Engineering Science and Technology, an International Journal 29 (2022) 101037

HOSTED BY

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

Visible and infrared image fusion using an efficient adaptive transition region extraction technique

Bikash Meher [a], Sanjay Agrawal [a], Rutuparna Panda [a,*], Lingraj Dora [b], Ajith Abraham [c]

[a] Department of Electronics & Telecommunication Engineering, VSS University of Technology, Burla, India
[b] Department of Electrical and Electronics Engineering, VSS University of Technology, Burla, India
[c] Machine Intelligence Research Labs, Washington, USA

ARTICLE INFO

Article history:
Received 4 January 2021
Revised 4 June 2021
Accepted 29 June 2021
Available online 10 July 2021

Keywords:
Image fusion
Infrared image
Visible image
Object region extraction
Transition region extraction

ABSTRACT

In order to track a targeted environment, concealed weapon detection, navigation and military require various imaging modalities, for instance, visible image (VI) and infrared (IR) image. These modalities provide additional details. Complementary information from these images need to be fused into a single image for improved situational awareness. Hence, an ideal fused image should assimilate the essential bright information from the IR image and retain much of the original visual information from the VI. To achieve this, a region based image fusion technique using an efficient adaptive transition region extraction (ATRE) strategy is suggested in this paper. For the first time, the transition region extraction based approach is brought into the context of visible and infrared image fusion. This method is beneficial because it overcomes the problems of noise sensitivity, poor contrast and blurring effects associated with the conventional pixel-based methods. The proposed ATRE technique is used to efficiently extract the bright object regions from the IR image and retain much of the visual background regions from the VI. An adaptive parameter is introduced for accurate segmentation. A region mapping process is followed to get the fused image. Our technique is tested on standard fusion datasets. Image inspection and objective fusion indices are utilized to validate the results. They are compared with conventional and current pixel based and region based fusion techniques. The outcomes reveal that the suggested technique is comparable or better than state-of-the-art fusion techniques.
© 2021 Karabuk University. Publishing services by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Image fusion is a promising research subject in the field of image processing. It is a technique for combining reciprocal and redundant details from multiple images, either of the same view or of a different modality, into a single image. The fused image obtained may yield an explicit visual perception and applied in advanced image processing applications. With the invention of advanced imaging devices for capturing images, many researchers are attracted and applied the image fusion techniques to many applications i.e. surveillance, disease diagnosis, remote sensing etc. More specifically, the IR image and VI fusion techniques are extensively utilized in many applications such as military surveillance, object recognition, detection, image enhancement, remote sensing etc. It is especially important in military technology for automatic target detection and localization. The sensors used in the VI capture reflected lights from the object with rich appearance information. However, the images captured by the visible sensors are influenced by many impairments such as bad weather condition, poor illumination, fog and night time. On the other hand, the IR sensors capture images using the principle of thermal radiation. IR images are unaffected by the above mentioned disturbances. Instead, they have low resolution and poor details. Thus, a good image is obtained by combining the complementary information of both the IR and visible images using various image fusion techniques.

The image fusion is performed in three ways i.e. pixel level, feature level and decision level. Numerous studies are reported using pixel level image fusion techniques [1–7]. The pixel level image fusion techniques are simple and easy to implement. However, they have several shortcomings such as misregistration, blurring effect etc. These shortcomings can be eliminated by the use of region based approaches which belong to feature level image fusion. In region based approach, the regions (i.e. group of correlated pixels) are considered for fusion instead of individual pixels. The decision level image fusion is the highest level of fusion. This

* Corresponding author.
E-mail address: rpanda_etc@vssut.ac.in (R. Panda).
Peer review under responsibility of Karabuk University.

https://doi.org/10.1016/j.jestch.2021.06.017
2215-0986/© 2021 Karabuk University. Publishing services by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

method is based on the outputs of initial object detection and classification. Generally, an initial decision is made from the feature level fusion. The outcome of the initial decision is taken as the input for the decision level fusion. Many image fusion processes have incorporated region based approach due to its advantages over pixel based approach. A thorough study on region based image fusion techniques is reported in Meher et al. [8]. Further, a detail review on IR and VI fusion is undertaken in [9,10]. In this paper, we have developed a region based fusion technique for IR and VI fusion and compared our results with various pixel based and region based fusion techniques.

In region based fusion techniques, the segmentation task is vital. The researchers have used various techniques for segmentation. In recent years, transition region based thresholding has been successfully investigated for image binarization. The conventional gradient based transition region extraction techniques are highly affected by noise. The transition region descriptor (a key component in the process) affects the region extraction and then thresholding. The local entropy (LE) (a descriptor) [11] considers only the frequency of gray level variations. This results in inaccurate classification of non-transition regions with frequent but minor gray level variations into the transition regions. To overcome this problem, a modified descriptor (modified local entropy (MLE)) considering both the frequency and the degree of gray level variations is implemented in Li et al. [12]. This concept of image segmentation has not been used for fusion before.

This has motivated us to develop a region based IR and VI fusion scheme using the idea of transition region extraction. The MLE concept of image binarization is used for the object region extraction. In this context, we suggest a novel efficient adaptive transition region extraction (ATRE) method for fusion. It is to be noted that the efficiency of the proposed method does not mean fusion time efficiency. Here, we have extracted the object region from the IR image introducing the ATRE based segmentation approach. The ATRE approach is used to determine the threshold values. In Li et al. [12], the threshold values are determined from a randomly chosen coefficient in the range [0,1]. However, we have suggested an adaptive coefficient to determine the threshold value which plays a key role in segmentation. The coefficient is expressed in terms of the maximum and the average value of the transition region descriptor matrix. This results in an accurate delineation of the transition region leading to exact object region and background region extraction for fusion. Then the background regions from both the VI and IR image is extracted using the inverted binarized image obtained from ATRE. Next a suitable patch based fusion rule is applied to get the fused background image. Finally, the output image is obtained by using region mapping. The proposed method is experimented with a number of test images from standard database [13]. The results are compared with state-of-the-art traditional and modern pixel level and region level fusion techniques. It is observed that our results are encouraging and may set the path for future research in this area. The major contributions of this work are: i) to use the idea of transition region extraction for object region delineation for fusion, ii) an adaptive coefficient is introduced to determine the threshold values for accurate segmentation.

The rest of this paper is structured as follows: The related work is discussed in Section 2. The suggested work is described in Section 3 including the object region extraction and fusion rules. The result comparisons and discussions are given in Section 4. Lastly, the conclusion is given in Section 5.

2. Related work

Many researchers have presented the fusion processes for VI and IR images. Most of the methods used are pixel based. The commonly used pixel based methods for VI and IR image fusion includes – multi-scale transform (MST), saliency based methods, sparse representation, neural network, subspace, hybrid models etc. [10].

The authors in Liu et al. [14] suggested the fusion of VI and IR images based on discrete wavelet transform (DWT). The DWT provides a good time–frequency representation compared to the pyramid transform. However, the DWT produces pseudo-Gibbs effect because of the down sampling process at every decomposition level. Further, it lacks shift invariance and directionality properties. This is overcome by the use of dual tree complex wavelet transform (DT-CWT). Lewis et al. [15] employed the DT-CWT to get a multi resolution decomposition of the source images. However, this transform cannot detect the curves and edges of the images in fusion. The curvelet transform (CVT) is used to solve the problem of DT-CWT method. Quan et al. [16] used CVT for the decomposition of the VI and IR image and obtained two groups of coefficients i.e. high frequency and low frequency components. However, the visual quality of the fused image is degraded due to the shift variant character of the curvelets. Naidu [17] used multiscale singular value decomposition (MSVD) for the fusion of VI and IR image, which does not have fixed set of basis vectors. Bavirisetti and Dhuli [18] proposed a new edge preserving image fusion technique for VI and IR images. The input images are decomposed into two layers: approximation and detail layers using anisotropic diffusion based fusion (ADF). However, this method suffers from blocky effects or artefacts. The authors in Bavirisetti et al. [19] suggested a fusion scheme based on the fourth order partial differential equation (FPDE) and principal component analysis (PCA). It is observed that the MST based fusion methods suffer from the following difficulties. The determination of number of decomposition levels is difficult. The size of the decomposition levels is a compromise between getting the spatial details and sensitivity to noise and transform artefacts. Further, the problems of choosing the MST and the predefined fusion rules are always there.

The researchers have also used saliency based methods for the same. In Bavirisetti and Dhuli [20], the authors suggested a two-scale image fusion based on visual saliency (TSIFVS) algorithm for the fusion of IR and visible images. However, the dimension of the mean and median filters used to find the visual saliency affects the performance of their method. In Zhan et al. [21], the authors suggested multimodal image seamless fusion (MISF) for fusion. However, their method produces gradient reversal artefacts.

In Jin et al. [22], the authors proposed a VI and IR image fusion using a pixel based hybrid technique stationary wavelet transform-discrete cosine transform-spatial frequency (SWT-DCT-SF). They used discrete cosine transform (DCT) and local spatial frequency (LSF) in discrete stationary wavelet transform (SWT) domain. The selection of window size is also very important for DCT and may affect the quality of the output image. In Ma et al. [23], the authors proposed a procedure utilizing gradient transfer fusion (GTF) and total variation minimization. However, the authors did not take into account the intensity information of the VI, which could lead to a low dynamic range and detail loss.

In Zhang et al. [24], the authors used pixel based infrared feature extraction and visual information preservation (IFEVIP) method to extract the important bright features from the IR image and merged with the original visual features of VI. However, their algorithm is suited to low light IR image and VI pairs only. Further, the contrast of the fused image is degraded as the IR features are extracted with much background information. It is to be noted that all the above mentioned techniques are based on pixel level fusion.

Mitianoudis and Stathaki [25] proposed the region based image fusion using ICA. The source images are segmented into two regions: (i) active and (ii) non-active. The active regions contain

the detail information and the non-active regions contain the background information. The authors in [25] extended their work to a more sophisticated region-based image fusion [26]. The method may not be suitable for multi-modality images as the several modality images have diverse texture characteristics. We have also suggested a region based method for fusion using the idea of transition region extraction for object region delineation.

Recently, deep learning methods have been widely used in the direction of visible and infrared image fusion [27–32]. The authors in [27] used deep learning for IR and VI fusion. They used deep learning for feature extraction from the detail part of the images. Ma et al. [28] proposed FusionGAN for the problem on hand. They further improvised their work in [29]. The authors in [30] proposed a fusion method via dual-discriminator conditional generative adversarial network (DDcGAN) improve the losses in the thermal and visible images. Li and Wu [31] proposed a novel deep learning architecture, which is constructed by encoding network and decoding network. The encoding network extracts the features and the decoding network gives the fused image. Zhao et al. [32] proposed a novel auto-encoder (AE) based network for the fusion of VI and IR image. It is observed that most of the methods used deep learning for feature extraction and classification. It is known that a large number of input data is required for training deep networks. Hence, the authors used patches and the deep networks to generate the large number of inputs. A comparison of the proposed method and the deep learning based methods is shown later in the results section.

3. Proposed fusion scheme

The VI and the IR image contain different image features. The VI sensor detects the textural information. The IR sensor detects objects that are not perceptible to the human eye. The VI contains clear background region. The IR image contains clear object region.

