# Image Fusion schemes using ICA bases

Nikolaos Mitianoudis, Tania Stathaki

*Communications and Signal Processing group, Imperial College London, Exhibition Road, SW7 2AZ London, UK*

## Abstract

*Image fusion* is commonly described as the task of enhancing the perception of a scene by combining information captured by different modality sensors. The *pyramid decomposition* and the *Dual-Tree Wavelet Transform* have been employed as analysis and synthesis tools for image fusion by the fusion community. Using various fusion rules, one can combine the important features of the input images in the transform domain to compose an enhanced image. In this study, the authors demonstrate the efficiency of a transform constructed using *Independent Component Analysis* (ICA) and *Topographic Independent Component Analysis* bases for image fusion. The bases are trained offline using images of similar context to the observed scene. The images are fused in the transform domain using novel *pixel-based* or *region-based* rules. An unsupervised adaptation ICA-based fusion scheme is also introduced. The proposed schemes feature improved performance compared to approaches based on the wavelet transform and slightly increased computational complexity.

*Key words:* Image Fusion, Independent Component Analysis, Topographic ICA.
*PACS:*

## 1 Introduction

The need for data fusion in current image processing systems is increasing mainly due to the increase of image acquisition techniques [1]. Current technology in imaging sensors offers a wide variety of different information that can be extracted from an observed scene. This information is jointly combined to provide an enhanced representation in many cases of experimental sciences. The automated procedure of conveying all the meaningful information from the input sensors to a composite image is the topic of this article. *Fusion systems* appear to be an essential preprocessing stage for a number of applications, such as aerial and satellite imaging, medical imaging, robot vision and vehicle or robot guidance [1].

Let $I_1(x,y), I_2(x,y), \dots, I_T(x,y)$ represent $T$ images of size $M_1 \times M_2$ capturing the same scene. Each image has been acquired using different instrument modalities or capture techniques. Consequently, each image has different characteristics, such as degradation, thermal and visual characteristics. These images need not be perfect, otherwise fusion would not be necessary. This imperfection can appear in the form of imprecision, ambiguity or incompleteness. However, the source images should offer complementary and redundant information about the observed scene [1]. In addition, each of these images should contain information that might be useful for the composite image and is not provided by the other input images. In other words, there is no potential in fusing an image that has mainly degraded information compared to the other input images. Although the fusion system will most probably be able to reject the misleading information, it is not conceptually valid to present the system with no beneficial information, as the performance might be degraded and the computational complexity increased.

In this scenario, we usually employ multiple sensors that are placed relatively close and are observing the same scene. The images acquired by these sensors, although they should be similar, are bound to have some translational errors, i.e. miscorrespondence between several points of the observed scene. *Image registration* is the process of establishing point-by-point correspondence between a number of images, describing the same scene [7]. In the opposite case that the sensors are arbitrarily placed, all input images need to be registered. In this study, the input images $I_i(x,y)$ are assumed to have negligible registration problems, which implies that the objects in all images are geometrically aligned.

The process of combining the important features from these $T$ images to form a single enhanced image $I_f(x,y)$ is usually referred to as *image fusion*. Fusion techniques are commonly divided into *spatial domain* and *transform domain* techniques [8]. In spatial domain techniques, the input images are fused in the spatial domain, i.e. using localised spatial features. Assuming that $g(\cdot)$ represents the "fusion rule", i.e. the method that combines features from the input images, the spatial domain techniques can be summarised, as follows:

$$I_f(x,y) = g(I_1(x,y), \dots, I_T(x,y)) \tag{1}$$

The main motivation behind moving to a transform domain is to work in a framework, where the image's salient features are more clearly depicted than in the spatial domain. It is important to understand the underlying image structure for fusion rather than fusing image pixels independently. Most transformations used in image processing are decomposing the images into important local components, i.e. unlocking the basic image structure. Hence, the choice of the transformation is very important. Let $\mathcal{T}\{\cdot\}$ represent a transform operator and $g(\cdot)$ the applied fusion rule. Transform-domain fusion techniques can then be outlined, as follows:

$$I_f(x,y) = \mathcal{T}^{-1}\{g(\mathcal{T}\{I_1(x,y)\},\dots,\mathcal{T}\{I_T(x,y)\})\} \tag{2}$$

The fusion operator $g(\cdot)$ describes the merging of information from the different input images. Many fusion rules have been proposed in the literature [18,20,22]. These rules can be categorised, as follows:

- *Pixel-based rules*: the information fusion is performed in a pixel-by-pixel basis either in the transform or spatial domain. Each pixel $(x,y)$ of the $T$ input images is combined with various rules to form the corresponding pixel $(x,y)$ in the "fused" image $I_T$. Several basic transform-domain schemes were proposed [18], such as:
  · *fusion by averaging*: fuse by averaging the corresponding coefficients in each image ("mean" rule).

$$\mathcal{T}\{I_f(x,y)\}) = \frac{1}{T}\sum_{i=1}^{T}\mathcal{T}\{I_i(x,y)\} \tag{3}$$

  · *fusion by absolute maximum*: fuse by selecting the greatest in absolute value of the corresponding coefficients in each image ("max-abs" rule)

$$\mathcal{T}\{I_f(x,y)\}) = \mathrm{sgn}(\mathcal{T}\{I_i(x,y)\}) \max_i |\mathcal{T}\{I_i(x,y)\}| \tag{4}$$

  · *fusion by denoising (hard/soft thresholding)*: perform simultaneous fusion and denoising by thresholding the transform's coefficients (sparse code shrinkage [12]).
  · *high/low fusion*, i.e. combining the "high-frequency" parts of some images with the "low-frequency" parts of some other images.

  The different properties of these fusion schemes will be explained later on. For a more complete review on pixel-based fusion methods, one can have always refer to Piella [20], Nikolov et al [18] and Rockinger et al [22].
- *Region-based fusion rules*: in order to exploit the image structure more efficiently, these schemes group image pixels to form contiguous regions, e.g. objects and impose different fusion rules to each image region. In [15], Li et al created a binary decision map to choose between the coefficients using a majority filter, measuring activity in small patches around each pixel. In [20], Piella proposed several activity level measures, such as the absolute value, the median or the contrast to neighbours. Consequently, she proposed a region-based scheme using a local correlation measurement to performs fusion of each region. In [14], Lewis et al produced a joint-segmentation map out of the input images. To perform fusion, they measured *priority* using *energy*, *variance*, or *entropy* of the wavelet coefficients to impose weighting on each region in the fusion process along with other heuristic rules.

In this study, the application of *Independent Component Analysis* (ICA) and *Topographic Independent Component Analysis* bases as an analysis tool for image fusion in both noisy and noiseless environments is examined. The performance of the proposed framework in image fusion is compared to traditional fusion analysis tools, such as the *wavelet transform*. Common pixel-based fusion rules are tested together with a proposed "weighted-combination" scheme, based on the $\mathcal{L}_1$-norm. A region-based approach that segments and fuses active and non-active areas of the image is introduced. Finally, an adaptive unsupervised scheme for image fusion in the ICA domain using *sparsity* is presented.

The paper is structured, as follows. In section 2, we introduce the basics of the Independent Component Analysis technique and how it can be used to generate analysis/synthesis bases for image fusion. In section 3, we describe the general method for performing image fusion using ICA bases. In section 4, the proposed pixel-based weighted combination scheme and a combinatory region-based scheme are introduced. In section 5, we describe an unsupervised adaptive fusion scheme in the ICA framework. In section 6, several issues concerning the reconstruction of the fused image from the ICA representation are discussed. In section 7, the proposed transform and fusion schemes is benchmarked using common fusion testbed. Finally, in section 8, we outline the advantages and disadvantages of the proposed schemes together with some suggestions about future work.

## 2 ICA and Topographic ICA bases

Assume an image $I(x,y)$ of size $M_1 \times M_2$ and a window $W$ of size $N \times N$, centered around the pixel $(x_0, y_0)$. An "image patch" is defined as the product between a $N \times N$ neighbourhood centered around pixel $(x_0, y_0)$ and the window $W$.

$$I_w(k,l) = W(k,l)I(x_0 - \lfloor N/2 \rfloor + k, y_0 - \lfloor N/2 \rfloor + l), \qquad \forall\ k,l \in [0, N-1] \tag{5}$$

where $\lfloor \cdot \rfloor$ represents the lower integer part and $N$ is odd. For the subsequent analysis, we will assume a rectangular window, i.e.

$$W(k,l) = 1, \qquad \forall\ k,l \in [0, N-1] \tag{6}$$

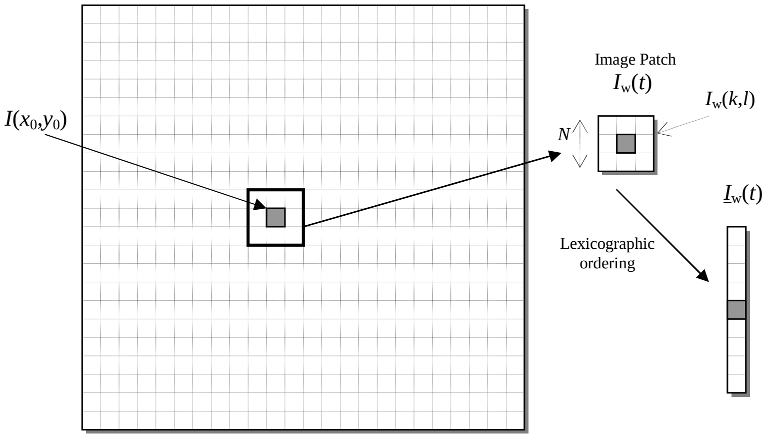


Fig. 1. Selecting an image patch $I_w$ around pixel $(x_0, y_0)$ and the lexicographic ordering.

### *2.1 Definition of bases*

In an effort to understand the underlying structure of an image, it is common practice in image analysis to express an image as the synthesis of several *basis* images. These bases are chosen according to the image features that need to be highlighted with this analysis. A number of basis have been proposed in literature so far, such as cosine bases, complex cosine bases, Hadamard bases and wavelet bases. In this case, the bases are well-defined in order to serve some specific analysis tasks. However, one can estimate non-standard bases by training with a population of similar content images. The bases are estimated after optimising a cost function that defines the bases' desired properties.

The $N \times N$ image patch $I_w(k,l)$ can be expressed as a linear combination of a set of $K$ basis images $b_j(k,l)$, i.e.

$$I_w(k,l) = \sum_{j=1}^{K} u_j b_j(k,l) \tag{7}$$

where $u_j$ are scalar constants. The two-dimensional (2D) representation can be simplified to an one-dimensional (1D) representation, by employing *lexicographic ordering*, in order to facilitate the analysis. In other words, the image patch $I_w(k,l)$ is arranged into a vector $\underline{I}_w$, taking all elements from matrix $I_w$ in a row-wise fashion. The vectors $\underline{I}_w$ are normalised to zero mean, to avoid the possible bias of the local grayscale levels. Assume that we have a population of patches $I_w$, acquired randomly from the original image $I(x,y)$. These image patches can then be expressed in lexicographic ordering, as follows:

$$\underline{I}_w(t) = \sum_{j=1}^{K} u_j(t)\underline{b}_j = [\underline{b}_1\ \underline{b}_2 \ldots \underline{b}_K] \begin{bmatrix} u_1(t) \\ u_2(t) \\ \ldots \\ u_K(t) \end{bmatrix} \tag{8}$$

where $t$ represents the $t$-th image patch selected from the original image. The whole procedure of image patch selection and lexicographic ordering is depicted in figure 1. Let $B = [\underline{b}_1\ \underline{b}_2 \ldots \underline{b}_K]$ and $\underline{u}(t) = [u_1(t)\ u_2(t) \ldots u_K(t)]^T$. Then, equation (8) can be simplified, as follows:

$$\underline{I}_w(t) = B\underline{u}(t) \tag{9}$$

$$\underline{u}(t) = B^{-1}\underline{I}_w(t) = A\underline{I}_w(t) \tag{10}$$

In this case, $A = B^{-1} = [\underline{a}_1\ \underline{a}_2 \ldots \underline{a}_K]^T$ represents the *analysis* kernel and $B$ the *synthesis* kernel. This "transformation" projects the observed signal $\underline{I}_w(t)$ on a set of basis vectors $\underline{b}_j$. The aim is to estimate a finite set of basis vectors that will be capable of capturing most of the signal's structure (energy). Essentially, we need $N^2$ bases for a *complete* representation of the $N^2$-dimensional signals $\underline{I}_w(t)$. However, with some redundancy reduction mechanisms, we can have efficient *overcomplete* representations of the original signals using $K < N^2$ bases.

The estimation of these $K$ vectors is performed using a population of training image patches $\underline{I}_w(t)$ and a criterion (cost function), which is going to be optimised in order to select the basis vectors. In the next paragraphs, we will estimate bases from image patches using several criteria.

