J. Vis. Commun. Image R. 25 (2014) 855–863

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Bandlet-based sparsity regularization in video inpainting

Ali Mosleh [a], Nizar Bouguila [b,*], A. Ben Hamza [b]

[a] Department of Electrical and Computer Engineering, Concordia University, Montréal, QC H3G 2W1, Canada
[b] Concordia Institute for Information Systems Engineering, Concordia University, Montréal, QC H3G 2W1, Canada

ARTICLE INFO

Article history:
Received 6 July 2012
Accepted 13 January 2014
Available online 24 January 2014

Keywords:
Bandlets
Inpainting
Patch fusion
Regularization
Video completion
Spatio-temporal flows
Video sequence
Missing information

ABSTRACT

We present a bandlet-based framework for video inpainting in order to complete missing parts of a video sequence. The framework applies spatio-temporal geometric flows extracted by bandlets to reconstruct the missing data. First, a priority-based exemplar scheme enhanced by a bandlet-based patch fusion generates a preliminary inpainting result. Then, the inpainting task is completed by a 3D volume regularization algorithm which takes advantage of bandlet bases in exploiting the anisotropic regularities. The method does not need extra processes in order to satisfy visual consistency. The experimental results demonstrate the effectiveness of our proposed video completion technique.

1. Introduction

Missing parts in still images and video sequences may be caused by damages or deliberately undesired object removal from the images or the video frames. The image/video inpainting problem has attracted a great attention in the past few years due to its powerful ability in fixing and restoring damaged saptial/spatio-temporal data. In this paper, we focus on video inpainting as a technique to recover missing data in some specified regions of videos. Due to the large dimensionality of video data coupled with its saptio-temporal consistency which must be preserved, video inpainting can be considered as a challenging task even though large amount of data can be highly desirable to fill-in the missing regions.

One can refer to [1] for detailed mathematical interpolation models specialized in image inpainting. The pioneering work in digital inpainting [2] employs non-linear partial differential equations (PDEs) as an interpolation platform to perform image and video frame inpainting. The concepts of PDEs and interpolation in inpainting have been employed in many techniques, including [3] which derives a third-order PDE based on Taylor expansion to propagate the border isophotes to the missing regions. An explicit extension of the technique introduced in [2] is presented in [4] which applies Navier-Stoke equations. This approach applies ideas from classical fluid dynamics to continuously propagate isophote lines of the image from the exterior into the inpainting zone. As another technique, the proposed video inpainting scheme in [5] benefits from discrete *p-Laplacian* regularization on a weighted graph. Despite their promising results, the PDE and interpolation based methods perform frame-by-frame completion that neglects the continuity across consecutive frames unless PDEs are adapted in a 3D scheme [6]. Moreover, these methods are appropriate only for narrow and small missing regions.

The concept of priority in image inpainting introduced in [7] has been adopted in various video inpainting approaches. In these techniques, a correct order of filling-in process leads to a high performance in the completion task. Important properties, such as availability, trackability and motion vectors of the pixels, and geometric properties contribute to the calculation of the priority of the missing regions to be filled-in first. For instance, the method introduced in [8] performs moving object segmentation to separate the background and foreground of the video. Hence, the search space is reduced for completion of partially occluded moving objects [9]. In this method a motion confidence value is used to find the priority of the filling-in area in order to maintain the temporal consistency in the foreground completion task. For the background completion step, the image inpainting technique introduced in [7] is adopted. Modifications based on analysis of continuities on stationary and non-stationary videos are carried out to find the best priority in [10]. Then, in [11] the technique is further improved for various

* Corresponding author. Permanent address: Office EV-007-632, Concordia Institute for Information Systems Engineering, Concordia University, 1515 St. Catherine Street West, EV.007.632, H3G 2W1 Montreal, Canada.
E-mail addresses: mos_ali@encs.concordia.ca (A. Mosleh), bouguila@ciise.concordia.ca (N. Bouguila), hamza@ciise.concordia.ca (A.B. Hamza).

