Optimal Contrast Correction for ICA-based Fusion of Multimodal Images

Nikolaos Mitianoudis, *Member, IEEE,* and Tania Stathaki

Abstract

In this paper, the authors revisit the previously proposed Image Fusion framework, based on self-trained Independent Component Analysis (ICA) bases. In the original framework, equal importance was given to all input images in the reconstruction of the "fused" image's intensity. Even though this assumption is valid for all applications involving sensors of the same modality, it might not be optimal in the case of multiple modality inputs of different intensity range. The authors propose a method for estimating the optimal intensity range (contrast) of the fused image via optimisation of an image fusion index. The proposed approach can be employed in a general fusion scenario including multiple sensors.

Index Terms

Multi-modal Image Fusion, Independent Component Analysis (ICA).

I. Introduction

Modern technology has enabled the development of low-cost, wireless sensors of various modalities that can be deployed to monitor a scene. One can create a wireless network consisting of these spatially distributed autonomous sensors in order to monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants, at different locations. The development of wireless sensor networks is strongly motivated motivated by military and civilian application areas, including battlefield surveillance, environment and habitat monitoring, health-care applications, home automation and traffic control [1].

In this study, the case of multi-modal imaging sensors of known position, that are employed to monitor a scene, will be investigated. The information provided by multimodal sensors can be quite diverse. Each image has been obtained using different instruments or acquisition techniques, allowing each image to have different characteristics, such as degradation, thermal and visual characteristics. Multimodal sensors are increasingly being employed in military applications [2]. Therefore, an operator or a computer vision system needs to examine the information

Manuscript received 7 February, 2008; revised 3 June 2008. This work has been funded by the UK MoD Data and Information Fusion Defence Technology Centre (DIF-DTC) AMDF cluster project.

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, SW7 2AZ London, UK (e-mail: n.mitianoudis@imperial.ac.uk, tel: +44 (0)207 594 6229, fax: +44 (0)207 594 6234).

provided by the individual sensors simultaneously, in order to exploit all the provided information and process the observed scene. Nonetheless, this process of analysing all input images simultaneously is rather impossible for a human operator or computationally very expensive for a computer vision system. If there existed a mechanism to extract all the useful information from the input images to form a new composite one, the analysis system would need to process a single image only. *Image Fusion* can be considered as the process of combining visual information, obtained from various imaging sources, into a single representation, in order to facilitate the information inference.

In this study, the input images are assumed to have negligible registration problems, i.e. correct point-by-point correspondence between the input images [3]. Let $x_1(i,j),\ldots,x_T(i,j)$ represent T input sensor images of size $M_1 \times M_2$ capturing the same scene, where i,j refer to the pixel coordinates in the image. As already mentioned, the process of combining the important features from the original T images to form a single enhanced image $f(i,j)$ is referred to as *Image Fusion*. Fusion techniques can be divided into *spatial domain* and *transform domain* techniques [4]. 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:

$$f(i,j) = g(x_1(i,j),\ldots,x_T(i,j)) \tag{1}$$

The main motivation behind moving to a transform domain is to work in a framework, where the image's salient features are more efficiently identified than in the spatial domain. 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:

$$f(i,j) = \mathcal{T}^{-1}\{g(\mathcal{T}\{x_1(i,j)\},\ldots,\mathcal{T}\{x_T(i,j)\})\} \tag{2}$$

Several transformations have been proposed for image fusion, including the *Dual-Tree Wavelet Transform* [4], [5], [6], *Pyramid Decomposition* [7] and *self-trained Independent Component Analysis bases* [8]. All these transformations project the input images onto localised bases, modelling sharp and abrupt transitions (edges) and therefore, transform the image into a more meaningful representation that can be used to detect and emphasize salient features, which is crucial for performing image fusion. In essence, these transformations can discriminate between salient information (strong edges and other high activity patterns) and constant background or weak edges and also evaluate the quality of the provided salient information. Consequently, one can employ the information provided in the transform domain and select the required information from the input images to construct the "fused" image, following the criteria presented earlier on.

In an earlier work [8], [9], the authors proposed a self-trained Image Fusion framework based on Independent Component Analysis, where the analysis transformation is estimated from a selection of images of similar content. The analysis framework projects the images into localised patches of small size. The local mean value of the patches is subtracted and stored in order to reconstruct the local means of the fused image. In [8], an average of the stored means was used to reconstruct the fused image. In [10], [11], [12], it was demonstrated that this choice might not be optimal in several multi-modal cases and proposed an exhaustive search solution of the optimum performance

in terms of the Piella and Heijmans Fusion Quality index [13] for the case of two input sensors. In this paper, the authors examine and provide a complete solution to this problem for the general T sensor fusion scenario, based on the Fusion Quality Index of [13].

The combination of the means of the input images defines the contrast of the fused image. Contrast is related to the local differences in an image intensity that make an object (or its representation in an image) distinguishable from other objects and the background. In a multimodal fusion scenario, the images have different intensity range. This intensity range is represented by the local mean information of the extracted patches. Balancing or combining these means to construct the means of the fused image is equivalent to adjusting the difference between the input images that highlight different parts of the image, i.e. the contrast in essence. The objective of this work is to find an optimal contrast setting for the fused image in the ICA-fusion framework.

II. Introduction to Image Fusion using ICA bases

Assume an image $x(i,j)$ of size $M_1 \times M_2$. An "image patch" x_w is defined as an $N \times N$ neighbourhood centered around the pixel (i_0, j_0). Assume that there exists a population of patches x_w, acquired randomly from the image $x(i,j)$. Each image patch $x_w(k,l)$ is arranged into a vector $\underline{x}_w(t)$, using lexicographic ordering (see Figure 1). The vectors $\underline{x}_w(t)$ are normalised to zero mean, producing unbiased vectors. These vectors can be expressed as a linear combination of K basis vectors $\underline{b}_j$ with weights $u_i(t), i = 1, \ldots, K$:

$$\underline{x}_w(t) = \sum_{k=1}^{K} u_k(t)\underline{b}_k = [\underline{b}_1 \ \underline{b}_2 \ldots \underline{b}_K] \begin{bmatrix} u_1(t) \\ u_2(t) \\ \cdots \\ u_K(t) \end{bmatrix} \tag{3}$$

where t represents the t-th image patch selected from the original image. The coefficients $u_i(t)$ can be represented as the projections of the input patches on the trained bases, i.e. $u_i(t) =< \underline{x}_w(t), \underline{b}_i >$, where $< \underline{a}, \underline{b} >$ corresponds to the inner product of vectors $\underline{a}$ and $\underline{b}$. Equation (3) can be expressed, as follows:

$$\underline{x}_w(t) = B\underline{u}(t) \tag{4}$$

$$\underline{u}(t) = B^{-1}\underline{x}_w(t) = A\underline{x}_w(t) \tag{5}$$

where $B = [\underline{b}_1 \ \underline{b}_2 \ldots \underline{b}_K]$ and $\underline{u}(t) = [u_1(t) \ u_2(t) \ldots u_K(t)]^T$. 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. The estimation of these basis vectors is performed using a population of training image patches $\underline{x}_w(t)$ and a criterion (cost function) that selects the basis vectors. Analysis/synthesis bases can be trained using *Independent Component Analysis* [14] (ICA) and Topographic ICA [15], as explained in more detail in [8]. The training procedure needs to be performed only once for similar-content images. In a similar fashion to [16], a rectangular window is assumed for the patch extraction procedure during the training and fusion phases.

