Conditional Random Field Model for Robust Multi-Focus Image Fusion

Odysseas Bouzos, Ioannis Andreadis, and Nikolaos Mitianoudis

Abstract

In this paper, a novel multi-focus image fusion algorithm based on Conditional Random Field optimization (mf-CRF) is proposed. It is based on an unary term that includes the combined activity estimation of both high and low frequencies of the input images, while a spatially varying smoothness term is introduced, in order to align the graph-cut solution with boundaries of focused and defocused pixels. The proposed model retains the advantages of both spatial-domain methods and multi-spectral methods and by solving an energy minimization problem, finds an optimal solution for the multi-focus image fusion problem. Experimental results demonstrate the effectiveness of the proposed method that outperforms current state-of-the-art multi-focus image fusion algorithms in both qualitative and quantitative comparisons. Successful application of the mf-CRF model in multi-modal image fusion (visible-infrared and medical) is also presented here.

Index Terms

Conditional Random Field (CRF), multi-focus image fusion, Energy minimization, Independent Component Analysis (ICA)

I. Introduction

Since modern optical lenses exhibit limited Depth-of-Field, only objects within a certain distance range from the camera sensor can always be in focus. Parts of the scene that exist outside the focal plane of the camera sensor, are not in good focus or are blurred when captured, thus cannot provide trustworthy information about these parts of the observed scene. Fortunately, multi-focus fusion techniques can be used to merge images, captured with different focal settings, into a single composite image with extended

The authors are with the Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece.

E-mail: obouzos@ee.duth.gr, iandread@ee.duth.gr, nmitiano@ee.duth.gr

Manuscript received August 1, 2018; revised XXXX XX, 2018.

depth-of-field. The main goal of the multi-focus image fusion algorithms is to preserve regions that are focused and discard the respective defocused parts in the input images.

Several multi-focus image fusion algorithms have been developed lately, which can be classified according to the recent survey of Li et al. [1] into four major groups: *the multi-scale decomposition based methods*, *the sparse representation based methods*, *the methods which perform the fusion directly to the image pixels or in other transform domains* and *the methods that are a combination of different transforms*.

In multi-scale decomposition based methods, a multi-scale transform is applied to the input images in order to obtain their multi-scale representations. Then, a specific fusion rule is applied to the multi-scale representations to obtain a fused multi-scale representation. Finally, the fused image is obtained by applying the corresponding inverse multi-scale transform on the fused representation. Both the selection of the multi-scale decomposition method and the fusion strategy of the multi-scale representations greatly affect the quality of the fused image. Typical multi-scale decomposition based methods include: Laplacian Pyramids (LP) [2], Gradient Pyramids (GP) [3], Curvelet Transform (CVT) [4], Discrete Wavelet Transform (DWT) [5] and the Non-Subsampled Contourlet Transform (NSCT) [6]. Li et al. [7] made a study on multi-scale decomposition based methods, where the NSCT [6] has shown to achieve the best performance. Nonetheless, the main drawback of multi-scale decomposition based methods is their sensitivity to sensor noise [8].

Recently, experts have shown interest towards algorithms based on sparse representations (SR) [9]. SR-based methods use an overcomplete trained dictionary that leads to a sparser decomposition of the input images, compared to previous orthogonal transforms. Some of the recently proposed SR-based algorithms include the Robust Sparse Representation (RSR) algorithm [10], the Multi-Task Sparse Representation (MRSR) algorithm [10] and the Adaptive Sparse Representation - (ASR) algorithm [11]. According to [12], the Convolutional sparse representation (CSR) method [13] and the Convolutional Neural Network - CNN fusion method [14] can also be considered SR-based methods. The main drawback of SR-based algorithms is their low computational efficiency [15], due to the large overcomplete dictionaries, which are employed and the related forward/backward domain transforms. Either the forward or the backward transforms has to be performed using a series of orthogonal projections, known as Orthogonal Matching Pursuit (OMP) or derivatives of this mechanism, which is computationally expensive.

There are many methods that are not based on multi-scale decomposition or sparse representations. These methods can be classified to two general classes: *pixel domain methods* and *methods performed in other transform domains.* Pixel-domain methods apply fusion rules directly on the raw images, according to some local clarity information metrics and as a result, they tend to have less loss of original information, compared to transform-domain methods. These methods can be categorized as *pixel-based, block-based* and *region-based* and their goal is to select pixels, blocks or regions respectively, with higher clarity information. Typical measures for clarity estimation include local variance, spatial frequency and Laplacian energy. Unfortunately, fused images, produced by block-based methods, tend to suffer from blocking effects, since there might exist blocks that contain focused and defocused pixels simultaneously. The same drawback holds with the region-based methods. Traditional spatial-domain fusion methods suffer from wrong decisions in sub-regions [16], resulting to artifacts in the fused image. Recently, Bai et al. [17] used a quad-tree structure approach to decompose the source images into blocks with optimal sizes. Pixel-domain methods, such as Image Matting (IM) [18] and DSIFT [19], firstly calculate the activity levels of source images, which are later refined at a post-processing step, by making full use of the spatial correlation of adjacent pixels.

Apart from fusion methods that use multi-scale decomposition or sparse representation, different transform-domain based methods have been successfully applied for the image fusion problem. Typical applications include the multispectral domain transforms: Principal Component Analysis (PCA) [20] and Independent Component Analysis (ICA) [21]. Sun et al. [22] applied gradient-based image fusion using a Markov Random Field (MRF) fusion model. Fusion was then performed in the gradient domain and a Poisson equation was solved in order to force the gradients of the fused image to be close to the fused gradients. The major drawback of the fusion methods in other transform domain is the Gibbs phenomenon, due to poor image decomposition performance, which leads to spatial distortion, blocking and ringing artifacts. Performance can also decline in the case of small misregistration errors between the input images, since these methods use shift-variant dictionaries that are very sensitive to mis-registration. This can introduce visible artifacts in the fused image.

Since different transform domains have different advantages and limitations, various fusion methods that combine the advantages of different transform domains have been introduced. Li et al. [23] proposed a hybrid multiresolution approach that combined wavelets and contourlets. Experiments demonstrated that their hybrid method performed better than either the wavelet or contourlet domain when used independently. Liu et al. [15] proposed an image fusion framework based on a combination of multi-scale transform and sparse representation.

In this paper, we propose a novel CRF-based algorithm that uses a combination of different domains and preserves their respective advantages. An unary energy term is extracted by a combined activity estimation in both the multi-spectral and spatial domain, making the algorithm robust against both noise and against mis-registration of input images. A spatially varying smoothness term is then used to align the graph-cut solution to the border of focused and defocused pixels.

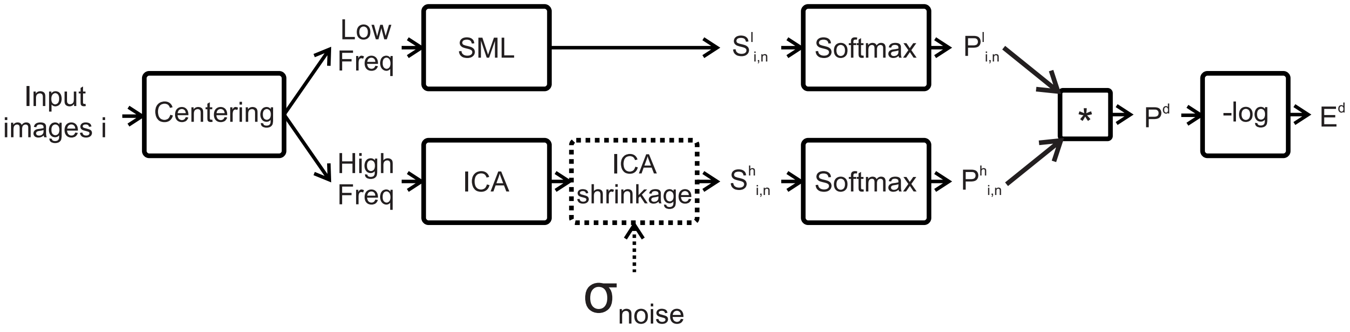


Fig. 1. Block diagram of data term extraction for clean and possibly noisy input images.

The main contributions of this paper are as follows:

1) A multi-focus CRF based algorithm (mf-CRF) for decision fusion in the spatial domain is introduced. Both multispectral and spatial domain activity estimations are used to formulate a unary energy term, while a novel smoothness term for consistency of the decision map is introduced. The smoothness term uses joint gradient information of input images and input image difference to align the graph cut solution to the border of focused and defocused input pixels and achieve consistency of the decision map.
2) Fused images obtained by mf-CRF are robust against input noise, since the algorithm retains the advantages of multispectral domain methods and preserves original information by fusing input images in the spatial domain and using weights equal to one.
3) Seamless fused images can be obtained through the introduced priors used in the smoothness term. Moreover, a global optimum solution to the multi-focus image problem can be achieved when two input images are used, by solving the energy minimization equation.

To the best of our knowledge, this is the first time that graph cuts are used in a multi-domain image fusion approach, preserving advantages of both spatial and multi-spectral domains. In addition, we exhibit the importance of Conditional Random Field optimization for the problem of image fusion, especially for the multi-focus image fusion problem.