### *2.1.1 Principal Component Analysis (PCA) bases*

One of the transform's targets might be to analyse the image patches into uncorrelated components. *Principal Component Analysis* (PCA) can identify uncorrelated vector bases [10], assuming a linear generative model, as in (9). In addition, PCA can be used for dimensionality reduction to identify the $K$ most important basis vectors. This is performed by eigenvalue decomposition of the data correlation matrix $C = \mathcal{E}\{\underline{I}_w \underline{I}_w^T\}$, where $\mathcal{E}\{\cdot\}$ represents the expectation operator. Assume that $H$ is a matrix containing all the eigenvectors of $C$ and $D$ a diagonal matrix containing the eigenvalues of $C$. The eigenvalue at the $i$-th diagonal element should correspond to the eigenvector at the $i$-th column of $H$. The rows of the following matrix $V$ provide an orthonormal set of uncorrelated bases, which are called PCA bases.

$$V = D^{-0.5} H^T \tag{11}$$

The above set forms a *complete* set of bases, i.e. we have as many bases as the dimensionality of the problem ($N^2$). As PCA has efficient energy compaction properties, one can form a reduced (*overcomplete*) set of bases, based on the original ones. The eigenvalues can illustrate the significance of their corresponding eigenvector (basis vector). We can order the eigenvalues in the diagonal matrix $D$, in terms of decreasing absolute value. The eigenvector matrix $H$ should be arranged accordingly. Then, we can select the first $K < N^2$ eigenvectors that correspond to the $K$ most important eigenvalues and form reduced versions of $\hat{D}$ and $\hat{H}$. The reduced $K \times N^2$ PCA matrix $\hat{V}$ is calculated using (11) for $\hat{D}$ and $\hat{H}$. The input data can be mapped to the PCA domain via the transformation:

$$\underline{z}(t) = \hat{V} \underline{I}_w(t) \tag{12}$$

The size of the overcomplete set bases $K$ is chosen so that the computational load of a complete representation can be reduced. However, the overcomplete set should be able to provide an almost lossless representation of the original image. Therefore, the choice of $K$ is usually a trade-off between computational complexity and image quality.

### *2.1.2 Independent Component Analysis (ICA) bases*

A stricter criterion than uncorrelatedness is to assume that the basis vectors or equivalently the transform coefficients are *statistically independent. Independent Component Analysis* (ICA) can identify statistically independent basis vectors in a linear generative model [13]. A number of different approaches have been proposed to analyse the generative model in (9), assuming statistical independence between the coefficients $u_i$ in the transform domain. Statistical independence can be closely linked with the nonGaussianity. The *Central Limit Theorem* states that the sum of several independent random variables tends towards a Gaussian distribution. The same principal holds for any linear combination $I_w$ of these independent random variables $u_i$. The Central Limit Theorem also implies that a combination of the observed signals in $I_w$ with minimal Gaussian properties can be one of the independent signals. Therefore, statistical independence and nonGaussianity can be interchangeable terms.

A number of different techniques can be used to estimate independent coefficients $u_i$. Some approaches estimate $u_i$ by minimising the *Kullback-Leibler*

(KL) divergence between the estimated coefficients $u_i$ and *several probabilistic priors* on the coefficients. Other approaches minimise the *mutual information* conveyed by the estimated coefficients or perform approximate diagonalisation of a *cumulant tensor* of $\underline{I}_w$. Finally, some methods estimate $u_i$ by estimating the directions of the most nonGaussian components using *kurtosis* or *negentropy*, as nonGaussianity measures. More details on these techniques can be found in tutorial books on ICA, such as [3,13].

In this study, we will use an approach that optimises negentropy, as a non-Gaussianity measurement to identify the independent components $u_i$. This is also known as FastICA and was proposed by Hyvärinen and Oja [9]. According to this technique, PCA is used as a preprocessing step to select the $K$ most important vectors and orthonormalise the data using (12). Consequently, the statistical independent components can be identified using orthogonal projections $\underline{a}_i^T \underline{z}$. In order to estimate the projecting vectors $\underline{a}_i$, we have to minimise the following non-quadratic approximation of negentropy:

$$J_G(\underline{q}_i) = \Big(\mathcal{E}\{G(\underline{q}_i^T \underline{z})\} - \mathcal{E}\{G(v)\}\Big)^2 \tag{13}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, $v$ is a Gaussian variable of zero mean and unit variance and $G(\cdot)$ is practically any non-quadratic function. A couple of possible functions were proposed in [11]. In our analysis, we will use:

$$G(x) = \frac{1}{\alpha} \log \cosh \alpha x \tag{14}$$

where $\alpha$ is a constant that usually is bounded to $1 \leq \alpha \leq 2$. Hyvärinen and Oja produced a fixed-point method, optimising the above definition of negentropy, which is also known as the *FastICA* algorithm.

$$\underline{q}_i^+ \leftarrow \mathcal{E}\{\underline{q}_i \phi(\underline{q}_i^T \underline{z})\} - \mathcal{E}\{\phi'(\underline{q}_i^T \underline{z})\}\underline{q}_i, \qquad 1 \leq i \leq K \tag{15}$$

$$Q \leftarrow Q(Q^T Q)^{-0.5} \tag{16}$$

where $\phi(x) = -\partial G(x)/\partial x$. We randomly initialise the update rule in (15) for each projecting vector $\underline{q}_i$. The new updates are then orthogonalised, using the symmetric orthogonalisation scheme in (16). These two steps are iterated, until $\underline{q}_i$ have converged.

### *2.1.3 Topographical Independent Component Analysis (TopoICA) bases*

In practical applications, one can frequently observe clear violations of the independence assumption. It is possible to find couples of estimated components that they are clearly dependent on each other. This dependence structure can be very informative about the actual image structure and it would be useful to estimate it [11].

Hyvärinen et al [11] used the residual dependency of the "independent" components, i.e. dependencies that could not be cancelled by ICA, to define a *topographic* order between the components. Therefore, they modified the original ICA model to include a topographic order between the components, so that components that are near to each other in the topographic representation are relatively strongly dependent in the sense of higher-order correlations or mutual information. The proposed model is usually known as the *Topographic ICA* model. The topography is introduced using a neighbourhood function $h(i,k)$, which expresses the proximity between the $i$-th and the $k$-th component. A simple neighbourhood model can be the following:

$$h(i,k) = \begin{cases} 1, \text{ if } |i-k| \leq L \\ 0, \text{ otherwise} \end{cases} \tag{17}$$

where $L$ defines the width of the neighbourhood. Consequently, the estimated coefficients $u_i$ are no longer assumed independent, but can be modelled by some generative random variables $d_k, f_i$ that are controlled by the neighbourhood function and shaped by a nonlinearity $\phi(\cdot)$ (similar to the one in the FastICA algorithm). The topographic source model, proposed by Hyvärinen et al [11], is the following:

$$u_i = \phi\left(\sum_{k=1}^{K} h(i,k) d_k\right) f_i \tag{18}$$

Assuming a fixed-width neighbourhood $L \times L$ and a PCA preprocessing step, Hyvärinen et al performed Maximum Likelihood estimation of the synthesis kernel $B$ using the linear model in (9) and the topographic source model in (18), making several assumptions for the generative random variables $d_k$ and $f_i$. Optimising an approximation of the derived log-likelihood, they formed the following gradient-based Topographic ICA rule:

$$\underline{q}_i^{+} \leftarrow \underline{q}_i + \eta \mathcal{E}\{\underline{z}(\underline{q}_i^T \underline{z}) r_i\}, \qquad 1 \leq i \leq K \tag{19}$$

$$Q \leftarrow Q(Q^T Q)^{-0.5} \tag{20}$$

where $\eta$ defines the learning rate of the gradient optimisation scheme and

$$r_i = \sum_{k=1}^{K} h(i,k) \phi\left(\sum_{j=1}^{K} h(j,k)(\underline{q}_i^T \underline{z})^2\right) \tag{21}$$

As previously, we randomly initialise the update rule in (19) for each projecting vector $\underline{q}_i$. The new updates are then orthogonalised and the whole procedure is iterated, until $\underline{a}_i$ have converged. For more details on the definition and derivation of the Topographic ICA model, one can always refer to the original work by Hyvärinen et al [11].

Finally, after estimating the matrix $Q$, using the ICA or the topographic ICA algorithm, the analysis kernel is given by multiplying the original PCA bases matrix $\hat{V}$ with $Q$.

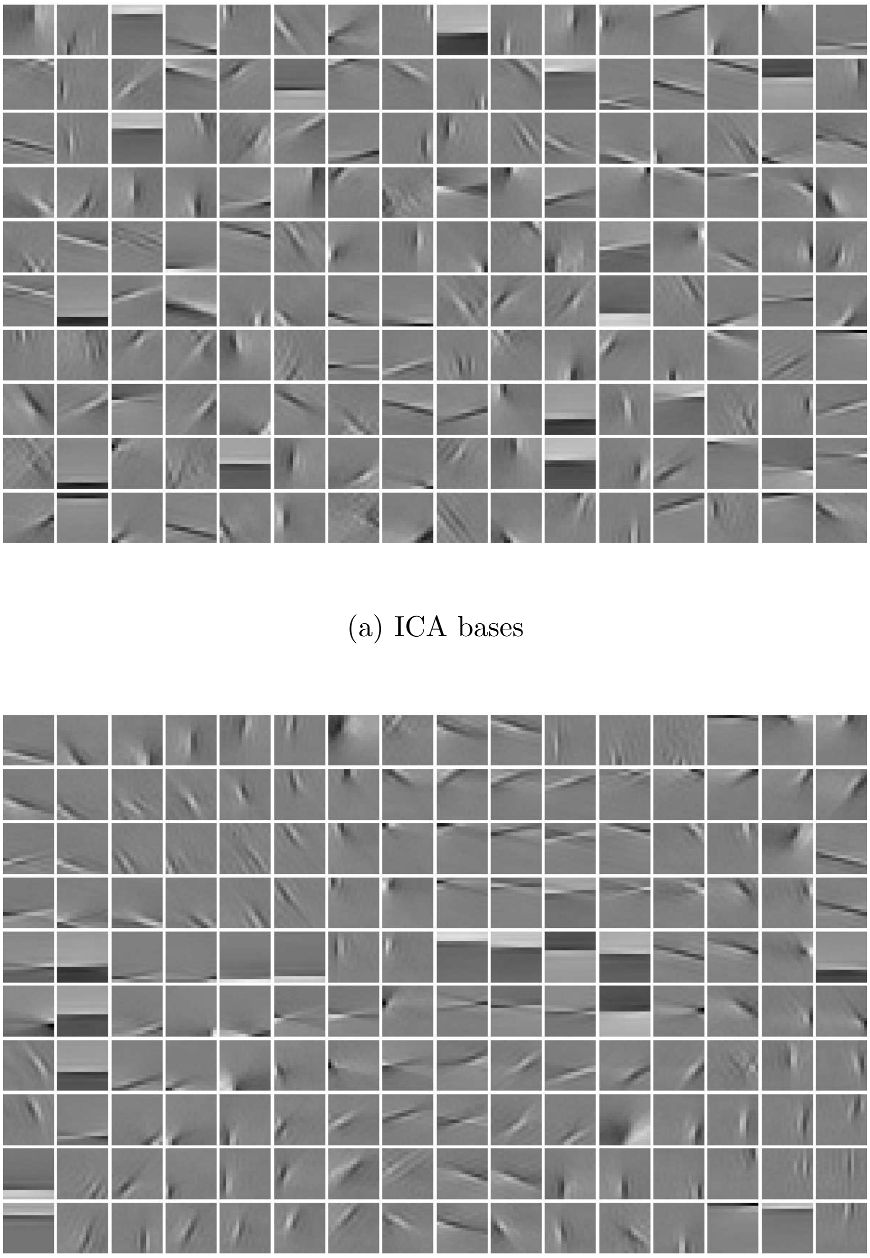
$$A \leftarrow Q\hat{V} \tag{22}$$

### *2.2 Training ICA bases*

In this paragraph, we describe the training procedure of the ICA and topographic ICA bases more thoroughly. The training procedure needs to be completed only once for each data type. After we have successfully trained the desired bases for each image type, the estimated transform can be used for fusion of similar content images.

We select a set of images with similar content to the ones that will be used for image fusion. A number of $N \times N$ patches (usually around 10000) are randomly selected from the training images. We apply lexicographic ordering to the selected images patches and normalise them to zero mean. We perform PCA on the selected patches and select the $K < N^2$ most important bases, according to the eigenvalues corresponding to the bases. It is always possible to keep the complete set of bases. Then, we iterate the ICA update rule in (15) or the topographical ICA rule in (19) for a chosen $L \times L$ neighbourhood until convergence. After each iteration, we orthogonalise the bases using the scheme in (16).