http://dx.doi.org/10.1016/j.jvcir.2014.01.007
1047-3203/

camera motions by keeping the track of similar regions. The priority is determined based on the trackability of the pixels in the method introduced in [12]. The highest priority fragment around the boundary of the missing region is completed using a graph-cut fragment updating instead of copying just a similar texture from the undamaged region. In [13] a priority-based method considers the video completion task as a global search optimization in order to find the best match. The whole video is considered as a volume and a multi-scale scheme is employed to reduce the computation time. Motion layer segmentation is the key step in the method proposed in [14]. Each separate layer is completed using the image inpainting method, and then all the layers are combined in order to restore the final video. A two phase sampling and alignment video inpainting technique is introduced in [15]. The method predicts motion data in the foreground, then missing moving foreground pixels are reconstructed by spatio-temporal alignment of the sampled data. Then, the background inpainting is done by 3D tensor voting as an extension of the still image repairing technique introduced in [16]. The methods in [17,18] proposed to inpaint videos by transferring sampled motion fields from the available parts of the video. The latter method tracks patches containing missing regions in the adjacent frames by employing a global motion estimation scheme. In [19,20] 3D patch-based probability models with potential applications in video inpainting are introduced. The probability model introduced in [19] is an alternative for motion models such as optical-flow. A sparsity-based prior for a variational Bayesian model is defined for video sequences. The damaged portion of a video can be treated in this Bayesian framework as an inpainting task. A learning strategy in [20] on the video's 3D space–time patches leads to video epitomes. Epitomes are viewed as a set of 3D arrays of probability distributions applied for video reconstruction. Although the preliminary results of video inpainting using these methods are promising, they need more improvements to be able to deal with large missing portions.

Maintaining the visual consistency along with handling the large dimensionality of videos in the inpainting process is an important fact. No wonder we see complicated steps in the state-of-the-art techniques, such as segmentation of different motion layers or objects, foreground/background separation, tracking, optical-flow mosaics computation and so onto cope with spatio-temporal consistency. In this paper we propose an approach that takes advantage of the bandlets sparse representation to reconstruct missing data visually pleasingly. Image sparse representation methods were introduced for spatial inpainting problems [21–23]. In such methods, missing pixels are inferred by adaptively updating the sparse representation (e.g. wavelets, DCT, etc). Although these approaches are very challenging to be adapted to video completion that deals with unsound and damaged estimated motion vectors, they yield satisfactory results in the case of image inpainting. Apparently, employing an efficient sparse representation can enhance the inpainting results. The main motivation behind employing the bandlet domain is due to its effective capability in capturing the geometric properties of an image as an efficient sparse representation [24]. The captured geometric features are used in our technique to firstly blend the results of patch matching in order to keep the visual consistency. Secondly, the overall bandlet geometry of the frames can be a good prior if we consider the video inpainting as an ill-posed linear problem. The obtained overall geometry is used for sparse regularization to reconstruct the video. In our method, making distinction between static camera videos and sequences containing camera motions is not needed. Besides, there is no segmentation, tracking or complex motion estimation as applied in many of the previously discussed methods to facilitate the inpainting process. This is the main difference with our previous work [25] that relies on an accurate background/foreground segmentation in order to treat videos captured by static and moving cameras in different fashions by patch matching rather than bandlet-based patch fusion and 3D regularization.

The rest of this paper is organized as follows. Section 2 describes the idea behind the bandlet transform capability in reconstructing missing regions. Then, the proposed bandlet-based video inpainting method is presented in Section 3. In Section 4, the experimental results are provided. Finally, Section 5 concludes this paper.

2. Using bandlets in inpainting

The bandlet framework can achieve an effective geometric representation of texture images. It is essential in sparse regularization and spatial or spatio-temporal data reconstruction for digital inpainting purposes.

Although geometric regularity along image edges is an anisotropic regularity, conventional wavelet bases can only exploit the isotropic regularity on square domains. An image can be differentiable in the direction of the tangent of an edge curve even though the image may be discontinuous across the curve. Bandlet transform [26] exploits such anisotropic regularity. Bandlet bases construct orthogonal vectors elongated in the direction of the maximum regularity of a function. The earlier bandlet bases [27,28] have been improved by a multi-scale geometry defined over wavelet coefficients [29,30]. Indeed, bandlets are anisotropic wavelets warped along the geometric flow.

Considering the Alpert transform as a polynomial wavelet transform adapted to an irregular sampling grid, one can obtain vectors that have vanishing moments on this irregular sampling grid. This is the principal need to approximate warped wavelet coefficients. Only a few vectors of Alpert basis can efficiently approximate a vector corresponding to a function with anisotropic regularity. This *bandletization* using wavelet coefficients is defined as

$$b^k_{j,l,n}(x) = \sum_p a_{l,n}[p]\psi^k_{j,p}(x), \tag{1}$$

where j and k represent wavelet scale and orientation, respectively. The $a_{l,n}[p]$ are the coefficients of the Alpert transform where l is the scale and n is the index of the Alpert vector. In essence, $a_{l,n}[p]$ are the coordinates of the bandlet function $b^k_{j,l,n}$. These coefficients strictly depend on the local geometric flow. Bandlet coefficient are generated by inner products $\left\langle f, b^k_{j,l,n}\right\rangle$ of the image f with the bandlet functions $b^k_{j,l,n}$. The set of wavelet coefficients are segmented in squares S for polynomial flow approximation of the geometry. For each scale 2^j and orientation k, the segmentation is carried out using a recursive subdivision in dyadic squares. A square S should be further subdivided into four sub-squares, if there is still a geometric directional regularity in the square. Apparently, only for the edge squares, the adaptive flow is needed to be computed to obtain the bandlet bases. The geometry of an image evolves through scales. Therefore, for each scale 2^j of the orientation k a different geometry Γ^k_j is chosen. The set of all geometries $\{\Gamma^k_j\}$ represents the overall geometry of an image. Each member of this set is in fact a geometry value associated to one segmentation square S. For details about bandlets the reader is referred to [26].