II. mf-CRF Fusion algorithm

In the following section, we present a novel multi-focus CRF based model (mf-CRF) that preserves the advantages of multi-spectral algorithms (robust against input images with noise) and overcomes their limitations that lead to loss of original information, by working in the spatial domain. A spatially varying smoothness prior, which is sensitive to the joint contrast and difference of input images is introduced, and is used to assign pairwise constraints between adjacent pixels to achieve spatial consistency of the decision map. Another advantage of the proposed method is that it can handle images contaminated by noise. Equally, it can work with either fixed or adaptive dictionaries. Moreover, the mf-CRF can directly process multiple image inputs (more than 2).

For simplicity, we firstly assume that we have two input images, and we thus formulate the multi-focus fusion problem as a binary energy minimization problem. A binary decision map with labels $\ell_n \epsilon \{0, 1\}$ is used to reconstruct the fused image. When $\ell_n = 0$ the first input image is selected to represent the fused pixel at the spatial location n, similarly, when $\ell_n = 1$ the second input image is selected to represent the pixel at n in the fused image. In section III, we extend this model to multiple input images. A binary decision map with weights equal to one is preferred over a decision map with weights different to one, since fused images obtained by the former approach have high contrast and preserve original information, while fused images obtained by the latter approach have loss of original information and have lower contrast.

The fused image F for two input images I_1, I_2 is computed as:

$$F_n = (1 - \ell_n) I_{1,n} + \ell_n I_{2,n}. \tag{1}$$

We estimate the labels ℓ of the decision map according to the energy minimization equation:

$$\ell_{1..N} = \underset{\ell_{1..N}}{\operatorname{argmin}} \left[\sum_{n=1}^{N} E_n^d\left(\ell_n\right) + \sum_{(m,n)\epsilon C} E_{m,n}^s\left(\ell_m, \ell_n\right) \right] \tag{2}$$

where the data term E^d is the negative log likelihood of the activity levels at each spatial location n, while the smoothness term E^s is used to impose pairwise spatial consistency between the labels of adjacent pixels m, n of the decision map that belong in clique C equal to the $N - 8$ neighborhood. The $\alpha - \beta$ *swap* algorithm [24] based on graph-cuts is used to solve efficiently the energy minimization problem.

A. *Pipeline*

The block diagram for the estimation of the unary term E^d is depicted in Fig. 1. Firstly, centering is applied to the input images and both low and high frequency content from either image are extracted. Low and high frequency is performed using simple 1st-order filters with cut-off frequency at $\pi/2$. The created 1D 1st-order low-pass and high-pass filters are applied firstly row-wise and then column-wise. The high frequency content is converted to the transform domain and the sum of absolute transform coefficients is used as an estimator of their activity level. The *Sum of Modified Laplacian* (SML) [25] is employed to measure activity of low frequency contents. In order to draw safer activity level estimations and preserve the advantages of both spatial domain and multi-spectral domain, SML and ICA have been

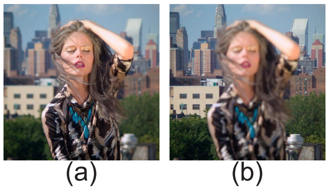


Fig. 2. Source input images, (a) Foreground is well focused, (b) Background is well focused.

employed. More precisely, SML uses a $[3 \times 3]$ window and when solely used, produces a very noisy activity level estimation decision map (Fig. 6a) and a noisy fused image (Fig. 7a), which however has good discrimination and accuracy near object boundaries mainly due to the small window. On the other hand ICA uses a $[7 \times 7]$ window. Due to the larger window the decision map (Fig. 6b) has less noise, high coherence but the activity level estimation is not very accurate near object boundaries since the larger window is more prone to include both focused and defocused pixels at the same time (Fig. 6b) and thus the fused image also suffers from blocking artifacts near the boundaries (Fig. 7b). ICA is also robust against noise since transform domain shrinkage algorithm can remove the noisy transform coefficients, while SML is not robust against additive gaussian noise. By using both SML for the low frequency and ICA for the high frequency activity estimations, the proposed mf-CRF preserves successfully the advantages of both pixel based and transform domain based methods. Finally, the *Softmax* function [26] is then used to convert the activity estimation of both frequency bands into probabilities, and the negative log-likelihood of the combined probability is used to form the unary E^d term.

B. Unary term E^d

Following the centering pre-processing step, activity estimation of both high and low frequencies of input images $I_{1...N}$ is used to form the unary term.

Fig. 2 shows two multi-focus input images that will be used to demonstrate the steps for the computation of the unary term E^d.

1) Probability of high frequency activity P^h: The high frequency content H is obtained after centering the input images and is estimated by dividing the images in $[7 \times 7]$ blocks. At each block, the ICA decomposition [21], [27] is performed. At each spatial location n, an activity level indicator S_n^h is computed. In the case of clean input images, S_n^h is equal to the the L1-norm of the block's transform coefficients. In the case of noisy input images, corrupted with Additive White Gaussian noise with known σ^2 variance, *transform-domain shrinkage* [27] is first used to filter noisy coefficients and then S_n^h is computed as the L1-norm of the block's denoised transform coefficients. In order to demonstrate the

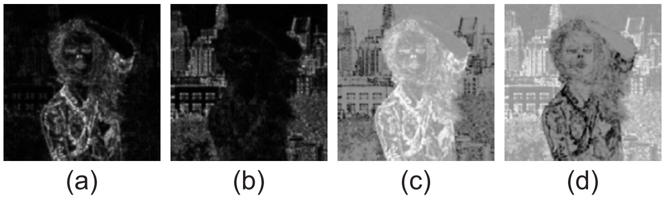


Fig. 3. Activity of high frequency S^h, probability of high frequency P^h. Darker colors correspond to smaller values, whereas brighter colors to higher values. (a) S_1^h, (b) S_2^h, (c) P_1^h, (d) P_2^h.

effectiveness of the mf-CRF model, a small shift-variant dictionary of fixed size is used to measure the activity of high frequency contents in transform domain. More precisely, we use an ICA dictionary of 48 bases. More details about ICA dictionaries can be found at [27]. ICA dictionary bases are edge sensitive [21] and are used as estimators of activity level for every patch of the high frequency contents of input images.

$$S_n^h = \sum_{b=1}^{N_b} |c_n^b| \tag{3}$$

where S_n^h is the activity measure of the transform coefficients at the spatial location n, c_n^b is the transform coefficient for the b basis at n and S_n^h equals to L1-norm of the transform coefficients. Fig. 3a, 3b depict the activity level estimations of S^h. Darker intensities indicate lower values, while brighter intensities higher values of activity level S^h. In Fig. 3a, S_1^h has higher values for the focused part of the girl which is well-focused, while in Fig. 3b S_2^h has higher values for the background. Ambiguous regions, such as the region above the girl's hand, have low activity levels S^h in both Fig. 3a, 3b.

The joint probability of activity estimation for high frequency contents P^h, given the label ℓ of the decision map at each spatial location n, is calculated as the softmax function of the S_n^h:

$$P^h\left(S_{1,n}^h, S_{2,n}^h \mid \ell_n\right) = \begin{cases} \dfrac{\exp(S_{1,n}^h)}{\sum_{i=1}^2 \exp(S_{i,n}^h)} & \text{if } \ell_n = 0 \\ \dfrac{\exp(S_{2,n}^h)}{\sum_{i=1}^2 \exp(S_{i,n}^h)} & \text{if } \ell_n = 1 \end{cases} \tag{4}$$

where $S_{i,n}^h$ corresponds to the activity measure of high frequency h of image i at location n. Fig. 3c, 3d show the probabilities of high frequency P^h. Darker intensities indicate lower probabilities while brighter intensities higher probabilities. The well-focused parts of the input images have higher probabilities and brighter intensities, while the defocused regions have lower probabilities and darker intensities. Ambigious regions such as the region above the girl's hand have the approximately the same probabilities in both images.

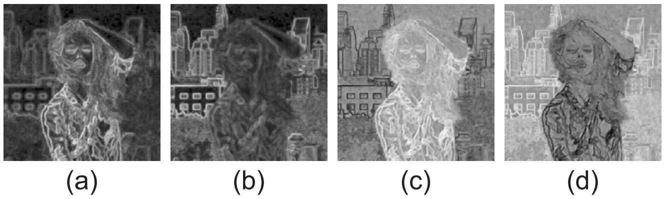


Fig. 4. Activity of low frequency S^l, probability of low frequency P^l. Darker colors correspond to smaller values, whereas brighter colors to higher values. (a) S_1^l, (b) S_2^l, (c) P_1^l, (d) P_2^l.

2) Probability of low frequency activity: In this work, we use an activity level indicator for the low frequency content. In order to measure the low-frequency activity level, we assign S_n^l to the Sum of Modified Laplacian (SML) [25]. For the low-frequency content, we define the saliency estimator $S_{i,n}^l$ for the low frequency l and spatial location n for the i-th image.

