}_{x_i x_i} \), \( i = 1, 2 \). According to (1), the probability density functions (PDF) of \( \hat{x}_i \) conditioned on \( x_i \) can be written as

\[
p(\hat{x}_i | x_i) = (2\pi)^{-S/2} |\hat{P}_{x_i x_i}|^{-1/2}
\exp \left\{ -\frac{1}{2} [\hat{x}_i - H_i x_i]^T \hat{P}_{x_i x_i}^{-1} [\hat{x}_i - H_i x_i] \right\}, \quad i = 1, 2, \tag{13}
\]
where \( S \) denotes the number of pixels within the vector \( \tilde{x}_i \). The log likelihood function is thus
\[
L(x_i|\tilde{x}_i) = -\frac{1}{2} \log |\hat{P}_{x_i}x_i| - \frac{S}{2} \log (2\pi) \\
-\frac{1}{2} \left\{ \tilde{x}_i - H_i x_i \right\}^T \hat{P}_{x_i}^{-1} \left\{ \tilde{x}_i - H_i x_i \right\}.
\]
(14)
In factor analysis, \( \hat{P}_{x_i}x_i \) is the diagonal matrix under the white Gaussian noise assumption. The E-step of EM computes the expected log likelihood function, i.e.,
\[
E_P = E[L(x_i|\tilde{x}_i)|\tilde{x}_i], \quad i = 1, 2.
\]
(15)
In the M-step, \( \hat{x}_i \) and \( \hat{P}_{x_i}x_i \) are obtained by maximizing (15). More specifically, the partial derivatives of the expected log likelihood function are set as zero. With the constant term ignored, we have,
\[
\frac{\partial E_P}{\partial x_i} = \hat{x}_i \hat{P}_{x_i}^{-1} - H_i \hat{x}_i \hat{P}_{x_i}^{-1} H_i^T = 0, \quad i = 1, 2,
\]
(16)
\[
\frac{\partial E_P}{\partial \hat{P}_{x_i}x_i} = -\frac{1}{2} [\tilde{x}_i \tilde{x}_i^T - 2H_i \tilde{x}_i \tilde{x}_i^T + H_i \tilde{x}_i \tilde{x}_i^T H_i^T] + \frac{1}{2} \hat{P}_{x_i}x_i = 0, \quad i = 1, 2.
\]
(17)
Solving these equations generates
\[
\hat{x}_i = (H_i^T \hat{P}_{x_i}^{-1} H_i)^{-1} H_i^T \hat{P}_{x_i}^{-1} \tilde{x}_i, \quad i = 1, 2,
\]
(18)
\[
\hat{P}_{x_i}x_i = \hat{x}_i \hat{x}_i^T - 2H_i \hat{x}_i \hat{x}_i^T + H_i \hat{x}_i \hat{x}_i^T H_i^T, \quad i = 1, 2.
\]
(19)
Based on \( x_i \) and \( \hat{P}_{x_i}x_i \), \( i = 1, 2 \), we now compute \( \hat{x} \) and \( \hat{P}_{xx} \) using CI. For a covariance matrix \( P \), the covariance ellipse is the locus of the points
\[
B_P(g) = \{ x : (x - \bar{x})^T P^{-1} (x - \bar{x}) = g \},
\]
(20)
where \( g \) is a constant and \( \bar{x} \) is the mean of \( x \). The covariance ellipse is a convenient way of visualizing the relative size of covariance matrices. If \( P_1 < P_2 \), \( B_{P_1}(g) \subset B_{P_2}(g) \). From the geometric interpretation of (10), the covariance ellipse of \( \hat{P}_{xx} \) always lies within the intersection of the covariance ellipses of \( \hat{P}_{x_1}x_1 \) and \( \hat{P}_{x_2}x_2 \), for any possible choice of \( \hat{P}_{x_1}x_1 \), as shown in Figure 1. Hence, a method which finds a \( \hat{P}_{xx} \) which encloses the intersection region must be consistent even if there is no knowledge about \( P_{xx} \) [14, 15]. When \( P_{xx} \) is unknown, in order to obtain an upper bound of (9), the inequality
\[
E[\{(\sqrt{\gamma} M_1 (x_1 - \hat{x}_1) - \sqrt{\gamma} M_2 (x_2 - \hat{x}_2)) \}^2] \geq 0
\]
(21)
is utilized, where \( \gamma > 0 \) is a scalar. It follows that
\[
\gamma M_1 P_{x_1}x_1 M_1^T + \frac{1}{\gamma} M_2 P_{x_2}x_2 M_2^T \\
\geq M_1 P_{x_1}x_1 M_1^T + M_2 P_{x_2}x_2 M_2^T.
\]
(22)