A number of $N \times N$ patches (approximately 10000) are randomly selected from similar-content training images. We perform Principal Component Analysis (PCA) [17] on the selected patches in order to select the $K < N^2$ most

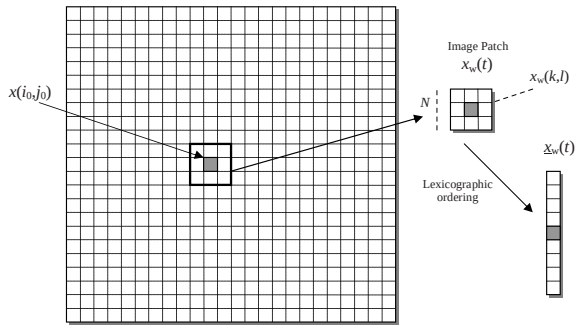


Fig. 1. Segmenting an image for the extraction of local bases.

important bases. Subtracting the mean of the input patches introduces a linear dependence between the estimated bases and therefore the effective number of available bases is N^2-1. The ICA update rule in [14] or the topographical ICA rule in [15] for a chosen $L \times L$ neighbourhood is then iterated until convergence. In each iteration, the bases are orthogonalised using a symmetric decorrelation scheme [14]. In contrast to [10], [11], sample patches from all multimodal inputs are selected to train the ICA bases.

A. Fusion in the ICA domain

After estimating an ICA or Topographic ICA transform $\mathcal{T}\{\cdot\}$, Image fusion using ICA or Topographic ICA bases is performed following the approach depicted in the generic diagram of Figure 2. Every possible $N \times N$ patch is extracted from each image $x_k(i,j)$ and is consequently re-arranged to form a vector $\underline{x}_k(t)$. These vectors $\underline{x}_k(t)$ are normalised to zero mean and the subtracted local mean $MN_k(t)$ is stored for the reconstruction process. Each of the input vectors $\underline{x}_k(t)$ is transformed to the ICA or Topographic ICA domain representation $\underline{u}_k(t)$, using equation (5). Optional denoising in the ICA representation is also possible via sparse code shrinkage of the coefficients in the ICA domain [16]. The corresponding coefficients $\underline{u}_k(t)$ from each image are then combined to construct a composite image representation $\underline{u}_f(t)$ in the ICA domain. The next step is to move back to the spatial domain, using the synthesis kernel B, and synthesise the image $f(i,j)$ by averaging the image patches $\underline{u}_f(t)$ in the same order that were extracted during the analysis step.

In contrast to the proposed framework in [8], Cvejic et al [12], [11], [10] proposed to train different sets of ICA bases for each sensor of different modality and thus, analyse the input sensor images using different ICA bases for each modality. However, the bases sets that will be produced by the different ICA training procedures will have no correspondence to each other. Consequently, it is not viable to combine projections on different bases sets, in order to form the "fused" image, as claimed in [10], [11]. This is similar to attempting to fuse images analysed by wavelet decomposition using a different wavelet family for each input image. In addition, one has to select one of the possible synthesis kernels to transform the fused image to the spatial domain, which will denote preference on a specific sensor modality. The ICA bases training mechanism is an adaptive procedure that can

extract interesting local features from all the presented patches. There is no specific need for the training to be performed independently for each sensor type. Thus, interesting bases from all different modality inputs will be extracted by iterating on the complete training dataset.

B. Various fusion rules using ICA bases

Some basic rules that can be used for image fusion in the ICA-bases domain are described in this section. Fusion by the *absolute maximum* rule has been used widely by the image fusion community. This rule selects the greatest in absolute value of the corresponding ICA-domain coefficients in each image ("max-abs" rule). This process seems to convey all the information about the edges to the fused image, however, the intensity information in constant background areas seems to be slightly distorted. In contrast, fusion by the *averaging* rule averages the corresponding coefficients ("mean" rule). This process seems to preserve the correct contrast information, however, the edge details seem to get smoother. A *Weighted Combination* (WC) pixel-based rule can be established using the ICA framework [8]. The fused image coefficients are constructed using a "weighted combination" of the input transform coefficients, i.e.

$$\underline{u}_f(t) = \sum_{k=1}^{T} w_k(t)\underline{u}_k(t) \tag{6}$$

To estimate the contributions $w_k(t)$ of each image to the "fused" image, the mean absolute value ($\mathcal{L}_1$-norm) of each patch (arranged in a vector) in the transform domain can be employed as an activity indicator. The $\mathcal{L}_1$-norm is preferred because it fits a more general sparse profile of the ICA coefficients, denoted by a Laplacian distribution.

$$E_k(t) = ||\underline{u}_k(t)||_1 \qquad k = 1, \ldots, T \tag{7}$$

The weights $w_k(t)$ should emphasise sources with 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.

$$w_k(t) = E_k(t) / \sum_{k=1}^{T} E_k(t) \tag{8}$$

In some patches, where $\sum_{k=1}^{T} E_k(t)$ might be very small, one can use the "max-abs" or "mean" fusion rule to avoid numerical instability. A *regional* approach can also be established, by dividing the observed area into areas of "low" and "high" activity. The areas of "high" activity contain salient information and can be fused using a "max-abs" or a "weighted-combination" fusion rule and the areas of "low-activity" contain background information and can be fused using the "mean" rule. A heuristic approach to differentiate between a "low" and a "high" activity region is to use the $\mathcal{L}_1$-norm based $E_k(t)$ measurement. Another regional approach can be to use alternative segmentations of the observed scene, based on the input sensor images and consequently fuse the different regions independently [10], [11]. In this work, the weighted combination rule will be used for ICA-based fusion for simplicity.

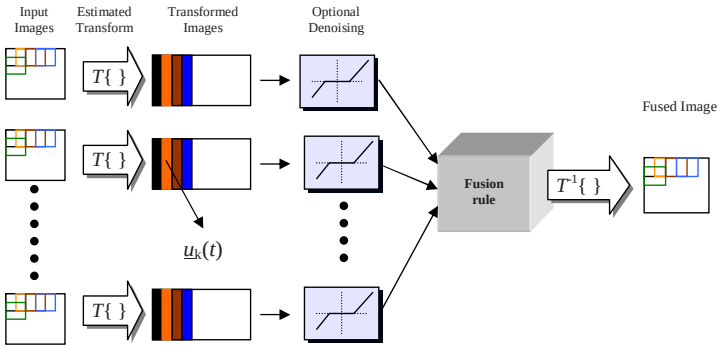
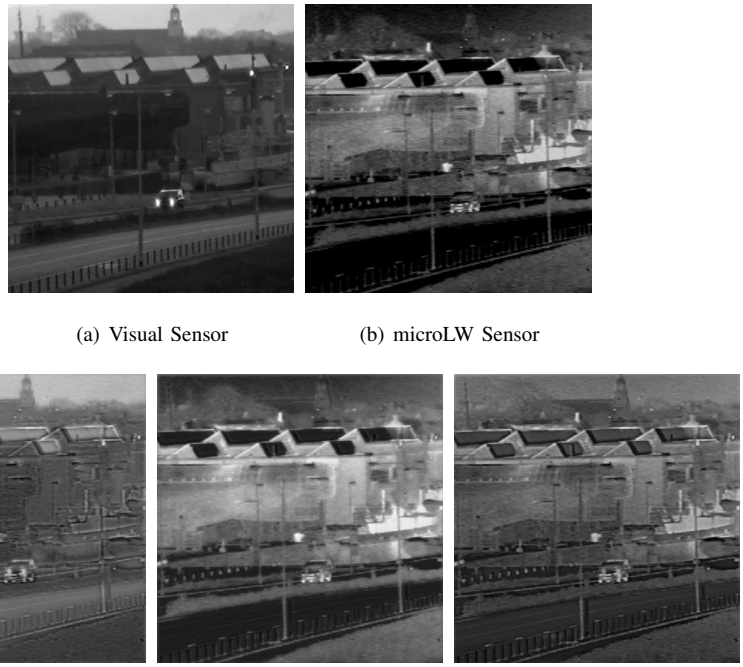


Fig. 2. The proposed fusion system using ICA / Topographical ICA bases.