Figure 1 shows the schematic block diagram of the suggested technique. The object region is extracted from the IR image using the proposed ATRE technique. Then a segmentation threshold is found using the proposed method. Next, the IR image is binarized using this threshold to get the object region. Next, the binarized image is inverted to get the background regions from both the IR image and the VI. Both the regions are decomposed into patches. Then the fused background region is found by using a suitable fusion rule. Finally, it is mapped to the object region of IR image to obtain the output image. In this paper, the input images are assumed to be registered.

3.1. Object and background region extraction using ATRE

The transition region is present between the object and the background in an image. It is characterized on the basis of region, boundary and variations of gray levels. The change of gray level plays an important role in the determination of the transition region. The gray level of pixels changes frequently and intensively in the transition region, which brings rich information for its description. To get the transition region accurately, both the occurrence and degree of gray level variations are required. To find the gray level variations, a transition region descriptor (i.e. LE) is suggested in Yan et al. [11]. To explain the concept of LE, let I represent an image having L intensity levels $[0, 1, ..., L-1]$, of dimension $M \times N$. Let n_i denotes the count of pixels having intensity level i. Assume $U = \{(m,n) : m = 1, 2, ..., M; n = 1, 2, ..., N\}$ represent the size of the image. Let $f(m,n)$ be the gray level of a pixel at(m,n). The entropy (E) of an image is stated as [33,34]

$$E = -\sum_{i=0}^{L-1} P_i \log P_i \tag{1}$$

where $P_i = n_i/(M \times N)$ represent the probability of occurrence of gray level i in the image. Now the local entropy $(LE(W_k))$ of a pixel can be defined by taking a neighbourhood W_k of window size $u \times v$ within the image and is expressed as

$$LE(W_k) = -\sum_{r=0}^{L-1} P_r \log P_r \tag{2}$$

where $P_r = n_r/(u \times v)$ denotes the probability of occurrence of gray level r existing in the neighbourhood W_k, n_r represents the number of pixels with gray level r in the neighbourhood. It is studied from the literature that this process is computationally intensive as it computes every gray level' s probability of occurrence in its neighbourhood. In order to decrease the computational intensity, a similar parameter, local complexity (LC) as in [35], is utilized to represent the frequency of gray level variations as,

$$LC(m,n) = C(W_k) = \sum_{r=0}^{L-1} sgn(r) \tag{3}$$

where $sgn(r) = \begin{cases} 1, & \text{if } \exists f(x,y) = r \\ 0, & \text{else} \end{cases}$ and (x,y) denotes the pixel location in the neighbourhood W_k.

It is to be noted that both LE and LC considers only the occurrence of intensity level variations. It does not consider the degree of the changes. Hence, local variance (LV) is employed to define the degree gray level variations and for a neighbourhoodW_k, it can be expressed as

$$LV(m,n) = \sigma^2(W_k) = \frac{1}{u \times v - 1} \sum_{x=1}^{u} \sum_{y=1}^{v} \left(f(x,y) - \bar{f} \right)^2 \tag{4}$$

Note that $\bar{f}$ is the mean intensity level ofW_k. Accordingly, when the neighborhood window moves within the image (pixel by pixel) we get each pixel' s LC and LV matrices. To find the gray level variations more accurately, a new transition region descriptor is suggested by using the normalized local complexity (NLC) and normalized local variance (NLV) computed as [12],

$$NLC(m,n) = \frac{LC(m,n) - \min\limits_{\forall(x,y)} LC(x,y)}{\max\limits_{\forall(x,y)} LC(x,y) - \min\limits_{\forall(x,y)} LC(x,y)} \tag{5}$$

$$NLV(m,n) = \frac{LV(m,n) - \min\limits_{\forall(x,y)} LV(x,y)}{\max\limits_{\forall(x,y)} LV(x,y) - \min\limits_{\forall(x,y)} LV(x,y)} \tag{6}$$

The new transition region descriptor (S) is thus formed using both the normalized factors as

$$S(m,n) = \beta \times NLC(m,n) + (1-\beta) \times NLV(m,n) \tag{7}$$

where $\beta$is a weight factor which balances the contributions of NLV and NLC. From the equation, it is obvious that the transition region descriptor is equal to NLC when $\beta = 1$ and is equal to NLV when $\beta = 0$. Hence, the value of β should be in the range [0, 1]. The $S(m,n)$ value obtained from Eqn. (7) for each pixel at location (m,n) will form an image matrix S. The transition region pixels have higher $S(m,n)$ values as compared to the non-transition region pixels. The threshold S_T for the transition region extraction is obtained as:

$$S_T = \gamma \times S_{max} \tag{8}$$

where $S_{\max} = \max\limits_{\forall(m,n)} S(m,n)$ and γ is a random coefficient between [0,1]. The value of γ plays an important role in determining the

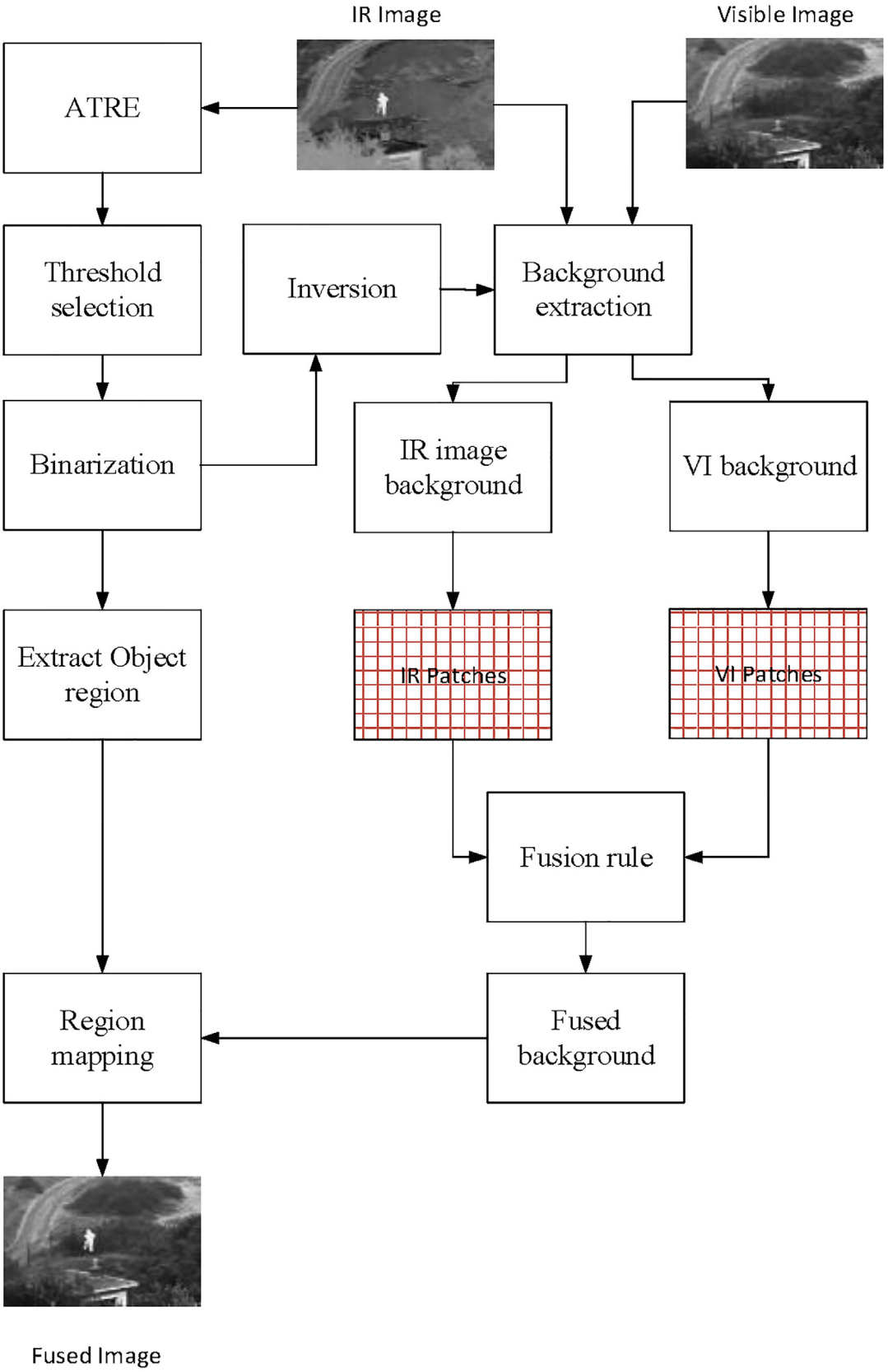


Fig. 1. Block diagram of the suggested fusion scheme.

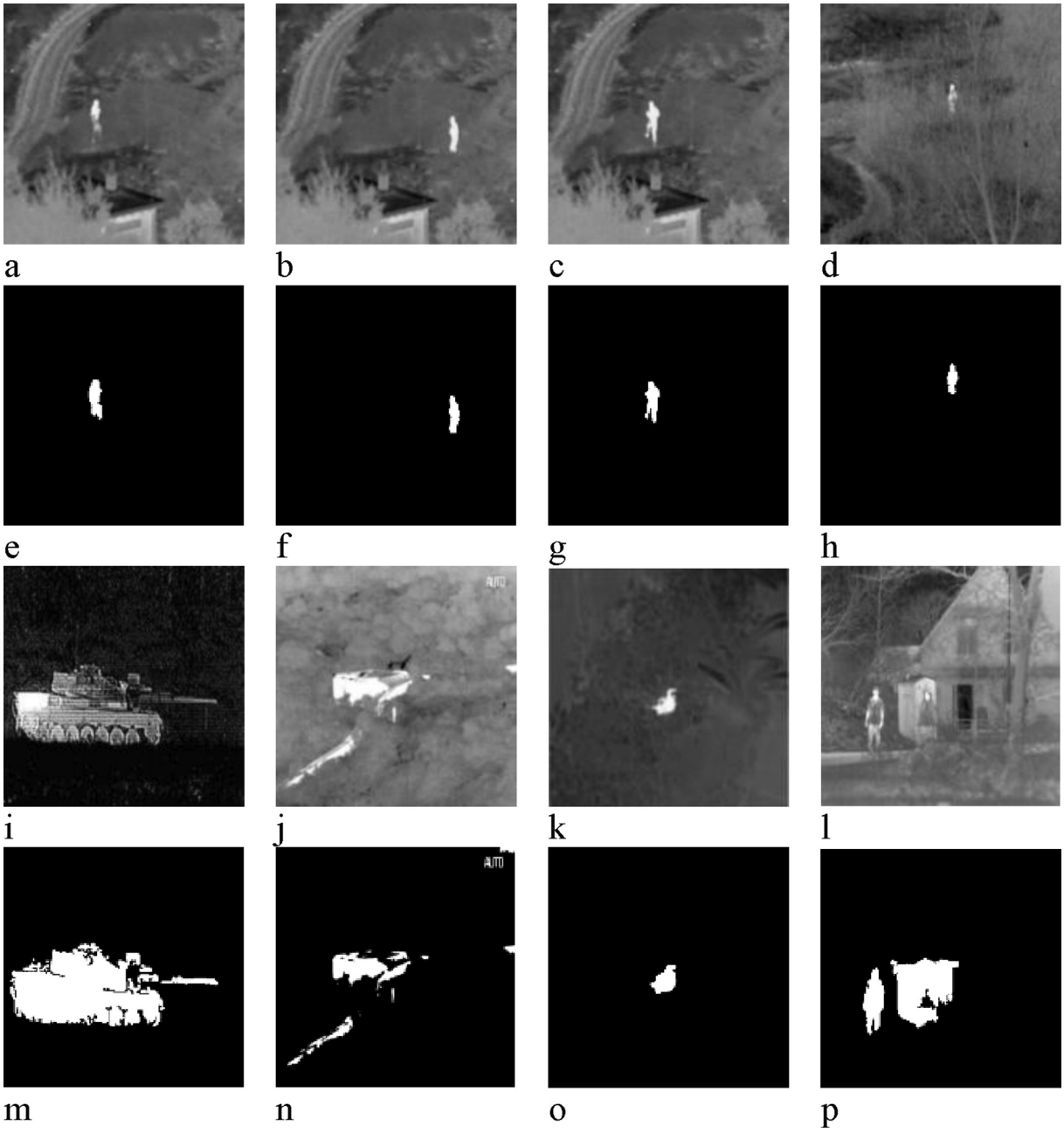


Fig. 2. Examples of object region extraction from different IR images.(a-d, i-l) input IR images, (e-h, m-p) extracted object regions.