From (4) and (22), a consistent estimate \( \hat{P}_{xx} \) for \( P_{xx} \) can be obtained as
\[
\hat{P}_{xx} = (1 + \gamma) M_1 \hat{P}_{x_1}x_1 M_1^T + (1 + \frac{1}{\gamma}) M_2 \hat{P}_{x_2}x_2 M_2^T.
\]
(23)
The value of \( \gamma \) is then chosen to minimize the trace of \( \hat{P}_{xx} \). Note that
\[
\text{trace} (\hat{P}_{xx}) = \text{trace}(M_1 \hat{P}_{x_1}x_1 M_1^T) + \text{trace}(M_2 \hat{P}_{x_2}x_2 M_2^T) + \gamma \text{trace}(M_1 \hat{P}_{x_1}x_1 M_1^T) + \frac{1}{\gamma} \text{trace}(M_2 \hat{P}_{x_2}x_2 M_2^T) \\
\geq \text{trace}(M_1 \hat{P}_{x_1}x_1 M_1^T) + \text{trace}(M_2 \hat{P}_{x_2}x_2 M_2^T) + 2 \sqrt{\text{trace}(M_1 \hat{P}_{x_1}x_1 M_1^T) \text{trace}(M_2 \hat{P}_{x_2}x_2 M_2^T)},
\]
(24)
where the equality holds when
\[
\gamma = \sqrt{\frac{\text{trace}(M_2 \hat{P}_{x_2}x_2 M_2^T)}{\text{trace}(M_1 \hat{P}_{x_1}x_1 M_1^T)}}.
\]
(25)
For a particular choice, i.e., if \( M_1 = \omega_1 P_{xx} \hat{P}_{x_1}x_1 \), \( M_2 = \omega_2 P_{xx} \hat{P}_{x_2}x_2 \), \( \gamma = \frac{\omega_2}{\omega_1} \), can be further written as
\[
\hat{P}_{xx} = \omega_1 P_{xx} \hat{P}_{x_1}x_1 \hat{P}_{x_1}x_1^{-1} \hat{P}_{xx} + \omega_2 P_{xx} \hat{P}_{x_2}x_2 \hat{P}_{x_2}x_2^{-1} \hat{P}_{xx} \\
= \hat{P}_{xx} (\omega_1 \hat{P}_{x_1}x_1^{-1} + \omega_2 \hat{P}_{x_2}x_2^{-1}) \hat{P}_{xx},
\]
(26)
which is satisfied by
\[
\hat{P}_{xx} = \omega_1 \hat{P}_{x_1}x_1^{-1} + \omega_2 \hat{P}_{x_2}x_2^{-1}.
\]
(27)
Meanwhile, (6) becomes
\[
\hat{x} = \omega_1 \hat{P}_{xx} \hat{P}_{x_1}x_1 \hat{x}_1 + \omega_2 \hat{P}_{xx} \hat{P}_{x_2}x_2 \hat{x}_2
\]
(28)
with nonnegative coefficients \( \omega_1 \) and \( \omega_2 \) obeying
\[
\omega_1 + \omega_2 = 1.
\]
(29)
The weighting coefficients, \( \omega_1 \) and \( \omega_2 \), are usually chosen to minimize the trace of \( \hat{P}_{xx} \). The minimizing process requires optimization of a nonlinear cost function which is convex with respect to \( \omega_1 \) and \( \omega_2 \) [9]. The minimized trace is usually obtained by an exhaustive search of \( \omega_1 \) and \( \omega_2 \) within the range of \([0,1]\). This process has a high computational burden. So, a suboptimal non-iterative linear algorithm is necessary. It is well known that the trace of the covariance matrix \( \hat{P}_{x_i}x_i \), \( i = 1, 2 \), provides the uncertainty measure of the estimation on \( \hat{x}_i \), \( i = 1, 2 \). If \( \text{trace}(\hat{P}_{x_i}x_i) \ll \text{trace}(\hat{P}_{x_2}x_2) \), it infers that \( \hat{x}_2 \) has a much larger estimation error than \( \hat{x}_1 \). We thus have \( \hat{x} \approx \hat{P}_{xx} \hat{P}_{x_1}x_1 \hat{x}_1 \), which implies \( \omega_1 \approx 1 \) and \( \omega_2 \approx 0 \). Similarly, if \( \text{trace}(\hat{P}_{x_1}x_1) = \text{trace}(\hat{P}_{x_2}x_2) \), it is expected that \( \hat{x} \approx \hat{P}_{xx} \hat{P}_{x_2}x_2 \hat{x}_2 \). In other words,
that compared to the visual image, the fused image improves the contrast between the square-like object and the background. At the mean time, the fused image reserves all the background information which is not available in the IR image, but is carried by the visual image.