Some examples from trained ICA and topographic ICA bases are depicted in figure 2. We randomly selected 10000 $16 \times 16$ patches from natural landscape images. Using PCA, we selected the 160 most important bases out of the 256 bases available. In figure 2(a), we can see the ICA bases estimated using FastICA (15). In figure 2(b), the set of the estimated Topographic ICA bases using the rule in (19) and a $3 \times 3$ neighbourhood for the topographic model are depicted. The estimated bases feature an ordering based on similarity and correlation and thus offer a more structured and meaningful representation.



(a) ICA bases

(b) Topographic ICA bases

Fig. 2. Comparison between ICA and the topographical ICA bases trained on the same set of image patches. We can observe the spatial correlation of the bases, introduced by “topography”.

### *2.3 Properties of the ICA bases*

Let us explore some of the properties of the ICA and the Topographical ICA bases and the transforms they constitute. Both transforms are *invertible*, i.e.

they guarantee perfect reconstruction. Using the symmetric orthogonalisation step $Q \leftarrow Q(Q^TQ)^{-0.5}$, the estimated bases remain orthogonal in the ICA domain, i.e. the transform is *orthogonal*.

We can examine the estimated example set of ICA and Topographical ICA bases in figure 2. The ICA and topographical ICA basis vectors seem to be closely related to wavelets and Gabor functions, as they all represent localised edge features. However, the ICA bases have more degrees of freedom than wavelets [11]. The Discrete Wavelet transform has only two orientations and the Dual-Tree wavelet transform can give six distinct sub-bands at each level with orientation $\pm 15^o$, $\pm 45^o$, $\pm 75^o$. In contrast, the ICA bases can get arbitrary orientations to fit the training patches. On the other hand, the ICA bases do not offer a multilevel representation as the wavelet or pyramid decomposition, but only focus on localised features.

One basic drawback of the ICA-based transformations is that they are not *shift invariant* by definition. This property is generally mentioned to be very important for image fusion in literature [18]. Piella [20] comments that the fusion result will depend on the location or orientation of objects in the input sources in the case of misregistration problems or when used for image sequence fusion. As we assume that the observed images are all registered, the lack of shift invariance should not necessarily be a problem. In addition, Hyvarinen et al proposed to approximate shift invariance in these ICA schemes, by employing a *sliding window* approach [12]. This implies that the input images are not divided into distinct patches, but instead every possible $N \times N$ patch in the image is analysed. This is similar to the *spin cycling* method, proposed by Coifman and Donoho [4]. This will also increase the computational complexity of the proposed framework. The sliding window approach is only necessary for the fusion part and not for the estimation of bases.

The basic difference between ICA and topographic ICA bases is the "topography", as introduced in the latter bases. The introduction of some local correlation in the ICA model enables the algorithm to uncover some connections between the independent components. In other words, topographic bases provide an ordered representation of the data, compared to the unordered representation of the ICA bases. In an image fusion framework, "topography" can identify groups of features that can characterise certain objects in the image. One can observe the ideas comparing figures 2(a) and 2(b). Topographic ICA seems to offer a more comprehensive representation compared to the general ICA model.

Another advantage of the ICA bases is that the estimated transform can be tailored to the application field. Several image fusion applications work with specific types of images. For example, military applications work with images of airplanes, tanks, ships etc. Biomedical applications employ Computed

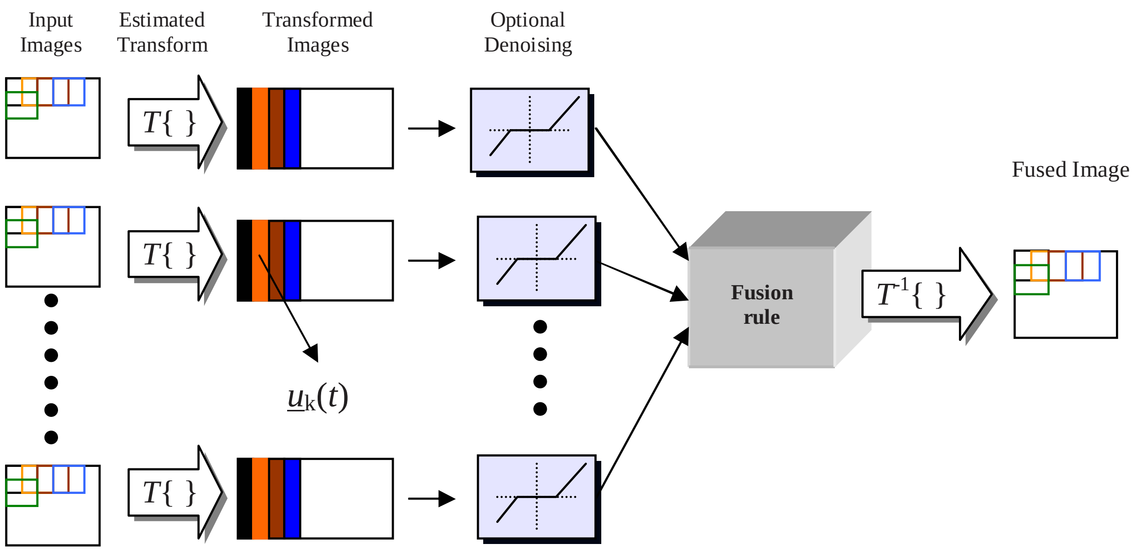


Fig. 3. The proposed fusion system using ICA / Topographical ICA bases.

Tomography (CT), Positron Emission Tomography (PET), ultra-sound scan images etc. Consequently, one can train bases for specific application areas using ICA. These bases should be able to analyse the trained data types more efficiently than a generic transform.

## 3 Image fusion using ICA bases

In this section, we describe the whole procedure of performing image fusion using ICA or Topographical ICA bases, which is summarised in figure 3 [17,16]. We assume that a ICA or Topographic ICA transform $\mathcal{T}\{\cdot\}$ is already estimated, as described in section 2.2. Also, let $I_k(x,y)$ be $T$ $M_1 \times M_2$ registered sensor images that need to be fused. From each image we isolate every possible $N \times N$ patch and using lexicographic ordering, we form the vector $\underline{I}_k(t)$. The patches' size $N$ should be the same as the one used in the transform estimation. Therefore, each image $I_k(x,y)$ is now represented by a population of $(M_1 - N)(M_2 - N)$ vectors $\underline{I}_k(t)$,$\forall$ $t \in [1,(M_1-N)(M_2-N)]$. These vectors are normalised to zero mean and the subtracted means of each vector $MN_k(t)$ are stored in order to be used in the reconstruction of the fused image. Each of these representations $\underline{I}_k(t)$ is transformed to the ICA or Topographic ICA domain representation $\underline{u}_k(t)$. Assuming that $A$ is the estimated analysis kernel, we have:

$$\underline{u}_k(t) = \mathcal{T}\{\underline{I}_k(t)\} = A\underline{I}_k(t) \tag{23}$$

Once the image representations are in the ICA domain, one can apply a "hard" threshold on the coefficients and perform optional denoising (sparse code shrinkage), as proposed by Hyvärinen et al [12]. The threshold can be determined by supervised estimation of the noise level in constant background areas of the image. Then, one can perform image fusion in the ICA or Topographic ICA domain in the same manner that is performed in the wavelet or dual-tree wavelet domain. The corresponding coefficients $\underline{u}_k(t)$ from each image are combined in the ICA domain to construct a new image $\underline{u}_f(t)$. The method $g(\cdot)$ that combines the coefficients in the ICA domain is called "fusion rule":

$$\underline{u}_f(t) = g\left(\underline{u}_1(t), \ldots, \underline{u}_k(t), \ldots, \underline{u}_T(t)\right) \tag{24}$$

Many of the proposed rules for fusion, as they were analysed in the introduction section and in literature [20,18], can be applied to this framework. The "max-abs" and the "mean" rules can be two very common options. However, one can use more efficient fusion rules, as will be presented in the next section. Once the composite image $\underline{u}_f(t)$ is constructed in the ICA domain, one can move back to the spatial domain, using the synthesis kernel $B$, and synthesise the image $I_f(x, y)$ by averaging the image patches $I_f(t)$ in the same order they were selected during the analysis step. The whole procedure can be summarised as follows:

(1) Segment all input images $I_k(x, y)$ into every possible $N \times N$ image patch and transform them to vectors $\underline{I}_k(t)$ via lexicographic ordering.
(2) Move the input vectors to the ICA / Topographic ICA domain, and get the corresponding representation $\underline{u}_k(t)$.
(3) Perform optional thresholding of $\underline{u}_k(t)$ for denoising.
(4) Fuse the corresponding coefficient using a fusion rule and form the composite representation $\underline{u}_f(t)$.
(5) Move $\underline{u}_f(t)$ to the spatial domain and reconstruct the image $I_f(x, y)$ by averaging the overlapping image patches.

## 4 Pixel-based and Region-based fusion rules using ICA bases

In this section, we describe two proposed fusion rules for ICA bases. The first one is an extension of the "max-abs" pixel-based rule, which we will refer to as the *Weight Combination* (WC) rule. The second one is a combination of the WC and the "mean" rule in a region-based scenario.

### *4.1 A Weight Combination (WC) pixel-based method*

An alternative to common fusion methods, is to use a "weighted combination" of the transform coefficients, i.e.

$$\mathcal{T}\{\underline{I}_f(t)\} = \sum_{k=1}^{T} w_k(t)\mathcal{T}\{\underline{I}_k(t)\} \tag{25}$$

There are several parameters that can be employed in the estimation of the contribution $w_k(t)$ of each image to the "fused" one. In [20], Piella proposed several *activity measures*. Following the general ideas proposed in [20], we propose the following scheme. As each image is processed in $N \times N$ patches, we can use the mean absolute value ($\mathcal{L}_1$-norm) of each patch (arranged in a vector) in the transform domain, as an activity indicator in each patch.

$$E_k(t) = ||\underline{u}_k(t)||_1 \qquad k = 1, \dots, T \tag{26}$$

The weights $w_k(t)$ should emphasise sources that feature more intense activity, as represented by $E_k(t)$. Consequently, the weights $w_k(t)$ for each patch $t$ can be estimated by the contribution of the $k$-th source image $\underline{u}_k(t)$ over the total contribution of all the $T$ source images at patch $t$, in terms of activity. Hence, we can choose:

$$w_k(t) = E_k(t) / \sum_{k=1}^{T} E_k(t) \tag{27}$$

There might be some cases, where $\sum_{k=1}^{T} E_k(t)$ is very small, denoting small edge activity or constant background in the corresponding patch. As this can cause numerical instability, the "max-abs" or "mean" fusion rule can be used for those patches. Equally, a small constant can be added to alleviate this instability.

### *4.2 Region-based Image fusion using ICA bases*

In this section, the analysis of the input images in the estimated ICA domain will be employed to perform some regional segmentation in order to fuse these regions using different rules, i.e. perform *region-based* image fusion. During, the proposed analysis methodology, we have already divided the image in small $N \times N$ patches (i.e. regions). Using the splitting/merging philosophy of region-based segmentation [23], a criterion is employed to merge the pixels corresponding to each patch in order to form contiguous areas of interest.

One could use the energy activity measurement, as introduced by (26), to infer the existence of edges in the corresponding frame. As the ICA bases tend to focus on the edge information, it is clear that great values for $E_k(t)$, correspond to increased activity in the frame, i.e. the existence of edges. In contrast, small values for $E_k(t)$ denote the existence of almost constant background or insignificant texture in the frame. Using this idea, we can segment the image in two regions: i) "active" regions containing details and ii) "non-active" regions containing background information. The threshold that will be used to characterise a region as "active" or "non-active" can be set heuristically to $2\text{mean}_t\{E_k(t)\}$. Since the aim here is to create the most accurate edge-detector, we can allow some tolerance around the real edges of the image. As a result, we form the following segmentation map $m_k(t)$ from each input image:

$$m_k(t) = \begin{cases} 1, \text{ if } E_k(t) > 2\text{mean}_t\{E_k(t)\} \\ 0, \text{ otherwise} \end{cases} \tag{28}$$

The segmentation map of each input image is combined to form a single segmentation map, using the logical OR operator. As mentioned earlier, we are not interested in forming a very accurate edge detection map, but instead it is important to ensure that our segmentation map contains most of the strong edge information.

$$m(t) = \text{OR}\{m_1(t), m_2(t), \dots, m_T(t)\} \tag{29}$$

Once the image has been segmented into "active" and "non-active" regions, we can fuse these regions using different pixel-based fusion schemes. For the "active" region, we can use a fusion scheme that preserves the edges, i.e. the "max-abs" scheme or the weighted combination scheme and for the "non-active" region, we can use a scheme that preserves the background information, i.e. the "mean" or "median" scheme. Consequently, this could form a more accurate fusion scheme that looks into the actual structure of the image itself, rather than fuse information generically.

## 5 A general optimisation scheme for image fusion

In this section, the focus is placed on defining an unsupervised image fusion approach based on the minimisation of a formulated cost function involving several source images. The main aim is to achieve visual improvements over the original source images, such that certain specific features in the original source images can be detected visually or through various models in the fused image.