threshold and hence the transition region. Instead of taking it a random value, we have computed the threshold S_T as

$$S_T = \gamma_a \times S_{max} \tag{9}$$

where $\gamma_a = (S_{max} - S_{mean})/(S_{max} + S_{mean})$is an adaptive parameter introduced, depending on the type of images. Note that $S_{mean} = \underset{\forall(m,n)}{mean}\ S(m,n)$. Then the transition region is extracted as follows:

$$TR(m,n) = \begin{cases} 1 \text{ if } S(m,n) \geqslant S_T \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

The final segmentation threshold $T_f$is then calculated as the average of the gray levels in the transition region using the formula as given below:

$$T_f = (\sum_m \sum_n TR(m,n) \times f(m,n))/\sum_m \sum_n TR(m,n) \tag{11}$$

Table 1
Objective performance metrics comparison for '*Bunker*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.3198	6.7134	2.1252	0.8873	0.2085
	DT-CWT	0.6359	7.0776	2.1233	0.9051	0.2230
	CVT	0.6044	7.0937	2.1180	0.8987	0.2335
	MSVD	0.3896	6.7381	2.1093	0.8924	0.2181
	ADF	0.5668	6.8880	2.1068	0.8359	0.1781
	FPDE	0.5040	6.8197	2.1200	0.8520	0.1761
Saliency based	TSIFVS	0.5843	7.2115	2.1339	0.8839	0.3446
	MISF	0.7123	7.5056	**2.7169**	**0.9162**	0.2464
Subspace based	ICA-Region	0.3286	6.6782	2.2191	0.8934	0.2039
	ICA-Textr-std	0.5722	6.5874	2.2047	0.8890	0.2525
Hybrid method	SWT-DCT-SF	0.6274	7.2127	2.2717	0.9046	0.2427
Other methods	GTF	0.5962	6.9413	2.0655	0.8897	0.1774
	IFEVIP	0.6291	7.0707	2.3282	0.8526	0.2835
ATRE	Proposed	**0.7250**	**7.5108**	2.7045	0.9095	**0.3607**

Table 2
Objective performance metrics comparison for '*Tank*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.2075	7.2756	2.1852	0.7953	0.1930
	DT-CWT	0.4910	7.4132	2.1587	0.8183	0.1856
	CVT	0.4445	7.4097	2.1515	0.8087	0.1878
	MSVD	0.1710	7.2433	2.1728	0.7985	0.1851
	ADF	0.3116	7.3459	2.1267	0.8216	0.1546
	FPDE	0.4396	7.2984	2.1145	0.7667	0.1227
Saliency based	TSIFVS	0.6193	7.4686	2.1692	0.7993	0.2360
	MISF	0.6521	7.9576	**2.5816**	**0.8320**	0.0516
Subspace based	ICA-Region	0.3820	7.2294	2.2327	0.8111	0.1905
	ICA-Textr-std	0.6515	7.4549	2.1629	0.8104	0.2395
Hybrid method	SWT-DCT-SF	0.5070	7.9495	2.2806	0.8192	0.1625
Other methods	GTF	0.5419	6.3637	2.2139	0.7827	0.0993
	IFEVIP	0.6625	7.8245	2.4755	0.8011	0.2124
ATRE	Proposed	**0.7577**	**7.9583**	2.5578	0.8298	**0.2856**

Table 3
Objective performance metrics comparison for '*Nato_camp1*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.3541	6.2599	2.1189	0.8695	0.3149
	DT-CWT	0.4683	6.4918	2.1127	0.8959	0.3309
	CVT	0.4251	6.5466	2.1074	0.8860	0.3528
	MSVD	0.3518	6.2689	2.1174	0.8664	0.3179
	ADF	0.4643	6.2759	2.1226	0.8688	0.2678
	FPDE	0.3972	6.3237	2.1096	0.8621	0.2599
Saliency based	TSIFVS	0.4626	6.6400	2.1152	0.8760	**0.4613**
	MISF	0.5490	6.8869	2.2500	0.9016	0.2635
Subspace based	ICA-Region	0.5020	6.3172	2.2642	0.8897	0.3515
	ICA-Textr-std	0.5493	6.3913	2.2646	0.8946	0.3823
Hybrid method	SWT-DCT-SF	0.4532	6.7909	2.1732	0.8823	0.3779
Other methods	GTF	0.5817	6.6379	2.0625	0.8812	0.2393
	IFEVIP	0.4637	6.7728	2.2187	0.8806	0.4049
ATRE	Proposed	**0.5979**	**7.1797**	**3.0281**	**0.9115**	0.2522

The image is binarized using T_f to extract the object region from the different IR images. The binarized image is then inverted to extract the background regions from both the IR image and the VI. The extracted object region for different IR images is illustrated in Fig. 2. It is observed that the proposed approach accurately extracts the object regions from the different IR images. Because of the difference between the intensity levels of the object and the background region, the transition region extraction technique successfully delineates the object region from the background.

3.2. Fusion rule

The background regions from both the IR image and the VI are fused using a patch based fusion rule. The regions from both the IR image and VI are decomposed into patches. The energy of each patch is calculated and compared to get the fused background region. Let the background region of the IR image be denoted by I_{IR_B}, the background region of the VI be denoted by I_{VI_B}, the fused background region be denoted as I_{F_B} and the fused image by I_F. To

Table 4
Objective performance metrics comparison for '*Nato_camp2*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.3500	6.2629	2.1226	0.8771	0.3129
	DT-CWT	0.4679	6.5157	2.1075	0.8796	0.3301
	CVT	0.4275	6.5607	2.1019	0.8763	0.3513
	MSVD	0.3460	6.2689	2.1209	0.8745	0.3148
	ADF	0.4624	6.2845	2.1244	0.8735	0.2698
	FPDE	0.3857	6.3495	2.1105	0.8624	0.2507
Saliency based	TSIFVS	0.4656	6.6310	2.1172	0.8848	**0.4601**
	MISF	**0.5515**	6.9891	2.2590	0.9080	0.2908
Subspace based	ICA-Region	0.5037	6.3247	2.2651	0.8984	0.3529
	ICA-Textr-std	0.5503	6.3975	2.2690	0.8997	0.3854
Hybrid method	SWT-DCT-SF	0.4583	6.7993	2.1758	0.8887	0.3820
Other methods	GTF	0.3960	6.6940	2.3428	0.8853	0.2472
	IFEVIP	0.4633	6.7753	2.2067	0.8906	0.4099
ATRE	Proposed	0.5210	**7.1473**	**2.8637**	**0.9137**	0.3681

Table 5
Objective performance metrics comparison for '*Sandpath*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.3043	6.1005	2.0953	0.8441	0.2754
	DT-CWT	0.5532	6.4813	2.0974	0.8792	0.2325
	CVT	0.4993	6.5302	2.0950	0.8677	0.2476
	MSVD	0.2961	6.1019	2.0941	0.8428	0.2711
	ADF	0.5562	6.3294	2.0966	0.8446	0.1942
	FPDE	0.5110	6.2463	2.0955	0.8369	0.1968
Saliency based	TSIFVS	0.4813	6.6249	2.1007	0.8482	0.3727
	MISF	0.6563	7.1347	2.8492	**0.9006**	0.2554
Subspace based	ICA-Region	0.3339	6.1036	2.1555	0.8620	0.2876
	ICA-Textr-std	0.4324	6.2324	2.1859	0.8645	0.3187
Hybrid method	SWT-DCT-SF	0.4915	6.7592	2.1494	0.8697	0.2808
Other methods	GTF	0.5273	6.5381	2.0214	0.8511	0.1788
	IFEVIP	0.4194	6.6274	2.1972	0.8561	0.2318
ATRE	Proposed	**0.6595**	**7.1454**	**2.9105**	0.8954	**0.3805**

Table 6
Objective performance metrics comparison for '*Gun*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.4430	6.5648	2.2364	0.9231	0.3072
	DT-CWT	0.6753	6.7990	2.2104	0.9372	0.6087
	CVT	0.6455	6.8250	2.2014	0.9309	0.5898
	MSVD	0.4621	6.5736	2.2641	0.9189	0.3682
	ADF	0.5929	6.5587	2.2500	0.9319	0.3448
	FPDE	0.5627	6.5847	2.2617	0.9009	0.3932
Saliency based	TSIFVS	0.6623	6.9472	2.2098	0.9181	0.7670
	MISF	0.7363	7.1166	2.5836	**0.9402**	0.8357
Subspace based	ICA-Region	0.5864	6.7816	2.1780	0.9321	0.4721
	ICA-Textr-std	0.6652	6.8420	2.1916	0.9315	0.5927
Hybrid method	SWT-DCT-SF	0.6745	7.0751	2.3695	0.9364	0.7071
Other methods	GTF	0.6025	6.2678	2.1397	0.9221	0.4400
	IFEVIP	0.6175	7.0818	2.4475	0.9307	0.6623
ATRE	Proposed	**0.7369**	**7.1479**	**2.8871**	0.9377	**0.8443**

obtain the fused background region,I_{IR_B} and$I_{VI_B}$are partitioned into patches. The energy of each patch in both the regions is calculated as:

$$E_{IR_B}(P_{IR_{B_i}}) = \sum_{(u,v)\in w} P^2_{IR_{B_i}} \tag{12}$$

$$E_{VI_B}(P_{VI_{B_i}}) = \sum_{(u,v)\in w} P^2_{VI_{B_i}} \tag{13}$$

where P_i is the i^{th} patch of the image, $E_{IR_B}$is the energy of the background image patch from the IR image, $E_{VI_B}$is the energy of the background image patch from the VI. Here we have taken the patch size of $w = 3 \times 3$. It is to be noted that patch size of 5×5, 7×7 can also be used. However, larger patch size may introduce blocking artefacts. It is noteworthy to mention here that the local energy of the infrared object is significantly higher than other areas. The fused background image $I_{F_B}$is obtained by utilizing the maximum rule based on the comparison of $E_{IR_B}$and E_{VI_B}as illustrated below

Table 7
Objective performance metrics comparison for '*Tree*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.3955	5.6953	2.0064	0.8476	0.3878
	DT-CWT	0.4997	5.8111	2.0286	0.8720	0.2878
	CVT	0.4659	5.8434	2.0293	0.8650	0.3154
	MSVD	0.3799	5.6977	2.1410	0.8441	0.2644
	ADF	0.4036	5.6803	2.0065	0.8513	0.3650
	FPDE	0.5160	5.7391	2.1364	0.8438	0.2029
Saliency based	TSIFVS	0.5079	5.9292	2.1389	0.8585	0.4199
	MISF	0.5501	6.3664	2.3297	0.8828	0.5058
Subspace based	ICA-Region	0.4472	6.2988	2.0207	0.8606	0.5421
	ICA-Textr-std	0.4637	6.3655	2.0190	0.8629	**0.6000**
Hybrid method	SWT-DCT-SF	0.4669	6.2062	2.2101	0.8622	0.3044
Other methods	GTF	0.4616	5.6830	2.1862	0.8657	0.2689
	IFEVIP	0.5734	6.1491	**2.8952**	0.8763	0.1865
ATRE	Proposed	**0.5891**	**6.7818**	2.6045	**0.8831**	0.2527

Table 8
Objective performance metrics comparison for '*Two men in front of house*' image.