The image inpainting problem may be formulated as follows. An image I contains a set of missing pixels indicated by Ω and a source ($\Phi = I \setminus \Omega$) area. The goal is finding an image $\acute{I}$ such that $\acute{I}(x)$ is equal to $I(x)$ for the pixels that belong to Φ, i.e., $\acute{I}(x) = I(x)\ \forall x \notin \Omega$ while the overall geometry of $\acute{I}$ has the same geometrical regularity as that of I in Φ. In the presence of additive noise ω we have the image f with missing pixels as $f = \theta I + \omega$ where

$$\theta I(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ I(x) & \text{if } x \in \Phi. \end{cases} \tag{2}$$

A sparsity-based regularization solution for the inverse problem $f = \theta I + \omega$ was proposed in [31] as

$$\acute{I} = \arg\min_g \frac{1}{2}\|f - \theta g\|^2 + \lambda \sum_k |\langle g, \psi_k \rangle|. \tag{3}$$

This minimization has been used with the orthogonal wavelet bases ψ_k for denoising [31] where the value of λ is chosen based on the level of noise and can be set to 1 for a noise-free image. Considering the bandlets as anisotropic wavelets warped along the geometry flow, we substitute the conventional wavelet bases of Eq. (3) with the bandlet bases introduced in Eq. (1) as

$$\acute{I} = \arg\min_g \frac{1}{2}\|f - \theta g\|^2 + \lambda \sum_{j,l,n,k} \left| \left\langle g, b^{\Gamma}_{k,j,l,n} \right\rangle \right|. \tag{4}$$

where similar to Eq. (1), k and j are the number of orientations and scales of the wavelets, and l, n are the sampling grid parameters in the Alpert transform employed in the bandlet transform. As discussed in the next section, our video inpainting scheme is subject to reconstructing the missing part of the frames generated due to occlusions and/or undesired object removal. Therefore, we avoid the noise level in the above equations (i.e., $\omega = 0$) and rewrite Eq. (4) as

$$\acute{I} = \arg\min_g \sum_{j,l,n,k} \left| \left\langle g, b^{\Gamma}_{k,j,l,n} \right\rangle \right|. \tag{5}$$

This equation is indeed minimizing the ℓ^1 norm of the bandlet image representation by which we achieve a solution for the spatial inpainting problem. In the next section, we utilize this idea to develop a 3D video volume regularization algorithm as well as the effectiveness of bandlets for blending the matching results of a best match search approach in the video completion task.

3. Spatio-temporal video completion

An important task of video completion is to fill in large missing regions produced by object occlusion or undesired object removal. The large missing region completion cannot be carried out well by simply applying PDE, regularization, or other interpolation based methods. On the other hand, in the exemplar-based methods, finding a reliable area around the missing parts and also finding a proper match in the source frames toward the end of the process reduces the accuracy of the results. Therefore, a video inpainting technique is proposed here that benefits from both an exemplar-based patch matching and a sparsity regularization scheme. The process starts looking for best candidates that match a patch Ψ_p on the border of the missing region. The N best retained matching patches in the whole sequence (Fig. 1) are then fused and the resulting data replaces the missing part of the border patch. In case there is no proper match for the border patch, i.e., $N = 0$, the border patch is kept unchanged for a further process by the 3D video volume regularization to generate the final inpainting result.

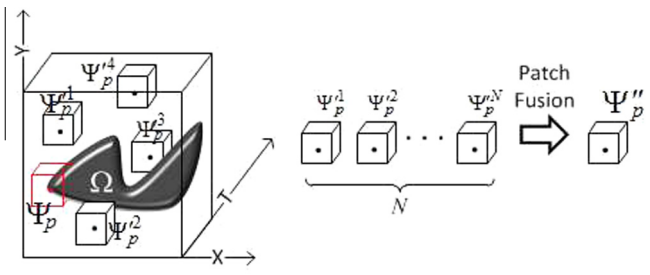

Fig. 1. Fusion strategy of patch matching results in a 3D volume video. Ψ_p lies on the missing region border. $\acute{\Psi}^1_p, \ldots, \acute{\Psi}^N_p$ are the N most similar patches to Ψ_p and Ψ''_p is the patch fusion result.