Practical usage of this algorithm includes the confirmation of a particular target in military purposes, when several different source images are obtained from different sensors under different conditions [16].

The minimisation of a cost function involves the estimation of a set of optimal parameters that will minimise the output value of the cost function. This concept can thus be incorporated into the process of image fusion to obtain a set of optimal coefficients that can be used to produce a fused image of better quality than each of the original source images.

Let us assume that we are interested in the $N \times N$ patches around pixel $(x_0, y_0)$ in the input sensor image $I_1, \ldots, I_T$. These patches are lexicographically ordered, as described in the previous section, to form the vectors $\underline{I}_1, \ldots, \underline{I}_T$. We also assume that an ICA transform $\mathcal{T}\{\cdot\}$ has been trained, using patches of similar content images. In this case, we will be using a complete representation, i.e. $K = N^2$, although any overcomplete representation may also be used. The input patches in the transform domain are denoted by $\underline{u}_i = \mathcal{T}\{\underline{I}_i\}$. The fused image $\underline{u}_f$ in the transform domain can be given by the following linear combination:

$$\underline{u}_f = w_1 \underline{u}_1 + w_2 \underline{u}_2 + \ldots + w_T \underline{u}_T \tag{30}$$

where $w_1, \ldots, w_T$ are scalar coefficients that denote the mixing of each input sensor patch in the transform domain. We denote $\underline{w} = [w_1 \ w_2 \ \ldots \ w_T]^T$. All elements of vector $\underline{u}_i$ will contribute in the formation of the fused image, according to the weight $w_i$. Let us now define:

$$\underline{x}(n) = [u_1(n) \ u_2(n) \ \ldots \ u_T(n)]^T \qquad \forall \ n = 1, \ldots, N^2 \tag{31}$$

Hence, the fusion procedure can be equivalently described by the following product:

$$u_f(n) = \underline{w}^T \underline{x}(n) \qquad \forall \ n = 1, \ldots, N^2 \tag{32}$$

The problem of fusion can now be described as an optimisation problem of estimating $\underline{w}$, so that the fused image follows certain properties, described by the cost function. A logical assumption is that the fusion process should enhance *sparsity* in the ICA domain. In other words, the fusion should emphasize the existence of strong coefficients in the transform, whilst suppress small values. We will approach the problem of estimating $\underline{w}$, using a ML estimation approach, assuming several probabilistic priors, that describe sparsity.

The connection between *sparsity* and ICA representations has been investigated thoroughly by Olshausen [19]. The basis functions that emerge when adapted to static, whitened natural images under the assumption of statistical independence, resemble the Gabor-like spatial profiles of cortical simple-cell receptive fields. That is to say that the functions become spatially localised, oriented and bandpass. Because all of these properties emerge purely from the objective of finding sparse, independent components for natural images, the results suggest that the receptive fields of V1 neurons have been designed under the same principle. Therefore, the actual non-distorted representation of the observed scene in the ICA domain should be more sparse than the distorted or different sensor input. Consequently, an algorithm that maximises the sparsity of the fused image in the ICA domain can be justified.

### *5.1 Laplacian priors*

Assuming a Laplacian model for $u_f(n)$, we can perform Maximum Likelihood (ML) estimation of $\underline{w}$. The Laplacian probability density function is given below:

$$p(u_f) \propto e^{-\alpha|u_f|} \tag{33}$$

where $\alpha$ is a parameter that controls the width (variance) of the Laplacian. The likelihood expression for ML estimation can be given by:

$$\begin{aligned} L_n &= -\log p(u_f|\theta_n) \\ &\propto -\log e^{-\alpha|u_f|} = \alpha|u_f| \\ &= \alpha|\underline{w}^T \underline{x}(n)| \end{aligned} \tag{34}$$

Maximum Likelihood estimation can be performed by maximising the cost function $J(\underline{w}) = \mathcal{E}\{L_n\}$. Hence, the optimisation problem to be solved is the following:

$$\max_{w} \mathcal{E}\{\alpha|\underline{w}^T \underline{x}|\} \tag{35}$$

$$\text{subject to} \quad \underline{e}^T \underline{w} = 1 \tag{36}$$

$$\underline{w} > 0 \tag{37}$$

where $\underline{e} = [1\ 1\ \ldots\ 1]^T$. To begin evaluate the solutions to this problem, we can firstly calculate the first derivative:

$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}} \mathcal{E}\{\alpha|\underline{w}^T \underline{x}|\} = \alpha \mathcal{E}\{\text{sgn}(\underline{w}^T \underline{x})\underline{x}\} \tag{38}$$

To solve the above optimisation problem, one has to consult methods for constraints optimisation. Using the Lagrange multipliers method for equality constraints and the Kuhn-Tucker conditions for inequality constraints is definitely going to increase the computational complexity of the algorithm. In addition, the available data points for the estimation of the expectation are limited to $N^2$. Therefore, we propose to solve the unconstrained optimisation problem using a gradient ascent method and impose the constraints at each stage of the adaptation. Consequently, the proposed algorithm can be summarised, as follows:

(1) Initialise $\underline{w} = \underline{e}/T$. This implies the mean fusion rule, i.e. equal importance to all input patches.

(2) Update the weight vector, as follows:

$$\underline{w}^+ \leftarrow \underline{w} + \eta\mathcal{E}\{\mathrm{sgn}(\underline{w}^T\underline{x})\underline{x}\} \tag{39}$$

where $\eta$ represents the learning rate

(3) Apply the constraints, using the following update rule:

$$\underline{w}^+ \leftarrow |\underline{w}|/(\underline{e}^T|\underline{w}|) \tag{40}$$

(4) Iterate steps $2, 3$ until convergence.

Effectively, equation (40) ensures that the weights $w_i$ remain always positive and they sum up to one, as it is essential not to introduce any sign or scale deformation during the estimation of the fused image.

### *5.2 Verhulstian priors*

The main drawback of using Laplacian priors is the use of the $\mathrm{sgn}(u)$ function in the update algorithm, that has a discontinuity at $u \to 0$ and therefore may cause numerical instability and errors during the update. Usually, this problem is alleviated by thresholding $u$ by a small constant, so that $u$ never gets zero values. Therefore, one can use alternate probabilistic priors that denote sparsity, such as the *generalised Laplacian* or the *Verhulstian* distribution. In the section, we will examine the use of Verhulstian priors in the ML estimation of the fused image.

The *Verhulstian* probability density function can be defined, as follows:

$$p(u) = \frac{e^{-\frac{u-m}{s}}}{s\left(1 + e^{-\frac{u-m}{s}}\right)^2} \tag{41}$$

where $m$, $s$ are parameters that control the mean and the standard deviation of the density function. In our case, we will assume zero mean and therefore $m = 0$. We can now derive the log-likelihood function for ML estimation:

$$
\begin{aligned}
L_n &= -\log \frac{e^{-\frac{u_f}{s}}}{s\left(1+e^{-\frac{u_f}{s}}\right)^2} \\
&= \frac{u_f}{s} + \log s + 2\log\left(1+e^{-\frac{u_f}{s}}\right) \\
&= \frac{1}{s}\underline{w}^T\underline{x} + \log s + 2\log\left(1+e^{-\frac{1}{s}\underline{w}^T\underline{x}}\right) \qquad (42)
\end{aligned}
$$

Maximum Likelihood estimation can be performed in a similar fashion to Laplacian priors, by maximising the cost function $J(\underline{w}) = \mathcal{E}\{L_n\}$. Again, a gradient ascent algorithm is employed, as explained in the previous section with a correcting step that will constrain the solutions in the solution space, permitted by the optimisation problem. The gradient is calculated, as follows:

$$
\begin{aligned}
\frac{\partial J(\underline{w})}{\partial \underline{w}} &= \frac{\partial}{\partial \underline{w}}\mathcal{E}\left\{\frac{1}{s}\underline{w}^T\underline{x} + \log s + 2\log\left(1+e^{-\frac{1}{s}\underline{w}^T\underline{x}}\right)\right\} \\
&= \mathcal{E}\left\{\frac{1}{s}\underline{x} - \frac{1}{s}\underline{x}\frac{2e^{-\frac{1}{s}\underline{w}^T\underline{x}}}{1+e^{-\frac{1}{s}\underline{w}^T\underline{x}}}\right\} \\
&= \frac{1}{s}\mathcal{E}\left\{\frac{1-e^{-\frac{1}{s}\underline{w}^T\underline{x}}}{1+e^{-\frac{1}{s}\underline{w}^T\underline{x}}}\underline{x}\right\} \qquad (43)
\end{aligned}
$$

We can now perform the same algorithm as introduced for Laplacian priors, the only difference being that in equation (39), we have to replace the gradient with that of equation (43). Consequently, the algorithm can be outlined as follows:

(1) Initialise $\underline{w} = \underline{e}/T$. This implies the mean fusion rule, i.e. equal importance to all input patches.

(2) Update the weight vector, as follows:

$$\underline{w}^+ \leftarrow \underline{w} + \eta\mathcal{E}\left\{\frac{1-e^{-\frac{1}{s}\underline{w}^T\underline{x}}}{1+e^{-\frac{1}{s}\underline{w}^T\underline{x}}}\underline{x}\right\} \qquad (44)$$

where $\eta$ represents the learning rate

(3) Apply the constraints, using the following update rule:

$$\underline{w}^+ \leftarrow |\underline{w}|/(\underline{e}^T|\underline{w}|) \qquad (45)$$

(4) Iterate steps $2, 3$ until convergence.

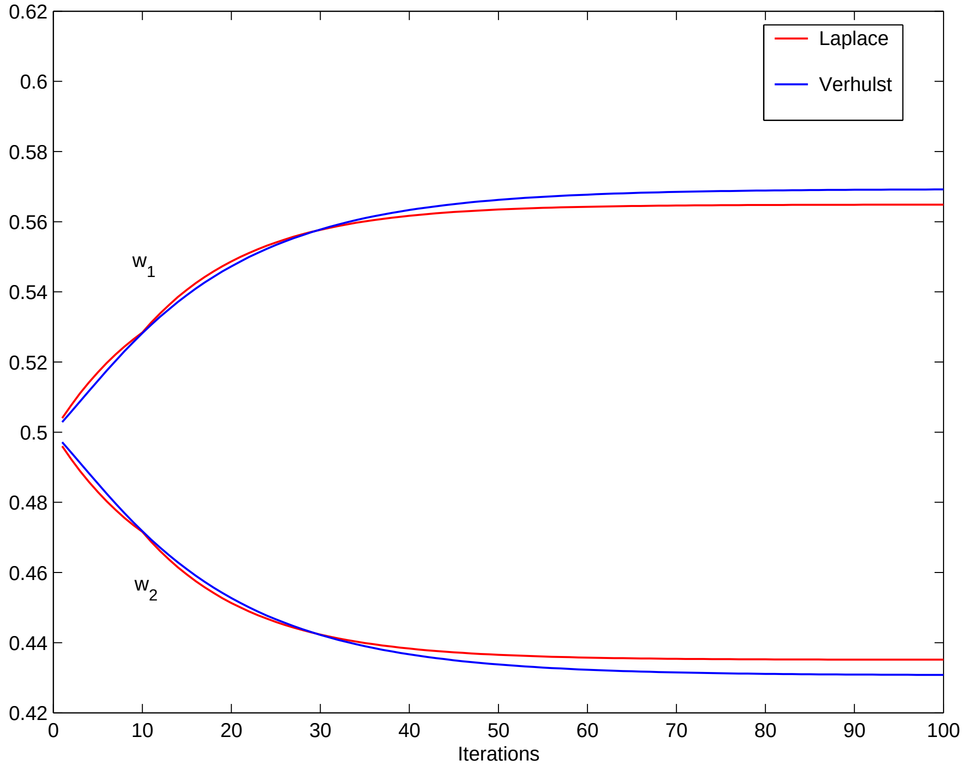


Fig. 4. Typical convergence of the ML-estimation fusion scheme using Laplacian and Verhulstian priors.