The discrete version of the *Modified Laplacian* (ML) is computed by the following formula:

$$\begin{aligned} ML\left(x,y\right) = &\left|2f\left(x,y\right) - f\left(x-s,y\right) - f\left(x+s,y\right)\right| \\ &+ \left|2f\left(x,y\right) - f\left(x,y-s\right) - f\left(x,y+s\right)\right| \end{aligned} \tag{5}$$

where x, y are pixel coordinates and s is a variable spacing between the pixels used to compute the derivatives. The sum of modified laplacian (SML) for a small window can then be calculated by:

$$SML\left(x,y\right) = \sum_{i=x-N}^{x+N} \sum_{j=y-N}^{y+N} ML\left(i,j\right), \quad ML\left(i,j\right) > T \tag{6}$$

where N defines the size of the window and T is a threshold value.

$$S_{i,n}^l = SML\left(L_{i,n}\right) \tag{7}$$

where $L_{i,n}$ is the low-frequency content for the input image i at the spatial location n. In our experiments, SML is applied to a window of size $[3 \times 3]$ and a threshold value $T = 0$ and $s = 1$. Fig. 4a, 4b include the low frequency activity estimations S^l for the input images, darker intensities indicate lower activity levels, while brighter intensities indicate higher activity estimations. The softmax function is again used to convert the saliency information to probability.

$$P^l\left(S_{1,n}^l, S_{2,n}^l \mid \ell_n\right) = \begin{cases} \dfrac{\exp(S_{1,n}^l)}{\sum_{i=1}^2 \exp(S_{i,n}^l)} & \text{, if } \ell_n = 0 \\ \dfrac{\exp(S_{2,n}^l)}{\sum_{i=1}^2 \exp(S_{i,n}^l)} & \text{, if } \ell_n = 1 \end{cases} \tag{8}$$

where $S_{i,n}^l$ corresponds to the activity measure of low frequency l of image i at location n. Fig. 4c, 4d include the probabilities P^l for the low frequency, darker intensities indicate lower activity levels, while

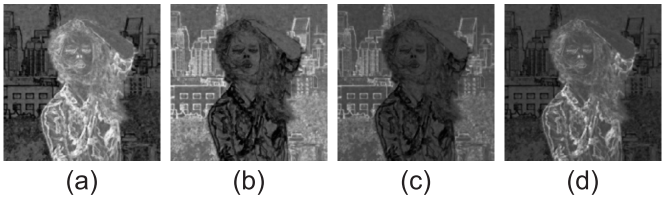


Fig. 5. Combined probabilities P^d and energy data term E^d: (a) P^d_1, (b) P^d_2, (c) E^d_1, (d) E^d_2. Darker colors correspond to smaller values, while brighter colors to higher values.

brighter intensities indicate higher activity estimations. Well-focused regions have brighter intensities than defocused regions. Ambiguous regions also have approximately same gray values in both images.

3) Data Term: The joint probability distribution for the unary data term over the label field is given by the product of the probability density function for the low and high frequencies, i.e. $P^d = P^h P^l$. The unary term E^d is then defined as the negative log likelihood of P^d, i.e. $E^d = -\log(P^d)$. Fig. 5a, 5b depict the combined probabilities P^d of the input images while Fig. 5c, 5d show the unary term E^d for the two input images. Darker intensities indicate lower values, while brighter intensities denote higher values.

C. Smoothness energy term

Here, we introduce the spatially varying penalty term $f\left(\Delta I, \nabla I\right)$ for neighboring pixels having different labels, which depends on both the image difference and the gradients of input images. It combines two priors about both the fused image F^g gradients and the joint contrast of both input images C.

The prior of fused image gradients F^g is used to achieve a seamless fusion result, which can be obtained in the case that the difference of pixels around the segmentation cut is small. The segmentation cut corresponds to a change of selected input image to represent adjacent pixels, thus the F^g prior aims to minimize the difference of input images when adjacent pixels p, q have different labels. Thus, F^g_{pq} corresponds to the euclidean distance of potential gradients $\sum \nabla_{pq} F$ and is approximated, as follows:

$$F^g_{pq} = \sqrt{\left(J_{1,p} - J_{2,q}\right)^2 + \left(J_{2,p} - J_{1,q}\right)^2} \tag{9}$$

$$J_{i,n} = \|I_{i,n}\| \tag{10}$$

where $J_{i,n}$ is the L2-norm of the RGB values of I_i image at location n. The joint contrast term C is used to align the segmentation cut with the boundaries of focused and out-of-focus regions, which are

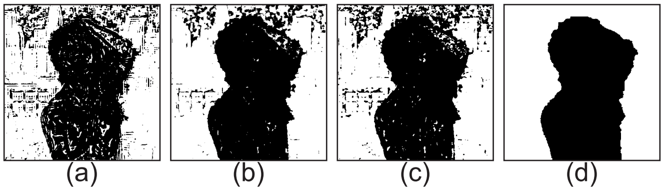


Fig. 6. Decision maps obtained by the labels a) $\ell = \operatorname{argmax}\left(P^l\right)$, b) $\ell = \operatorname{argmax}\left(P^h\right)$ c) $\ell = \operatorname{argmin}\left(E^d\right)$, d) $\ell = \operatorname{argmin}\left(\text{mf-CRF}\right)$.

likely to appear near object boundaries, where both input images have high contrast and is likely to have a cut solution.

The proposed term C, which depends on both input images gradients, is given by:

$$C_{pq} = \sqrt{\left(\nabla_{pq} J_1\right)^2 + \left(\nabla_{pq} J_2\right)^2} \tag{11}$$

$$\nabla_{pq} J_i = J_i(p) - J_i(q) \tag{12}$$

Hence, the smoothness term E^s is computed as:

$$E^s\left(\ell_p, \ell_q\right) = \operatorname{dis}(p,q)^{-1} \frac{F^g_{pq} + c}{C_{pq} + c} I_{\ell_p \neq \ell_q} \tag{13}$$

where dis$(\cdot)$ is the Euclidean distance of neighboring pixels p, q, I is an indicator function that equals to 1 for different labels and c is a small constant to ensure stability.

Fig. 6c demonstrates the labels that directly minimize the unary term, which result to a very noisy label estimation and Fig. 6d shows the labels that minimize the proposed mf-CRF model. The extracted Decision Map from the mf-CRF model has label consistency, without being noisy and the graph cut solution follows the shape prior of the model, and the boundary between the focused and defocused pixels. Fig. 7 includes the fused images obtained by the labels that minimize the unary term (Fig. 7c) and by the labels that minimize the mf-CRF model (Fig. 7d) for the input source images Fig. 2.

The fused image obtained by the mf-CRF model has high visual quality, without artifacts around the graph-cut solution due to the introduced smoothness term.

D. Energy Minimization and $\alpha - \beta$ swap

The energy minimization of (2) can be solved efficiently using graph cut optimization. Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a weighted graph, which consists of vertices $\mathcal{V}$ and edges $\mathcal{E}$. An s-t cut C on a graph $\mathcal{G}$ with two

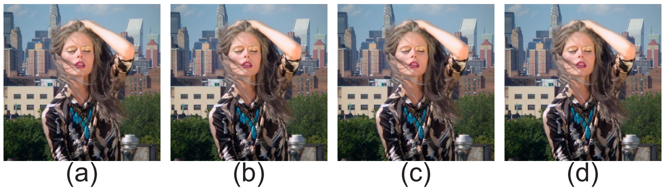


Fig. 7. Fused images obtained by application of the labels a) $\ell = \operatorname{argmax}\left(P^{l}\right)$, b) $\ell = \operatorname{argmax}\left(P^{h}\right)$ c) $\ell = \operatorname{argmin}\left(E^{d}\right)$, d) $\ell = \operatorname{argmin}\left(\text{mf-CRF}\right)$.

terminals s, t is a partitioning of the nodes in the graph into two disjoint subsets S and T such that the source s is in S and the sink t is in T. The cost of the cut $\mathcal{C}$, denoted $|\mathcal{C}|$, equals to the sum of its edge weights. The minimum cut problem is to find the cut with smallest cost. Boykov et al. [24] proposed two move-making algorithms: a) the $\alpha-$expansion and b) the $\alpha\beta-$swap in order to solve the minimum cut problem. Both algorithms were tested in our experiments resulting to the same graph-cut solution. In the mf-CRF approach, we chose to use the $\alpha\beta-$swap. The swap algorithm has one possible move for every pair of labels $\alpha, \beta\ \epsilon\mathcal{L}$. An $\alpha\beta$-swap move allows a random variable, whose current label is α or β, to take either label α or label β. The transformation function $T_{\alpha\beta}\left(\cdot\right)$ for an $\alpha\beta$-swap transforms the current label x_i^c of a random variable x_i, as follows:

$$x_i^n = T_{\alpha\beta}\left(x_i^c, t_i\right) = \begin{cases} x_i^c & \text{if } x_i^c \neq \alpha \text{ and } x_i^c \neq \beta, \\ \alpha & \text{if } x_i^c = \alpha \text{ or } \beta \text{ and } t_i = 0, \\ \beta & \text{if } x_i^c = \alpha \text{ or } \beta \text{ and } t_i = 1. \end{cases} \tag{14}$$

