Fusion of Infrared and Visible-Light Videos Using Motion-Compensated Temporal Sub-Band Decompositions

Jonathan N. Gois
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Brazil
jonathan.gois@cefet-rj.br

Eduardo A. B. da Silva
Universidade Federal do Rio de Janeiro, PEE/COPPE/DEL/Poli, Brazil
eduardo@smt.ufrj.br

Carla L. Pagliari Marcelo M. Perez
Instituto Militar de Engenharia, PEE/PGED, Brazil
carla,perez@ime.eb.br

Abstract

The fusion of visible-light and infrared videos has applications in several areas, and is an active research topic. To this end, it is common to employ complex fusion methods that take into account spatial and/or temporal information in the videos. In this paper we propose a video fusion method that is based on a motion-compensated, two-band temporal sub-band decomposition. The alignment provided by the motion vectors, besides providing a reduction in the registration errors of the input images, allows the use of a simple fusion rule to the temporal sub-bands. The results indicate that the use of the proposed exploration of the temporal information alone is quite effective, and gives objective fusion quality results that compare favorably to more sophisticated methods based on complete spatiotemporal information.

1. Introduction

The analysis of the information conveyed by sensors from different spectral bands is important to different areas, such as military, security, aerospace, remote sensing and medical, among others. As the joint analysis of information captured by each sensor can be a difficult task, the methods of fusing images/videos promote the combination of the information generated by multiple sensors into a single output, with relevant information. Many image fusion methods for the infrared and visible-light bands employ multiscale techniques such as the Discrete Wavelet Transform (DWT), the Contourlet Transform and the Undecimated Wavelet transform (UWT) [6], to name a few. In general, multiscale image fusion methods decompose each spectral band input image into sub-bands using some multiscale transform. Then a fusion rule is applied to the corresponding input sub-bands, in order to generate each sub-band of the fused image. The desired fused image is then obtained by applying the inverse of the multiscale transform originaly employed for the input spectral images over the resulting fused sub-bands.

These techniques can be applied to videos on a frame-by-frame basis, ignoring the temporal redundancy among the video frames. However, temporal information is often used in the fusion process. An example is the work in [13], where a fusion rule is applied according to the temporal content after the spatial multiscaling decomposition in sub-bands. Zhang [9, 22, 29–31] proposed different methods using 3D transforms (where time is the third dimension) in order to produce fused videos with better subjective quality. The frame-based method introduced in [23] employs local contrast and color consistency to combine the scene details under different exposures. A framework based on generalized random walks obtains a globally optimal solution subject to two quality measures. A two-scale decomposition of an image into a base layer, that captures large scale variations in intensity, and a detail layer capturing small scale details is proposed in [14]. Then a guided filter-based weighted average technique is proposed to explore the spatial consistency for fusion of both layers, not considering the temporal correlation. In [27] a transform, the 3D shearlet, that incorporates the temporal dimension is proposed. In [28] the authors use the wavelet transform to decompose the image frames, previously processed by an optical flow motion-compensation approach. Each fused frame is obtained using a rule that takes into consideration the energy of the wavelet coefficients and the motion compensation information. Despite using motion vectors, the filtering step in [28] is performed only in the spatial dimension. The work in [26] uses an algorithm that combines the distribution of

gray levels with a texture filter and inter-frame correlation.

In this paper, we propose a novel infrared and visible-light video fusion method, the MCTSF (Motion-Compensated Temporal Sub-band decomposition video Fusion). A common drawback of frame-by-frame basis schemes is the temporal stability of the resulting fused frames, that may present fluctuations in appearance due to the variation of fusion parameters across the sequence of images. The novelty of the proposed method relies on using motion-compensated temporal filtering to efficiently process the relevant information. This way, the videos are analyzed and fused as a whole rather than in a frame-by-frame basis, where the temporal filtering is performed along the direction of motion using different filter banks.

The MCTSF is based on a two-band temporal sub-band decomposition of the visible and infrared videos. To increase the effectiveness of such a decomposition we use Motion-Compensated Temporal Filtering (MCTF) [16, 18], whereby neighboring frames are temporally aligned by motion compensation during filtering. This is achieved by performing a block-based motion estimation scheme in order to obtain the motion information. Using this information one employs MCTF to decompose the video frames into temporal high-pass and low-pass sub-bands. Finally, a simple fusion rule is applied. The workflow of the proposed method is depicted in Figure 1.

The remaining sections of this paper are organized as follows: Section 2 presents the motion-compensated temporal sub-band decomposition technique employed, while in Section 3 the proposed video fusion framework is introduced in detail. Section 4 presents and discusses the experimental results, and the conclusions are presented in Section 5.