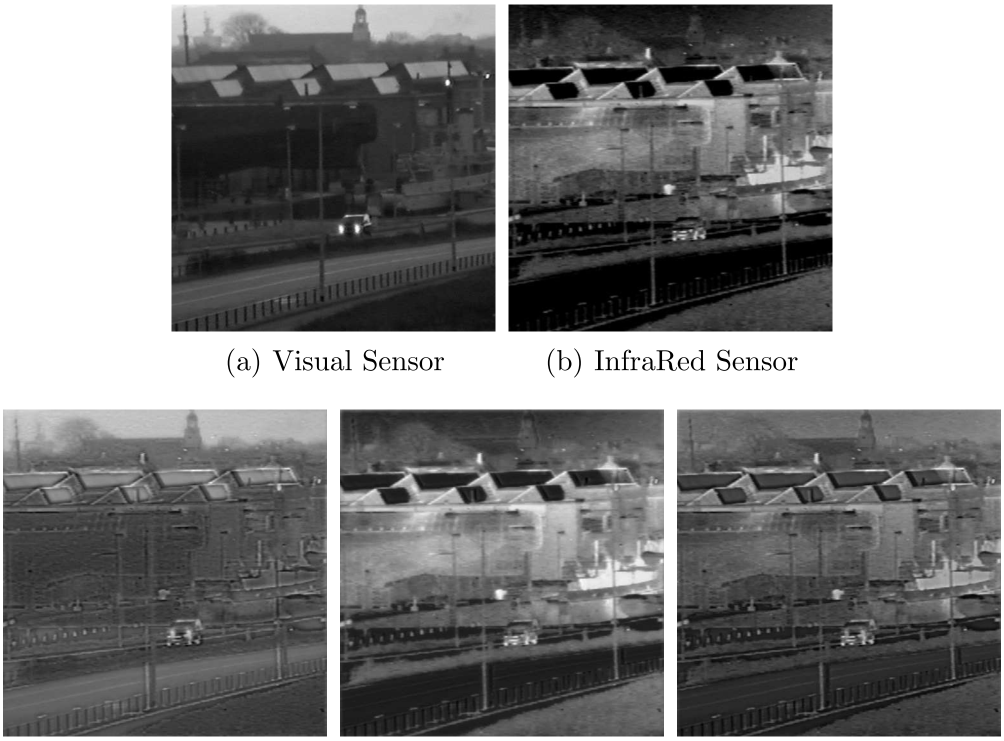
In figure 4, a typical convergence of the two ML-estimation schemes using the two proposed priors is shown. The algorithms converge smoothly after an average of $50 - 60$ iterations.

## 6 Reconstruction of the fused image

The above algorithms have provided a number of possible methods to estimate the fused image $\underline{u}_f(t)$ in the ICA transform domain. The next step is to estimate the spatial-domain representation of the image $I_f(x, y)$. To reconstruct the image in the spatial domain, the process described in Section 2 is inverted. The vectors $\underline{u}_f(t)$ are re-transformed to the local $N \times N$ patches $I_f(k, l)$. The local mean of each patch is restored using the stored patches means $MN_k(t)$. The patches are consequently averaged with 1-pixel overlap to create the grid in Figure 1, i.e. the fused image. This averaging usually creates an artificial “frame” around the reconstructed image, which occurs due to the reduced number of frames that are available around the image’s borders. To overcome this effect, one can pad with zeros the borders of the input sensors images before the fusion stage, so that the “framing” effect affects the zero-padded areas only.

The restoration of the patches' local means is a very important issue. Initially, all the patches were normalised to zero mean and the subtracted local intensity mean $MN_k(t)$ was stored to be used in the reconstruction of the fused image. Consequently, there exist $T$ local intensity values for each patch of the reconstructed image, each belonging to the corresponding input sensor. In the case of performing multi-focus image fusion, it is evident that the local intensities from all input sensors will be similar, if not equal, for all corresponding patches. In this case, the local means are reconstructed by averaging the $MN_k(t)$, in terms of $k$. In the case of multi-modal image fusion, the problem of reconstructing the local intensities of the fused image becomes more serious, since the $T$ input images are acquired from different modality sensors with different intensity range and values. The fused image is an artificial image, that does not exist in nature, and it is therefore difficult to find a criterion that can dictate the most efficient way of combining the input sensors intensity range. The details from all input images will be transferred to the fused image by the fusion algorithm, however, the local intensities will be selected to define the intensity profile of the fused image. In Figure 5, the example of a multi-modal fusion scenario is displayed: a visual sensor image is fused with an infrared sensor image. Three possible reconstructions of the fused image's means are shown: a) the contrast (local means) is acquired from the visual sensor, b) the contrast is acquired from the infrared image and c) an average of the local means is used. All three reconstructions contain the same salient features, since these are dictated by the ICA fusion procedure. Each of the three reconstructions simply gives a different impression of the fused image, depending on the prevailing contrast preferences. The average of the local means seems to give a more balanced representation compared to the two extremes. The details are visible in all three reconstructions. However, an incorrect choice of local means may render some of the local details, previously visible in some of the input sensors, totally invisible in the fused image and therefore deteriorate the fusion performance. In this chapter, we will use the average of the local means, giving equal importance to all input sensors. However, there might be another optimum representation of the fused image, by perhaps emphasising means from input sensors with greater intensity range.

An additional problem can be the creation of a "colour" fused image, as the result of the fusion process. Let us assume that one of the input sensors is a visual sensor. In most real-life situations the visual sensor will provide a colour input image or in other terms a number of channels representing the colour information provided by the sensor. The most common representation in Europe is the RGB (Red-Green-Blue) representation featuring 3 channels of the three basic colours. If the traditional fusion methodology is applied on this problem, a single channel "fused" image will be produced featuring only intensity changes in grayscale. However, most users and operators will demand a colour rather than a grayscale representation of the "fused" image. There are several surveillance applications, where a colour "fused" image is



(a) Visual Sensor

(b) InfraRed Sensor

(c) Means from Visual Sensor

(d) Means from InfraRed Sensor

(e) Average Means

Fig. 5. Effect of local means choice in the reconstruction of the fused image.

expected from a visual and an Infrared sensor [27]. Even in the case of a grayscale visual input sensor and other infrared, thermal sensors, the operator is more likely to prefer a synthetic colour representation of the "fused" image, rather than a grayscale one [26]. Therefore, the problem of creating a 3-channel representation of the "fused" image from $T$ channels available by the input sensors can be rather demanding.

A first thought would be to treat each of the visual color channels independently and fuse them with the input channels from the other sensors independently to create a three channel representation of the "fused" image. Although this technique seems rational and may produce satisfactory results in several cases, it does not utilise the dependencies between the colour channels that might be beneficial for the fusion framework [2]. Another proposed approach [2,27] was to move to another color space, such as the YUV color space that describes a colour image using one luminance and two chrominance channels [2] or the HSV color space that describes a colour image using Hue, Saturation and Intensity (luminance) channels. The two chrominance channels as well as the hue-saturation channels convey colour information solely, whereas the Intensity channel describes the image details more accurately. Therefore, the proposed strategy is to fuse the intensity channel with the other input sensor channels and create the intensity channel for the "fused" image. The chrominance/hue-saturation channels can be used to provide color information for the “fused” image. This scheme features reduced computational complexity as one visual channel is fused instead of the original three. In addition, as all these colour transformations are linear mappings from the RGB space, one can use Principal Component Analysis to define the principal channel in terms of maximum variance. This channel is fused with the other input sensors and the resulting image is mapped back to the RGB space, using the estimated PCA matrix. The above techniques are producing satisfactory results in the case of colour out-of-focus input images, since all input images have the same chrominance channels. In the case of multi-modal or multi-exposure images, these methods may not be sufficient and then one can use more complicated color channel combination and fusion schemes in order to achieve an enhance “fused” image [27]. These schemes may offer enhanced performance for selected applications only but not in every possible fusion scenario.

## 7 Experiments

In this section, we test the performance of the proposed image fusion schemes based on ICA bases. It is not our intention to provide an exhaustive comparison of the many different transforms and fusion schemes that exist in literature. Instead, a comparison with fusion schemes using *wavelet packets* analysis and the *Dual-Tree (Complex) Wavelet Transform* are performed. In these examples we will test the “fusion by absolute maximum” (maxabs), the “fusion by averaging” (mean), the Weighted Combination (weighted), the Region-based (Regional) fusion and the adaptive (Laplacian prior) fusion rules.

We present three experiments, using both artificial and real image data sets. In the first experiment, the *Ground Truth* image $I_{gt}(x, y)$ is available, enabling us to perform explicit numerical evaluation of the fusion schemes. We assume that the input images $I_i(x, y)$ are processed by the fusion schemes to create the “fused” image $I_f(x, y)$. To evaluate the scheme’s performance, we can use the following *Signal-to-Noise Ratio* (SNR) expression to compare the ground truth image with the fused image.

$$SNR_{(dB)} = 10 \log_{10} \frac{\sum_x \sum_y I_{gt}(x, y)^2}{\sum_x \sum_y (I_{gt}(x, y) - I_f(x, y))^2} \tag{46}$$

As traditionally employed by the fusion community, we can also use the *Image Quality Index* $Q_0$, as a performance measure [25]. Assume that $m_I$ represents the mean of the image $I(x, y)$ and all images are of size $M_1 \times M_2$. As $-1 \leq$

$Q_0 \leq 1$, the value of $Q_0$ that is closer to 1, indicates better fusion performance.

$$Q_0 = \frac{4\sigma_{I_{gt}I_f} m_{I_{gt}} m_{I_f}}{(m_{I_{gt}}^2 + m_{I_f}^2)(\sigma_{I_{gt}}^2 + \sigma_{I_f}^2)} \tag{47}$$

where

$$\sigma_I^2 = \frac{1}{M_1 M_2 - 1} \sum_{x=1}^{M_1} \sum_{y=1}^{M_2} (I(x,y) - m_I)^2 \tag{48}$$

$$\sigma_{IJ} = \frac{1}{M_1 M_2 - 1} \sum_{x=1}^{M_1} \sum_{y=1}^{M_2} (I(x,y) - m_I)(J(x,y) - m_J) \tag{49}$$

For the rest of the experiments, as the "ground truth" image is not available, two Image Fusion performance indexes will be used: one proposed by Piella [21] and one proposed by Petrovic and Xydeas [28]. Both indexes are widely used by the image fusion community to benchmark the performance of fusion algorithms. They both attempt at quantifying the amount of "interesting" information (edge information) that has been conveyed from the input images to the fused image. In addition, as Piella's index employs the Image Quality Index $Q_0$ to quantify the quality of information transfer between each of the input images and the fused image, it is bounded between $-1$ and 1.

The ICA and the topographic ICA bases were trained using 10000 $8 \times 8$ image patches that were randomly selected from 10 images of similar content to the ground truth or the observed scene. We used 40 out of the 64 possible bases to perform the transformation in either case. The local means of the fused image were reconstructed using an average of the means of the input sensor images. We compared the performance of the ICA and topographic ICA transforms (topoICA) with a Wavelet Packet decomposition [1] and the Dual-Tree Wavelet Transform [2]. For the Wavelet Packet decomposition (WP), we used Symmlet-7 (Sym7) bases, with 5 level-decomposition using Coifman-Wickerhauser entropy. For the Dual-Tree Wavelet Transform (DTWT), we used 4 levels of decomposition and the filters included in the package. In the next pages, we will present some of the resulting fusion images. However, the visual differences between the fused images may not be very clear in the printed version of this chapter, due to limitation in space. Consequently, the reader is prompted to acquire the whole set either by download [3] or via email to us.

[1] We used WaveLab v8.02, as available at http://www-stat.stanford.edu/∼wavelab/.

[2] DT-WT code available online by the Polytechnic University of Brooklyn, NY at http://taco.poly.edu/WaveletSoftware/

[3] http://www.commsp.ee.ic.ac.uk/∼nikolao/BookElsevierImages.zip

### *7.1 Experiment 1: Artificially distorted images*

In the first experiment, we have created three images of an "airplane" using different localised artificial distortions. The introduced distortions can model several different types of degradation that may occur in visual sensor imaging, such as motion blur, out-of-focus blur and finally pixelate or shape distortion, due to low bit-rate transmission or channel errors. This synthetic example can be a good starting point for evaluation, as there are no registration errors between the input images and we can perform numerical evaluation, as we have the ground truth image. We applied all possible combinations of transforms and the fusion rules (the "Weighted" and "Regional" fusion rules can not be applied in the described form for the WP and DTWT transforms). Some results are depicted in figure 7, whereas the full numerical evaluation is presented in table 1.

We can see that using the ICA and the TopoICA bases, we can get better fusion results both in visual quality and metric quality (PSNR, $Q_0$). We observe the ICA bases provide an improvement of $\sim 2-4$ dB, compared to the wavelet transforms, using the "maxabs" rule. The topoICA bases seem to score slightly better than the normal ICA bases, mainly due to better adaptation to local features. In terms of the various fusion schemes, the "max-abs" rule seems to give very low performance in this example using visual sensors. This can be explained, due to the fact that this scheme seems to highlight the important features of the images, however, it tends to lose some constant background information. On the other hand, the "mean" rule gives the best performance (especially for the wavelet coefficient), as it seems to balance the high detail with the low-detail information. However, the "fused" image in this case seems quite "blurry", as the fusion rule has oversmoothed the image details. Therefore, the high SNR has to be cross-checked with the actual visual quality and image perception, where we can clearly that the salient features have been filtered. The "weighted combination" rule seems to balance the pros and cons of the two previous approaches, as the results feature high $PSNR$ and $Q_0$ (inferior to the "mean" rule), but the "fused" images seem sharper with correct constant background information. In figure 6, we can see the segmentation map created by (18) and (19). The proposed region-based scheme manages to capture most of the salient areas of the input images. It performs reasonably well as an edge detector, however, it produces thicker edges, as the objective is to identify areas around the edges, not the edges themselves. The region-based fusion scheme produces similar results to the "Weighted" fusion scheme. However, it seems to produce better visual quality in constant background areas, as the "mean" rule is more suitable for the "non-active" regions. The adaptive system based on the Laplacian prior seems to achieve the maximum performance in the case of Topographic ICA bases, but not on the trained ICA bases, where it matches the "mean" rule performance.