One iteration of the algorithm involves performing swap moves for all α and β in the label set $\mathcal{L}$ successively in some order. The energy of a move t is the energy of the labeling x^n that the move t induces, that is, $E_m\left(t\right) = E\left(T\left(x^c, t\right)\right)$. The move energy is a pseudo-Boolean function $\left(E_m : \left\{0, 1\right\}^n \rightarrow R\right)$ and will be denoted by $E_m\left(t\right)$. At each step of the swap-move algorithm, the optimal move t^* (i.e., the move decreasing the energy of the labeling by the greatest amount) is computed. This is done by minimizing the move energy, that is, $t^* = \operatorname{argmin}_t E\left(T\left(x^c, t\right)\right)$. The pseudo-Boolean energy corresponding to a swap move is:

$$\begin{aligned} E\left(T_{\alpha\beta}\left(x^c, t\right)\right) = & \sum_{i \epsilon \mathcal{V}} \mathit{\Phi}_i\left(T_{\alpha\beta}\left(x_i^c, t_i\right)\right) \\ & + \sum_{(i,j) \epsilon \mathcal{E}} \Psi_{ij}\left(T_{\alpha\beta}\left(x_i^c, t_i\right), T_{\alpha\beta}\left(x_j^c, t_j\right)\right) \end{aligned} \tag{15}$$

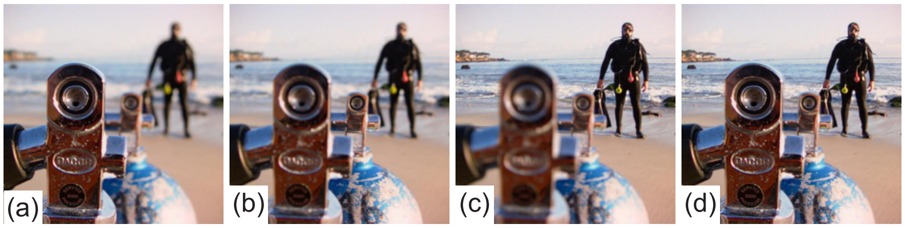


Fig. 8. Fusion result of more than two input images. a) Source 1, b) Source 2, c) Source 3, d) mf-CRF fused image.

where x^c is fixed and t is the unknown variable, $\Phi_i(x_i)$ the unary term and $\Psi_{ij}(x_i, x_j)$ the binary terms. The swap algorithm can be used whenever the following condition is satisfied:

$$\Psi_{ij}(\alpha,\alpha) + \Psi_{ij}(\beta,\beta) \leq \Psi_{ij}(\beta,\alpha) + \Psi_{ij}(\alpha,\beta), \forall\alpha,\beta\epsilon\mathcal{L} \tag{16}$$

E. Fusion of noisy multi-focus images

The proposed mf-CRF method can be extended to work competitively with noisy multi-focus images, since it preserves the advantages of the ICA domain, and the domain's robustness against noisy input images. More precisely, Fig. 1 demonstrates the extended pipeline that handles noisy input images.

Firstly, centering is applied to the noisy input images extracting low and high frequency contents. No further post-processing is required for the low-frequency contents, since they are treated as non-noisy, as shown in the pipeline Fig. 1, despite the white noise assumption. On the other hand, the high frequency contents, are assumed to include both information of the scene as well as noise. In order to handle efficiently the information of the high frequency contents and proceed to more accurate activity level estimation, a hard shrinkage algorithm is applied to the transform coefficients using the standard deviation of noise σ_n. Spatial ICA coefficients with absolute value below $2\sigma_n$ are set to zero [27]. Later, the activity level estimation for the high frequency contents is estimated as the summation of the respective absolute value of ICA coefficients after shrinkage.

The smoothness prior is extracted from the low-frequency content of the input images. In the mf-CRF approach, fusion of noisy input images is performed using the decision map obtained by solving the energy minimization problem with the $\alpha - \beta$ *swap* algorithm [24].

Detailed performance evaluation of the proposed mf-CRF against additive Gaussian noise is included in Section IV-E.

III. Multi-Focus image Fusion with more than two inputs

Most image fusion methods, in order to handle more than two input images, work in a pairwise fashion until all input images are fused (see Nejati et al. [16]). A main advantage of the proposed mf-CRF model is its ability to fuse more than two input images. The unary term is computed similarly to the case of two input images. The proposed Smoothness term E^s for N input images can be rewritten as follows:

$$E^s\left(\ell_p, \ell_q\right) = \text{dis}(p,q)^{-1} \frac{\sqrt{\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\Delta_{pq}^{ij} J\right)^2 + c}}{\sqrt{\sum_{k=1}^{N} \left(\nabla_{pq} J_k\right)^2 + c}} V_{pq} \tag{17}$$

$$V_{pq} = \frac{|\ell_p - \ell_q|}{N-1} \tag{18}$$

where labels $\ell_p, \ell_q \epsilon \{0, 1, \ldots, N-1\}$ and correspond to the labels of the first, second,. . . , N^{th} input image respectively, and c is a stability constant. In addition,

$$\Delta_{pq}^{ij} J = J_i(p) - J_j(q) \tag{19}$$

When the mf-CRF has multiple input images, images should be provided ordered from near to far focus. Adjacent pixels in the fused image are likely to be from the same image (same label) or from input images captured with slightly different focal length (labels that are close). Thus, the smoothness penalty for adjacent labels, is relative to the focal difference of the input images and the penalty for assigning labels to adjacent pixels is relative to the order difference of the input images. Adjacent pixels that are around edges of the input image are likely to belong to different objects which have different distance from the camera sensor, and thus the penalty is greatly reduced, allowing them to get more "distant" labels. In order to apply the mf-CRF algorithm to multiple input images, the images should be ordered either manually from near-focused to far-focused, or automatically using the focal length as found in the exif data of each input image.

In order to demonstrate the mf-CRF effectiveness, an example with three input images is shown in Fig. 8. As it can be seen, the fused image has no artifacts and contains all the focused regions from all input images.

IV. Experimental results

In order to evaluate the performance of the proposed algorithm, we need to benchmark some state-of-the-art multi-focus fusion algorithms. Both qualitative and quantitative comparisons are included using datasets that contain input pairs with RGB images and grayscale images respectively. An artificial dataset with manually added Gaussian blur and noise was also created, in order to assess towards the evaluation of compared methods against Additive Gaussian Noise. A short analysis of the proposed method's computational cost is also provided and discussed.

A. Datasets

The proposed algorithm is evaluated on two multi-focus image datasets and one artificial dataset. The first dataset consists of 20 RGB multi-focus image sets, obtained from the Lytro dataset [16]. The second dataset consists of 17 grayscale multi-focus image pairs, which were obtained from the dataset [28]. Lastly, an artificial dataset of 15 image pairs was created from 15 images from the Pascal Visual Object Classes (VOC) Challenge dataset [29]. All three datasets are available at: https://bit.ly/2M4hAzQ

B. Benchmark algorithms

The proposed algorithm is compared with 11 state-of-the-art multi-focus image fusion algorithms. More specifically, we used the following 3 multi-scale decomposition based methods: a) Non-Subsampled Contourlet Transform (NSCT) [6], b) Multi-scale Weighted Gradient-based Fusion algorithm (MWGF) [3], c) image fusion algorithm with guided filtering (GF) [30]. In addition, the following 3 sparse representation based (SR) methods were also used: a) Sparse Representation (SR) [9], b) Adaptive Sparse Representation (ASR) [11], c) multi-focus image fusion algorithm with a deep Convolutional Neural Network (CNN) [14]. The following 4 methods based on other domains were also used: a) the Dense SIFT image fusion algorithm (DSIFT) [19], b) the image matting algorithm for image fusion (IM) [18], c) the boundary-finding image fusion algorithm (BF) [28], d) the quadtree-based image fusion algorithm [17]. Finally, the following method, which uses a combination of different transforms, was also used: Non-Subsampled Contourlet Transform with Sparse Representation (NSCT-SR) [15].

C. Qualitative evaluation

The visual quality of fused images is used to assess the qualitative evaluation of multi-focus image fusion algorithms.