2. Motion-compensated temporal sub-band decompositions

Considering that a video fusion algorithm should be able to fuse moving objects, motion information is an important issue as it distinguishes a static area from a moving area in a sequence of images.

Motion estimation is the process that estimates the displacement of the objects on the scene over time. A widely used approach is to divide each frame that composes the video into blocks (as small as necessary to represent possible objects in the scene) to estimate the position of the same block in another frame. This process is accomplished by minimizing the sum of the absolute differences between the blocks. The displacement from the original position of the block to the position where it was estimated in the frame of interest is called the motion vector. Motion estimation can be disrupted by many factors, such as occlusion, perspective distortion, and illumination variations. In addition, in a block-based scheme it is assumed that all pixels belonging to the same block have the same displacement, which is not always true. Another alternative would be to use superpixel-based motion compensation, where the size of the approximated regions is adapted to the geometry and the characteristics of the objects. However, block matching using fixed-size 4×4 blocks provides motion vectors that are accurate enough when associated with motion-compensation using fractional pixel accuracy. Therefore, in this application, it is not necessary to pay the price of the increase in computational burden associated with the use of superpixels, and we decided to just use block matching with 4×4 blocks.

The main idea of this proposal is to use the motion vectors as a representation of temporal information, associated with temporal filtering using motion compensation to better process the relevant information.

While abrupt transitions of the pixel intensity values in the spatial domain may indicate the presence of discontinuities (*e.g.* edges) or fine details, in the temporal domain abrupt transitions may indicate movement within the scene. Therefore, an efficient video fusion method has to use the motion information in order to adequately treat these transitions. A motion compensation algorithm is able to identify the transitions due to motion. In order to use this motion information, the two-band sub-band decomposition is computed along the temporal axis using motion-compensated temporal filtering (MCTF) [3, 18]. MCTF became popular in the context of video coding using 3D discrete wavelet transforms [15]. In such methods, in order to exploit inter-frame redundancy the temporal part of the transform must compensate for motion between frames, and this is done using MCTF.

Note that, in the proposed video fusion framework videos are analyzed and fused as a whole rather than in a frame-by-frame basis. As video alignment is fundamental in pixel-level video fusion methods (as is the case of the proposed MCTSF), a synchronized and registered video database should be used. Then, synchronization is a necessary pre-processing step, as the sequences are usually acquired by two different capture systems: visible-light (VIS) and infrared (IR). This implies that the temporal sub-bands are computed by applying the MCFT scheme to each individual sequence (VIS and IR), that belongs to a synchronized pair.

2.1. Motion Compensated Temporal Filtering

In motion-compensated temporal filtering (MCTF), the frames are temporally filtered in the direction of motion. In a nutshell, it is as if the pixels in each frame are aligned by motion compensation prior to filtering. Note that MCTF does not employ a temporal recursive structure, that is, the pixels from a filtered past frame are not employed to compute the current filtered frame.

Since usually the visible-light videos are sharper than the infrared ones, we perform the motion estimation only on the visible-light sequences. A 4×4 block size with half-pixel precision motion estimation [2] is employed. For all 4×4 blocks from the current frame, the corresponding 4×4 blocks are searched in the reference frame, generating the motion vectors. For a pixel at spatial coordinates (m, n) in frame k, the motion vector, $\mathbf{d}_k(m, n) =$

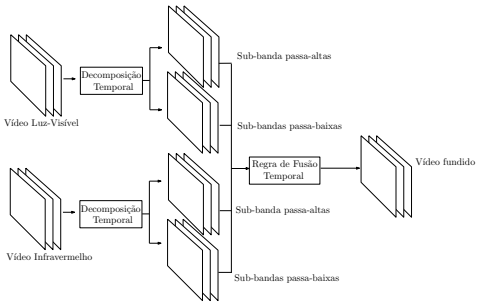
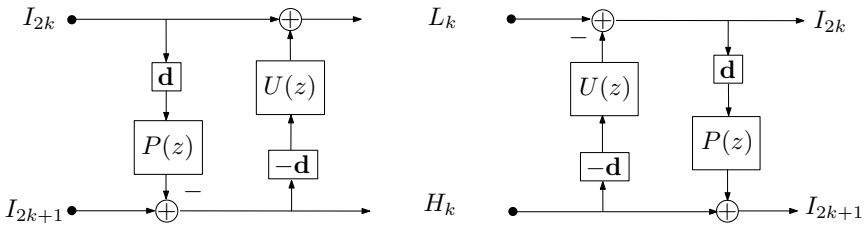


Figure 1: Algorithm workflow, where the first step is the temporal decomposition using MCTF.