Table 1
Performance comparison of several combinations of transforms and fusion rules in terms of $PSNR$ (dB)/$Q_0$ using the “airplane” example.

| | WP (Sym7) | DT-WT | ICA | TopoICA |
|---|---|---|---|---|
| Max-abs | 14.03/0.8245 | 13.77/0.8175 | 16.28/0.9191 | 17.49/0.9354 |
| Mean | 23.19/0.9854 | 23.19/0.9854 | 20.99/0.9734 | 21.21/0.9752 |
| Weighted | - | - | 21.18/0.9747 | 21.41/0.9763 |
| Regional | - | - | 21.17/0.9746 | 21.42/0.9764 |
| Laplacian | - | - | 20.99/0.9734 | 21.73/0.9782 |

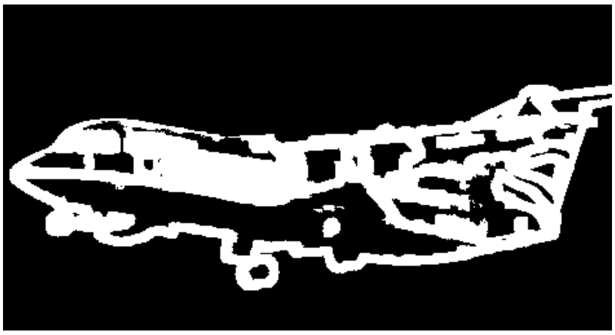


Fig. 6. Region mask created for the region-based image fusion scheme. The white areas represent “active” segments and the black areas “non-active” segments.

### *7.2 Experiment 2: Out-of-focus image fusion*

In the second experiment, we use the “Clocks” and the “Disk” examples, which are real visual sensor example provided by Lehigh Image Fusion group [6]. In these examples, there are two registered images with different focus points, observing two complicated scenes. In the first image of each set, the focus is on left part and in the second image the focus is on the right part of the image. The ground truth image is not available, which is common in many multi-focus examples. Therefore, SNR-type measurements are not available in this case. Instead, the Piella fusion index [21] and the Petrovic fusion index [28] were used and are depicted in Table 2 for various combinations of fusion rules and transform domains. In Figures 8, 9 the resulting fused images for different configurations of the two experiments are depicted.

Here, we can see that the ICA and TopoICA bases perform slightly better than wavelet-based approaches in the first example and a lot better in the second example. Also, we can see that the “maxabs” rule performs slightly better than any other approach, with almost similar performance from the “Weighted” scheme. The reason is that the three images have the same colour information, however, most parts of each image are blurred. Therefore, the “maxabs” that identifies the greatest activity, in terms of edge information, seems more suitable for a multi-focus example. The "maxabs" simply strengthens the existence of edges in the fused image and can therefore in an out-of-focus situation can excel in restoring these blurred parts of the input images.

Table 2
Performance comparison of several combinations of transforms and fusion rules for out-of-focus datasets, in terms of the Piella/Petrovic indexes.

| | WP (Sym7) | DT-WT | ICA | TopoICA |
|---|---|---|---|---|
| | Clocks dataset | | | |
| Max-abs | 0.8727/0.6080 | 0.8910/0.6445 | 0.8876/0.6530 | 0.8916/0.6505 |
| Mean | 0.8747/0.5782 | 0.8747/0.5782 | 0.8523/0.5583 | 0.8560/0.5615 |
| Weighted | - | - | 0.8678/0.6339 | 0.8743/0.6347 |
| Regional | - | - | 0.8583/0.5995 | 0.8662/0.5954 |
| Laplacian | - | - | 0.8521/0.5598 | 0.8563/0.5624 |
| | Disk dataset | | | |
| Max-abs | 0.8850/0.6069 | 0.8881/0.6284 | 0.9109/0.6521 | 0.9111/0.6477 |
| Mean | 0.8661/0.5500 | 0.8661/0.5500 | 0.8639/0.5470 | 0.8639/0.5459 |
| Weighted | - | - | 0.9134/0.6426 | 0.9134/0.6381 |
| Regional | - | - | 0.9069/0.6105 | 0.9084/0.6068 |
| Laplacian | - | - | 0.8679/0.5541 | 0.8655/0.5489 |

### *7.3 Experiment 3: Multi-modal image fusion*

In the third experiment, we explore the performance in multi-modal image fusion. In this case, the input images are acquired from different modality sensors to unveil different components in the observed scene. We have used some surveillance images from TNO Human Factors, provided by L. Toet [24]. More of these can be found in the Image Fusion Server [5]. The images are acquired by three kayaks approaching the viewing location from far away. As a result, their corresponding image size varies from less than 1 pixel to almost the entire field of view, i.e. they are minimal registration errors. The first sensor (AMB) is a Radiance HS IR camera (Raytheon), the second (AIM) is an AIM 256 microLW camera and the third is a Philips LTC500 CCD camera. Consequently, we get three different modality inputs for the same observed scene. The third example is taken from the "UN Camp" dataset available from the Image Fusion Server [5]. In this case, the inputs consist of a grayscale visual sensor and an infrared sensor. The Piella fusion index [21] and the Petrovic fusion index [28] are measured and are depicted in Table 3 for various combinations of fusion rules and transform domains.

In this example, we can witness some minor effects of misregistration in the fused image. We can see that all four transforms seem to have included most salient information from the input sensor images, especially in the "maxabs" and "weighted" schemes. However, the ICA and the TopoICA bases approaches seem to excel in comparison to the dual-tree wavelet transform and the wavelet packet approaches. The "fused image" constructed using the proposed framework seems to be sharper and less blurry compared to the other approaches, especially in the case of the "maxabs" and "weighted" schemes. These observations can be verified in Figures 10, 11 and 12, where some of the produced fused images are depicted for various configurations. The other proposed schemes offer reasonable performance in all multi-modal examples, but not the optimal.

## 8 Conclusion

The authors have introduced the use of ICA and Topographical ICA bases for image fusion applications. These bases seem to construct very efficient tools, which can compliment common techniques used in image fusion, such as the Dual-Tree Wavelet Transform. The proposed method can outperform wavelet approaches. The Topographical ICA bases offer more accurate directional selectivity, thus capturing the salient features of the image more accurately. A weighted combination image fusion rule seemed to improve the fusion quality over traditional fusion rules in several cases. In addition, a region-based approach was introduced. At first, segmentation into "active" and "non-active" areas is performed. The "active" areas are fused using the pixel-based weighted combination rule and the "non-active" areas are fused using the pixel-based "mean" rule. An adaptive fusion rule based on the sparsity of the coefficients in the ICA-domain was also introduced. Sparsity was modelled using either Laplacian or Verhulstian prior with promising results. The proposed framework was tested with an artificial example, two out-of-focus examples and three multi-modal, outperforming current state-of-the-art approaches based on the wavelet transform.

The proposed schemes seem to increase the computational complexity of the image fusion framework. The extra computational cost is not necessarily introduced by the estimation of the ICA bases, as this task is performed only once. The bases can be trained offline using selected image samples and then employed constantly by the fusion applications. The increase in complexity comes from the "sliding window" technique that is introduced to achieve shift invariance. Implementing this fusion scheme in a more computationally effi-

Table 3
Performance comparison of several combinations of transforms and fusion rules for multimodal datasets, in terms of the Piella/Petrovic indexes.

| | WP (Sym7) | DT-WT | ICA | TopoICA |
|---|---|---|---|---|
| | Multimodal-1 dataset | | | |
| Max-abs | 0.6198/0.4163 | 0.6399/0.4455 | 0.6592/0.4507 | 0.6646/0.4551 |
| Mean | 0.6609/0.3986 | 0.6609/0.3986 | 0.6591/0.3965 | 0.6593/0.3967 |
| Weighted | - | - | 0.6832/0.4487 | 0.6861/0.4528 |
| Regional | - | - | 0.6523/0.3885 | 0.6566/0.3871 |
| Laplacian | - | - | 0.6612/0.3980 | 0.6608/0.3983 |
| | Multimodal-2 dataset | | | |
| Max-abs | 0.5170/0.4192 | 0.58022/0.4683 | 0.6081/0.4759 | 0.6092/0.4767 |
| Mean | 0.6028/0.420 | 0.6028/0.4207 | 0.6056/0.4265 | 0.6061/0.4274 |
| Weighted | - | - | 0.6252/0.4576 | 0.6286/0.4632 |
| Regional | - | - | 0.5989/0.4148 | 0.5992/0.4133 |
| Laplacian | - | - | 0.6071/0.4277 | 0.6068/0.4279 |
| | ‘‘UN Camp’’ dataset | | | |
| Max-abs | 0.6864/0.4488 | 0.7317/0.4780 | 0.7543/0.4906 | 0.7540/0.4921 |
| Mean | 0.7104/0.4443 | 0.7104/0.4443 | 0.7080/0.4459 | 0.7081/0.4459 |
| Weighted | - | - | 0.7361/0.4735 | 0.7429/0.4801 |
| Regional | - | - | 0.7263/0.4485 | 0.7321/0.4508 |
| Laplacian | - | - | 0.7101/0.4475 | 0.7094/0.4473 |

cient framework than MATLAB will decrease the time required for the image analysis and synthesis part of the algorithm.

For future work, the authors would be looking at evolving to a more autonomous fusion system, exploring the nature of “topography”, as introduced by Hyvärinen et al, and form more efficient activity detectors, based on topographic information. In addition, they would be looking at more sophisticated methods for the selection of intensity or colour range of the fused image in the case of multi-modal or colour image fusion.

## Acknowledgements

This work was supported by the Data Information Fusion Phase-I project 6.4 and the Phase-II AMDF cluster project of the Defence Technology Centre, UK.

## References

[1] I. Bloch and H. Maitre. Data fusion in 2d and 3d image processing: An overview. In *Proc. X Brazilian symposium on Computer Graphics and Image Processing*, pages 122–134, 1997.

[2] L. Bogoni, M. Hansen, and P. Burt. Image enhancement using pattern-selective color image fusion. In *Proc. Int. Conf on Image Analysis and Processing*, pages 44–49, 1999.

[3] A. Cichocki and S.I. Amari. *Adaptive Blind Signal and Image Processing. Learning algorithms and applications.* John Wiley & Sons, 2002.

[4] R.R. Coifman and D.L. Donoho. Translation-invariant de-noising. Technical report, Department of Statistics, Stanford University, Stanford, California, 1995.

[5] The Image fusion server. http://www.imagefusion.org/.

[6] Lehigh fusion test examples. http://www.eecs.lehigh.edu/spcrl/if/toy.htm.

[7] A. Goshtasby. *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications.* John Wiley & Sons, 2005.

[8] P. Hill, N. Canagarajah, and D. Bull. Image fusion using complex wavelets. In *Proc. 13th British Machine Vision Conference*, Cardiff, UK, 2002.

[9] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[10] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[11] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13, 2001.

[12] A. Hyvärinen, P. O. Hoyer, and E. Oja. Image denoising by sparse code shrinkage. In S. Haykin and B. Kosko, editors, *Intelligent Signal Processing.* IEEE Press, 2001.

[13] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[14] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, and C.N. Canagarajah. Region-based image fusion using complex wavelets. In *Proc. 7th International Conference on Information Fusion*, pages 555–562, Stockholm, Sweden, 2004.

[15] H. Li, S. Manjunath, and S. Mitra. Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing*, 57(3):235–245, 1995.

[16] N. Mitianoudis and T. Stathaki. Adaptive image fusion using ICA bases. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

[17] N. Mitianoudis and T. Stathaki. Pixel-based and region-based image fusion schemes using ICA bases. *Elsevier Information Fusion*, 8(2):131–142, 2007.

[18] S.G. Nikolov, D.R. Bull, C.N. Canagarajah, M. Halliwell, and P.N.T. Wells. Image fusion using a 3-d wavelet transform. In *Proc. 7th International Conference on Image Processing And Its Applications*, pages 235–239, 1999.

[19] B.A. Olshausen. *Sparse Codes and Spikes.* In: Probabilistic Models of the Brain: Perception and Neural Function. R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, Eds., MIT Press, 2002.

[20] G. Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 4:259–280, 2003.

[21] G. Piella. New quality measures for image fusion. In *7th International Conference on Information Fusion, Stockholm, Sweden*, 2004.