Fig. 9a and 9b show the source images for the ‘Clock’ image set found in the grayscale dataset. Fig. 9 demonstrates the fusion results along with magnifications of the region around the boundary of focused and defocused pixels, which are included to assess the fusion visual quality. Sparse representation methods SR and ASR, multi-scale decomposition method NSCT and the transform combination method NSCT-SR lose part of the clock border near the digit ‘8’, due to the fusion rules that are used. The DSIFT fused image loses part of the clock border near the digits ‘8’ and ‘9’. The Quadtree fused image also suffers from loss of information near the digit ‘8’, since the algorithm fails to find blocks of optimal size to decompose the magnified region. The GF-fused image loses original information for the clock border near the digit ‘8’, since it cannot provide correct weights near the decision boundary. Fusion results of IM, BF, CNN, and mf-CRF provide visual results of higher quality, since they do not produce artifacts

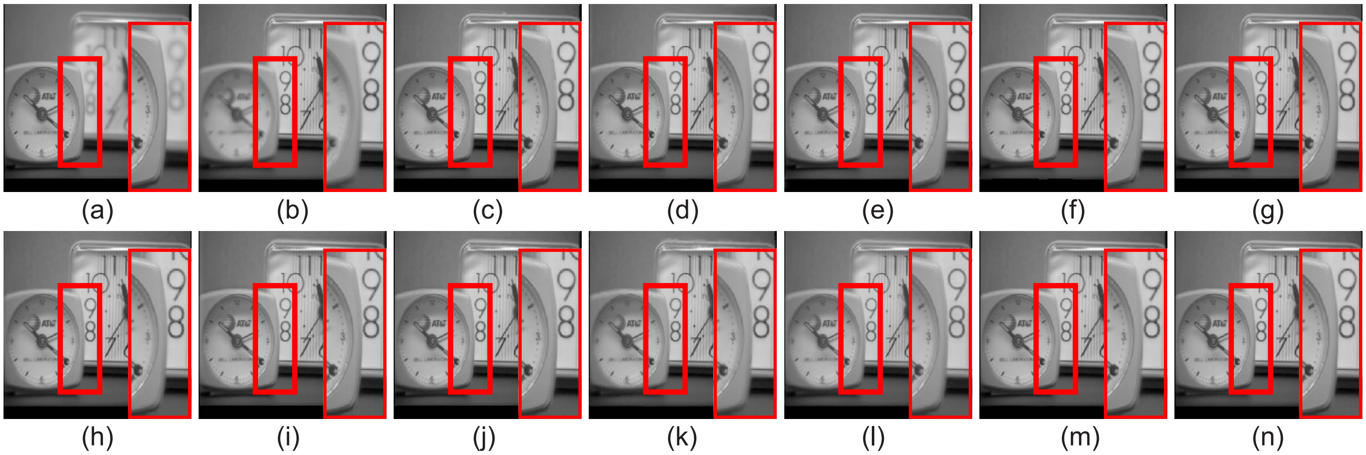
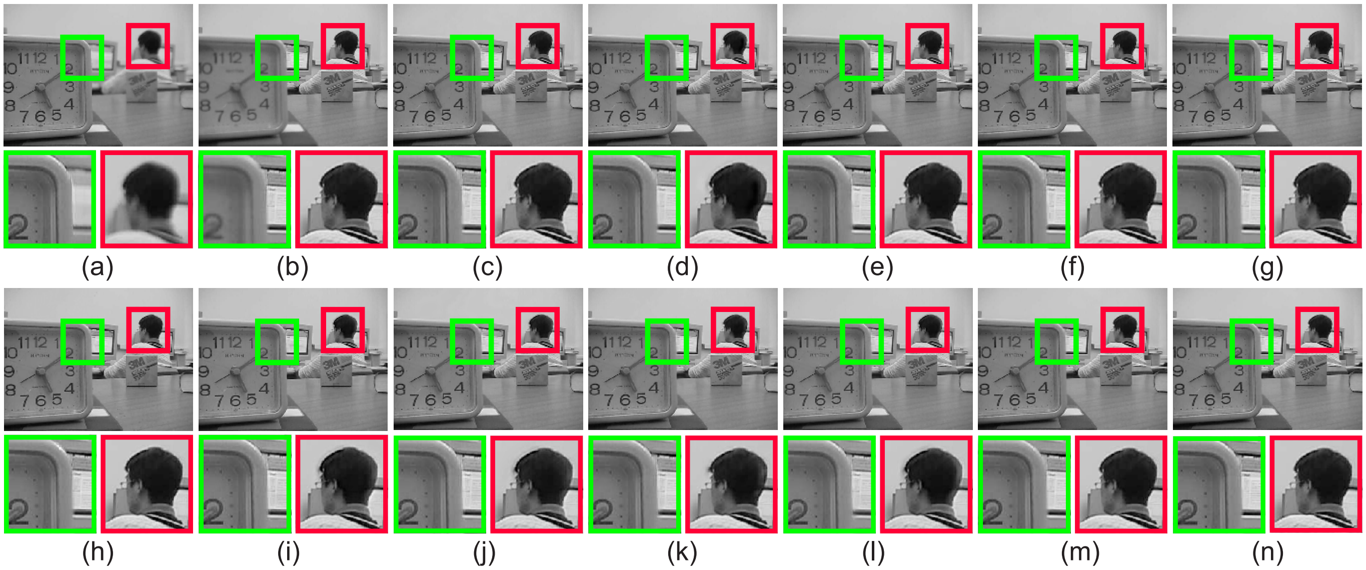


Fig. 9. Fusion results for the ‘Clock’ set. a) Source 1 (Foreground is well-focused), b) Source 2 (Background is well-focused.) c) DSIFT, d) GF, e) MWGF, f) Quadtree, g) BF, h) IM, i) NSCT, j) NSCT-SR, k) SR, l) ASR, m) CNN, n) mf-CRF.

near the clock border. The CNN method does not produce a crisp border, since it does not use weights equal to one resulting to fused images of lower contrast and loss of original information. Similarly, the fused image of IM has edge-blurring, due to the used weighted fusion approach with weights different to one. The BF method also has blurred regions, found at the lower left part of digit ‘8’ and also near the lower right part of the border clock, which are mainly the result of the weighted fusion with weights different to one, resulting to loss of original information. On the other hand, the fused result of the proposed mf-CRF method has higher visual quality against all other compared methods. The proposed smoothness prior can effectively align the cut with the clock border, which is preserved in the fused image best. A seamless fusion result is achieved without losing part of original information, since the weights used are equal to one.

Fig. 10a and 10b show the source images for the ‘Lab’ image set found in the grayscale dataset. Fig. 10 demonstrates the fusion results along with magnifications of two selected regions around the clock border and the student’s head. Transform-domain methods (NSCT, NSCT-SR, SR, ASR) produce artifacts around the head, since they are not robust against mis-registration in input images. Moreover, the edge of the clock is not sharp, due to loss of original information. The fused image obtained by the DSIFT algorithm produces artifacts in the left side of the head and also loses part of the original clock border. The image produced by GF does not have a sharp clock edge, while also produces artifacts near the student’s head. The methods BF, IM, MWGF, Quadtree, CNN as well as the proposed method provide visually satisfying results, which do not have artifacts. The Quadtree method produces some artifacts around the clock border, since the method cannot find blocks of optimal size for that region. The BF, IM,



(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l) (m) (n)

Fig. 10. Fusion results for the ‘Lab’ set. a) Source 1 (Near focused), b) Source 2 (Far - focused.) c) DSIFT, d) GF, e) MWGF, f) Quadtree, g) BF, h) IM, i) NSCT, j) NSCT-SR, k) SR, l) ASR, m) CNN, n) mf-CRF.

CNN, MWGF methods lose original information and produce images of lower contrast, since they use weighted fusion of input images, where weights can be different to one. The proposed mf-CRF algorithm can identify successfully the focused and defocused areas, providing a robust graph-cut solution for fusing input images with mis-registration. Magnified regions of the fused image obtained by mf-CRF have better visual quality than the respective magnified regions of all compared algorithms.

Fig.11a and 11b show the source images for the ‘Golfer’ image set found in the dataset with RGB image pairs, while Fig. 11 shows the fused images along with magnifications of two selected regions for each of the fused images. The Multi-scale based method (NSCT), the sparse-representation based methods (SR, ASR and NSCT-SR), which use a combination of transform domains, all exhibit a loss of original information, since the fused images have lower contrast compared to the pixel-domain methods (DSIFT, GF, Quadtree), the sparse-representation method (CNN) and the proposed mf-CRF, which uses a combination of pixel and multispectral domains. When larger dictionaries are used (ASR, NSCT-SR, SR), the reconstruction is more accurate and fused images have higher contrast compared to the NSCT. The BF method fails to identify the focused and defocused regions, in the red magnified region producing an out-of-focus result. The GF, MWGF, IM, CNN methods produce images of lower contrast compared to the DSIFT, Quadtree, mf-CRF algorithms, since they produce blurred results near the decision boundaries in both selected regions, due to using weights different to one. The DSIFT, Quadtree and mf-CRF methods produce better visual results in both magnified regions and produce fused images of better contrast. The