III. Local Means reconstruction considerations on the original framework

The next step is to estimate the spatial-domain representation $f(i,j)$ of the fused image. To reconstruct the image in the spatial domain, the process described in Section II is inverted. The vectors $\underline{u}_f(t)$ are re-transformed to the local $N \times N$ patches $u_f(k,l)$. The local mean of each patch is restored using the stored patches means $MN_k(t)$. There exist T local intensity values for each patch of the reconstructed image, each belonging to the corresponding input sensor. In the case of 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. Therefore, the local means are reconstructed by averaging $MN_k(t)$, 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 3, the example of a multi-modal fusion scenario is displayed: a visual sensor image is fused with a micro Long-Wave (microLW) 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 microLW 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 inappropriate 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.



(a) Visual Sensor (b) microLW Sensor

(c) Means from Visual Sensor (d) Means from microLW Sensor (e) Average of Means

Fig. 3. Effect of local means choice in the reconstruction of the fused image.

IV. A NOVEL APPROACH FOR AUTOMATED CONTRAST CORRECTION

In this section, an automated mechanism to calculate the optimal means for the local patches of the fused image in the ICA bases framework for N ($N > 2$) input sensors will be described. In the initial design of the algorithm [8], the local means of the fused image were created by averaging the corresponding local means of the input sensor images. In [10], [11], the authors demonstrated that this might not be an optimal solution to the problem, where optimality is defined in terms of the Piella and Heijmans index [13][1]. They proposed a method to estimate a weighted averaging of the corresponding local means in a two-sensor fusion scenario. In the two-sensor case, the proposed 2D optimisation problem can be reduced to an 1D problem, since the second weight q_2 can be expressed as a function of the first weight q_1 via $q_2 = 1 - q_1$. The 1D optimisation was solved by numerically assessing the Piella index for quantised values of $q_1 \in [0, 1]$ with a step value of 0.1 to infer approximately the value of q_1 that maximises the Piella index. This approach relies on a valid concept to solve the above mentioned problem, however, the proposed implementation in [10], [11] exhibits several drawbacks. More specifically, this approach can not be easily expanded in a multiple sensor scenario, since the exhaustive search for the optimum in the (N-1)-D half unit cube will be rather computationally expensive. In addition, the approach in [10], [11] assumes the existence of a single global optimum of the cost function based on the Piella Index, an assumption for which no theoretical justification was provided.

[1]For simplicity we will refer to this index as the Piella Index for the rest of the document

A. Optimising Piella's Index

Assume that $m_{x_1}, m_{x_2}, \ldots, m_{x_T}$ are the means of the input sensors images, m_f is the mean of the fused image and $q_1, q_2, \ldots, q_T$ are the weights that will be used to estimate m_f via the equation:

$$m_f = q_1 m_{x_1} + q_2 m_{x_2} + \cdots + q_T m_{x_T} \tag{9}$$

The Wang and Bovik Image Quality Index [18] of an estimated image f in relation to the original image x is given by the following formula:

$$Q(x,f) = \frac{2\sigma_{xf}}{\sigma_x^2 + \sigma_f^2} \frac{2m_x m_f}{m_x^2 + m_f^2} \tag{10}$$

where σ_{xf} represents the correlation between the two images and σ_f, σ_x represent the standard deviations of the two images respectively. Let $Q_\sigma(x,f) = \frac{2\sigma_{xf}}{\sigma_x^2+\sigma_f^2}$ and $Q_m(x,f) = \frac{2m_x m_f}{m_x^2+m_f^2}$. It is straightforward to see that the Image Quality Index can be factorised into the term Q_σ that is dependent on correlation/variance figures and the term Q_m that is dependent on mean values. More specifically,

$$Q(x,f) = Q_\sigma(x,f) Q_m(x,f) = Q_\sigma(x,f) \frac{2m_x m_f}{m_x^2 + m_f^2} \tag{11}$$

The first version of the Piella Index is based on the Wang and Bovik index. The Piella Index simply segments the input sensor images and the fused image into multiple overlapping patches and estimates the contribution of each input patch to the fused image in terms of the Image Quality Index. These scores are weighted according to the local information quality λ_i of each patch, (e.g. local variance), in order to emphasise the information-transfer scores of patches with strong local activity. Note that λ_i is normalised to the total local information quality (saliency) of the corresponding input patches. This implies that $\sum_{i=1}^{T} \lambda_i = 1$. Therefore for a single patch, the Piella Index is defined as:

$$\begin{aligned} Q_p^n &= \lambda_1 Q(x_1,f) + \lambda_2 Q(x_2,f) + \cdots + \lambda_T Q(x_T,f) \\ &= \sum_{i=1}^{T} \lambda_i Q(x_i,f) \end{aligned} \tag{12}$$

The next step is to estimate the expected value of the Piella index for all the extracted image patches, i.e.,

$$Q_p = \mathcal{E}\{\sum_{i=1}^{T} \lambda_i Q(x_i,f)\} \tag{13}$$

The expectation in the above equation will be approximated by an average of all patches, producing the first version of the Piella Index. This is similar to attributing equal probability for each patch, i.e. assuming a uniform prior and consequently the expectation is approximated by the sample mean. There is also a second version of Piella's index, where this expectation is approximated by a weighted sum of the individual terms. The imposed weights are based on the importance of each frame, in terms of maximum input sensor saliency, compared to the total maximum saliency of the samples. A third version was also proposed in [13], by estimating the second version of the index for the edge maps of the input sensors and fused images and multiplying it with the original second index version [13]. In this optimisation the first version is considered for simplicity.