(a) Decomposition (b) Reconstruction

Figure 2: Decomposition/Reconstruction of frames $2k$ and $2k+1$ using one lifting step. The variables L_k and H_k are the low-pass and high-pass sub-bands, respectively. The blocks labeled $\mathbf{d}$ and $-\mathbf{d}$ indicate that one is performing motion-compensated temporal filtering (Eq. (2)) using sets of symmetric motion vectors for $P(z)$ and $U(z)$.

$(d_k^x(m,n), d_k^y(m,n))$, is computed to minimize

$$D_{\mathcal{B},k} = \sum_{(m,n)\in\mathcal{B}} |I_k(m,n) - I_{k-1}(m - d_k^x(m,n), n - d_k^y(m,n))|, \tag{1}$$

where $I_k(m,n)$ represents the intensity value of the pixel located at (m,n) in video frame of index k and $\mathcal{B}$ is the region corresponding to the block for which the motion vector is computed, and $d_k^x(m,n)$ and $d_k^y(m,n)$ are the displacements in the x (line) and y (column) directions, respectively. It is assumed that the motion vectors of all pixels inside a block are equal.

After motion estimation, the motion vectors of each pixel are available and it is possible to perform temporal filtering along the direction of motion using a filter $h(k)$. The output of the filtering process at pixel (m,n) from frame k can be expressed as:

$$v(m,n,k) = \sum_l h(l) I_{k-l}(n - d_{k-l}^x(m,n), m - d_{k-l}^y(m,n)), \tag{2}$$

where the summation index l varies according to the support of the temporal filter $h(k)$. Note that the filter $h(l)$ can be non-causal, that is, l can assume negative values.

We are interested in building a temporal filter bank using filtering operations such as the one in Eq. (2). The goal is to first apply the analysis filter bank to both the visible-light and infrared videos, then to merge the corresponding sub-bands using the chosen fusion rule to finally apply the synthesis filter bank to obtain the fused video. However, the analysis and synthesis filter banks cannot be mere extensions of conventional filter banks for which the convolutions are made according to Eq. (2). This is so because the motion-compensation process is inherently non-linear, and therefore filter banks based on Eq. (2) would also be non-linear, which can create difficulties for their perfect reconstruction. In addition, consider that, due to an occlusion or mismatch in the motion estimation process, the pixel (m,n) from frame $k-1$ has no match in frame k. In this case, the summation in Eq. (2) cannot be computed. In the next section we introduce the lifting scheme [24], that presents a solution to such problems. For example, filter banks implemented using the lifting scheme can incorporate non-linearities such as the ones introduced by motion compensation while keeping their perfect reconstruction property. In fact, it is one of the main tools enabling the existence of motion-compensated perfect reconstruction filter banks.

2.2. Lifting Scheme

The lifting scheme has been proposed in [24] as a way to design and/or implement filter banks that inherently have the perfect reconstruction property. One important characteristic of the lifting scheme is that it guarantees perfect reconstruction, even in the case the filter banks are non-linear. The main building blocks of a lifting scheme are illustrated in Figure 2. Figure 2(a) corresponds to an analysis building block and Figure 2(b) corresponds to a synthesis building block. The two polyphase components of the signal, I_{2k} and I_{2k+1}, are input to the analysis building block, that presents at its output the low pass (L_k) and high-pass (H_k) sub-bands. The synthesis building block reconstructs the two polyphase components given as inputs the L_k and H_k sub-bands. Each building block consists of a cascade of two steps: a predict step, using filter $P(z)$ and an update step using filter $U(z)$. In the predict step, motion-compensated

temporal filtering is used to predict the samples of the second polyphase component (I_{2k+1}), using the samples of the first polyphase component (I_{2k}). The residual of this prediction is output as H_k. In the update step, a filtered version of the H_k component is used to update I_{2k}, generating the L_k component. In the synthesis process, the same update step is used, but this time the filtered version of H_k is subtracted from L_k in order to generate I_{2k}. Then, the same predict step is used for computing the prediction of I_{2k+1} using I_{2k}, that is added to H_k to reconstruct I_{2k+1}. In Figure 2, the two blocks $\mathbf{d}$ and $-\mathbf{d}$ indicate that the filtering operations that follows them is actually a motion-compensated temporal filtering according to Eq. (2). The opposite signs in the predict and update steps indicate that the motion vectors used in one case are the symmetric of the ones in the other case.

In [4] it is shown that any perfect reconstruction filter bank can be decomposed as a cascade of (predict, update) pairs. Therefore, using the scheme in Figure 2 one can perform motion-compensated sub-band decomposition and reconstruction for a general perfect reconstruction filter bank.

3. Proposed Video Fusion Algorithm

Motion-compensated temporal filtering is widely used in wavelet-based scalable video coding schemes, where the lifting implementation of these decompositions represents a resourceful tool for spatiotemporal optimizations [11]. These schemes often use the *Haar* filter bank for MCTF, as well as other filter banks such as the Cohen-Daubechies-Feauveau 5-3 (also known as Le Gall 5-3). In this paper, we implement the MCTSF using these two filter banks.