A 3D patch centered at p on the border $\partial\Omega$ of the source Φ and missing Ω regions is denoted by Ψ_p as depicted by red in Fig. 1. We search for the best match of Ψ_p in the Φ of the whole frames. The best match $\acute{\Psi}_p$ is found using sum of squared differences (SSD)

$$\acute{\Psi}_p = \arg\min_{\Psi_q \in \Phi} SSD(\Psi_q, \Psi_p), \tag{6}$$

$$SSD(\Psi_q, \Psi_p) = \sum_{(x,y,t)} \|\Psi_p(x,y,t) - \Psi_q(x,y,t)\|^2, \tag{7}$$

where for each *RGB* pixel located at (x, y) in the source region (Φ) of frame t we have a vector containing 5 elements (R, G, B, u, v). Considering (Y_x, Y_y, Y_t) as spatial and temporal derivatives of gray-scale video Y, $u = Y_t/Y_x$ and $v = Y_t/Y_y$ represent instantaneous motions in x and y directions respectively [13]. The motion information is involved in the space–time patch matching in order to preserve the motion consistency.

Unlike many of the exemplar-based methods, we do not simply replace the missing portion Ω of Ψ_p by the corresponding pixels in $\acute{\Psi}_p$. Instead, the best N matches $B_p = \left\{\acute{\Psi}^1_p, \acute{\Psi}^2_p, \acute{\Psi}^3_p, \ldots, \acute{\Psi}^N_p\right\}$ are fused using the bandlet transform as described in Section 3.1, then the fusing result pixels are copied into the missing part Ω of Ψ_p. The idea behind using several top similar patches instead of a single patch in image inpainting was presented in [32,23] by using nonlocal means and a linear blending of the patches spatially, respectively. The reason for employing a fusion framework in our video completion scheme stems from the fact that, for other border patches $\Psi_{\dot{p}}$ spatio-temporally near Ψ_p that have many pixels in common with Ψ_p the resulting set $B_{\dot{p}}$ would have many matching patches in common with B_p of Ψ_p. Therefore, their results of fusion can be very similar. Consequently, the results of inpainting for spatio-temporally close regions become reasonably consistent both spatially and temporally.

The value of N is determined using a threshold value τ. If SSD of a patch $\acute{\Psi}_p$ and Ψ_p is lower than τ, B saves $\acute{\Psi}_p$. The value of τ should not be too large to filter out many patches and at the same time it should not be too small to keep so many of them. Based on our observations we choose 0.85 as a good value for this threshold. This value may vary depending on the patch size. Also, N should not be too large to avoid unnecessary fusions. In our experiments N is limited to $N \leqslant 10$. It is worth noting that the number obtained for N indicates the degree of reliability of the best matching patches found for Ψ_p. A lower value of N means Ψ_p is not frequently repeated in the entire frames and consequently the obtained matches are not quite reliable for Ψ_p. This case happens frequently in inpainting of scenes captured by a static camera where the goal is reconstructing the missing region after a stationary object removal. Therefore, we leave a border patch Ψ_p intact once the length N of its B_p set is 0 (i.e., $\forall \Psi_q \in \Phi,\ SSD(\Psi_p, \Psi_q) > \tau$).

The priority of filling-in process is very important in the exemplar patch matching. We give the highest priority to a border patch Ψ_p that contains more reliable pixels, lies on the continuation of textures and also lies on the moving regions of the video comparing to other patches. The reliability of pixels in the border patch is measured by the confidence value given by

$$C(p) = \left(\sum_{q \in \Psi_p \cap \Phi} C(q) \right) / |\Psi_p|. \tag{8}$$

This parameter is adopted from [7] for the 3D patches, where $|\Psi_p|$ is the volume size of Ψ_p. In this equation and the equations that appear hereafter, $\Psi_p \cap \Phi$ indicates pixels of the border patch Ψ_p that lie in the source pixels Φ of the video. In the initialization, the confidence value is set to 1 for the pixels in the source region and 0 for the pixels in the missing area, i.e. $C(p) = 0 \forall p \in \Omega$ and $C(p) = 1 \forall p \in \Phi$. A patch centered at p on the border $\partial\Omega$ with already more filled-in pixels has a larger confidence than those of other

patches. The number of edge pixels can be used to measure the structural information contained in the patch. This is obtained by means of the already computed spatial derivatives Y_x and Y_y. Suppose $\acute{Y}_x$ and $\acute{Y}_y$ represent 0-1 maps of thresholded horizontal and vertical derivatives of the entire frames, respectively. Instead of manually defining threshold values to generate these two binary maps, Otsu's method can be used to find a proper threshold. Then, the structural data value of Ψ_p is defined as

$$D(p) = \left(\sum_{q \in \Psi_p \cap \Phi} \acute{Y}_x(q) \vee \acute{Y}_y(q) \right) / |\Psi_p|. \tag{9}$$

Similarly, $\acute{Y}_t$ that contains 0-1 maps of temporal derivatives is used to determine the motion data value of a border patch,

$$M(p) = \left(\sum_{q \in \Psi_p \cap \Phi} \acute{Y}_t(q) \right) / |\Psi_p|. \tag{10}$$