[22] O. Rockinger and T. Fechner. Pixel-level image fusion: The case of image sequences. *SPIE Proceedings*, 3374:378–388, 1998.

[23] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, Analysis and Machine Vision.* Brooks/Cole Publishing Company, 2nd edition, 1999.

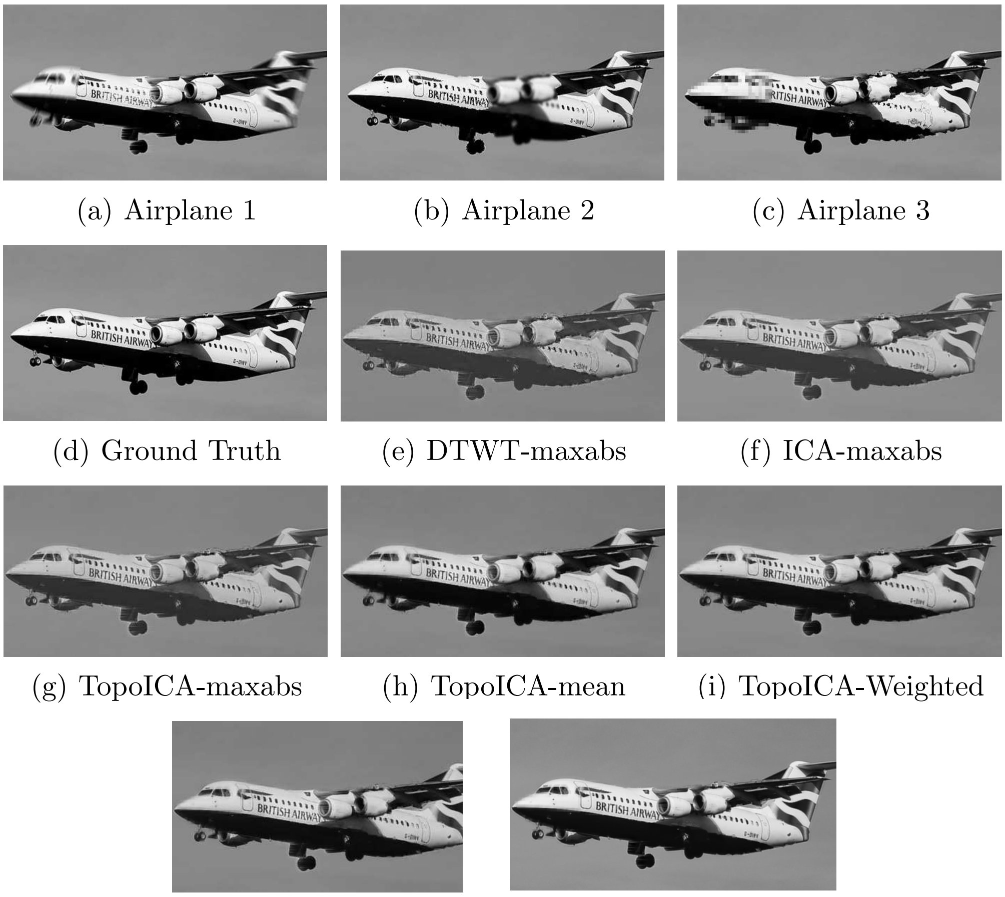
[24] A. Toet. Targets and backgrounds: Characterization and representation viii. *The International Society for Optical Engineering*, pages 118–129, 2002.

[25] Z. Wang and A.C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[26] A.M. Waxman, M. Aguilar, D.A. Fay, D.B. Ireland, J.P. Racamato Jr., W.D. Ross, J.E. Carrick, A.N. Gove, M.C. Seibert, E.D. Savoye, R.K. Reich, B.E. Burke, W.H. McGonagle, and D.M. Craig. Solid-state color night vision: Fusion of low-light visible and thermal infrared imagery. *Lincoln Laboratory Journal*, 11(1):41–60, 1998.

[27] Z. Xue and R.S. Blum. Concealed weapon detection using color image fusion. In *Proc. Int. Conf on Information Fusion*, pages 622– 627, 2003.

[28] C. Xydeas and V. Petrovic. Objective pixel-level image fusion performance measure. In *In Sensor Fusion IV: Architectures, Algorithms and Applications , Proc. SPIE, vol. 4051*, pages 88 – 99, Orlando, Florida,, 2000.



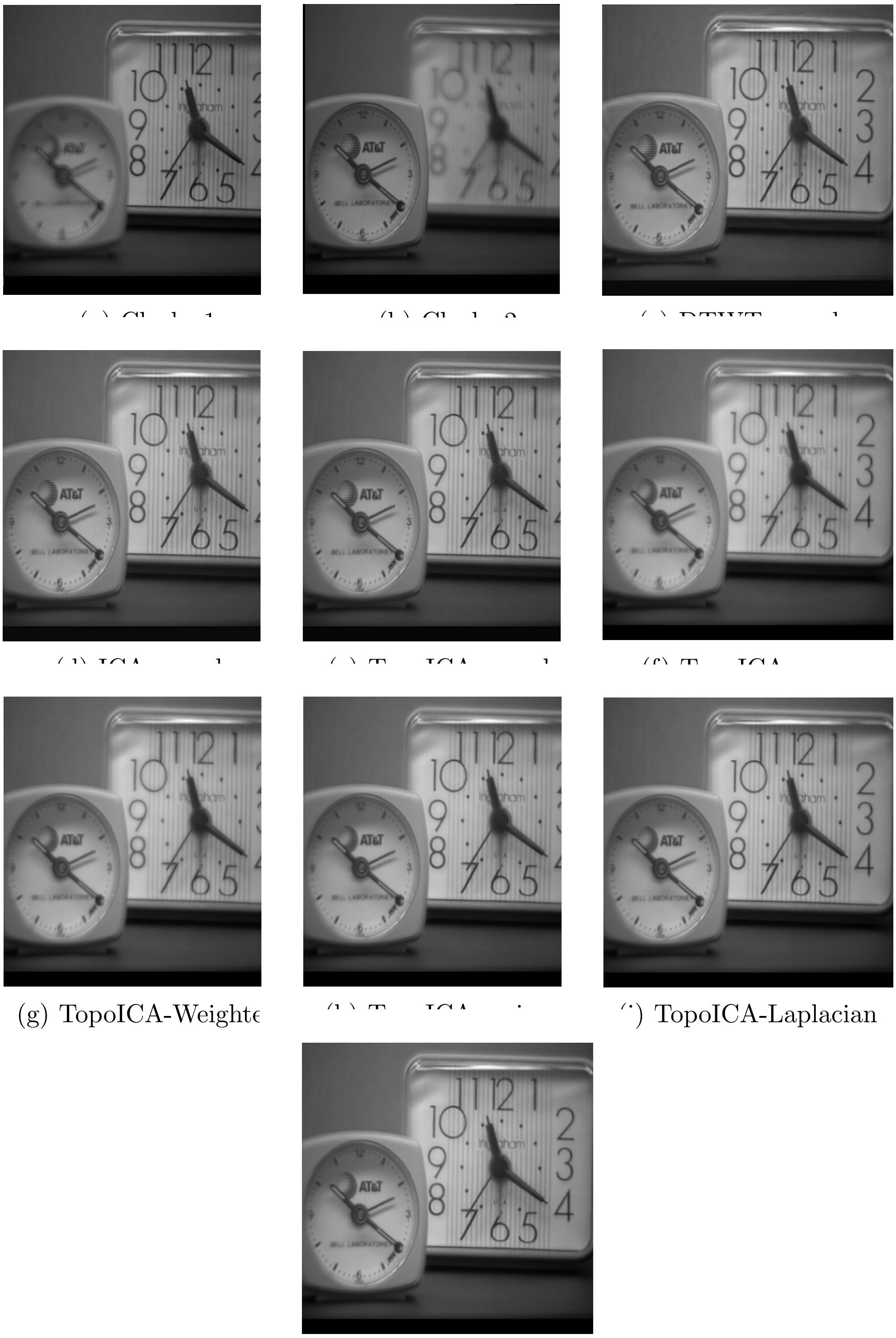
(a) Airplane 1 (b) Airplane 2 (c) Airplane 3

(d) Ground Truth (e) DTWT-maxabs (f) ICA-maxabs

(g) TopoICA-maxabs (h) TopoICA-mean (i) TopoICA-Weighted

(j) TopoICA-regional (k) TopoICA-Laplacian

Fig. 7. Three artificially-distorted input images and various fusion results using various transforms and fusion rules.



(a) Clocks 1

(b) Clocks 2

(c) DTWT [illegible]

(d) ICA [illegible]

(e) TopoICA [illegible]

(f) TopoICA [illegible]

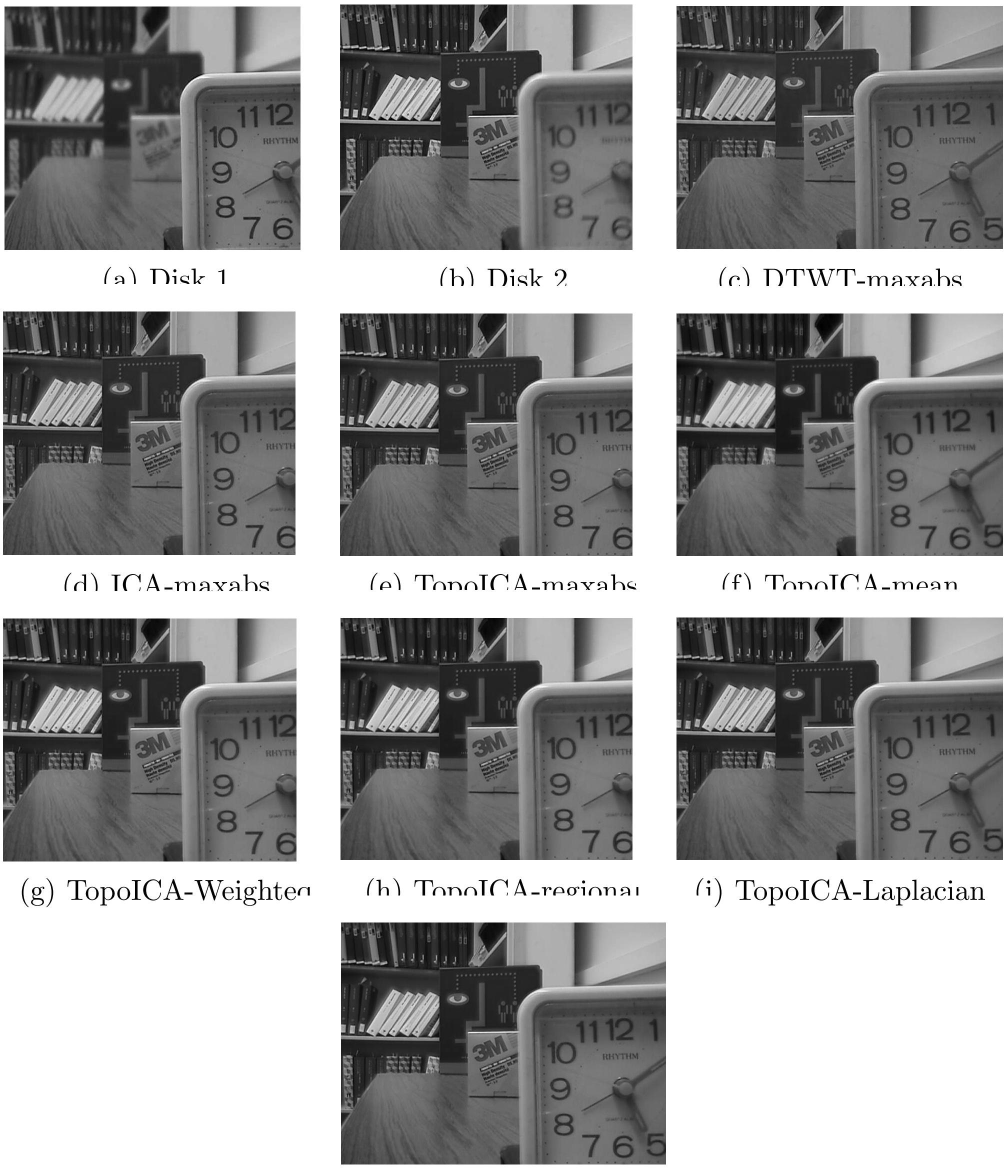
(g) TopoICA-Weighte

(h) TopoICA [illegible]

(i) TopoICA-Laplacian

(j) TopoICA-Verhulstian

Fig. 8. The “Clocks” data-set demonstrating several out-of-focus examples and various fusion results with various transforms and fusion rules.