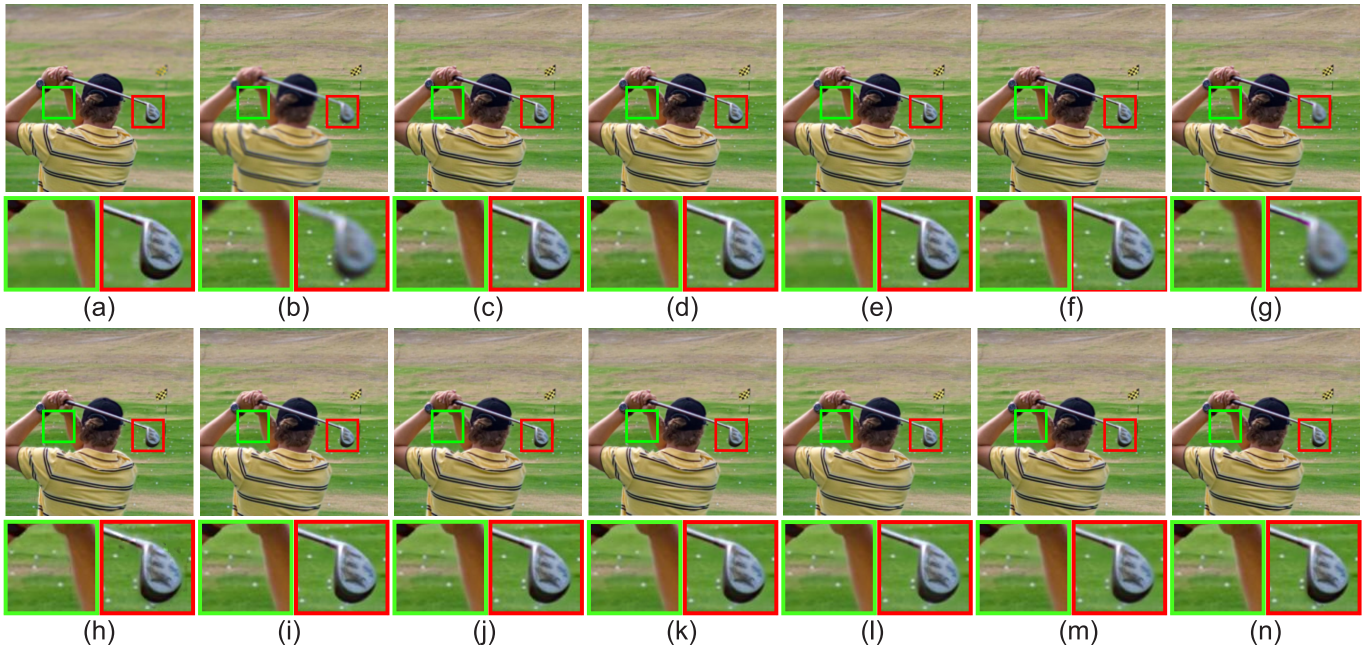


Fig. 11. Fusion results for the 'Golfer' set. a) Source 1 (Foreground is well focused), b) Source 2 (Background is well focused.) c) DSIFT, d) GF, e) MWGF, f) Quadtree, g) BF, h) IM, i) NSCT, j) NSCT-SR, k) SR, l) ASR, m) CNN, n) mf-CRF.

DSIFT method has blurred pixels around the club. The Quadtree method also has blurred pixels in the fused image, since it fails to find blocks of optimal size, especially near the boundaries of focused and defocused pixels. The fused image obtained by the mf-CRF has better visual quality than the compared methods for the input image pair 'golfer' that features perfect registration. More fusion results for the proposed mf-CRF method can be found online at: https://bit.ly/2M4hAzQ, along with all figures of the paper.

D. Objective Evaluation

Liu et al. [31] studied 12 popular image fusion metrics and classified them in four categories: (1) information theory based metrics, (2) image feature based metrics, (3) image structural similarity based metrics and (4) human perception inspired metrics. In this paper, a representative metric from each of the above categories is used to evaluate objectively the performance of the proposed mf-CRF method. In order to guarantee the objectivity of the presented evaluation results, all metrics were implemented by third-parties and the parameters used by each fusion algorithm are set to the values described in the respective publications. Larger values in each of the metrics, indicate better fusion result.

Tables I and II include the objective performance of different fusion methods using four metrics that belong to all four categories of [31]. The average scores on 20 RGB image pairs and 17 grayscale image

TABLE I

MEAN SCORE OF METRICS FOR 20 RGB SETS.

Methods	MI	Q_G	Q_Y	Q_{CB}
DSIFT	8.9315 (1,10)	0.7633 (2,8)	0.9877 (0,0)	0.8092 (4,8)
GF	8.2421 (0,0)	0.7616 (2,1)	0.9821 (0,1)	0.7976 (0,0)
MWGF	8.4569 (0,0)	0.7497 (0,0)	0.9873 (0,3)	0.7972 (0,0)
Quadtree	8.9297 (1,7)	0.7628 (0,1)	0.9884 (0,5)	0.8089 (6,7)
BF	8.8953 (1,1)	0.7583 (0,0)	0.9896 (4,7)	0.8091 (1,2)
IM	8.6090 (0,0)	0.7575 (0,0)	0.9834 (0,0)	0.7969 (0,0)
NSCT	7.1986 (0,0)	0.7502 (0,0)	0.9649 (0,0)	0.7527 (0,0)
NSCT-SR	7.4092 (0,0)	0.7511 (0,0)	0.9673 (0,0)	0.7602 (0,0)
SR	8.3321 (0,0)	0.7581 (0,1)	0.9760 (1,0)	0.7827 (0,0)
ASR	7.1310 (0,0)	0.7510 (0,0)	0.9691 (0,0)	0.7264 (0,0)
CNN	8.6420 (0,0)	0.7629 (4,4)	0.9875 (1,1)	0.8084 (1,0)
mf-CRF	**8.9506** (17,2)	**0.7641** (12,5)	**0.9896** (14,3)	**0.8098** (8,3)

pairs are shown in Table I and Table II respectively and the best score appears in bold. A parenthesis that includes the number of times each method achieves the first (left) and second (right) best scores according to each metric is also shown in the table.

Firstly, every method is compared, according to the *Mutual information* (MI) [32] metric of the information theory category. Pixel-domain based methods (DSIFT, MWGF, BF, IM), the multi-scale decomposition method GF, the sparse representation based CNN and the proposed mf-CRF exhibit greater performance compared to the multiscale method NSCT, the sparse representation methods SR and ASR and the method based on transform combination NSCT-SR. The sparse representation methods SR, ASR and multi-scale method NSCT have higher loss of original information compared to pixel-domain based methods. Moreover, methods that incorporate weighted fusion of original input images (CNN, IM, GF, MWGF) tend to have lower performance, since they lose original information, despite the binary-hard decision fusion methods (mf-CRF, DSIFT, Quadtree), which preserve the original information.

In both cases of RGB and grayscale image sets, the proposed mf-CRF fusion method performs very well, compared to the other methods and achieves first and second best scores in most benchmarks.

The proposed method exhibits the highest and most frequently best score in terms of gradient-based

TABLE II

MEAN SCORE OF METRICS FOR 17 GRAYSCALE SETS.

Methods	MI	Q_G	Q_Y	Q_{CB}
DSIFT	8.6267 (0,5)	0.7405 (0,2)	0.9799 (0,1)	0.7851 (5,1)
GF	7.5732 (0,0)	0.7361 (0,1)	0.9715 (0,0)	0.7584 (0,0)
MWGF	7.8171 (0,0)	0.7231 (0,0)	0.9761 (0,0)	0.7695 (0,1)
Quadtree	8.6604 (6,4)	0.7400 (1,7)	0.9818 (0,3)	0.7777 (1,6)
BF	8.6147 (1,3)	0.7397 (1,2)	0.9886 (6,7)	0.7884 (2,6)
IM	8.3264 (0,0)	0.7353 (0,2)	0.9756 (0,1)	0.7683 (0,0)
NSCT	6.2947 (0,0)	0.7074 (0,0)	0.9439 (0,0)	0.7284 (0,0)
NSCT-SR	6.3474 (0,0)	0.7072 (0,0)	0.9416 (0,0)	0.7285 (0,0)
SR	7.7775 (0,0)	0.7241 (0,0)	0.9529 (0,0)	0.7402 (0,0)
ASR	6.3790 (0,0)	0.7192 (0,0)	0.9541 (0,0)	0.7057 (0,0)
CNN	8.3190 (0,0)	0.7390 (5,0)	0.9848 (3,0)	0.7832 (2,0)
mf-CRF	**8.7034** (10,5)	**0.7422** (10,3)	**0.9889** (8,5)	**0.7891** (7,3)

fusion performance index (Q_G) [33]. The pixel-domain based methods (DSIFT, Quadtree, BF), the sparse representation based (CNN), the multi-scale (GF) and the transform combination (mf-CRF) seem to preserve the strong edges that are transferred from input images to their fused results, compared to the multiscale (NSCT) , the sparse representation based (SR, ASR) and the transform combination (NSCT-SR).

In order to compare the image structural similarity of the fused images and the original input images, we use the Yang's metric Q_Y [34]. Since the proposed method uses data-driven smoothness priors, the structural features of the input images can be strongly preserved in the fused results. Both the mf-CRF and the Boundary Finding methods have the highest performance among the methods, and the proposed method slightly outperforms in terms of the average score. Transform-domain fusion methods, have lower scores compared to pixel-domain methods, since there is loss of original structural features.