A high value of D means that the patch is placed on the continuation of a highly textured region. Also, a large value of M indicates a large number of moving pixels with large motion vectors in the border patch. The priority of a border patch is obtained as follows

$$P(p) = C(p) \times D(p) \times M(p). \tag{11}$$

A border patch Ψ_p with the highest $P(p)$ is chosen from the whole frames to be filled-in first. Once the patch matching is carried out, the confidence value is updated as $\acute{C}(p) = \alpha C(p)$ where $0 < \alpha < 1$. The derivative matrices $\acute{Y}_x$, $\acute{Y}_y$ and $\acute{Y}_t$ are also updated by copying the derivative values of $\acute{\Psi}_p$ into the corresponding locations in $\Psi_p \cap \Omega$. Then the process is repeated for a new highest priority border patch until there is no border patch unprocessed.

The resulting video sequence containing unfixed regions (i.e. those with unreliable matches) are passed to the sparsity regularization inpainting stage for further processes as discussed in Section 3.2.

3.1. Patch fusion

Multi-scale decomposition (MSD) based image fusion schemes, especially wavelet-based ones, have a great performance compared to regular methods [33]. However, as discussed in Section 2, due to its capability to capture more complicated geometric flows and structural information in images, the bandlet transform is much more appropriate than wavelet transform for analysis and synthesis of edges and textures [34]. Hence, we design a fusion scheme based on bandlets to blend the best patch search results.

Fig. 2 shows the proposed image fusion scheme. Consider I_1 to I_M as M images of a single scene captured from M different sources (e.g., cameras, sensors, etc.), the bandlet transform is applied on each I_i to obtain the geometric features Γ_i in the form of real numbers and bandlet coefficients C of each image. Now we need to generate a fused set of geometry flows and bandlet coefficients.

The fused geometry flow set Γ_F is computed as follows

$$\Gamma_F = \left(\sum_{i=1}^{M} j_i \Gamma_i \right) \Big/ \left(\sum_{i=1}^{M} j_i \right), \tag{12}$$

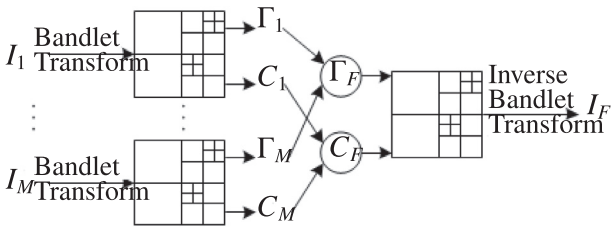

Fig. 2. Bandlet-based fusion framework for M source images.

Fig. 3. Fusion result for 3 different source images of Barbara. (a)–(c) Source images. (d) Resulting fused image.

where j_i is 0 if mean μ_i of the values of Γ_i is lower than a threshold σ. The value of σ is chosen as the mean of all $\mu_1, \mu_2, \ldots, \mu_M$. Indeed, this thresholding leads to applying only the highly structurally similar source images to produce the fused geometry. The most similar Γ of the source images are selected and their mean value generates Γ_F. The fused bandlet coefficients' set is calculated as

$$C_F = \left(\sum_{i=1}^{M} C_i \right) / M. \tag{13}$$

It is worth mentioning that the bandlet coefficients C and the geometric features Γ are produced for $l,\ n,\ j,\ k$ scales and orientations of Eq. (1).

The inverse bandlet transform is performed on Γ_F and C_F in order to generate the fused image from the M source images. Fig. 3(d) shows an example of the bandlet based fusion result for 3 source images, where Barbara's image is manually blurred and the resulting images are considered as the source images depicted in Fig. 3(a)–(c).

Now consider the set B_p of the N best matching patches obtained for Ψ_p in the proposed video inpainting technique in Section 3. Each $\acute{\Psi}_p^i$ of B_p has a size of $X \times Y \times T$. The corresponding spatial planes of patches in B_p are fused using the aforementioned fusion method to produce the resulting inpainting patch Ψ_p'', i.e.,

$$\begin{aligned}
\Psi_p''(t_1) &= \text{fuse}\left(\Psi_p'^1(t_1), \Psi_p'^2(t_1), \ldots, \Psi_p'^N(t_1)\right) \\
&\vdots \\
\Psi_p''(t_i) &= \text{fuse}\left(\Psi_p'^1(t_i), \Psi_p'^2(t_i), \ldots, \Psi_p'^N(t_i)\right) \\
&\vdots \\
\Psi_p''(t_T) &= \text{fuse}\left(\Psi_p'^1(t_T), \Psi_p'^2(t_T), \ldots, \Psi_p'^N(t_T)\right)
\end{aligned} \tag{14}$$

where $\Psi_p(t_i)$ represents all the $X \times Y$ pixels at time index $t_i (1 \leqslant t_i \leqslant T)$ in the patch Ψ_p. This fusion scheme takes more structural information into account than simply copying the source (Φ) pixels of the best match $\acute{\Psi}_p^1$ to produce the final inpainting result. Besides, as mentioned earlier such patch fusion strategy followed the introduced search process, helps gain more visual consistency.