The next step is to optimise Q_p in terms of $\underline{q}$ in order to estimate the mean m_f and essentially the weights $\underline{q}$. This adaptation will not necessarily affect the energy, i.e. the local activity of the patch in the fused image and its comparison to the activity input sensor patches. The bias (i.e. the local mean) is the only factor that is affected by the adaptation of m_f. Let $A_{\sigma_{x_i}} = \lambda_i Q_\sigma(x_i, f)$, then (13) can be expressed as:

$$Q_p(\underline{q}) = \mathcal{E}\left\{\sum_{i=1}^{T} A_{\sigma_{x_i}} \frac{2m_{x_i}m_f}{m_{x_i}^2 + m_f^2}\right\} \tag{14}$$

The objective is to estimate $\underline{q} = [q_1\ q_2\ \ldots\ q_T]^T$ by maximising Q_p. The derivative of Q_p in terms of $\underline{q}$ is given by

$$\frac{\partial Q_p}{\partial \underline{q}} = \begin{bmatrix} \frac{\partial Q_p}{\partial q_1} \\ \frac{\partial Q_p}{\partial q_2} \\ \ldots \\ \frac{\partial Q_p}{\partial q_T} \end{bmatrix} \tag{15}$$

The term $\partial Q_p/\partial q_i$ can be expressed, as follows:

$$\frac{\partial Q_p}{\partial q_i} = \frac{\partial Q_p}{\partial m_f}\frac{\partial m_f}{\partial q_i} = m_{x_i}\frac{\partial Q_p}{\partial m_f} \tag{16}$$

Consequently,

$$\frac{\partial Q_p}{\partial \underline{q}} = \frac{\partial Q_p}{\partial m_f}\begin{bmatrix} m_{x_1} \\ m_{x_2} \\ \ldots \\ m_{x_T} \end{bmatrix} = \frac{\partial Q_p}{\partial m_f}\underline{m}_x \tag{17}$$

$$\frac{\partial Q_p}{\partial m_f} = \mathcal{E}\left\{\sum_{i=1}^{T} A_{\sigma_{x_i}} m_{x_i} \frac{m_{x_i}^2 - m_f^2}{(m_{x_i}^2 + m_f^2)^2}\right\} \tag{18}$$

Performing gradient ascent on the proposed cost function yields the following update rule for the weight vector $\underline{q}$:

$$\underline{q}^+ \leftarrow \underline{q} + \eta\mathcal{E}\left\{\underline{m}_x \sum_{i=1}^{T} A_{\sigma_{x_i}} m_{x_i} \frac{m_{x_i}^2 - m_f^2}{(m_{x_i}^2 + m_f^2)^2}\right\} \tag{19}$$

where η denotes the learning rate. The above rule is iterated until convergence and consequently the estimated weights are employed to reconstruct the means of the fused image. To avoid extreme situations or deformations in the weights q_i during the adaptation, the following restriction is imposed during each step:

$$\underline{q}^+ \leftarrow |\underline{q}|/([\ 1\quad 1\quad \ldots\quad 1\ ]|\underline{q}|) \tag{20}$$

This restriction ensures that the weights remain positive during the adaptation and their summation is restricted to unity.

B. Uniqueness of solution ?

In [10], [11], the authors state that the cost function that was optimised numerically tends to have a single optimum, which is always a maximum. In this section, the validity of the assumption is investigated mathematically. Looking at (14), we can rewrite the cost function for a single patch, as follows:

$$g(m_f) = \sum_{i=1}^{T} \lambda_i \frac{m_{x_i} m_f}{m_f^2 + m_{x_i}^2} \qquad \forall m_f \in [m_{x_{min}}, m_{x_{max}}] \tag{21}$$

where $m_{x_{min}} = \min m_{x_i}$ and $m_{x_{max}} = \max m_{x_i}$. Following the investigations in Appendix A, it can be shown that the above cost function is not guaranteed to have a single maximum in the solution space. It is also shown that a sufficient condition for the above cost function to have a single maximum is $m_{x_{max}} \leq \sqrt{3} m_{x_{min}}$. In the case that the condition does not hold, it is dubious whether the cost function will meet the requirements of a single maximum. The physical interpretation of the above condition is that the corresponding input sensor patches should have similar means (i.e. intensity values) or at least within a margin of similarity. Of course, this is not always true especially in the case of several objects and their multimodal representation, e.g. a human will appear much more whiter in terms of intensity in a thermal representation than in a visual representation. Therefore, in theory, the above cost function is not guaranteed to have a single maximum in the solution space.

On the other hand, in the optimisation process, all the image patches are taken into account and are averaged to infer an estimate for the mean weights. Therefore, the final cost function is computed by averaging (21) for all input patches. If the required condition holds for the majority of input patches then the final cost function will feature a single maximum that needs to be estimated. In our experimentation with the complete "Dune", "Trees" and "Uncamp" datasets (see section V), no single case of multiple optima was encountered. In Figure 4, a typical example is depicted to demonstrate that the majority of pixels follows the sufficient condition that was derived previously and thus the final cost function features a single maximum. Nevertheless, no full assurance can be provided about the existence of a single optimum.

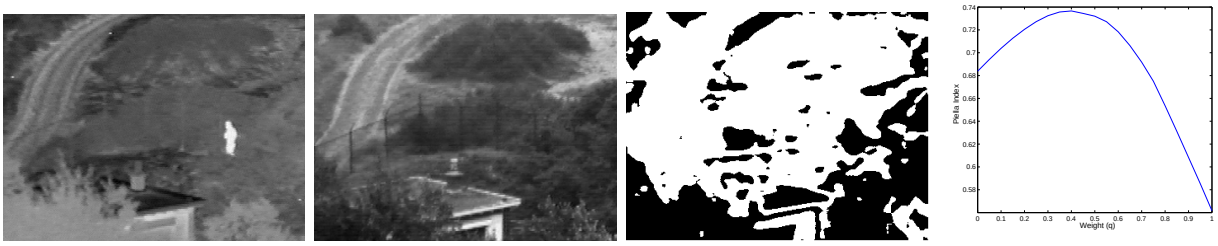
One can always devise a tactic to ensure the uniqueness of solution. The patches for which the condition does not hold are isolated at first. Essentially, one can add a constant c to these m_{x_i} and satisfy the required condition. The required constant for a given set of $m_{x_{max}}, m_{x_{min}}$ is calculated as follows:

$$\frac{m_{x_{max}} + c}{m_{x_{min}} + c} \leq \sqrt{3} \Rightarrow c \geq \frac{m_{x_{max}} - \sqrt{3} m_{x_{min}}}{\sqrt{3} - 1} \tag{22}$$

Then, we can calculate all the required constants c in order to ensure that all the patches satisfy the condition. The maximum of these constants c_{max} can be added to all m_{x_i} of all image patches ensuring a single optimum for all patches. Once the optimum means of the fused image are calculated, the added constant c_{max} can be subtracted to return the fused image to the required intensity range.

V. Experiments

To test the performance of the proposed scheme, a variety of multi-modal scenarios that exist in the literature are employed. The first step is to use the "Dune", "Trees" and "UNcamp" datasets of surveillance images from TNO



(a) InfraRed Image (b) Visual Image (c) Condition Check (d) Final Cost function

Fig. 4. Checking the single maximum condition in the case of the "Dune" dataset. The majority of patches (corresponding to every image pixel) satisfy the proposed condition and thus will contribute towards a single optimum for the total cost function.

Human Factors, provided by L. Toet [19] in the Image Fusion Server [20]. The datasets consist of two series of visual and infrared frames capturing a human subject walking through various areas. The above multimodal datasets are used to evaluate the performance of the adaptive scheme on finding the optimal means in terms of the Piella Index. We used the typical training procedure for the ICA framework, training 60 8×8 ICA bases for each dataset. The fusion method that is employed in the following experiments is the "weighted-combination" fusion rule. For performance comparison, the Dual-Tree Wavelet Transform (DT-WT) method using the "max-abs" rule will also be employed[2]. The Piella Index that is calculated in this section will constantly represent the second version of the Piella Index, as explained earlier.