Domain	Method	$Q^{PQ/F}$	E	MI	FMI	VIFF
MST based	DWT	0.4844	6.4563	2.1209	0.8801	0.2751
	DT-CWT	0.5386	6.7724	2.1311	0.9032	0.2755
	CVT	0.4999	6.7964	2.1248	0.8953	0.2945
	MSVD	0.3434	6.4590	2.1492	0.8766	0.2620
	ADF	0.5394	6.4579	2.1373	0.8873	0.2710
	FPDE	0.4585	6.5006	2.1440	0.8497	0.2258
Saliency based	TSIFVS	0.5071	6.9046	2.1295	0.8835	0.2686
	MISF	0.5975	7.1159	2.3507	0.9084	0.2180
Subspace based	ICA-Region	0.5156	6.4509	2.1142	0.8913	0.2998
	ICA-Textr-std	**0.6342**	6.5813	2.1170	0.8821	0.3595
Hybrid method	SWT-DCT-SF	0.5261	6.9326	2.2367	0.8989	0.2763
Other methods	GTF	0.3049	7.0580	2.2852	0.8805	0.1756
	IFEVIP	0.4892	6.6723	2.2824	0.8678	0.3363
ATRE	Proposed	0.4694	**7.1522**	**2.6588**	**0.9089**	**0.3969**

$$I_{F_B} = \begin{cases} I_{IR_B}(P_i) \text{ if } E_{IR_B}(P_i) > E_{VI_B}(P_i) \\ I_{VI_B}(P_i) \text{ Otherwise} \end{cases} \quad (14)$$

At last, the fused image $I_F$is obtained using the region mapping process between the object region of the IR image and the fused background region.

The *pseudocode* for the proposed method is given below:

Input: IR and visible images (assuming they are registered).

Initialize: patch size 3×3, $\beta = 0.3$.

Step 1: Apply ATRE approach to the IR image to extract the transition region.

Step 2: Find the final segmentation threshold T_f. Binarize the IR image using T_f.

Step 3: Extract the object region from the IR image using the binarized image.

Step 4: Extract the background regions from both the VI and the IR image using the inverted binarized image.

Step 5: Find the fused background region by merging the background region of the IR image and the background region of VI using the fusion rule in Eqn. (14).

Step 6: Find the output image by region mapping between the object region of the IR image and fused background region.

***Output*:** Fused image.

4. Results and discussions

In this work, different multimodal images (IR and visible) are used to carry out the proposed fusion method. The image pairs are publicly available in the image dataset [13]. We have selected eight standard image sets namely *Tank, Bunker, Nato_camp1, Nato_camp2, Sandpath, Gun, Tree,* and *Two men in front of house* for both qualitative and quantitative comparisons. The simulations are performed in MATLAB using core i5 processor with 8 GB RAM.

Generally, the performance of the output image is assessed in two methods i.e. qualitatively and quantitatively. Many performance metrics are reported in the literature for the assessment of fusion results. Usually, these metrics measure the amount of info conveyed from the source images to the output. In this work, we use the evaluation indices – Petrovic ($Q^{PQ/F}$) [36], entropy (E) [37], mutual information (MI) [38], feature mutual information (FMI) [39], and visual information fidelity for fusion (VIFF) [40]. The best in class results are displayed in bold. The details of the metrics are available in the respective literature.

In this study, we have initialized β and the neighbourhood size. The value of β is chosen as 0.3 after exhaustive experiments. It is employed to stabilize the contribution of local variance and complexity. A neighbourhood size of 3×3 is selected in our proposed scheme. The benefit of choosing this neighbourhood size is to decrease the blocking artefacts.

The visual results of the suggested method and the other methods for all the image pairs are shown in Figure (3) – (10). In all the figures, (a), (b) are the source images (VI and IR image respectively). The rest images from (c) – (p) are the output from various methods including the proposed technique. The visible images comprise the details of the background, while the IR images focus the objects i.e. bunker, tank, person etc. The objective of our suggested method is that the fused image should keep much of the thermal radiation information from the IR images along with the

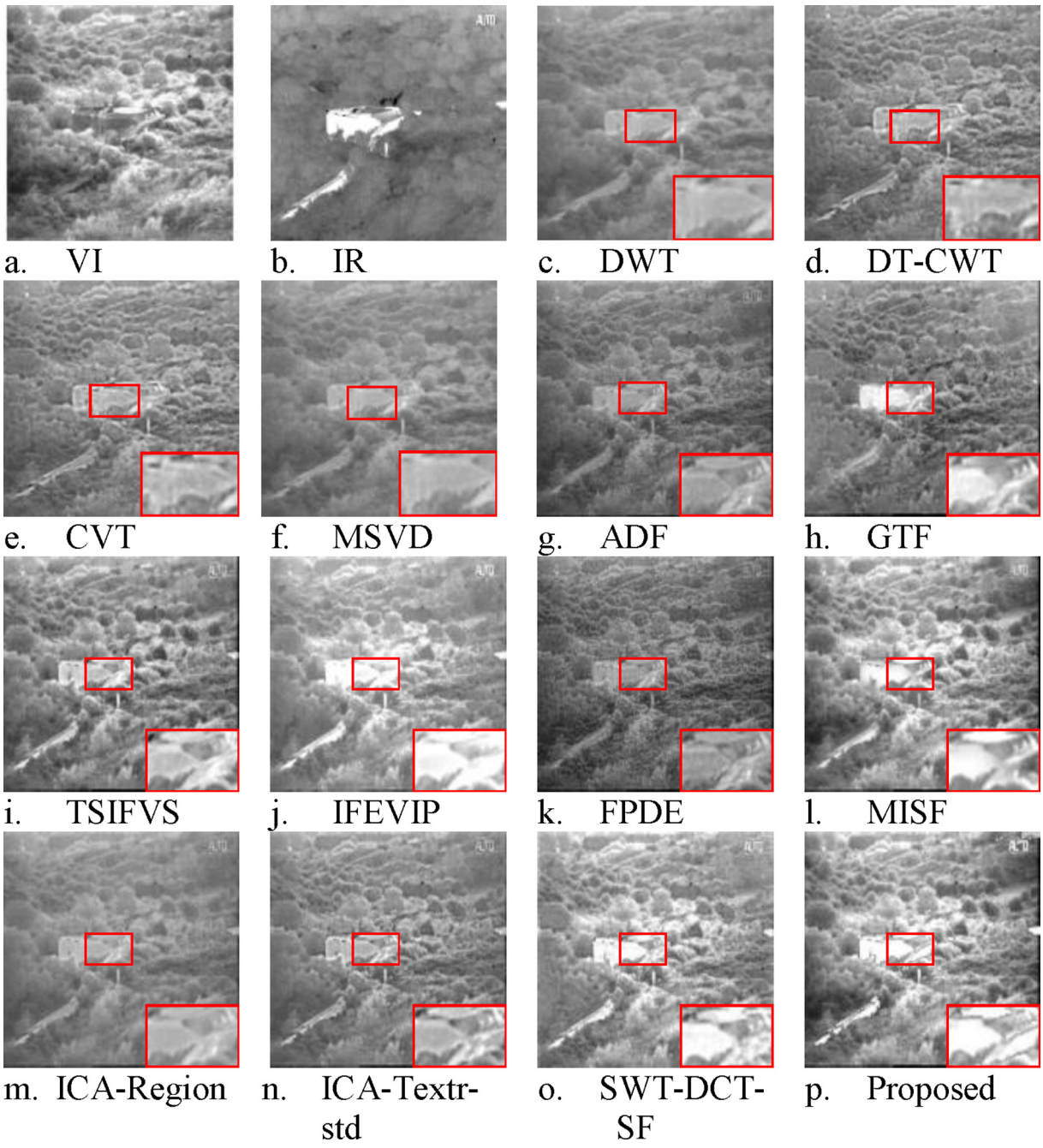


Fig. 3. Fusion results for *Bunker* images.

details of the background information from the visible images. The objective comparison of the different methods is depicted in Tables 1-8.

In Fig. 3, the bunker object is clearly visible in the fused images in (i), (j), (l), (o) and (p). In the image (h) the middle part of the bunker is visibly clear, however, the background looks blurry. Similarly, the other methods fail to retain the object info. In comparison to the other techniques, the output image found with the proposed technique highlights both the bunker and its background clearly. The contrast of the VI is also retained. The images with the rest of the methods are either dark or having poor contrast, especially in (c), (f), (g), (k) and (m). The reason may be the ATRE method used in the proposed technique effectively thresholds due to the adaptive parameter and extracts the object region in the IR image more accurately. Furthermore, most of the background information is retained in the fused image as we merged the background regions of the VI and IR image.

From Table 1, it is seen that the suggested technique is better than the other methods in terms of $Q^{PQ/F}$, E and VIFF. Nonetheless, the MI and FMI values of the proposed technique are close to the best value. As the bunker object is bigger in size, the proposed method is able to localize it efficiently. The adaptive nature of the technique accurately extracts the object region. The entropy obtained is highest as our approach is based on MLE concept for