The first point to be addressed is how the temporal decomposition will be performed. We start by defining the analysis filter bank. Initially, we use the *Haar* filter bank, which is composed of the following filters:

$$h = \sqrt{2}\begin{bmatrix}1 & 1\end{bmatrix} \tag{3}$$

$$g = \sqrt{2}\begin{bmatrix}1 & -1\end{bmatrix} \tag{4}$$

To use the lifting scheme, we need to decompose the filter bank in lifting steps. Using the decomposition method shown in [4], we have that only one (predict, update) step is needed. In this case, we have that $P(z) = 1/\sqrt{2}$ and $U(z) = \sqrt{2}$. Therefore, using the lifting scheme, we are able to write the sub-band decomposition as:

$$H_k(m,n) = I_{2k+1}(m,n) - \frac{1}{\sqrt{2}} I_{2k}(m - d^x, n - d^y), \tag{5}$$

$$L_k(m,n) = I_{2k}(m,n) + \sqrt{2} H_k(m + d^x, n + d^y)). \tag{6}$$

The reconstruction is performed using the following expressions:

$$I_{2k}(m,n) = L_k(m,n) - \sqrt{2} H_k(m + d^x, n + d^y), \tag{7}$$

$$I_{2k+1}(m,n) = H_k(m,n) + \frac{1}{\sqrt{2}} I_{2k}(m - d^x, n - d^y). \tag{8}$$

It is important to note that in Eqs. (5) to (8) both d^x and d^y depend on the pixel position (m,n). This has been dropped in order to make the notation less cumbersome.

On that account, we guarantee that the filter bank has perfect reconstruction despite the nonlinearities caused by the motion-compensated temporal filtering. However, there is one important point that has to be taken care of, which is the case of pixels in a past frame that have no corresponding pixels in the current frame. This may be caused, for example, by occlusions. Such pixels are referred to as unconnected pixels. Fortunately, the predict and the update steps may be changed just for the unconnected pixels using a rule that dispenses the use of the motion vectors. The motion information, that is used in both the analysis and synthesis filter banks, permits to determine if a pixel is connected or unconnected. This way, both the analysis and synthesis processes will be able to choose the correct filter for a given pixel. This implies that the whole process can have perfect reconstruction even if unconnected pixels are processed with a different filter. In this work we treat the unconnected pixels by following the rule described in [10], which, in the case of unconnected pixels, computes the sub-bands as:

$$H_k(m,n) = \frac{1}{\sqrt{2}}(I_{2k+1}(m,n) - I_{2k}(m,n)) \tag{9}$$

$$L_k(m,n) = \sqrt{2} I_{2k}(m,n), \tag{10}$$

that are reconstructed by:

$$I_{2k}(m,n) = \frac{1}{\sqrt{2}} L_k(m,n), \tag{11}$$

$$I_{2k+1}(m,n) = \sqrt{2} H_k(m,n) + I_{2k}(m,n)) \tag{12}$$

In the proposed method one of the two-band temporal sub-band decomposition is performed using the *Haar* lifting scheme for each video sequence (VIS and IR). After the decomposition of each sequence, we apply a simple fusion rule in which each pair of bands are combined using their pixel average, that is:

$$L_k^F(m,n) = (L_k^{VIS}(m,n) + L_k^{IR}(m,n))/2 \tag{13}$$

$$H_k^F(m,n) = (H_k^{VIS}(m,n) + H_k^{IR}(m,n))/2 \tag{14}$$

This simple fusion rule was chosen to highlight the effectiveness of this type of filtering for video fusion.

After computing the sub-bands of the fused video L_k^F and H_k^F, we apply, for the case of connected pixels, the reconstruction rules from Eqs. (7) and (8). For the case of unconnected pixels, we apply the reconstruction rules from Eqs. (11) and (12).

The other two-band sub-band decomposition used in this paper employs the CDF 5-3 filter bank. The motivation behind the usage of CDF 5-3 is that longer length filters can improve the exploitation of the pixels' correlation in the temporal domain [11].

The CDF 5-3 filter bank can be decomposed into lifting

steps using the procedure outlined in [4]:

$$P_1(z) = -0.200000; \quad (15)$$

$$U_1(z) = -0.208333 - 0.625000z^{-1}; \quad (16)$$

$$P_2(z) = +0.900000z + 1.500000; \quad (17)$$

$$U_2(z) = -0.069444z + 0.333333 - 0.069444z^{-1}. \quad (18)$$

However, this decomposition demands a two level lifting scheme. To facilitate the implementation, the sub-bands can be directly written as [10, 11]:

$$H_k(m,n) = I_{2k-1}(m,n) - 0.5(I_{2k}(m-d^x, n-d^y) + I_{2k-2}(m+d^x, n+d^y)), \quad (19)$$

$$L_k(m,n) = I_{2k}(m,n) + 0.25(H_{k+1}(m-d^x, n-d^y) + H_k(m+d^x, n+d^y)). \quad (20)$$

As the bidirectional prediction is approximated using the opposite direction of the forward motion vector, the algorithm complexity is reduced.