In Figure 5, several results of optimal contrast correction of a sample image (frame 1812) from the UNCamp dataset are depicted. Figures 5 (a), (b) depict the input images: one infrared sensor capture and one visual sensor capture. The Optimal Contrast algorithm, described in the previous sections, is initialised with $\underline{q} = [0.5\ 0.5]^T$ and the learning rate was set to $\eta = 2.5$. The learning rule of (19) is iterated until convergence producing the following result $\underline{q}_{opt} = [0.4156\ 0.5844]^T$. This implies that the algorithm has identified the optimal contrast in terms of the Piella index with relatively high accuracy. In Figure 5 (c), the original ICA framework output, assuming equal weights for the means, is depicted. In Figure 5 (d), the produced fused image using optimal means selection is depicted. Obviously, there is no much difference between the two results, since the estimated optimal point is very close to the assumption of equal weights. In Figure 5 (f), the produced fused image using the DT-WT transform and the "max-abs" rule is shown.

In Figure 6, a similar example using a sample image (frame 4904) from the Trees dataset is depicted. In this case, the algorithm converged at the weights $\underline{q}_{opt} = [0.3832\ 0.6168]^T$ and the resulting fused image featured a Piella Index of 0.7968 compared to that of 0.7920, achieved by the traditional ICA framework. Hence, the proposed optimisation offered improvement compared to the original scheme and still performed better than the DT-WT scheme with a Piella Index of $Q = 0.7758$. In Figure 7, the convergence plot of the proposed algorithm for this

[2]Code for the Dual-Tree Wavelet Transform available online by the Polytechnic University of Brooklyn, NY at http://taco.poly.edu/WaveletSoftware/

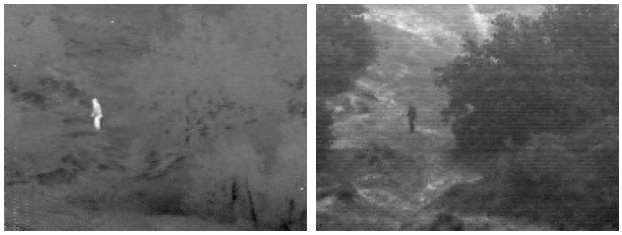
(a) InfraRed Image (b) Visual Image

(c) Equal Weights (d) Optimised Weights (e) DT-WT

Fig. 5. Optimal Contrast correction for the "UN camp" dataset. The fused image that features optimal mean selection is not essentially different to the original fused image, since the optimal weight selection, according to Piella's index, is close to assuming equal weights.

example is shown. The convergence plot was similar for all tested datasets and image sets in this experimental section. Consequently, the gradient-based rule features fast convergence to the optimal value with high accuracy, thus, avoiding the exhaustive search solution proposed in [10], [11].

Next, the aim is to visualise the shape of the cost function that was optimised for all images in the three datasets. The 2D problem can be reduced to an 1D problem, since the two weights can be represented by q and $1-q$. The above simplification is not necessary for our approach, as it is evident from the previous analysis, however, it will be employed in this example for visualisation purposes only. For all images of the three datasets, the Piella Index was evaluated for all values of q and the optimal Piella Index was estimated using both the proposed algorithm and numerical evaluation. In Figure 8, the proposed cost function is evaluated in terms of q for all frames and values of q for the three datasets. It is clear that the functions are smooth and feature a single optimum in all examined cases, which supports our statements and observations in the previous section. Consequently, no extra steps are needed for tackling multiple optima situations. In Figure 9, the achieved Piella Index using the proposed algorithm is compared to the optimal Piella Index that was numerically estimated from the cost function for all the three datasets. In the same figure, we plot the Piella Index achieved by the original ICA-based framework. It is clear that the proposed algorithm managed to identify the real optimum in performance in the majority of the cases with a small error margin. In most cases, the proposed algorithm outperformed the previous ICA-based framework. In Table I, the average Piella Indexes for the datasets "Dune", "Trees", "Uncamp" are depicted. The proposed approach outperformed the original ICA-based framework and the DT-WT-based framework. The second experiment was to employ two images that can demonstrate the usefulness of Image Fusion and the rectification in contrast to the



(a) InfraRed Image (b) Visual Image

(c) Equal Weights ($Q = 0.7920$) (d) Optimised Weights ($Q = 0.7968$)

(e) DT-WT ($Q = 0.7758$)

Fig. 6. Optimal Contrast correction for the "Trees" dataset. The optimised means featured enhanced performance compared to the traditional ICA framework and the DT-WT-based scheme and compared to the Dual-Tree Wavelet Transform (DTWT) scheme.

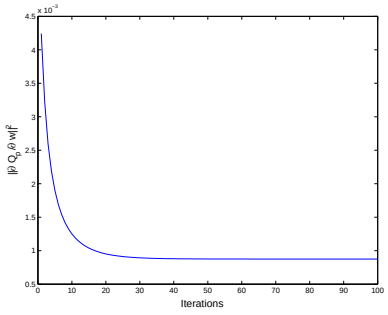
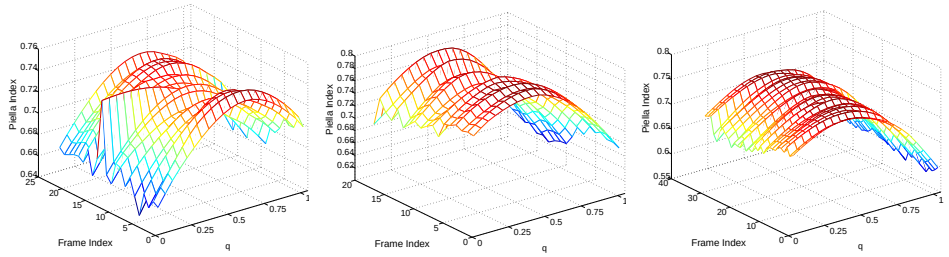


Fig. 7. Convergence of the adaptive scheme for contrast correction for multi-modal image fusion.



(a) Dune (b) Trees (c) Uncamp

Fig. 8. The estimated cost functions using the Piella Index for the datasets "Dune", "Trees" and "Uncamp". This verifies the statement that in practice most cost functions feature a single maximum.

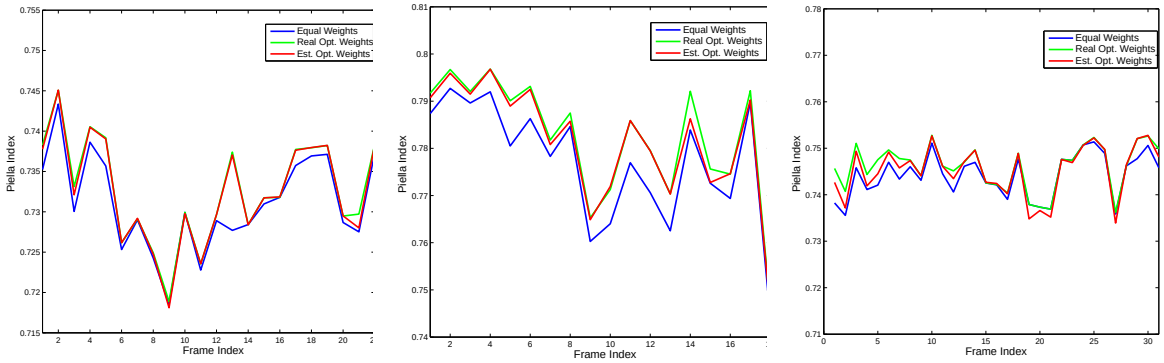
TABLE I

AVERAGE FUSION PERFORMANCE MEASUREMENTS USING PIELLA'S INDEX FOR THE FIVE DATASETS OF THIS EXPERIMENTAL SECTION. THE PROPOSED ICA-BASED SCHEMES ARE COMPARED WITH THE ORIGINAL ICA-BASED FRAMEWORK AND THE DUAL-TREE WAVELET FRAMEWORK.