Besides the subjective evaluation, the objective evaluation of the proposed method is also performed. As suggested by [16], the standard deviation (SD), the entropy (EN) and the objective edge-based measure (QE) are employed as the performance measures. For a fused image of size \( N \times M \), its standard deviation can be computed by

\[
SD = \sqrt{\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (f(i,j) - \bar{f})^2},
\]

where \( f(i,j) \) is the \((i,j)\)th pixel intensity value and \( \bar{f} \) is the sample mean of all pixel values of the image. SD is composed of two parts, the signal part and the noise part. This measurement will be more efficient in the absence of noise, when it represents the signal strength only. Entropy has often been used to measure the information content of an image. Using entropy, the information content of an image is

\[
EN = -\sum_{i=0}^{G} p(i) \log_2(p(i)),
\]

where \( G \) is the number of gray levels in the image’s histogram (which can be 255 for a typical 8-bit image) and \( p(i) \) is the normalized frequency of occurrence of each gray level, i.e., the histogram of the image. To sum up the self-information of each gray level from the image, the average information is estimated in the units of bits per pixel. It should be noted that entropy is also sensitive to noise and other unwanted rapid fluctuations. Objective edge based measure is obtained by evaluating the amount of edge information that is transferred from source images to the fused image. The definition of objective edge based measure can be written as

\[
QE = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \omega_{f_0}(i,j)w_{f_0}(i,j) + Q_{f_0,j}(i,j)w_{f_1}(i,j) + Q_{f_1,j}(i,j)w_{f_0}(i,j) + Q_{f_0,j}(i,j)w_{f_1}(i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{M} w_{f_0}(i,j) + w_{f_1}(i,j)},
\]

where \( \omega_{f_0}(i,j) \) and \( \omega_{f_1}(i,j) \) are the weights varying with the edge strength, \( Q_{f_0,j}(i,j) \) and \( Q_{f_1,j}(i,j) \) are the edge preservation values that are computed by the product of a sigmoid mapping function of the relative strength and orientation factors. The objective evaluation results of EM-CI are listed in Table 1. For the comparison purpose, we also list the results of pixel averaging and Laplacian Pyramid. It is shown that the proposed EM-CI fusion method is better than pixel averaging and the Laplacian pyramid method.

4 Experimental Results

Some real IR and visual images are used to evaluate the performance of the proposed EM-CI method. The first set of images were captured by Octec Ltd by using a Sony Camcorder and a LWIR sensor. As illustrated in Figure 2 (a) and (b), both the IR and visual image display men and buildings with a smoke screen. The fusion result by EM-CI is plotted in Figure 2 (c). As seen, the two people blocked by the smoke in the visual image are clearly displayed in the fused image. Meanwhile, the details of the house at the right side, which are not carried in the IR images, are also retained in the fused image. The second set of testing images are shown in Figure 3 (a) and (b). They were collected during a series of subjective image fusion evaluation trials in the period 1999-2002 as a part of the University of Manchester image fusion research programme. The fusion result by EM-CI is plotted in Figure 3 (c). It is seen
Figure 2: Experiment on the first set of testing images:
(a) the IR image (b) the visual image; (c) the fusion result by EM-CI.

Figure 3: Experiment on the second set of testing images:
(a) the IR image (b) the visual image; (c) the fusion result by EM-CI.
Table 1: Objective evaluation results of various fusion methods

<table>
<thead>
<tr>
<th>Fusion Methods</th>
<th>SD</th>
<th>EN</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM-CI</td>
<td>43.677</td>
<td>8.926</td>
<td>0.672</td>
</tr>
<tr>
<td>Laplacian Pyramid</td>
<td>40.158</td>
<td>6.388</td>
<td>0.628</td>
</tr>
<tr>
<td>Pixel Averaging</td>
<td>37.359</td>
<td>7.110</td>
<td>0.592</td>
</tr>
</tbody>
</table>

5 Conclusions

Covariance intersection, which combines multiple data sources through convex combinations for any degree of cross-correlation, is one of the most popular information fusion approach. In this paper, we propose a novel EM-CI approach for fusion of IR and visual images. More specifically, the ideal IR and visual images are estimated by EM along with the covariance matrices of the estimation error. Then, CI is applied to combine the two images and provide a consistent estimate of the fused scene. The simulation results show that the proposed method is able to achieve a better fusion performance than pixel averaging and the Laplacian pyramid method in terms of the three criteria, i.e., SD, EN, and QE.

References