Finally, Q_{CB} [35] is a human perception-inspired fusion metric, which measures the amount of contrast transferred from the source images to the fused result. The pixel domain methods (DSIFT, IM, BF, Quadtree), the sparse representation method (CNN) and the multi-scale based methods (GF, MWGF), preserve the original contrast and exhibit larger Q_{CB} values compared to the multi-scale based (NSCT),

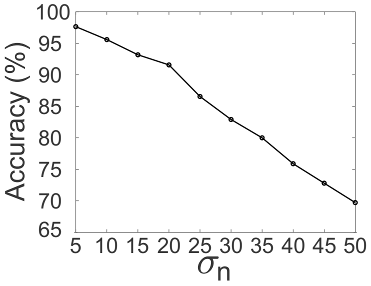


Fig. 12. Classification accuracy $(\%)$ vs σ_n.

the sparse representation based methods (SR, ASR) and transform combination method (NSCT-SR). Moreover, among the former methods, weighted-based fusion methods (IM, GF, MWGF, CNN) lose more contrast of the original images and thus have lower Q_{CB} score compared to DSIFT, Quadtree, BF, mf-CRF. The proposed mf-CRF fusion method, which uses a combination of pixel-domain and multi-spectral domain principles, has the largest average Q_{CB} score and is the most frequent to achieve fused images of higher contrast, compared to the other methods.

E. Noise Evaluation

In this section, we evaluate the performance of mf-CRF against additive Gaussian noise. For the first experiment, 15 images were selected from the VOC 2012 dataset [29], these images were labeled for segmentation competition. In this experiment, we used the provided segmentation maps, in order to create images with artificial PSF (Point Spread Function). According to [36], the PSF can be approximately modeled by a 2D Gaussian function. For each image, two new images were created using the segmentation maps. More precisely, the background pixels in the first artificial image were blurred with a Gaussian filter with variance σ_1^2 while the rest of the pixels were copied directly from the ground truth image. The second artificial image contained the pixels, that did not belong to the background, blurred with a Gaussian filter with variance σ_2^2, while the background pixels were copied directly from the ground truth image. The standard deviations were assigned random values for each dataset, $\sigma_1, \sigma_2 \epsilon \, [0.1, 2.5]$. As a result, 15 sets of two artificially out-of-focus input images with ground truth decision maps were made. Additive Gaussian noise with $\sigma_n \epsilon \, [5, 50]$ with step 5 was added to the two source images of each set, which were later provided to the mf-CRF method. For each σ_n, the accuracy of the predicted decision map was extracted from the comparison of the predicted decision map and the respective ground truth decision map. Moreover, in order to extract safe conclusions, the average accuracy over the 15 sets was computed and depicted in Fig. 12. The breakpoint of the mf-CRF is $\sigma_n = 20$. For $\sigma_n > 20$, the accuracy is reduced more significantly, which is probably due to the level of noise being greater than the activity levels of the input images.

A second experiment was conducted in order to evaluate all 12 compared fusion methods. All methods were tested against the aforementioned 15 sets created from the VOC-2012 segmentation competition [29]. Additive Gaussian noise with $\sigma_n = 20$ was added at the same spatial coordinates to the both input images and the ground truth image of each set. The pair of noisy input images was provided to all 12 compared methods and their fused results were compared to the noisy ground truth image. For evaluation purposes, PSNR and MSE were used to evaluate the performance of the compared fusion algorithms against Additive Gaussian Noise. Average PSNR values and average MSE values over the 15 sets were used to get more safe conclusion and are shown in Table III. A parenthesis that includes the number of times each method achieves the first (left) and second (right) best scores according to each metric is also shown. The proposed mf-CRF has the highest average PSNR value, and the highest PSNR score in all compared sets. Moreover, mf-CRF has the lowest average MSE value, and reaches the lowest MSE more frequently compared to the other methods. This experiment concludes that the proposed mf-CRF method is more robust against Additive Gaussian Noise than the other state-of-the-art multi-focus image fusion methods.

V. Computational efficiency

Table IV includes the average running time of the proposed and the compared methods. In this study, software implementations of the compared methods were provided by the original authors. All algorithms were executed in Matlab R 2017a on a Apple MacBook Pro with an 2,7 GHz Intel Core i5 processor and 8 GB RAM.

As shown in the table IV, the GF, QuadTree, BF, NSCT algorithms are the fastest multi-focus image fusion algorithms. The SR, NSCT-SR, ASR methods are the most computationally expensive, compared to the other algorithms, due to the very large and overcomplete dictionaries and the required forward/backward transform. The CNN method also has low computational efficiency compared to the rest of the compared spatial-domain methods. The MWGF algorithm also features higher execution time than the other spatial-domain methods, due to the EM algorithm used with the Laplacian mixture models. The proposed mf-CRF has average execution time compared to the methods of spatial domain, mainly due to the forward transform in ICA domain and SML estimation.

VI. Extension to other image fusion problems

In order to demonstrate the generalization of the mf-CRF algorithm to the visible-infrared fusion and the medical image fusion cases, examples are included in Fig. 13 and Fig. 14 respectively.

TABLE III

MSE OF METRICS FOR 15 SETS WITH ADDITIVE GAUSSIAN NOISE $\sigma_n = 20$.

Methods	PSNR	MSE
DSIFT	43.4738 (0,10)	0.7091 (1,10)
GF	41.2383 (0,0)	1.9121 (0,0)
MWGF	34.9378 (0,0)	5.6565 (0,0)
Quadtree	42.1046 (0,2)	1.0149 (0,3)
BF	36.9390 (0,0)	2.6380 (0,0)
IM	34.3958 (0,0)	3.6760 (0,0)
NSCT	41.3168 (0,0)	2.4063 (0,0)
NSCT-SR	41.3285 (0,2)	2.3944 (0,0)
SR	33.1029 (0,0)	11.5936 (0,0)
ASR	33.4684 (0,0)	11.4845 (0,0)
CNN	42.5759 (0,1)	1.3243 (0,1)
mf-CRF	**45.6510** (15,0)	**0.4487** (14,1)

TABLE IV

MEAN RUNNING TIME FOR LYTRO DATASET.

Method	DSIFT	GF	MWGF	QuadTree	BF	IM	NSCT	NSCT-SR	SR	ASR	CNN	mf-CRF
Time (s)	6.06	0.46	10.83	1.20	2.90	3.71	2.23	541.23	533.05	377.12	122.21	5.85

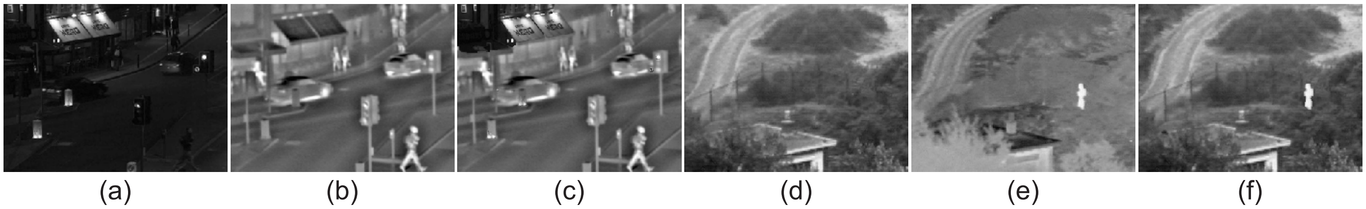


Fig. 13. Two applications of mf-CRF fusion in visible-infrared imaging. a) Source 1, b) Source 2, c) Fused, d) Source 1, e) Source 2, f) Fused

It can be seen in Fig. 13 that for both visible-infrared fusion cases, the mf-CRF can identify correctly the boundaries between the salient features of visible-infrared images. In order to preserve the best the salient features found in source images, binary image fusion is preferred. In the first visible-infrared example,

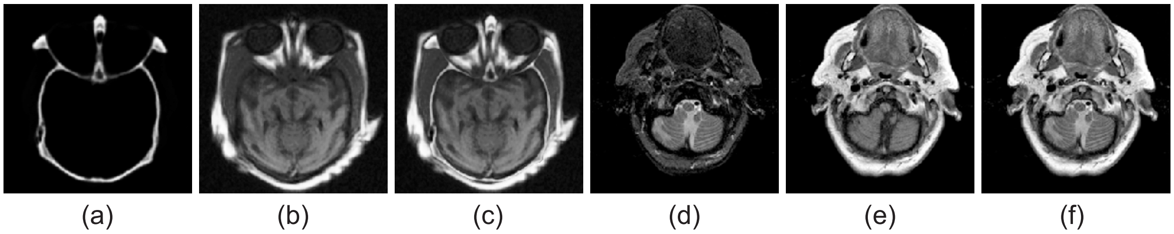


Fig. 14. Two applications of mf-CRF fusion in medical imaging. a) Source 1, b) Source 2, c) Fused, d) Source 1, e) Source 2, f) Fused

the fused image preserves all the information found in the infrared image and the shop inscription found in the visible source image. In the second example, the information found in the visible source image is retained in the fusion image, moreover, the human pose found in the infrared image is preserved during fusion. In both examples of visible-infrared fusion, seamless fusion results are achieved and the visual quality of the fused images is very satisfying.

For the case of medical image fusion, Miles et al. [37] proposed a spine image fusion via graph cuts. Fig. 14 includes two applications of the proposed mf-CRF model for the medical image fusion field. In both examples, the salient features of source images are well preserved in final fused images. Due to the applied smoothness prior, the fused images have high visual quality, with smooth transitions around the cut solution.