Method	ICA Equal Weights	ICA Opt. Weights	DT-WT
'Dune''	0.7311	0.7325	0.7156
''Trees''	0.7770	0.7814	0.7595
''Uncamp''	0.7441	0.7452	0.7317
Octet 1	0.8251	0.8354	0.8254
Octet 2	0.8176	0.8677	0.8602
Car Image	0.6822	0.6857	0.6392

previous ICA-based framework offered by the proposed approach. The two Octet image sets were employed, as provided by the ImageFusion Server [20]. These images, captured by Octec Ltd., show men and buildings with (Test Images 2, see Figure 11) and without (Test Images 1, see Figure 10) a smoke screen. They were captured with a Sony Camcorder and a Long Wave Infrared (LWIR) sensor. We employed the original ICA-based scheme, the proposed ICA-based scheme and the DT-WT framework to perform fusion of the two images. In Figures 10, 11, the fusion results of the three algorithms were depicted. The second example clearly demonstrates the problem of the original ICA-based framework. The equal weights on the input images' intensities will result in the foreground smoke of the visual image being transferred to fused image. This rather decreases the perception quality of the fused image. The optimal contrast approach weights the significance of the local intensities of the two images, resulting into a more balanced representation in the fused image, where the main desired targets are less hindered by the smoke in the visual input, although still visible due to its contrast and salient features. The Piella Indexes for the three methods and the two tests are shown in Table I.

A third example is used to demonstrate the efficiency of our algorithm in the case of more than two sensors. We used some surveillance images from TNO Human Factors, provided by L. Toet [19], obtained from the Image



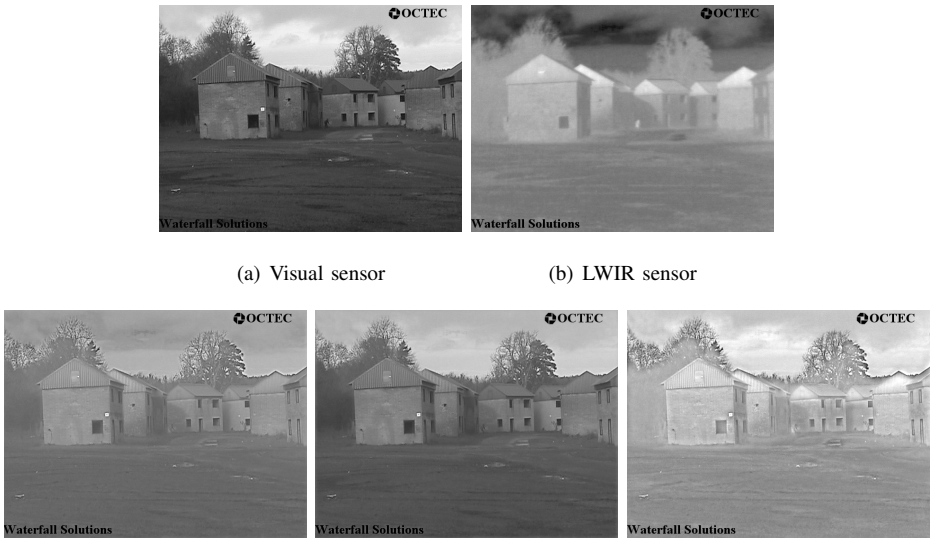
(a) Dune (b) Trees (c) Uncamp

Fig. 9. Comparison between the Piella Index achieved by the proposed algorithm and the estimated maximum Piella Index by exhaustive search for the datasets "Dune", "Trees" and "Uncamp".

Fusion Server [20]. 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. The three input sensor images are depicted in Figures 12 (a), (b), (c). To examine the nature of our cost function in this case, we evaluate the Piella Index, in terms of the normalised weights q_1, q_2 and $1 - q_1 - q_2$. The cost function, which is now a surface, is depicted in Figure 13. The function is again considerably smooth. There is a single maximum in the surface, nevertheless, it is not very well pronounced, giving a weak optimum. The proposed algorithm was initialised as previously and converged smoothly to the value of $q_{opt} = [0.6111\ 0.1289\ 0.2601]^T$. The Piella index for the estimated weights is $Q_p = 0.6857$. In Figure 12 (d), (e), we plot the fused image assuming equal weights and optimised weights respectively. The Piella index for the equal weights image is $Q_p = 0.6822$, which implies that the proposed approach has achieved improved performance compared to the original scheme and the DT-WT scheme. The convergence in the three-dimensional case was similar to the one depicted in Figure 7, which implies that the fusion using ICA bases of more than two input sensors is possible and efficient.

VI. Conclusion

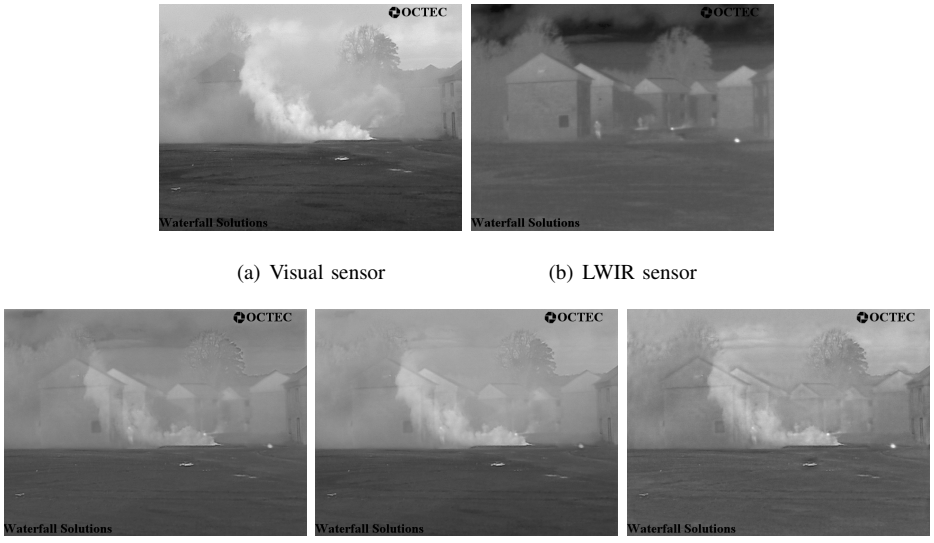
In this paper, the authors proposed an improvement to their previous ICA-based Image Fusion framework. In the original framework, the input sensor images are projected on localised patches of small size. The local mean value of the patches is subtracted and stored in order to reconstruct the local means of the fused image. Originally, an average of the stored means was used to reconstruct the fused image, nonetheless, it was demonstrated that this choice might not be optimal in several multi-modal cases in [10], [11]. In the same work an exhaustive search solution of the optimum performance in terms of the Piella index [13] was proposed for the case of two input sensors only. In this paper, the authors provide a generalised iterative solution of this problem using a detailed optimisation of the Piella index in the general case of T input sensors. The existence of a single solution to this optimisation