The reconstruction using these sub-bands can be performed using [10, 11]:

$$I_k(m,n) = L_k(m,n) - 0.25H_k(m+d^x, n+d^y), \quad (21)$$

$$I_{k+1}(m,n) = H_k(m,n) + 0.5I_k(m-d^x, n-d^y). \quad (22)$$

Unconnected pixels are processed in the same way as they were in the previous (*Haar*) filter bank (see Eqs (9) to (12)).

The filter banks used in this work were chosen for their relevance in the MCTF context. Other filter banks can be easily adapted to implement the fusion algorithm using the same steps shown in this Section.

4. Results

In this section we analyze the performance of the proposed Motion-Compensated Temporal Sub-band decomposition video Fusion algorithm (MCTSF). The objective results are summarized in Table 1 and the visual quality results are summarized in Figure 3.

The selected sequences present scenarios with different illumination conditions and nighttime footage, as well as different degrees of motion and texture. These videos challenge the joint processing of IR and visible-light spectral bands. The two OCTCVBS sequences (320×240 pixels at approximately 30 frames/sec (fps)) [5] are composed by color/IR images, registered using homography with manually-selected points. The videos display busy pathway intersections. The INO [1] Park Evening sequence pair (328×254 pixels) is composed by co-registered IR and visible videos, showing a parking lot in the evening. The other INO sequence, Group Fight (452×332 pixels), shows a fight between two groups of persons on a parking lot. Both sequences were captured at 10 fps. The EDEN [7] video pair (VIS/IR) exhibits a man dressed in camouflage walking through thick foliage. All videos, at 25fps, are registered and synchonized, presenting moving objects/people, shadows and highly textured regions. The VIS and IR synchronized sequences (320×240 pixels), from Dublin City University campus (DCUC), contain bike-racks, pedestrians, bicycles and vehicles [17]. The pixels alignment is performed by manually selecting the corresponding points in both spectra and computing the homography with least-squared error. No frame rate is informed in [17]. Both sequences were downsampled to 224×224 pixels to be compared with the results presented in [28].

Figure 3 pictures one visible-light (VIS), one infrared (IR) and one fused frame of each sequence, for the two MCTSF versions: MCTSF *Haar* and MCTSF CDF 5-3. All frames were cropped to better fit the page. Perceptual analysis shows that, in all sequences fused using the MCTSF *Haar*, the edges are well matched and the fusion results conveyed the information from both spectral bands.

Figures 3(a) and 3(b) picture the tenth frames of the VIS and IR sequences OCTBVS OSU:Dataset03 [5], respectively. The fused frames are depicted in Figures 3(c)–(d), for MCTSF versions *Haar* and CDF 5-3, respectively, were one can notice the correct matching between the two sequences, as well as a balance between the VIS and the IR data.

Figures 3(e)–(h) depict the frames resulting from the fusion of the first frames of the VIS and IR (Figures 3(e) and 3(f)) sequences OCTBVS OSU:Dataset06 [5], respectively. Likewise, the fused frame exhibits perfectly matched edges when fused using the MCTSF *Haar* (Figure 3(g)) and the MCTSF CDF 5-3 (Figure 3(h)) versions.

While the man near a car is barely seen in the VIS frame (Figure 3(i)) of INO Park Evening video, he is clearly visible in the IR frame (Figure 3(j)), with Figure 3(k) showing the 500$^{\text{th}}$ fused frame that precisely transports the information from both spectra. However, the alignment between the two videos is only maintained for the MCTSF *Haar* version (Figure 3(k)). Despite conveying both VIS and IR information, a great number of failed correspondences can be spotted in Figure 3(l).

Frame 530 of INO Group Fight shows a man crossing a parking lot, while two other men are standing up near a parked car. Figure 3(o) shows the fused frame presenting severe misalignment between the VIS and the IR data. In contrast, Figure 3(p) exhibits a good spatiotemporal alignment, showing that the shorter *Haar* filter bank is able to mitigate eventual registration problems. The longer CDF 5-3 filter bank, on the other hand, when applied to the temporal axis, tends to deteriorate the fused sequences when there is large motion between frames. One can clearly see that the objects that remained static, or that present a low degree of motion, are perfectly fused (*e.g.*, the two men in Figure 3(o)), while the man that is walking presents misaligned edges.