3.2. Spatio-temporal regularization using bandlets

Algorithm 1. Bandlet-based 3D video volume inpainting

1: $i = 0$ and $V^{i=0} = y$
2: **while** $|V^{(i+1)} - V^{(i)}| > \varepsilon$
3: Find $\acute{V}^i$ using Eq. (16)
4: **for** $z = 1 \rightarrow X \times Y \times T$ **do** [Update the estimate; $V^{i+1} = T_B(\acute{V}^i)$]
5: Bandlet transform on $\acute{V}_z^i$
6: Soft-thresholding Eq. (17) on $\acute{V}_z^i$ bandlet coefficients
7: Generate V_z^{i+1} by inverse bandlet transform
8: **end for**
9: $i \leftarrow i + 1$
10: **end while**

As a result of the N best patch matching strategy, the unreliable border patches (i.e. those that less likely have a match in the whole sequence or those less frequently are repeated in the frames) are recognized by the inpainting system. These kinds of patches remain unchanged in the first inpainting stage and are passed to the 3D regularization procedure introduced in the following paragraphs.

Considering the 2D minimization problem introduced in Eq. (5) as an exhaustive optimization, we adopt the *soft-theresholding* algorithm which has been used as a solution for multi-scale wavelet representation inverse problems such as denoising [35].

The overall geometry is supposed to be fixed for an estimate of the original video. The soft-theresholding function is carried out iteratively for the minimization of Eq. (5) for each plane in the 3D volume video. At each iteration, the estimate video V^{i+1} is updated as follows

$$V^{i+1} = T_B(\acute{V}^i), \tag{15}$$

$$\acute{V}^i = \begin{cases} V^i(x) & \text{if } x \in \Omega \\ y(x) & \text{if } x \in \Phi. \end{cases} \tag{16}$$

Pixels of the original video volume are represented by $y(x)$ in the above equation. T_B denotes the soft-thresholding function performed in the bandlet domain for each existing plane in $\acute{V}^i$ defined as

$$T_B(f_z) = \sum_{j,l,n,k} t_\lambda(\langle f_z, b_{j,l,n,k}\rangle).b_{j,l,n,k}, \tag{17}$$

where f_z denotes each existing plane in the video volume. For a 3D volume consisting of T frames of $X \times Y$ pixels, we consider T planes along the time, X planes along horizontal and Y planes along vertical directions. $t_\lambda(x) = \max\left(0, 1 - \frac{\lambda}{|x|}\right)x$ and the value of λ goes to 0 as the iteration number increases. $b_{j,l,n,k}$ represents the bandlet functions of various scales and orientations as in Eq. (1).

Algorithm 1 presents the details of the minimization procedure to inpaint a video volume. This algorithm stops once the difference between two consecutive estimates is less than a small value ε. One may think of applying this algorithm on each frame independently as the inpainting task. Obviously, in a video sequence the flow of motions and trajectories is very important and needs to be considered in the inpainting task to preserve the consistency. Fig. 4(a) displays the resulting video of the exemplar-based repair stage done on the original video of Fig. 7(c). This video contains black holes representing unfixed patches. Rotating the video volume around the Y axis, one can see the video volume T–Y planes. As seen for example in the T–Y plane of $X = 145$ in Fig. 4(b), pixels of the missing region do not only lie on the spatial geometric flows but also those along the time direction. As a consequence, in each iteration of Algorithm 1, the regularization is carried out on planes X–Y, T–X and T–Y denoted by $\acute{V}^i_z$. Due to limitations of a 3D illustration, the inpainting result for only a single frame is shown in Fig. 5.

4. Experimental results

Several video sequences, including some that are provided in [11,9] are used to evaluate the proposed video inpainting method.[1] This set of videos contains sequences captured by both static and moving camera. The resolution of each video sequence is 320×240. The intermediate results of the proposed two-stage video completion technique performed on a sample video sequence for one

[1] For sample video inpainting results visit: http://users.encs.concordia.ca/mos_ali/VideoInpainting/JVCIR.htm.