(a) Visual sensor (b) LWIR sensor

(c) Equal Weights (d) Optimised Weights (e) DT-WT

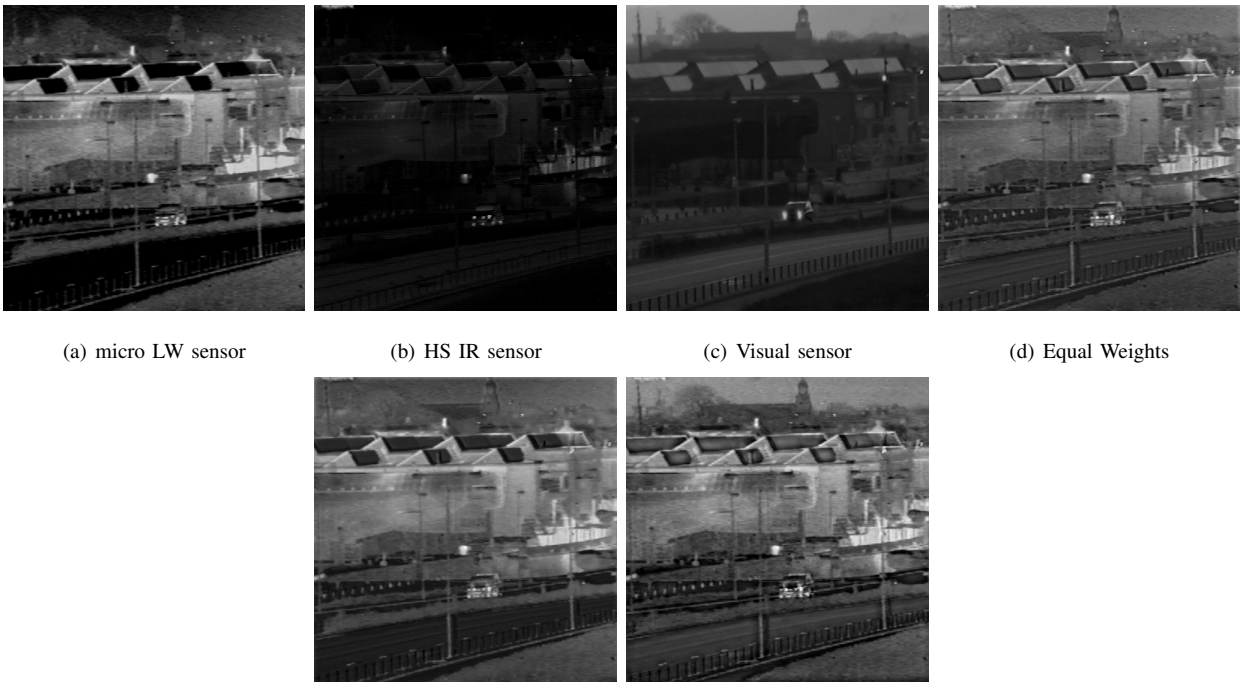
Fig. 10. The Octet 1 example of image fusion of a visual and a Long-Wave InfraRed sensor.



(a) Visual sensor (b) LWIR sensor

(c) Equal Weights (d) Optimised Weights (e) DT-WT

Fig. 11. The Octet 2 example of image fusion of a visual and a Long-Wave InfraRed sensor.



(a) micro LW sensor (b) HS IR sensor (c) Visual sensor (d) Equal Weights

(e) Optimised Weights (f) DT-WT

Fig. 12. Optimal Contrast correction for the TNO Human Factors dataset. The optimal contrast selection scheme chooses to emphasize the contrast of the two Infrared images to the low-contrast night visual image.

problem is investigated and conditions that can guarantee the existence of single maxima are presented. The proposed gradient-descent optimisation can identify the optimal value of contrast in maximum 70 iterations, providing an efficient and general solution for the case of T input sensors, removing the need of analytic evaluation of the whole solution space in [10], [11]. The proposed approach enhances the performance of the original ICA-based fusion framework, improving the perception of the produced "fused" image.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Gemma Piella for supplying the code for the Piella fusion index that was employed to verify the validity of our implementation.

APPENDIX

A SUFFICIENT CONDITION FOR THE EXISTENCE OF A SINGLE SOLUTION

The mean term Q_m in the Wang and Bovik metric can be represented by the following function:

$$f(x) = \frac{ax}{x^2 + a^2} \qquad x > 0 \tag{23}$$

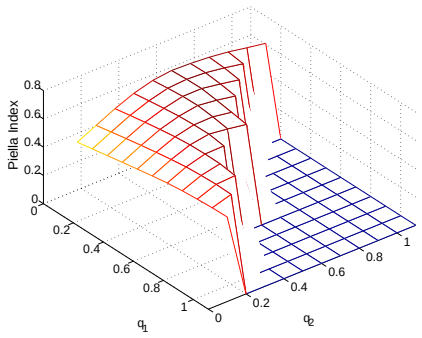


Fig. 13. Values of the Piella Index spanning the solution space of q_1, q_2 and $1 - q_1 - q_2$.

where $a > 0$. We estimate the first and the second derivative of $f(x)$, as they will be employed in the following analysis:

$$\frac{df(x)}{dx} = a\frac{a^2 - x^2}{(x^2 + a^2)^2} \qquad x > 0 \tag{24}$$

$$\frac{d^2f(x)}{dx^2} = -a\frac{x(3a^2 - x^2)}{(x^2 + a^2)^3} \qquad x > 0 \tag{25}$$

The first derivative has positive roots at $x = a$ and the second derivative has positive roots at $x = 0$ or $x = \sqrt{3}a$.

Let $g(x)$ be a function that resembles the Piella index for a single patch and $a_{min} = \min(a_i)$, $a_{max} = \max(a_i)$.

$$g(x) = \sum_{i=1}^{T} b_i \frac{a_i x}{x^2 + a_i^2} \qquad \forall x \in [a_{min}, a_{max}] \tag{26}$$

Also we have that $\sum_{i=1}^{T} b_i = 1$. The function $g(x)$ is continuous and infinitely differentiable. Therefore, we can calculate the first and second derivatives, as follows:

$$g'(x) = \frac{dg(x)}{dx} = \sum_{i=1}^{T} b_i a_i \frac{a_i^2 - x^2}{(x^2 + a_i^2)^2} \qquad \forall x \in [a_{min}, a_{max}] \tag{27}$$

$$g''(x) = \frac{d^2g(x)}{dx^2} = -\sum_{i=1}^{T} b_i a_i \frac{x(3a_i^2 - x^2)}{(x^2 + a_i^2)^3} \qquad \forall x \in [a_{min}, a_{max}] \tag{28}$$

- *Existence of optimums in* $[a_{min}, a_{max}]$

  The existence of optimums in the interval $[a_{min}, a_{max}]$ can be supported by simply verifying the validity of Bolzano theorem for this interval. More specifically,

$$g'(a_{min}) = \sum_{i=1}^{T} b_i a_i \frac{a_i^2 - a_{min}^2}{(a_{min}^2 + a_i^2)^2} \tag{29}$$

$$g'(a_{max}) = \sum_{i=1}^{T} b_i a_i \frac{a_i^2 - a_{max}^2}{(a_{max}^2 + a_i^2)^2} \tag{30}$$

  Since $a_i, b_i \geq 0$ and $a_{min} \leq a_i \leq a_{max}$ $\forall i$, it is straightforward to infer that $g'(a_{min}) > 0$ and $g'(a_{max}) < 0$. According to the Bolzano theorem, there exists at least one root of the equation $g'(x) = 0$ in the examined

interval. In other words, there exists at least one optimum (maximum or minimum) of the cost function $g(x)$ in $[a_{min}, a_{max}]$.