Frame 1 of the EDEN [7] VIS/IR video pair is pictured in Figures 3(q) and 3(r), respectively. Their associated fused frame using the MCTSF *Haar* version, displayed in Figure 3(t), shows data from both spectra, where the structural

information is preserved. However, despite also presenting VIS and IR data, Figure 3(s) shows misaligned edges, as well as a blurred VIS data (*e.g*. the foliage) in the fused image. Once again, the MCTSF *Haar* (Figure 3(t)) better preserves the texture information, yet displaying the IR data. The MCTSF *Haar* method produces well aligned edges in the fused video, while one can spot a few misaligned edges in the fused frame, generated by the MCTSF CDF 5-3, depicted in Figure 3(s).

Frame 330 of the DCUC [17] VIS/IR video pair is pictured in Figures 3(u) and 3(v), respectively. Despite both sequences being extremely noisy, both MCTSF CDF 5-3 and *Haar* versions, displayed in Figures 3(w)and 3(x), clearly show the IR information. The texture information from the VIS sequence appears as noisy as in the visible video. The fused video presents no halo effects when using any of the filter banks.

Table 1 shows different metrics to quantify the fusion results. One critical point of these metrics is the absence of a reference ground-truth, as only visible or IR sequences are available. Fused images/video should convey the "most relevant information", which is quite subjective and application dependent. Considering that there is a set of image/video fusion metrics that atempts to reflect the actual fusion quality, we provide the metrics used by the compared algorithms. As the details provided by the compared methods were not sufficient to be reproduced, we decided to publish the values of the metrics informed in the references.

Table 1 shows the results produced by the Real-time Fusion (RTF) [26], Generalized Random Walks (GRW) [23], the Guided Filter Fusion (GFF) [14], the Motion-Compensated Wavelet Transform (MCWT) [28], the tri-dimensional Dual Tree Complex Wavelet Transform (3D-DTCWT) [22,30] and the proposed MCTSF method. It presents the MCTSF *Haar* and CDF 5-3 results for 6 (six) selected sequences from the image/video fusion datasets OCTCVBS/OSU [5], INO [1], EDEN [7] and DCUC [17]. All sequences are composed by VIS and IR video pairs, presenting challenging characteristics for video fusion algorithms. The VIS/IR video pairs in Table 1 were selected from the chosen datasets to allow the comparison with the results presented in [14, 23, 26, 28, 30]. The fused videos using the MCTSF with their qualitative analysis, as well as the source code of the algorithm are available online at **http://www.smt.ufrj.br/~fusion/videofusion**, where one can assess the spatiotemporal alignment and stability by playing the videos.

Objective methods for quality analysis of video fusion have been well studied in the literature [19–21, 29], where most of them are based only on spatial information. The Mutual Information metric (MI) [8] measures the mutual information between the fused frames and their corresponding visible and infrared frames. The Structural Similarity index (SSIM) [25] is a quality assessment method based on the degradation of structural information. As it takes into account the high frequency content, it is well suited for the evaluation of image/video fusion algorithms, since it is able to spot edge misalignment. One exception is the Dynamic Quality metric (DQ) [19], which combines the spatial information preservation estimates obtained from the current frame with temporal information preservation estimates obtained from previous and subsequent frames. Note that as both MI and SSIM metrics are calculated between the fused and each source frame (VIS and IR), the values exhibited in Table 1 are averages over all frames. All the employed metrics have been reported to present good correlation with subjective tests [8].

The results presented in Table 1 show that the MCTSF *Haar* outperforms all the other methods when assessed by the SSIM metric, with the only exception being the the SSIM_{VIS} for the DCUC dataset fused by the MCTSF CDF 5-3 method. The RTF method presents the highest MI_{VIS} values for sequences OCTBVS OSU:Dataset03 and INO Group Fight, while the MCTSF *Haar* and CDF 5-3 outperform the other methods for the remaining sequences. The MI_{IR} highest values are presented by the RTF, GFF and MCTSF for different sequences. Table 1 shows that the MCWT method outperforms the MCTSF versions for the EDEN sequence, when the fused sequences are assessed by the DQ metric, while the MCTSF *Haar* and CDF 5-3 show the highest DQ values for all other sequences. In addition, all fused videos employing the MCTSF method preserve both texture (despite the different levels of spatial information) and infrared data, yet maintaining the spatiotemporal alignment.

Although presenting interesting results, the recent work of Hu *et al*. [12] employs a different methodology to calculate the metrics, preventing the realization of fair comparisons. Therefore, its results were not considered in the quantitative analysis of this paper.