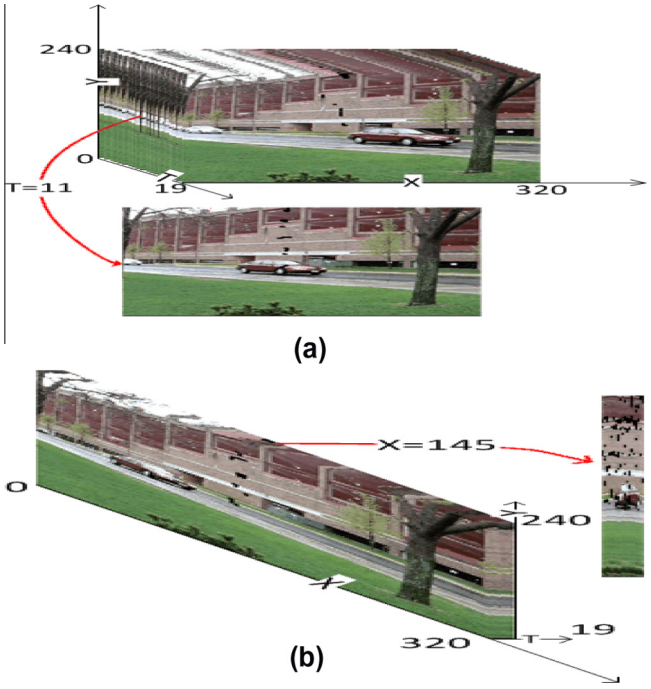

Fig. 4. A damaged video volume from different views. (a) X–Y planes view. (b) T–Y planes view.

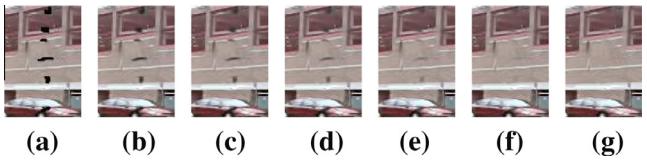

Fig. 5. Various iteration results of Algorithm 1 on the 11th frame of the video of Fig. 4. For a better illustration the images are cropped from left, right and bottom.

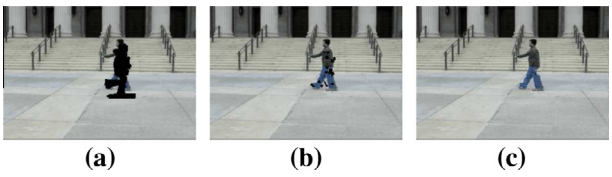

Fig. 6. The 2-stage proposed video completion method shown for a sample frame. (a) Original frame. (b) Stage 1 result: Exemplar-based patch fusion step (Section 3.1). (c) Stage 2 result: bandlet-based regularization on the result of stage 1 (Section 3.2).

of its frames are presented in Fig. 6. In the implementation, the following settings are used:

- The size of each patch is $9 \times 9 \times 5$ in the patch matching process.
- α is set to 0.5 for confidence update.
- τ is set to 0.85 to choose the N top matching patches.
- Gray-scale values of the RGB frames are found by $(R + G + B)/3$ whenever needed like instantaneous motion calculation.
- Considering a border patch Ψ_p centered at $p = (x, y, t)$, the search range is reduced to $x - 50 < x < x + 50$, $y - 50 < y < y + 50$ and $t - 7 < t < t + 7$ in the video sequence in order to avoid unnecessary search. This does not negatively

affect the patch matching result, since most likely the best patches for an arbitrary patch exist in its adjacent space and time locations.

The details of the bandlet transform applied in our technique are as follows:

- Number of scales j, on which geometry is computed, is set to 3.
- The introduced scale factor l by Alpert transform in the bandletization (Section 2) is set to 4.
- Orthogonal wavelets are used in the bandetization.
- In the wavelet transform, Daubechies wavelets are employed.
- A fixed size 8×8 segmentation is employed instead of the complex dyadic segmentation introduced in Section 2.

Fig. 7 depicts the results of our video inpainting scheme on different sequences. These videos are selected from TV, video games, and also captured by a digital camera. The objective in the sequence of Fig. 7(a) is to remove the stationary object and fill-in its missing region with proper data. Since the camera and the removed object are static, as discussed in Section 3, there is not much information about what was behind the object in the whole sequence. Therefore, the inpainting result is mostly produced by 3D regularization rather than patch matching. Other examples illustrated in Fig. 7 depict inpainting results of videos containing camera motions. In all cases, the proposed method performs the completion task quite well. In order to gain insight into the effect of each step of the proposed video completion scheme, several analyses are next presented as well as a comparison with two state-of-the-art methods.

4.1. Effects of patch fusion and 3D regularization

As mentioned before, the N best patch sorting and fusion results in a better performance in comparison to conventional patch replacement. We show this by means of a quantitative comparison.

A manual damage is generated on an original video sequence. Then, the damaged video is completed by the spatio-temporal video completion approach presented in Section 3. The completion is performed once without patch fusion, i.e, replacing the missing parts of a border patch by the corresponding pixels of the best matching patch. The spatio-temporal completion is carried-out once again by applying the introduced patch fusion technique. However, since the second stage of our proposed method (i.e., 3D regularization) is not applied in this experiment, we simply avoid the threshold τ (used to find N) and set $N = 5$. Then, for both cases, the difference of the completion result of the damaged video and the original video sequence is observed by computing the MSE value for the corresponding frames of the original and the completion result video sequences. Fig. 8(a) shows a frame of the video chosen for evaluation which is damaged as in Fig. 8(b) and then completed as in Fig. 8(c) and (d).