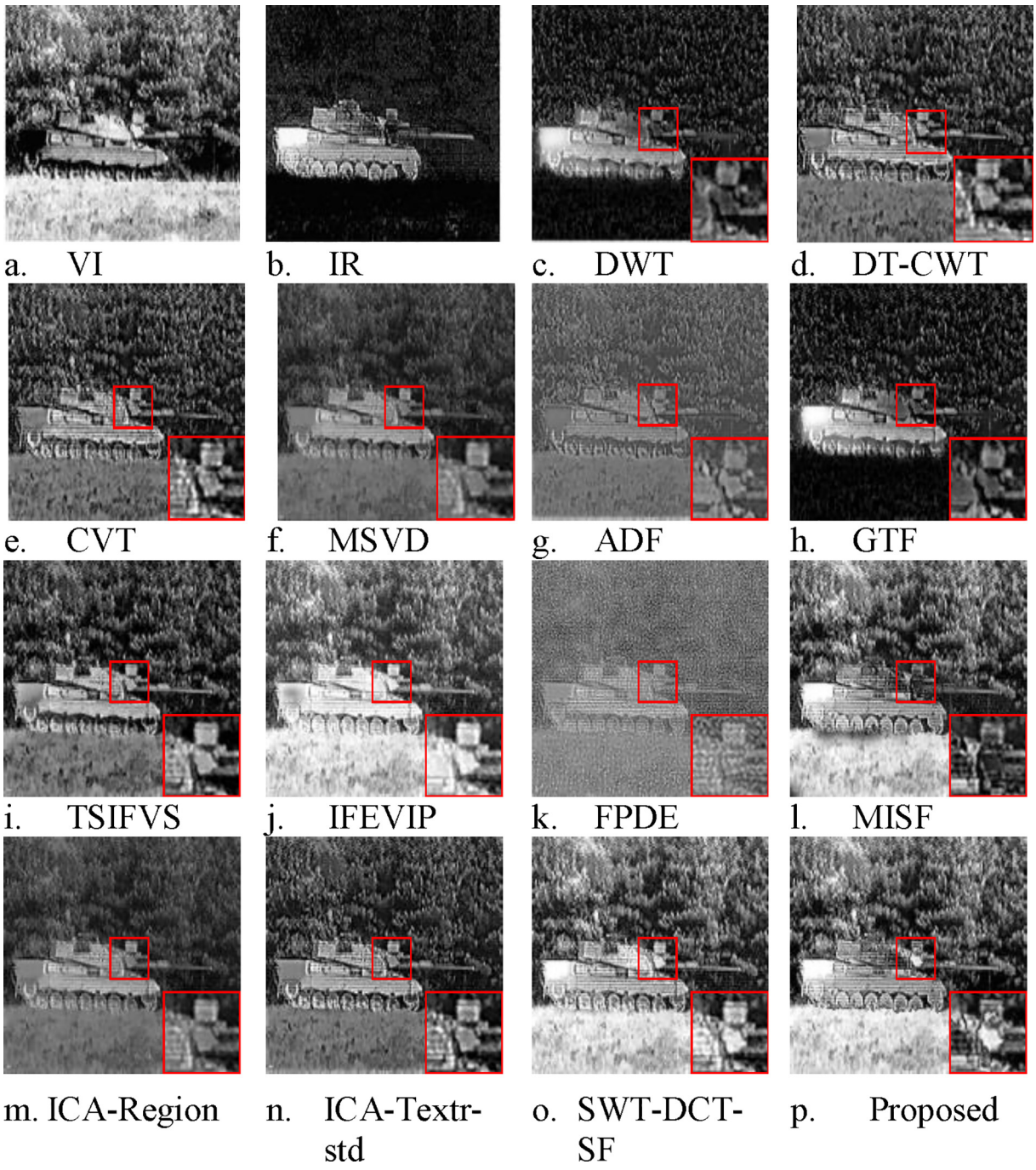


Fig. 4. Fusion results for *Tank* images.

segmentation. The information transferred to the output is also highest in the proposed method. Thus, the metrics show a better value as compared to the other methods.

It is observed in Fig. 4 that the visual performance is identical to that of the bunker image. The tank image and its background are not clearly visible in Fig. 4(g) and (k). In (c) and (h) the tank image looks clear, however, the background looks darker. The image obtained with ADF and FPDE is having poor contrast. However, the contrast of the fused image is retained in proposed method.

From Table 2, it is perceived that the suggested method does better than the other approaches in terms of $Q^{PQ/F}$, E and VIFF. The MISF method shows a high MI and FMI value. Still, the MI and FMI values obtained using our method is close. The reason may be the input images contain more texture and edge features.

From Fig. 5, it is seen that, the output images in (c), (f), (g) and (h) do not preserve the information of the source images. For instance, the person image is not detected clearly. The roof, tree and fence objects are also not clearly visible. The distinction between the object and the background is not clear. Some images are even having overlapped regions. On the other hand, the fused image in (p) from our proposed method looks visually clear.

From Table 3, it is observed that the suggested technique shows the best results for $Q^{PQ/F}$, E, MI, and FMI. However, the TSIFVS method leads in VIFF value. The reason may be the use of two-scale

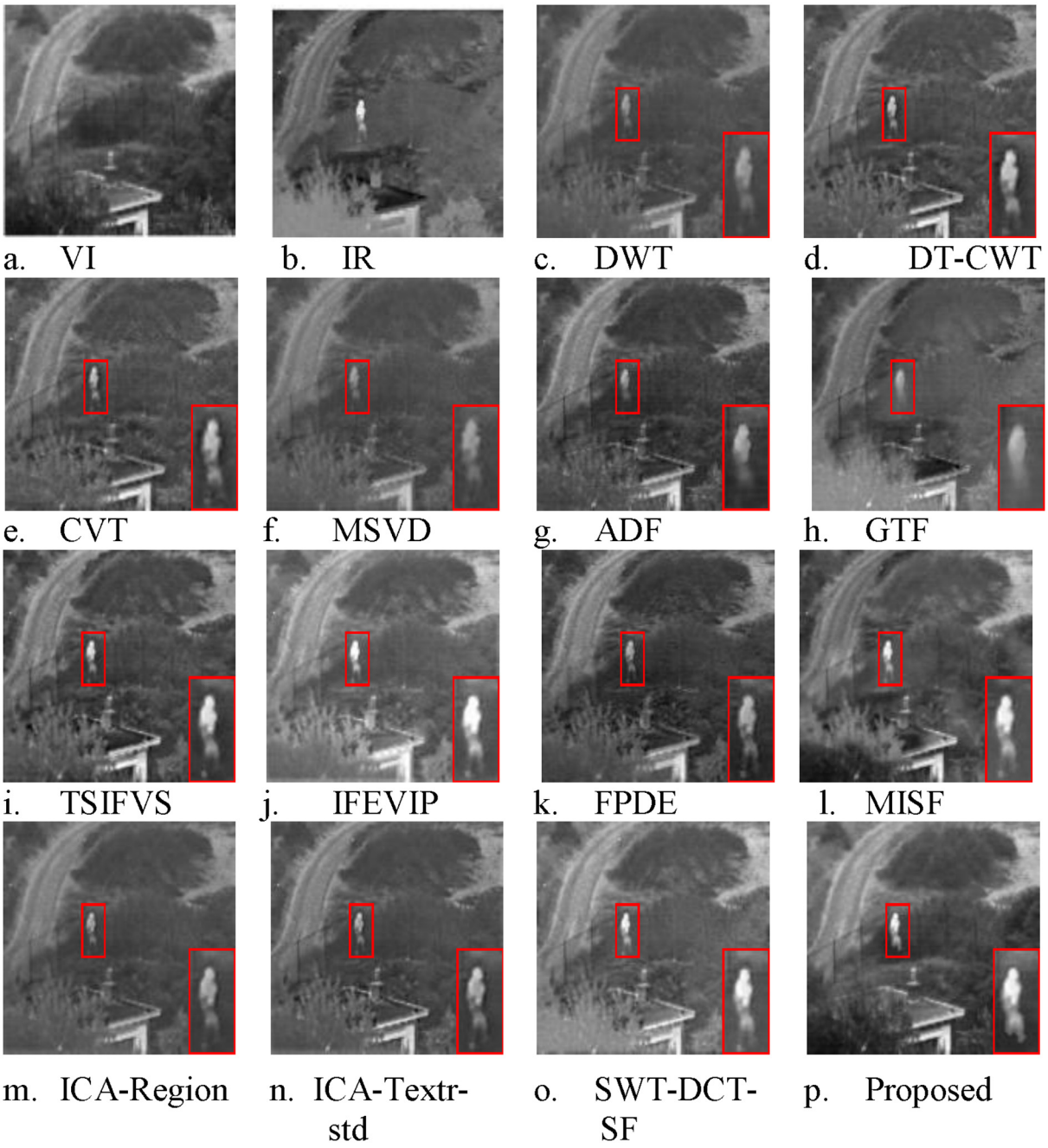


Fig. 5. Fusion results for *Nato_camp1* images.

decomposition technique, which reduces the distortion leading to a better value of VIFF.

In Fig. 6, a similar trend is observed. The person image is not properly traced in (f), (g) and (h). The roof, tree and fence images are also not clearly identifiable. However, in (c), (d), (e), (k), (m) and (n) the person is detected. The fused image (p) shows good result. The road, tree, fence and the roof are looking prominent. For Nato_camp2 image, the quantitative comparison is shown in Table 4. It is seen that the E, MI and FMI values are better for the suggested technique. It is interesting to note that the proposed method is close to the best value in case of $Q^{PQ/F}$value. In Fig. 7, the person image is not detected properly, especially in (c), (d), (e), (g), (h), (k), and (m). However, the person, tree and the path are clearly visible in the fused image (p). Moreover, the fused images obtained with other methods are having poor contrast.

The quantitative values in Table 5 show that the proposed technique outperforms in terms of $Q^{PQ/F}$, E, MI and VIFF. The reason may be the use of ATRE based approach for extracting the object region. The output images of Gun for different approaches are shown in Fig. 8. It is perceived that the gun image is clearly detected in our proposed method in (p). However, in (g), (h) and (m), the gun image is not properly detected. The boundary between the gun and the background is seen overlapped. The performance metrics $Q^{PQ/F}$, MI, E, and VIFF shown in Table 6 are higher in case of our method in comparison to the other methods excepting the value of FMI. Still, the value of FMI obtained with our method is very close to the best value.

In Fig. 9, the object i.e. the person in (p) is extracted properly using the suggested method. The visible quality of the fused image (p) is not clear as the background and the person image has less

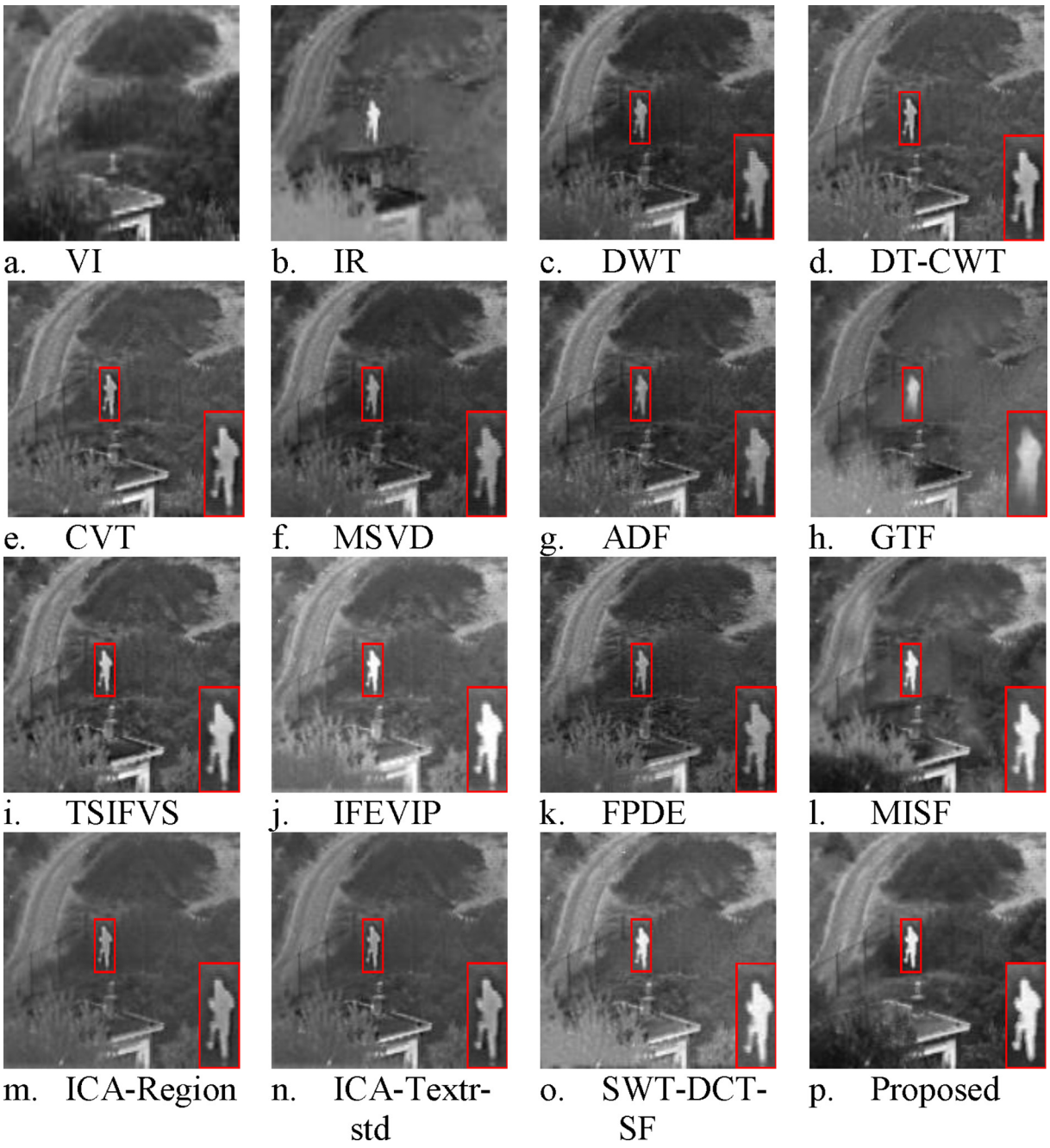


Fig. 6. Fusion results for *Nato_camp2* images.

intensity variations. Our proposed method retains most of the VI information. That is why during region mapping, the intensity of the object matches with the intensity of the background leading to an unclear image. This can be improved if we include some pre-processing operation (to widen the intensity difference) before the fusion. A possible solution may be to enhance the contrast of the object region.