5. Conclusions

The proposed video fusion method (MCTSF) is based on the use of motion-compensated temporal sub-band decompositions, using two different filter banks and simple fusion rules. Temporal filtering is carried out in the direction given by motion vectors that are computed by using half-pixel precision motion estimation on 4×4 blocks. Our results indicate that the temporal information introduced by the use of motion-compensated sub-band decompositions tends to reduce eventual registration errors, and is quite effective in producing good quality fusion results even when a very simple fusion rule is applied to the temporal subbands. The fused videos preserve both texture and infrared information, showing that the proposed method is competitive in comparison to state-of-the-art methods when considering the DQ and MI metrics. When evaluated by the SSIM image quality metric, the MCTSF even outperforms these state-of-the-art methods.

(a) VIS (b) IR (c) CDF 5-3 (d) Haar

(e) VIS (f) IR (g) CDF 5-3 (h) Haar

(i) VIS (j) IR (k) CDF 5-3 (l) Haar

(m) VIS (n) IR (o) CDF 5-3 (p) Haar

(q) VIS (r) IR (s) CDF 5-3 (t) Haar

(u) VIS (v) IR (w) CDF 5-3 (x) Haar

Figure 3: Results achieved using two different filter banks in MCTSF. Each line shows a sample frame from the sequences OTCBVS OSU:Dataset03, OTCBVS OSU:Dataset06, INO Park Evening, INO Group Fight, EDEN and DCUC respectively.

	Method	DQ	MI_{VIS}	MI_{IR}	$SSIM_{VIS}$	$SSIM_{IR}$
	RTF	-	**2.06**	1.20	0.75	0.51
	GRW	-	1.45	1.03	0.58	0.60
OTCBVS	GFF	-	1.90	0.95	0.75	0.47
	MCWT	0.3052	-	-	-	-
	3D-DTCWT	0.3401	-	-	-	-
(OSU:Dataset03)	**MCTSF** CDF 5-3	0.4185	1.46	1.29	0.60	0.56
	MCTSF *Haar*	**0.4207**	1.48	**1.31**	**0.99**	**0.99**
	RTF	-	-	-	-	-
	GRW	-	-	-	-	-
OTCBVS	GFF	-	-	-	-	-
	MCWT	-	-	-	-	-
	3D-DTCWT	0.3073	-	-	-	-
(OSU:Dataset06)	**MCTSF** CDF 5-3	**0.4155**	**1.38**	**0.96**	0.64	0.57
	MCTSF *Haar*	0.3178	1.01	0.77	**0.99**	**0.99**
	RTF	-	1.58	**2.45**	0.46	0.75
	GRW	-	1.36	1.64	0.41	0.74
INO	GFF	-	1.56	1.63	0.48	0.74
	MCWT	-	-	-	-	-
	3D-DTCWT	-	-	-	-	-
(Park Evening)	**MCTSF** CDF 5-3	0.3868	1.69	1.93	0.54	0.75
	MCTSF *Haar*	**0.3906**	**1.82**	2.11	**0.99**	**0.99**
	RTF	-	**1.87**	2.02	0.65	0.88
	GRW	-	1.43	1.58	0.69	0.75
INO	GFF	-	1.35	**2.36**	0.54	0.77
	MCWT	-	-	-	-	-
	3D-DTCWT	-	-	-	-	-
(Group Fight)	**MCTSF** CDF 5-3	0.4172	1.17	1.33	0.69	0.77
	MCTSF *Haar*	**0.4237**	1.32	1.41	**0.99**	**0.99**
	RTF	-	1.35	0.45	0.82	0.39
	GRW	-	1.21	0.33	0.76	0.61
EDEN	GFF	-	0.20	**3.48**	0.30	0.96
	MCWT	**0.4413**	-	-	-	-
	3D-DTCWT	0.3313	-	-	-	-
	MCTSF CDF 5-3	0.3692	1.25	0.41	0.62	0.59
	MCTSF *Haar*	0.3915	**1.84**	0.43	**0.99**	**0.99**
	RTF	-	-	-	-	-
	GRW	-	-	-	-	-
DCUC	GFF	-	-	-	-	-
	MCWT	0.3907	-	-	-	-
	3D-DTCWT	0.3753	-	-	-	-
	MCTSF CDF 5-3	0.3384	**0.7919**	0.2922	**0.4208**	0.3816
	MCTSF *Haar*	**0.3911**	0.7667	**0.7677**	0.4079	**0.6625**

Table 1: Comparative results of the proposed MCTSF with the RTF [26], the GRW [23], the GFF [14], the MCWT [28] and the 3D-DTCWT [22, 30]. The best result for each metric is given in bold face.

References

[1] http://www.ino.ca/en/video-analytics-dataset. [Accessed 01/09/2017].

[2] S. H. Chan, D. T. Võ, and T. Q. Nguyen. Subpixel motion estimation without interpolation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 722–725. IEEE, 2010.