- *Sufficient condition for the existence of a single maximum in* $[a_{min}, a_{max}]$

  In this section, a sufficient condition for the existence of a single maximum in the requested interval is introduced. The existence of optimums in the interval $[a_{min}, a_{max}]$ is shown in the previous paragraph. The cost function $g(x)$ will have a single maximum in the requested interval, if the function remains concave $\forall x \in [a_{min}, a_{max}]$. The curvature of a function can be determined from the second derivative of the function. More specifically, the function $g(x)$ will be concave, if it is shown that $g''(x) < 0$, $\forall x \in [a_{min}, a_{max}]$. First, the values of the second derivative at the two extreme points are evaluated:

$$g''(a_{min}) = -\sum_{i=1}^{T} b_i a_i \frac{a_{min}(3a_i^2 - a_{min}^2)}{(a_{min}^2 + a_i^2)^3} \tag{31}$$

  Again since $a_i, b_i \geq 0$ and $a_{min} \leq a_i < \sqrt{3}a_i$ $\forall i$, we can infer that $g''(a_{min}) < 0$.

$$g''(a_{max}) = -\sum_{i=1}^{T} b_i a_i \frac{a_{max}(3a_i^2 - a_{max}^2)}{(a_{max}^2 + a_i^2)^3} \tag{32}$$

  As previously, $a_i, b_i \geq 0$, however, the sign of $3a_i^2 - a_{max}^2$ can not be directly determined. Assuming that $3a_i^2 \geq a_{max}^2$ or equivalently $a_{max} \leq \sqrt{3}a_i$, $\forall i$ then it is straightforward to infer that $g''(a_{max}) < 0$ and also $g''(x) < 0$ for all $x \in [a_{min}, a_{max}]$, because

$$g''(x) = -\sum_{i=1}^{T} b_i a_i \frac{a_k(3a_i^2 - x^2)}{(x^2 + a_i^2)^3} \tag{33}$$

  In the case that $x \leq a_i$ the term $3a_i^2 - x^2$ will still remain positive. In the case that $x \geq a_i$, the term will still remain positive, due to the imposed condition, since $x^2 \leq a_{max}^2 \leq 3a_i^2$. This condition is equivalent to the condition $a_{max} \leq \sqrt{3}a_{min}$. Consequently, in the case that the condition $a_{max} \leq \sqrt{3}a_{min}$ holds, the cost function $g(x)$ has certainly a single maximum in the interval $[a_{min}, a_{max}]$. In the opposite case, there might also exist b_i, a_i for which $g''(x) < 0$ in the requested interval, however, there might be cases of curvature changes in that interval and thus existence of multiple optima (maxima and minima). The nominator of $g'(x)$ is a polynomial of degree $4T - 2$, that may generally have more than a single root in the requested interval and i.e. multiple optima, unless the sufficient condition holds.

REFERENCES

[1] K. Romer and F. Mattern, "The design space of wireless sensor networks," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 54–61, 2004.

[2] A. Mahmood, P.M. Tudor, W. Oxford, R. Hansford, J.D.B. Nelson, N.G. Kingsbury, A. Katartzis, M. Petrou, N. Mitianoudis, T. Stathaki, A. Achim, D. Bull, N. Canagarajah, S. Nikolov, A. Loza, and N. Cvejic, "Applied multi-dimensional fusion," *The Computer Journal*, vol. 50, no. 6, pp. 660–673, 2007.

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

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

[5] N. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," in *Proc. IEEE Digital Signal Processing Workshop*, Bryce Canyon UT, USA, 1998.

[6] 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*, 1999, pp. 235–239.

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

[8] N. Mitianoudis and T. Stathaki, "Pixel-based and Region-based image fusion schemes using ICA bases," *Information Fusion*, vol. 8, no. 2, pp. 131–142, 2007.

[9] 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.

[10] N. Cvejic, D. Bull, and N. Canagarajah, "Region-based multimodal image fusion using ICA bases," *IEEE Sensors Journal*, vol. 7, no. 5, pp. 743–751, 2007.

[11] N. Cvejic, J. Lewis, D. Bull, and N. Canagarajah, "Adaptive region-based multimodal image fusion using ica bases," in *Proc. Int. Conf on Information Fusion*, 2006.

[12] N. Cvejic, D. Bull, and N. Canagarajah, "Improving fusion of surveillance images in sensor networks using independent component analysis," *IEEE Trans. on Consumer Electronics*, vol. 53, no. 3, pp. 1029 – 1035, 2007.

[13] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *International Conference on Image Processing (ICIP)*, Barcelona, Spain, 2003, pp. 173–176.

[14] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

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

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

[17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, New York, 2001, 481+xxii pages.

[18] Z. Wang and A.C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[19] A. Toet, "Detection of dim point targets in cluttered maritime backgrounds through multisensor image fusion," *Targets and Backgrounds VIII: Characterization and Representation, Proceedings of SPIE*, vol. 4718, pp. 118–129, 2002.

[20] The Image fusion server, "http://www.imagefusion.org/," .

Nikolaos Mitianoudis received the diploma in Electronic and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 1998. He received a MSc in Communications and Signal Processing from Imperial College London, UK in 2000 and the PhD in Audio Source Separation using Independent Component Analysis from Queen Mary, University of London, UK in 2004. Currently, he is working as a Research Associate at Imperial College London, UK. His research interests include Independent Component Analysis, Audio Signal Processing, Image Fusion and Computer Vision.

Tania Stathaki was born in Athens, Hellas. In September 1991 she received the Masters degree in Electronics and Computer Engineering from the Department of Electrical and Computer Engineering of the National Technical University of Athens (NTUA) and the Advanced Diploma in Classical Piano Performance from the Orfeion Athens College of Music. She received the Ph.D. degree in Signal Processing from Imperial College in September 1994. She is currently a Senior Lecturer in the Department of Electrical and Electronic Engineering of Imperial College and the Image Processing Group leader of the same department. Previously, she was Lecturer in the Department of Information Systems and Computing of Brunel University in UK, Visiting Lecturer in the Electrical Engineering Department of Mahanakorn University in Thailand and Assistant Professor in the Department of Technology Education and Digital Systems of the University of Pireus in Greece. Her current research interests lie in the areas of image processing, data fusion, non-linear signal processing, signal modelling and biomedical engineering. Dr. Stathaki is the author of 90 journal and conference papers.