The $Q^{PQ/F}$, E and FMI indices are higher for our technique as compared to the other approaches as given in Table 7. In Fig. 10, it is seen that persons in the fused images (f), (g) and (h) are visually not prominent. The persons are clearly visible in the fused images (i) and (j). However, the background is not so prominent. There are some dark patches visible in the sky in (l). In (p) the persons and the house are visually clear as compared to the other methods. The window and the tree are also being prominently perceptible.

Similarly, from Table 8, it is seen that the proposed technique leads the E, MI, FMI and VIFF values. It is to be noted that most of the methods for comparison used pixel level fusion approach. The pixel-based methods introduce the blocking artifact in the fused images and the background is not clear. Further, the texture features of the VI is not properly transferred to the fused image. On the other hand, the region based approaches extract important region of interest by considering patches from both the IR and VI. In this process, the objects and the textural information are preserved resulting in better fused images.

A comparison of the proposed method with the recent deep learning based methods in shown in Table 9. The values indicated

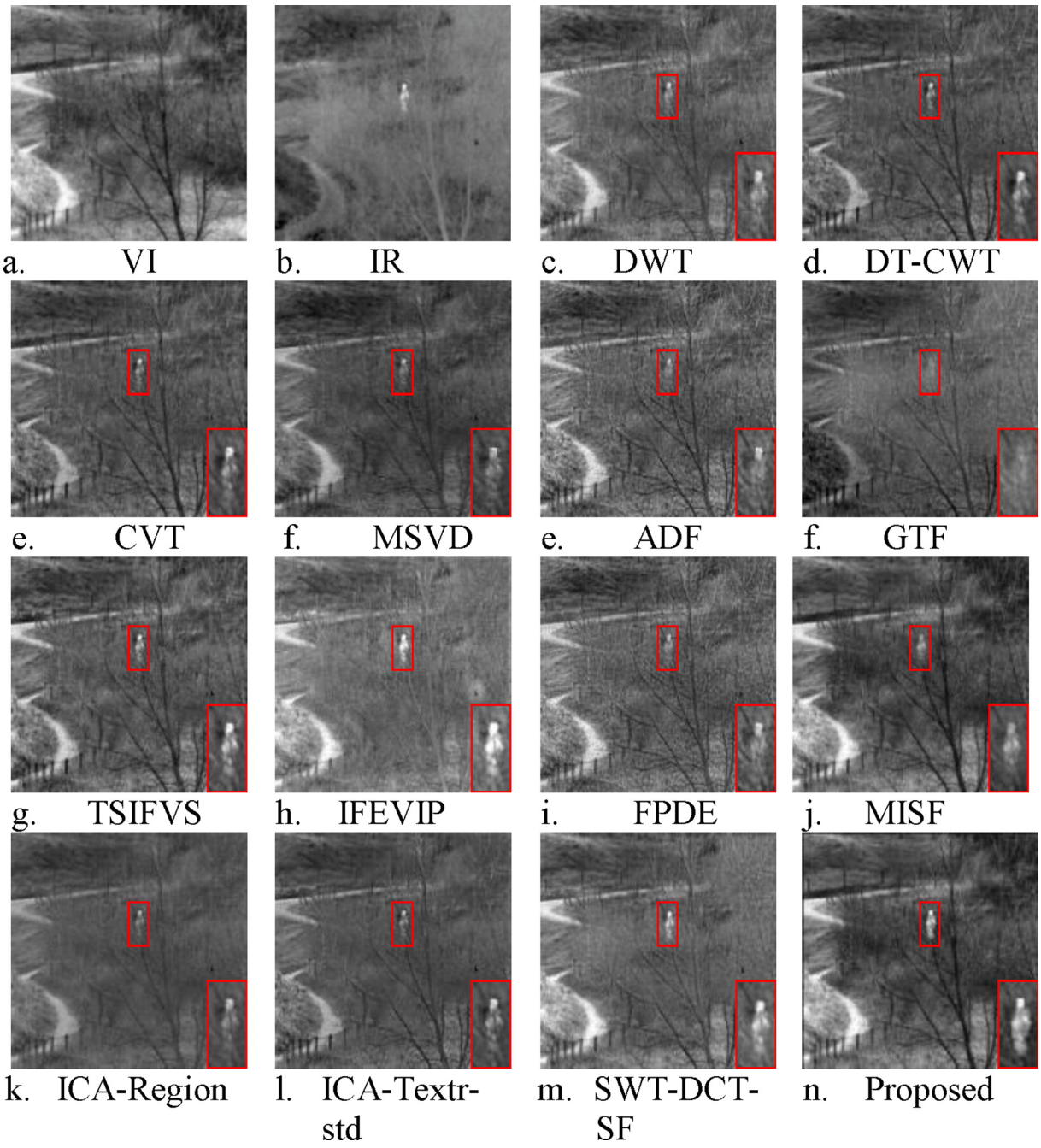


Fig. 7. Fusion results for *Sandpath* images.

for different metrics are average values. The methods using deep learning used different set of images for experiment. Hence, for a comparison we have shown the average values. The fusion methods use the deep networks for feature extraction and classification. The computational complexity and the hardware requirements employing deep network is very high. The deep networks require a large number of inputs for training. It is observed from the table that the DDcGAN method gives a better entropy value. Nonetheless, our proposed method gives values closer to the best in class values.

The experimental results on the selection of hyper-parameter β and patch size k are given in Table 10. For an illustration, we have considered the entropy value for different values of β from 0.1 to 0.9. A high value of entropy decides the β value. It is observed that the value of entropy is maximum at β = 0.3 for both the sample images. Hence, we have chosen the β value to be 0.3. Similarly, the entropy value is computed for k = 3, 5, 7. It is found maximum for $k = 3 \times 3$.

A graphical comparison of different performance metrics of various methods is illustrated in Fig. 11. It is to be noted that the average value of the metrics is displayed in the figure. The proposed method is able to obtain the largest average values on the four-evaluation metrics, i.e., E, $Q^{PQ/F}$, MI and FMI. Nonetheless, it gives comparable results in terms of VIFF. The largest E value demonstrates that the fused image obtained with the proposed method has more abundant information than the other comparing methods. The largest $Q^{PQ/F}$ value shows that more edge information is retained with our method. Similarly, the MI and FMI values

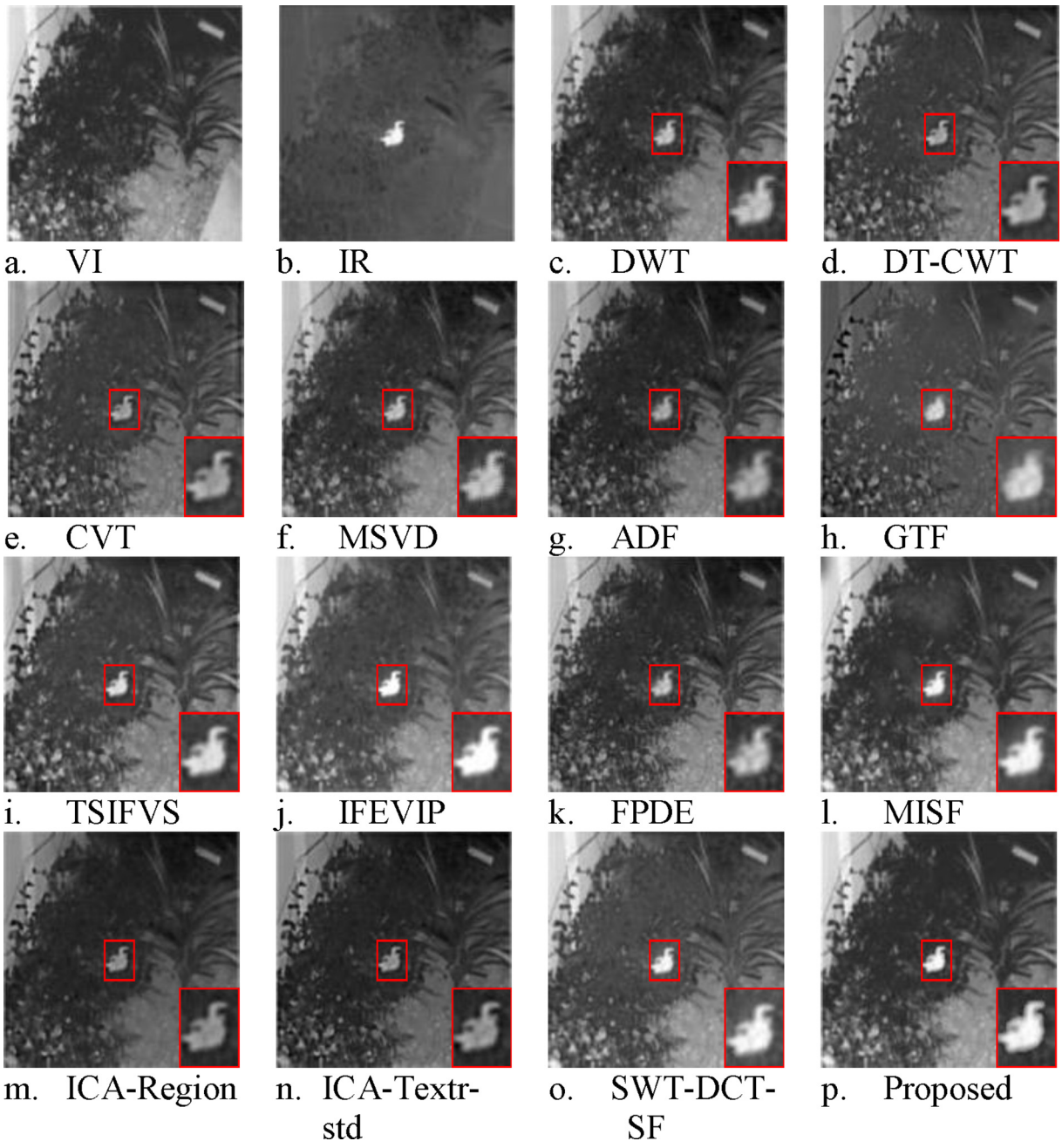


Fig. 8. Fusion results for *Gun* images.

obtained with our method outperforms the other methods demonstrating that it preserves the most useful information and features of the source images.