[3] S.-J. Choi and J. W. Woods. Motion-compensated 3-D subband coding of video. *IEEE Transactions on Image Processing*, 8(2):155–167, Feb 1999.

[4] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *Journal of Fourier analysis and applications*, 4(3):247–269, 1998.

[5] J. W. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(23):162 – 182, 2007. Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[6] A. Ellmauthaler, E. A. B. da Silva, C. L. Pagliari, and S. R. Neves. Infrared-visible image fusion using the undecimated wavelet transform with spectral factorization and target extraction. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2661–2664. IEEE, 2012.

[7] J. L. et al. The Eden Project multi-sensor data set. Technical report TR-UoB-WS-Eden-Project-Data-Set, University of Bristol and Waterfall Solutions Ltd, UK., 2006.

[8] M. H. et al. Image fusion metrics: evolution in a nutshell. In *Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on*. IEEE, 2013.

[9] Q. Z. et al. Similarity-based multimodality image fusion with shiftable complex directional pyramid. *Pattern recognition letters*, 32(13):1544–1553, 2011.

[10] A. Golwelkar. *Motion compensated temporal filtering and motion vector coding using longer filters*. PhD thesis, Rensselaer Polytechnic Institute, 2004.

[11] A. Golwelkar and J. W. Woods. Motion-compensated temporal filtering and motion vector coding using biorthogonal filters. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(4):417–428, 2007.

[12] H. M. Hu, J. Wu, B. Li, Q. Guo, and J. Zheng. An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels. *IEEE Transactions on Multimedia*, PP(99):1–1, 2017.

[13] J. Li, S. G. Nikolov, C. P. Benton, and N. E. Scott-Samuel. Motion-based video fusion using optical flow information. In *Information Fusion, 2006 9th International Conference on*, pages 1–8. IEEE, 2006.

[14] S. Li, X. Kang, and J. Hu. Image fusion with guided filtering. *IEEE Transactions on Image Processing*, 22(7):2864–2875, July 2013.

[15] N. Mehrseresht and D. Taubman. An efficient content-adaptive motion-compensated 3-D DWT with enhanced spatial and temporal scalability. *IEEE Transactions on Image Processing*, 15(6):1397–1412, June 2006.

[16] T. Naveen and J. Woods. Motion compensated multiresolution transmission of high definition video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 4(1):29–41, Feb 1994.

[17] C. O'Conaire, N. E. O'Connor, E. Cooke, and A. F. Smeaton. Comparison of fusion methods for thermo-visual surveillance tracking. In *2006 9th International Conference on Information Fusion*, pages 1–7, July 2006.

[18] J.-R. Ohm. Three-dimensional subband coding with motion compensation. *Image Processing, IEEE Transactions on*, 3(5):559–571, Sep 1994.

[19] V. Petrović, T. Cootes, and R. Pavlovic. Dynamic image fusion performance evaluation. In *Information Fusion, 2007 10th International Conference on*. IEEE, April 2007.

[20] V. Petrović and C. Xydeas. Objective image fusion performance characterisation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. IEEE, 2005.

[21] G. Piella. New quality measures for image fusion. In *Proceedings of the Seventh International Conference on Information Fusion*, pages 542–546, 2004.

[22] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6):123–151, Nov 2005.

[23] R. Shen, I. Cheng, J. Shi, and A. Basu. Generalized random walks for fusion of multi-exposure images. *IEEE Transactions on Image Processing*, 20(12):3634–3646, Dec 2011.

[24] W. Sweldens. Lifting scheme: a new philosophy in biorthogonal wavelet constructions. In *SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*, pages 68–79. International Society for Optics and Photonics, 1995.

[25] Z. Wang and A. C. Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 2002.

[26] J. Wu, H. M. Hu, and Y. Gao. A realtime fusion algorithm of visible and infrared videos based on spectrum characteristics. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3369–3373, Sept 2016.

[27] L. Xu, J. Du, and Z. Zhang. Image sequence fusion and denoising based on 3d shearlet transform. *Journal of Applied Mathematics*, 2014, 2014.

[28] L. Xu, J. Du, and Z. Zhang. Infrared-visible video fusion based on motion-compensated wavelet transforms. *IET Image Processing*, 9(4):318–328, 2015.

[29] Q. Zhang, Y. Chen, and L. Wang. Multisensor video fusion based on spatial–temporal salience detection. *Signal Processing*, 2013.

[30] Q. Zhang, S. Hua, R. S. Blum, and M. Chen. Video fusion performance assessment based on spatial-temporal phase congruency. *Signal Processing*, 2014.

[31] Q. Zhang, L. Wang, Z. Ma, and H. Li. A novel video fusion framework using surfacelet transform. *Optics Communications*, 2012.