Considering these examples, the mf-CRF method can be used to fuse medical images and visible-infrared image pairs, where binary fusion is preferred over the weighted fusion with weights different to one. For both fusion applications of the mf-CRF model, the unary term that is used to compute the most salient features has been the same as the one described in section II-B, also the smoothness term has been the same as described in section II-C in order to achieve seamless fusion result. The mf-CRF model can also be applied to other image fusion applications that require the most salient features of input images to be preserved in a seamless fused image result.

VII. Conclusion

This paper mainly presents a novel CRF-based model, suitable for multi-focus image fusion. The main novelty of the proposed mf-CRF method is its capability to preserve advantages of both multi-spectral and spatial-domain methods, while producing fused images of high visual quality, due to the smoothness term which aligns the graph-cut solution to the focused and defocused pixels.

The main contributions can be summarized: 1) A novel CRF model is introduced to perform multi-focus image fusion. 2) the mf-CRF approach preserves all advantages of pixel-domain and multi-spectral domain fusion algorithms, while by solving the proposed energy minimization problem, the method achieves globally optimal solution, when two input images are provided. The presented experimental results demonstrate that the proposed mf-CRF model outperforms state-of-the-art image fusion algorithms. 3) Applications of mf-CRF model to visible-infrared image fusion and medical image fusion are also provided to exhibit the generalization capability of the method.

The proposed mf-CRF model is competitive against state of the art methods, however, future work towards a dense-CRF based model could lead to improved segmentation of the focused and defocused pixels, and also fused images of higher visual quality.

ACKNOWLEDGMENT

This work was supported by the Eugenides Foundation, Greece grant. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Information Fusion*, vol. 33, pp. 100 – 112, 2017.

[2] B. Aiazzi, L. Alparone, A. Barducci, S. Baronti, and I. Pippi, "Multispectral fusion of multisensor image data by the generalized laplacian pyramid," in *IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, 1999, pp. 1183–1185.

[3] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," *Information Fusion*, vol. 20, pp. 60–72, 2014.

[4] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Information Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[5] H. Li, B. S. Manjunath, and S. K. Mitra, "Multi-sensor image fusion using the wavelet transform," in *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, vol. 1. IEEE, 1994, pp. 51–55.

[6] Q. Zhang and B.-l. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Processing*, vol. 89, no. 7, pp. 1334–1346, 2009.

[7] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Information Fusion*, vol. 12, no. 2, pp. 74–84, 2011.

[8] Z. Liu, Y. Chai, H. Yin, J. Zhou, and Z. Zhu, "A novel multi-focus image fusion approach based on image decomposition," *Information Fusion*, vol. 35, pp. 102–116, 2017.

[9] B. Yang and S. Li, "Multifocus Image Fusion and Restoration With Sparse Representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 884–892, 2010.

[10] Q. Zhang and M. D. Levine, "Robust Multi-Focus Image Fusion Using Multi-Task Sparse Representation and Spatial Context," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2045–2058, 2016.

[11] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Processing*, vol. 9, no. 5, pp. 347–357, 2015.

[12] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, “Deep learning for pixel-level image fusion: Recent advances and future prospects,” *Information Fusion*, vol. 42, pp. 158 – 173, 2018.

[13] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, “Image fusion with convolutional sparse representation,” *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882–1886, 2016.

[14] Y. Liu, X. Chen, H. Peng, and Z. Wang, “Multi-focus image fusion with a deep convolutional neural network,” *Information Fusion*, vol. 36, pp. 191–207, jul 2017.

[15] Y. Liu, S. Liu, and Z. Wang, “A general framework for image fusion based on multi-scale transform and sparse representation,” *Information Fusion*, vol. 24, pp. 147–164, 2015.

[16] M. Nejati, S. Samavi, and S. Shirani, “Multi-focus image fusion using dictionary-based sparse representation,” *Information Fusion*, vol. 25, pp. 72–84, 2015.

[17] X. Bai, Y. Zhang, F. Zhou, and B. Xue, “Quadtree-based multi-focus image fusion using a weighted focus-measure,” *Information Fusion*, vol. 22, pp. 105–118, 2015.

[18] S. Li, X. Kang, J. Hu, and B. Yang, “Image matting for fusion of multi-focus images in dynamic scenes,” *Information Fusion*, vol. 14, no. 2, pp. 147–162, 2013.

[19] Y. Liu, S. Liu, and Z. Wang, “Multi-focus image fusion with dense SIFT,” *Information Fusion*, vol. 23, pp. 139–155, 2015.

[20] P. S. Chavez Jr and A. Y. Kwarteng, “Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis,” *Photogrammetric Engineering and Remote Sensing*, vol. 55, no. 3, pp. 339–348, 1989.

[21] 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.

[22] J. Sun, H. Zhu, Z. Xu, and C. Han, “Poisson image fusion based on Markov random field fusion model,” *Information fusion*, vol. 14, no. 3, pp. 241–2535, 2013.

[23] S. Li and B. Yang, “Hybrid multiresolution method for multisensor multimodal image fusion,” *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1519–1526, 2010.

[24] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[25] S. K. Nayar and Y. Nakagawa, “Shape from focus,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 8, pp. 824–831, 1994.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*, 2006.

[27] A. Hyvärinen, J. Hurri, and P. O. Hoyer, “Independent Component Analysis BT - Natural Image Statistics: A Probabilistic Approach to Early Computational Vision,” A. Hyvärinen, J. Hurri, and P. O. Hoyer, Eds. London: Springer London, 2009, pp. 151–175.

[28] Y. Zhang, X. Bai, and T. Wang, “Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure,” *Information Fusion*, vol. 35, pp. 81–2535, 2017.

[29] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[30] S. Li, X. Kang, and J. Hu, “Image fusion with guided filtering,” *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[31] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, “Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Study,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 94–109, 2012.

[32] M. Hossny, S. Nahavandi, and D. Creighton, “Comments on ’Information measure for performance of image fusion’,” *Electronics Letters*, vol. 44, no. 18, pp. 1066–1067, 2008.

[33] C. S. Xydeas and V. Petrovic, “Objective image fusion performance measure,” *Electronics Letters*, vol. 36, no. 4, pp. 308–309, 2000.

[34] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, “A novel similarity based quality metric for image fusion,” *Information Fusion*, vol. 9, no. 2, pp. 156–160, 2008.

[35] Y. Chen and R. S. Blum, “A new automated quality assessment algorithm for image fusion,” *Image and Vision Computing*, vol. 27, no. 10, pp. 1421–1432, sep 2009.

[36] S. Chaudhuri and A. N. Rajagopalan, *Depth from defocus: a real aperture imaging approach*. Springer Science & Business Media, 2012.

[37] B. Miles, I. B. Ayed, M. W. K. Law, G. Garvin, A. Fenster, and S. Li, “Spine image fusion via graph cuts,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1841–1850, 2013.

Odysseas Bouzos received the Diploma Degree from the Department of Electrical and Computer Engineering, Democritus University of Thrace (DUTH), Greece, in 2013. He also received an MSc in High Dynamic Range image fusion from the same department in 2014. He is currently pursuing a PhD on Machine Learning techniques for image fusion at the same department. His research interests include Probabilistic Graphical Models, Deep Learning, Machine Learning, Computer Vision and Image Fusion. From 2015 until 2016 he was an intern at Centre for Robotics, MINES ParisTech, PSL Research University, Paris, France. From 2009 until 2010 he worked as a user of the ATLAS experiment at CERN - the European Organisation for Nuclear Research, .

Ioannis Andreadis received the Diploma Degree from the Department of Electrical and Computer Engineering, Democritus University of Thrace (DUTH), Greece, in 1983 [1978-1983] IKY Scholarship] and the M.Sc. [Electrical Power Systems Engineering] and Ph.D. [Machine Vision] Degrees from the University of Manchester Institute of Science and Technology, UK, in 1985 and 1989, respectively. He was awarded the IET Image Processing Premium in 2009. He received the best paper award (Computer Vision and Applications) in PSIVT 2007 as well the best paper award in EUREKA 2009. He is author of an Institute of Physics (IOP) Select Paper in 2010. He has supervised 13 PhD Theses, 20 MSc Theses and 70 Diploma dissertations. Professor Andreadis was a member of the Board of Governors of the European Commission Joint Research Center (JRC) from 2008-2010. He is a Fellow of the Institution of Engineering and Technology (2006), (IET - London, UK).

Nikolaos Mitianoudis (S'98 - M'04 - SM'11) received the diploma in Electronic and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 1998. He received the 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. Between 2003 and 2009, he was a Research Associate at Imperial College London, UK working on the Data Information Fusion-Defense Technology Centre project ”Applied Multi-Dimensional Fusion”, sponsored by General Dynamics UK and QinetiQ. From 2009 until 2010, he was an Academic Assistant at the International Hellenic University. Currently, he is an Assistant Professor in Audio and Image Processing at the Democritus University of Thrace, Greece. He is an Associate Editor of the IEEE Trans. on Image Processing (2018-2021). His research interests include Independent Component Analysis, Image Fusion, Computer Vision, Deep Learning, Music Information Retrieval and Blind Source Separation/Extraction.