5. Conclusion

In this paper, we have suggested a region based approach for the fusion of IR image and VI. The proposed approach shows its ability to identify the objects precisely. The benefit of using adaptive transition region extraction based segmentation is to outline the object region in the IR image clearly. The proposed ATRE approach is suitable for the determination of a better segmentation threshold value using the adaptive parameter (γ_a). The parameter (γ_a) adapts itself to change in the input images. The benefit of the region mapping approach is to integrate the object region with the background information efficiently. The experimental results

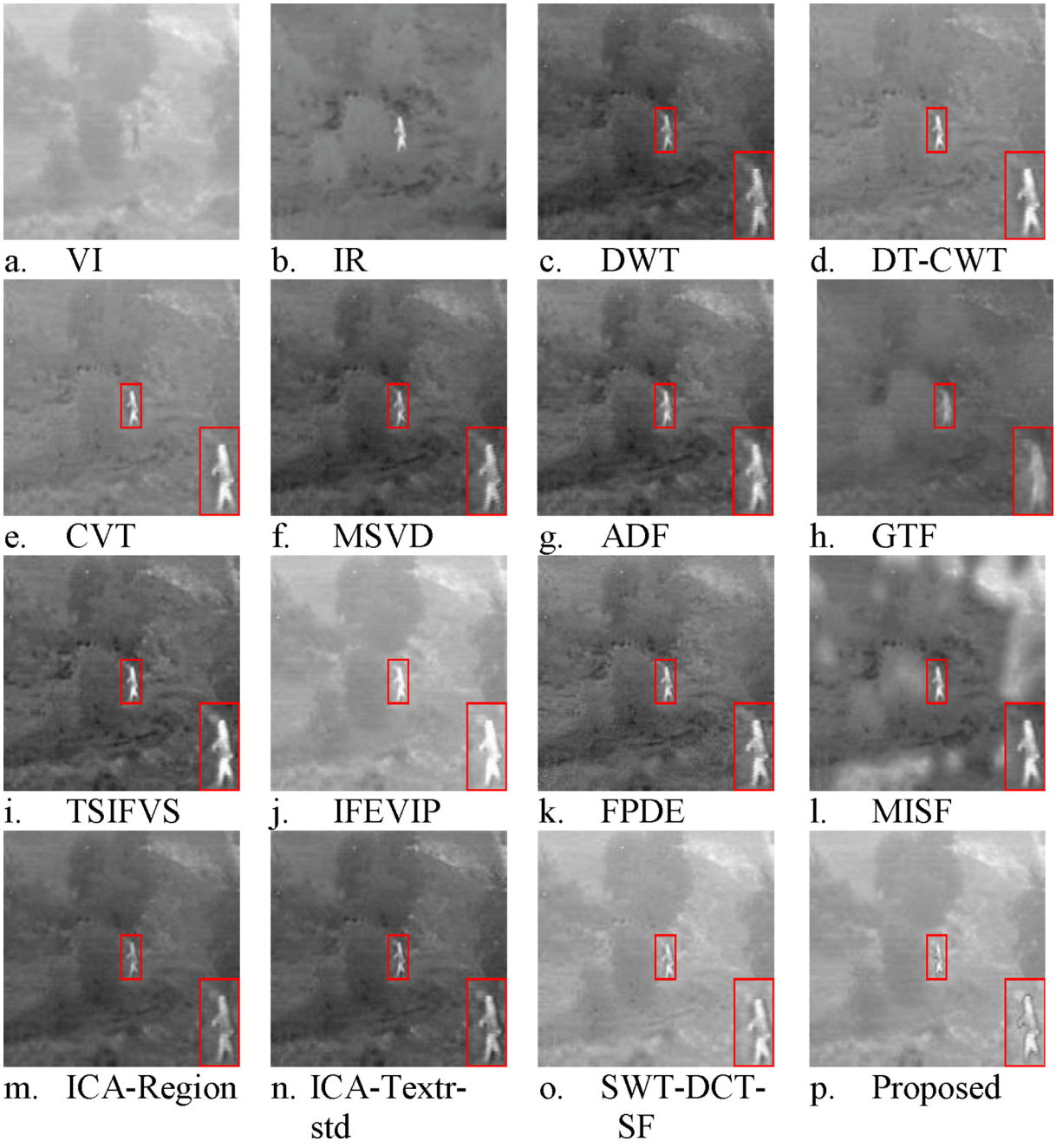


Fig. 9. Fusion results for *Tree* images.

reveal that the suggested technique has improved fusion results as compared to state-of-the-art fusion techniques. Although our method performed better, it has some limitations as well. Specifically, when the background near the object region in the VI is more

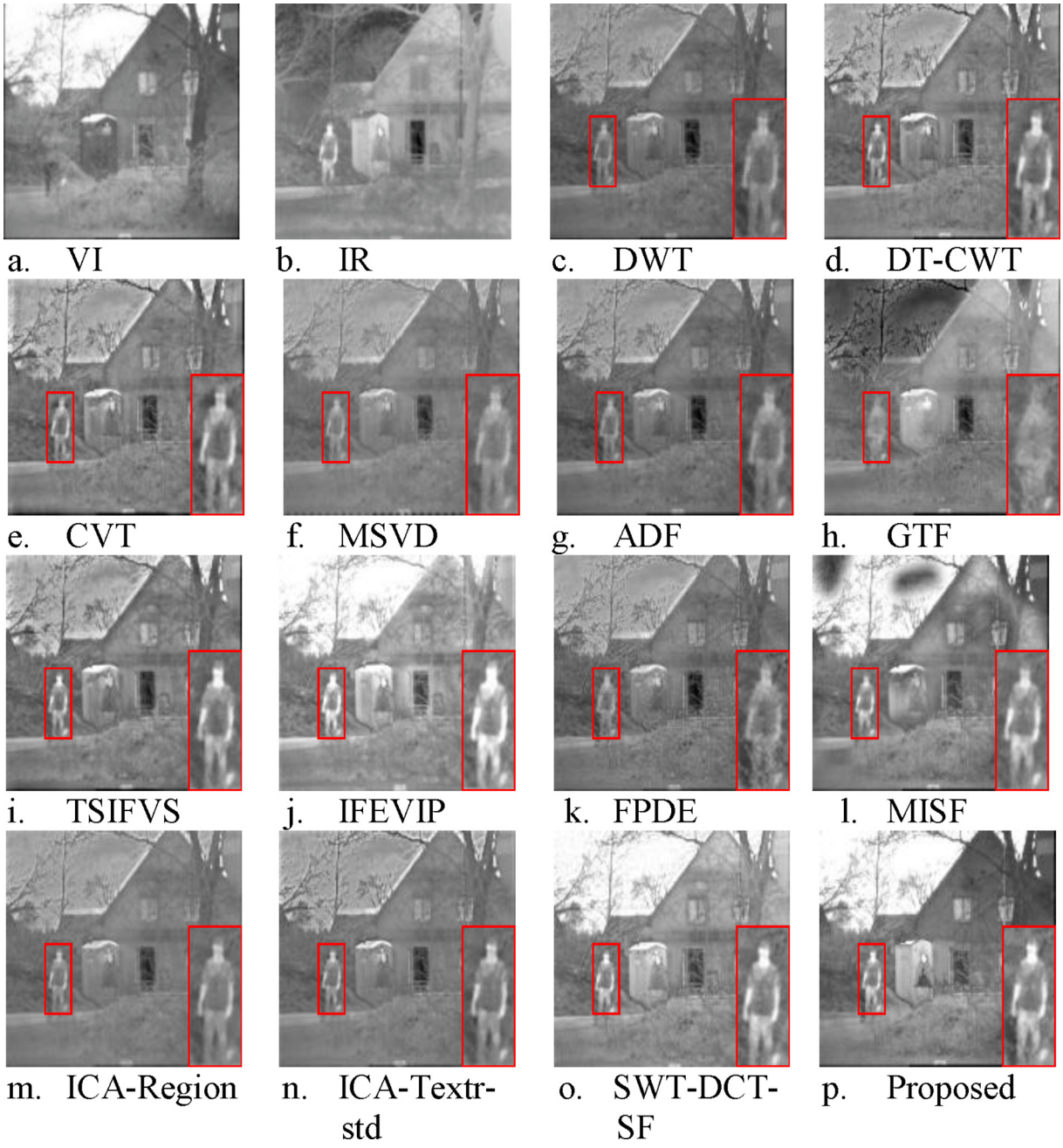


Fig. 10. Fusion results for *Two men in front of house* images.

Table 9
Comparison with deep learning based methods.

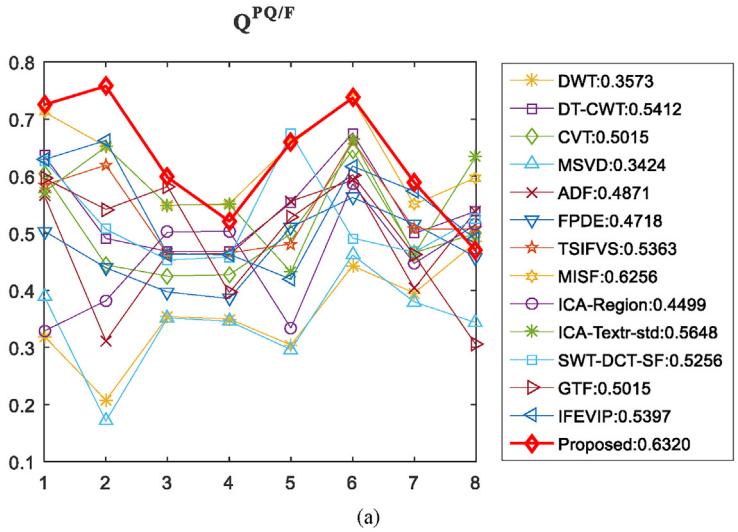
Method	E	VIFF	MI
FusionGAN [28]	6.8416	0.3162	2.3410
GAN [29]	7.0546	0.4034	–
DDcGAN [30]	7.3493	0.3192	–
Densefuse [30]	6.8248	0.3980	2.3020
DIDFuse [32]	7.0060	0.6230	2.3470
Proposed	7.1933	0.3683	2.7665

'-'indicates data unavailability.

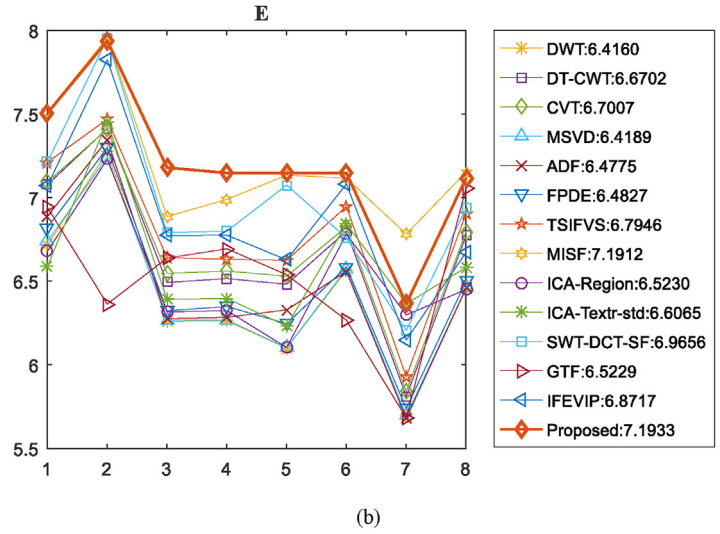
prominent than the corresponding region in the IR image, the object may not be traced well. In future studies, the proposed scheme may be experimented with video data.