(a) Disk 1

(b) Disk 2

(c) DTWT-maxabs

(d) ICA-maxabs

(e) TopoICA-maxabs

(f) TopoICA-mean

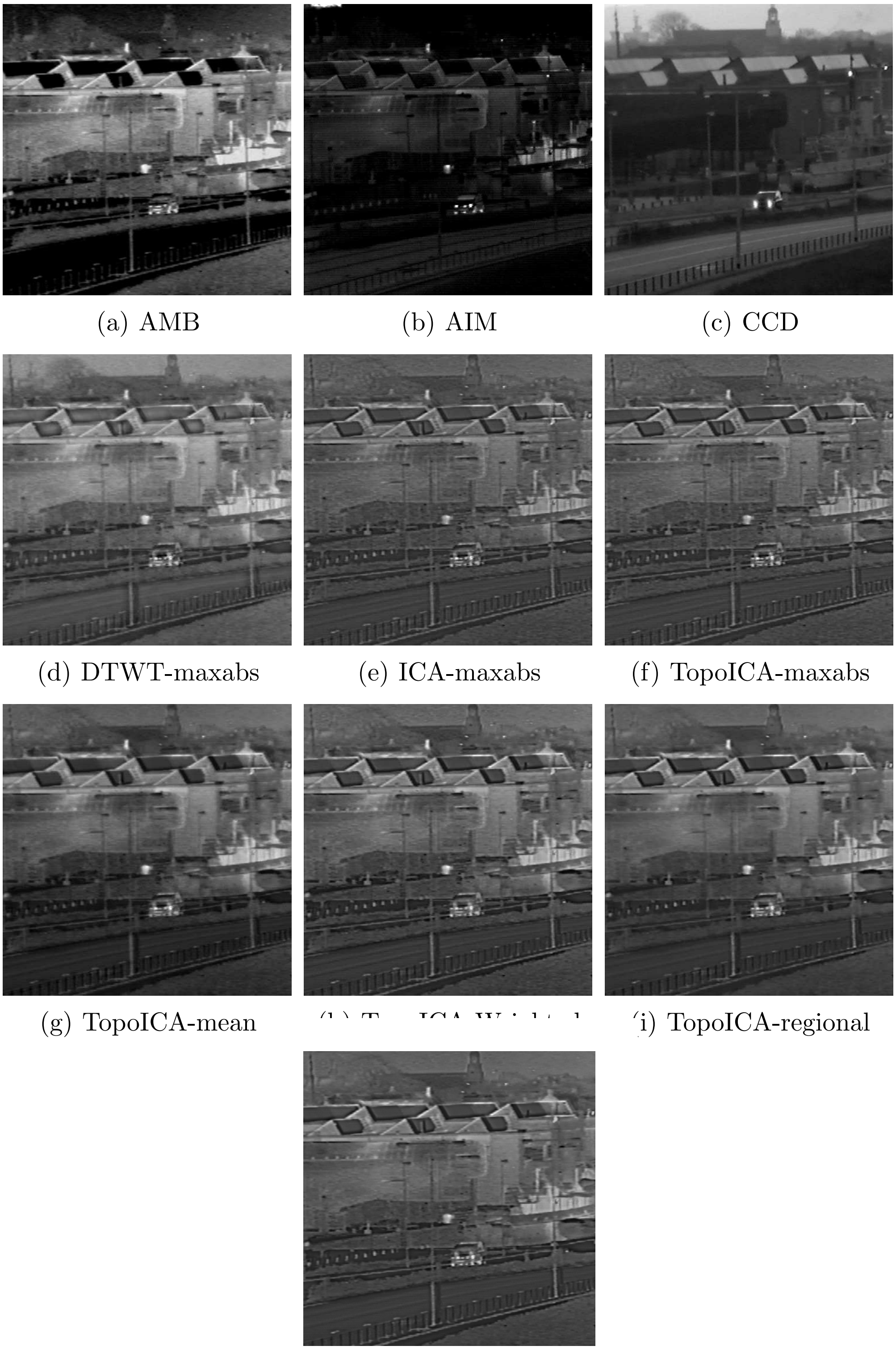
(g) TopoICA-Weighted

(h) TopoICA-regional

(i) TopoICA-Laplacian

(j) TopoICA-Verhulstian

Fig. 9. The “Disk” data-set demonstrating several out-of-focus examples and various fusion results with various transforms and fusion rules.



(a) AMB

(b) AIM

(c) CCD

(d) DTWT-maxabs

(e) ICA-maxabs

(f) TopoICA-maxabs

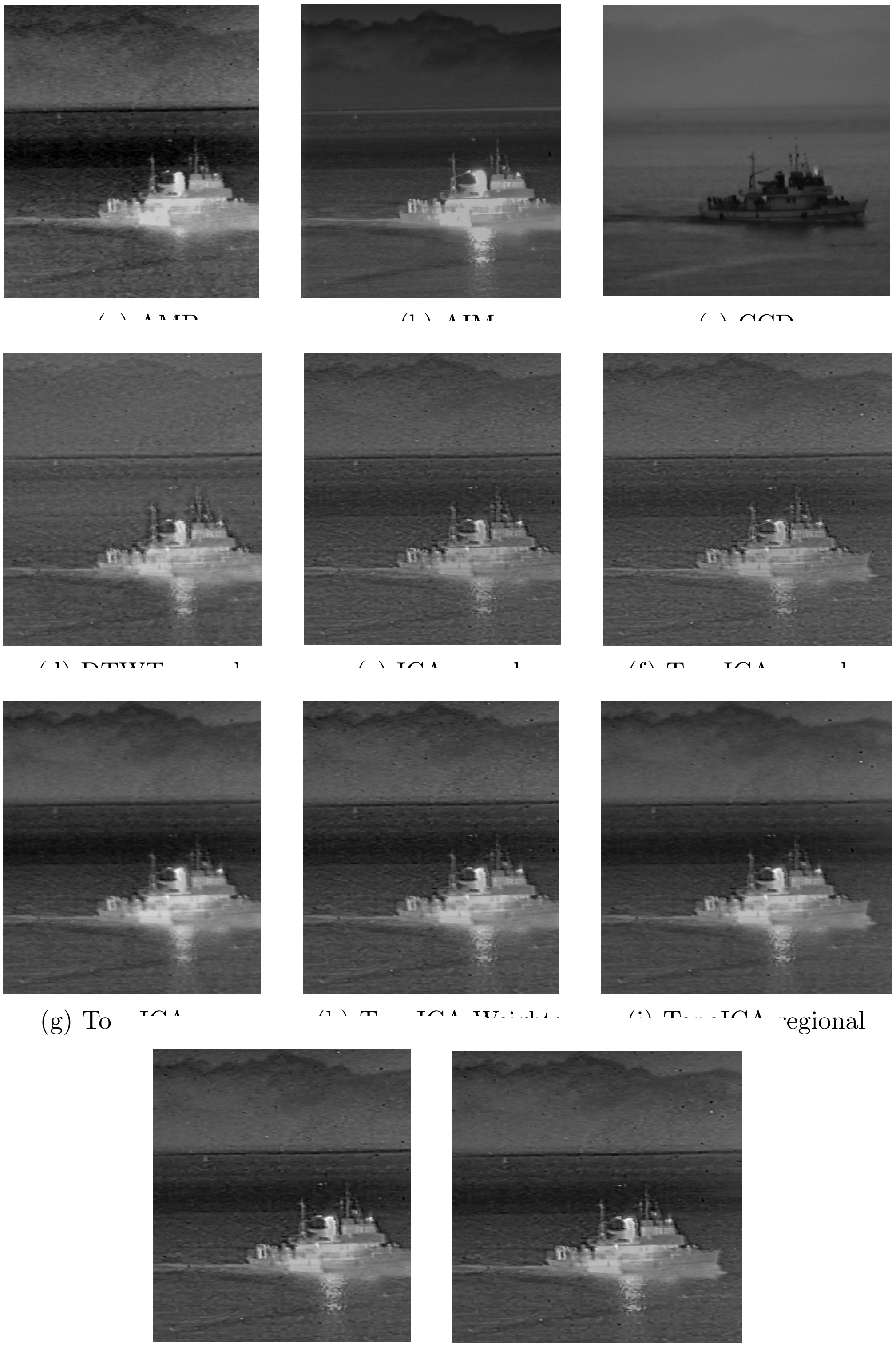
(g) TopoICA-mean

(h) TopoICA-Weighted

(i) TopoICA-regional

(j) TopoICA-Laplacian

Fig. 10. Multi-modal image fusion: Three images acquired through different modality sensors and various fusion results with various transforms and fusion rules.



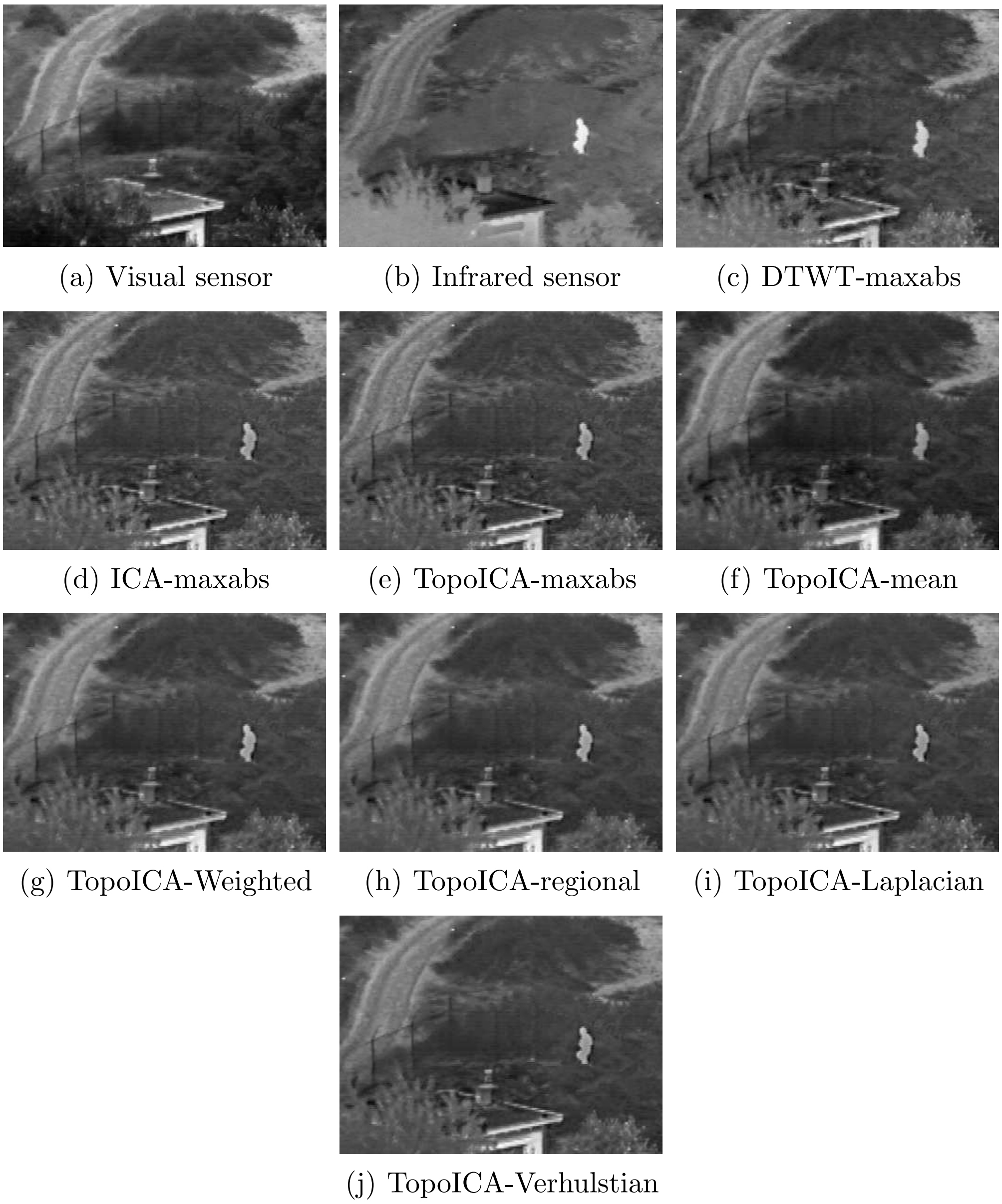
(a) AMR (b) AIM (c) CCD

(d) DTWT-[illegible] (e) ICA-[illegible] (f) TopoICA-[illegible]

(g) TopoICA (h) TopoICA-Weighted (i) TopoICA-regional

(j) TopoICA-Laplacian (k) TopoICA-Verhulstian

Fig. 11. Multi-modal image fusion: Three images acquired through different modality sensors and various fusion results with various transforms and fusion rules.



(a) Visual sensor (b) Infrared sensor (c) DTWT-maxabs

(d) ICA-maxabs (e) TopoICA-maxabs (f) TopoICA-mean

(g) TopoICA-Weighted (h) TopoICA-regional (i) TopoICA-Laplacian

(j) TopoICA-Verhulstian

Fig. 12. The “UN camp” dataset containing visual and infrared surveillance images fused with various transforms and fusion rules.