Table 10
Entropy (E) values for selection of β and patch size *k*.

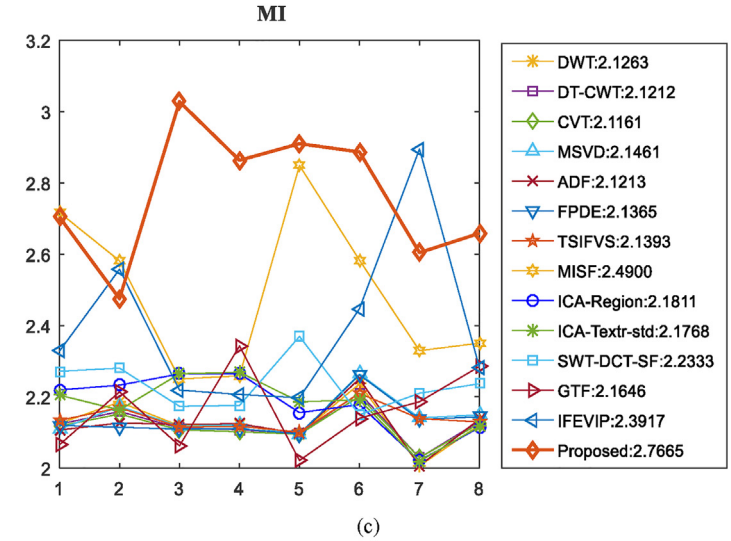
Parameters		Images	
		Sandpath	Two men in front of house
β	0.1	7.0886	6.9753
	0.2	7.0662	7.1384
	0.3	**7.1454**	**7.1522**
	0.4	7.0722	6.9007
	0.5	7.1437	7.1342
	0.6	7.0730	6.9253
	0.7	7.0665	7.1193
	0.8	7.0820	7.0235
	0.9	7.0957	7.0463
k	3×3	**7.1454**	**7.1522**
	5×5	6.5111	6.1299
	7×7	6.4897	6.1122



(a)

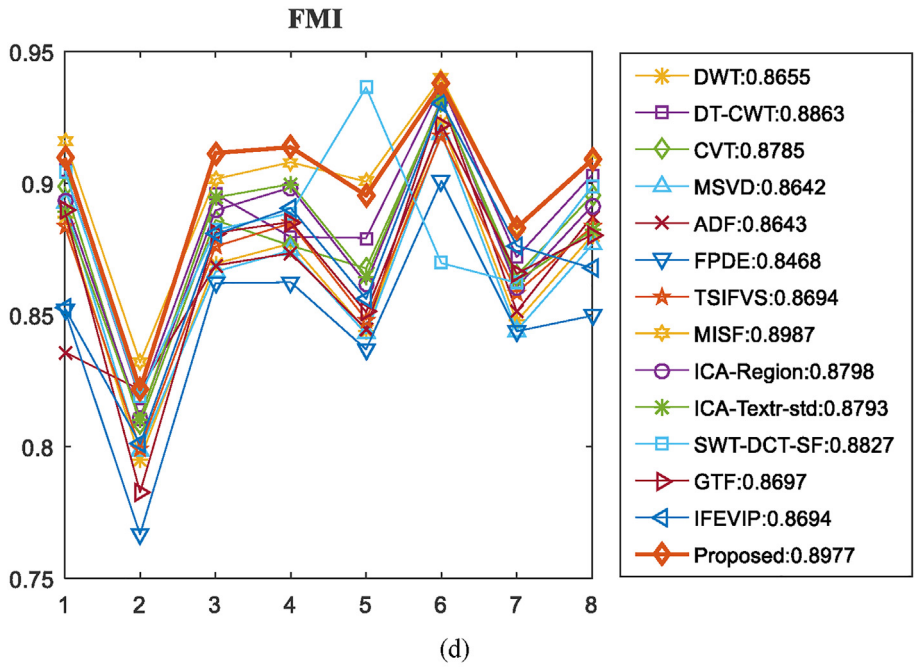


(b)

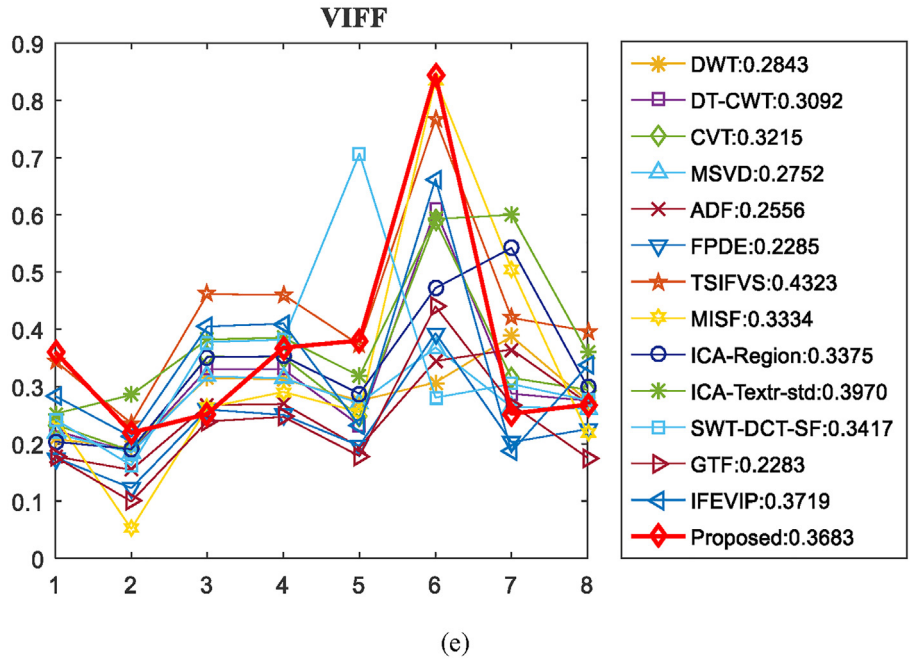


(c)

Fig. 11. Quantitative comparison of the performance metrics of different VI and IR image pairs, (a) $Q^{PQ/F}$, (b) E, (c) MI, (d) FMI, (e) VIFF.



(d)



(e)

Fig. 11 (*continued*)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: a survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[2] K. Padmavathi, C.S. Asha, V.K. Maya, A novel medical image fusion by combining TV-L1 decomposed textures based on adaptive weighting scheme, Eng. Sci. Technol. Int. J. 23 (1) (2020) 225–239.
[3] S. Li, B. Yang, Hybrid multiresolution method for multisensor multimodal image fusion, IEEE Sens. J. 10 (9) (2010) 1519–1526.
[4] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters, Inf. Fusion 30 (1) (2016) 15–26.
[5] J. Jinju, N. Santhi, K. Ramar, B. Sathya Bama, Spatial frequency discrete wavelet transform image fusion technique for remote sensing applications, Eng. Sci. Technol. Int. J. 22 (3) (2019) 715–726.
[6] B. Yang, Z.L. Jing, H.T. Zhao, Review of pixel-level image fusion, J. Shanghai Jiaotong Univ. 15 (2010) 6–12.
[7] G. Bhatnagar, Q.M.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in NSCT domain, IEEE Trans. Multimed. 15 (5) (2013) 1014–1024.
[8] B. Meher, S. Agrawal, R. Panda, A. Abraham, A survey on region based image fusion methods, Inf. Fusion 48 (2019) 119–132.
[9] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Phys. Technol. 85 (2017) 478–501.

[10] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (2019) 153–178.
[11] C. Yan, N. Sang, T. Zhang, Local entropy-based transition region extraction and thresholding, Pattern Recognit. Lett. 24 (16) (2003) 2935–2941.
[12] Z. Li, D. Zhang, Y. Xu, C. Liu, Modified local entropy-based transition region extraction and thresholding, Appl. Soft Comput. J. 11 (8) (2011) 5630–5638.
[13] http://figshare.com/articles/TNO_Image_Fusion_dataset/1008029.
[14] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[15] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inf. Fusion 8 (2007) 119–130.
[16] S. Quan, W. Qian, J. Guo, H. Zhao. Visible and infrared image fusion based on curvelet transform. In The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014) (pp. 828-832). IEEE.
[17] V.P.S. Naidu, Image fusion technique using multi-resolution singular value decomposition, Def. Sci. J. 61 (5) (2011) 479–484.
[18] D.P. Bavirisetti, R. Dhuli, Fusion of Infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform, IEEE Sens. J. 16 (2016) 203–209.
[19] D.P. Bavirisetti, G. Xiao, G. Liu, Multi-sensor image fusion based on fourth order partial differential equations, in: In: Proceedings of 20th IEEE International Conference on Information Fusion (Fusion), 2017, pp. 1–9.
[20] D.P. Bavirisetti, R. Dhuli, Two-scale image fusion of visible and infrared images using saliency detection, Infrared Phys. Technol. 76 (2016) 52–64.
[21] K. Zhan, L. Kong, B. Liu, Y.H. Multimodal image seamless fusion, J. Electr. Imaging 28(2) 2019 023027.
[22] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, S.J. Lee, K. He, Infrared and visual image fusion method based on discrete cosine transform and localspatial frequency in discrete stationary wavelet transform domain, Infrared Phys. Technol. 88 (2018) 1–12.
[23] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.
[24] Y. Zhang, L. Zhang, X. Bai, L. Zhang, Infrared and visual image fusion through infrared feature extraction and visual information preservation, Infrared Phys. Technol. 83 (2017) 227–237.
[25] N. Mitianoudis, T. Stathaki, Pixel-based and region-based image fusion schemes using ICA bases, Inf. Fusion 8 (2007) 131–142.
[26] N. Mitianoudis, S.A. Antonopoulos, T. Stathaki. Region-based ICA image fusion using textural information, In: Proceedings of 18th Int. Conf. Digit. Signal Process. (DSP), 2013, pp.1–6.
[27] H. Li, X.J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in: In 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 2705–2710.
[28] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[29] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.
[30] J. Ma, H. Xu, J. Jiang, X. Mei, X.P. Zhang, DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.
[31] H. Li, X.J. Wu, Densefuse: a fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623.
[32] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, J. Zhang. DIDFuse: Deep image decomposition for infrared and visible image fusion, arXiv preprint arXiv: 2020, 2003.09210.
[33] P. Thierry, A new method for grey-level picture thresholding using the entropy of the histogram, Signal Process. 2 (1980) 223–237.
[34] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[35] C.X. Yan, N. Sang, T.X. Zhang, Z.K. Image transition region extraction and segmentation based on local complexity, J. Infrared Millim. Waves. 2005 24, 312-316.
[36] C.S. Xydeas, V. Petrović, Objective image fusion performance measure, *Electron. Lett.* 36 (4) (2000) 308, https://doi.org/10.1049/el:20000267.
[37] W. Wang, F. Chang, A multi-focus image fusion method based on Laplacian pyramid, J. Comput. 6 (12) (2011) 2559–2566.
[38] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[39] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Comput. Electr. Eng. 37 (5) (2011) 744–756.
[40] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (2013) 127–135.