The plot indicated as "Exemplar Bandlet-Based Patch Fusion" in Fig. 9 shows mean square error (MSE) graph of all the 50 frames of the original video and the spatio-temporal completion result sequence using the bandlet based patch fusion. Obviously, the MSE value of the fusion-based completion for almost all the frames is lower than that of the conventional exemplar-based completion scheme labeled as "Exemplar-based" in Fig. 9. In order to show the performance of the proposed bandlet-based patch fusion technique in video completion tasks, the experiment is performed another time using another image fusion technique. A patch fusion scheme similar to Section 3.1 is considered for a well-known image fusion technique based on wavelets introduced in [36]. Then, the completion task is performed by means of the exemplar-based platform applying this fusion technique. Similar to the wavelet stage of the bandlet transform, Daubechies wavelets are employed in this wavelet-based patch fusion scheme. The resulting MSE values of all the generated frames using this method are presented as the "Exemplar-based Wavelet patch Fusion" plot in Fig. 9. The plots shown in Fig. 9 indicate visually pleasing completion results for the bandlet-based patch fusion scenario compared to simply replacing

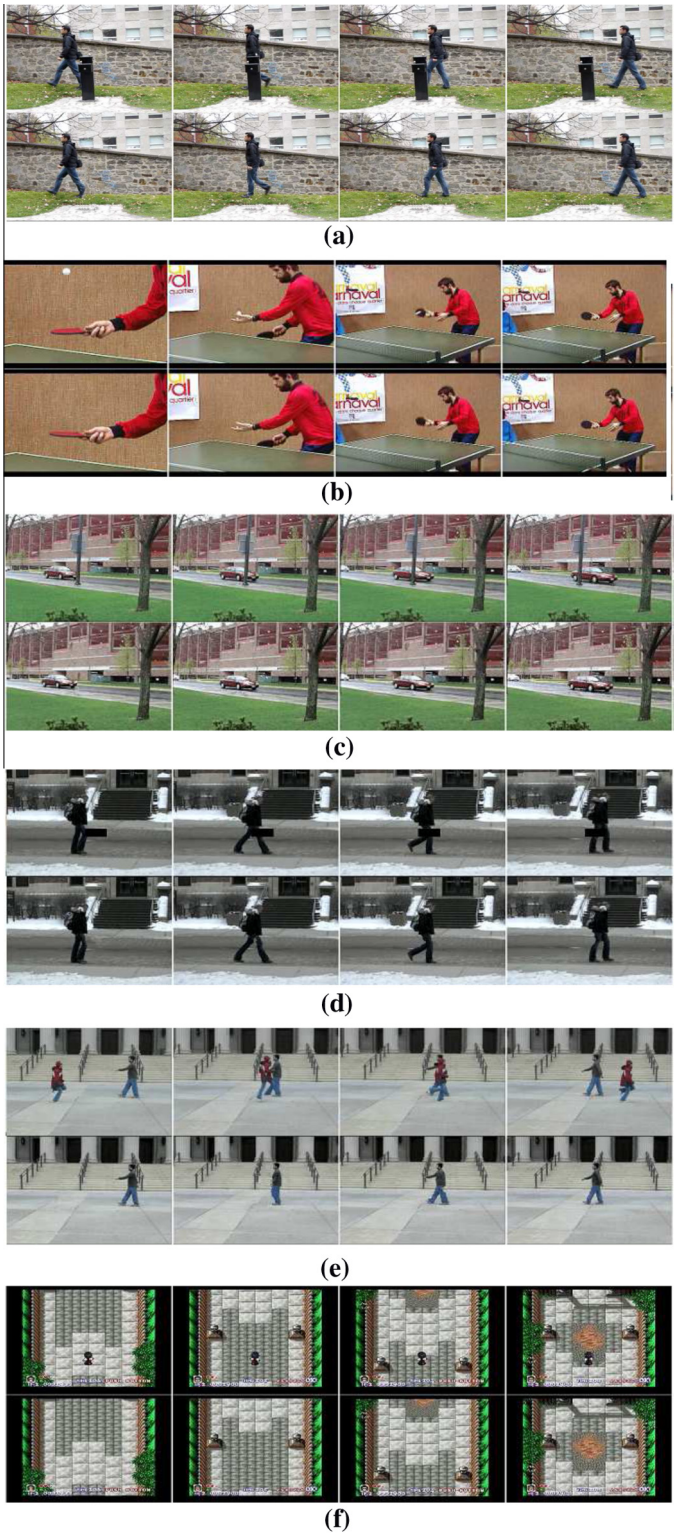

Fig. 7. Completion results for different video sequences. In each sub-figure, the top row shows the original frames and the bottom row demonstrates the corresponding inpainting results.

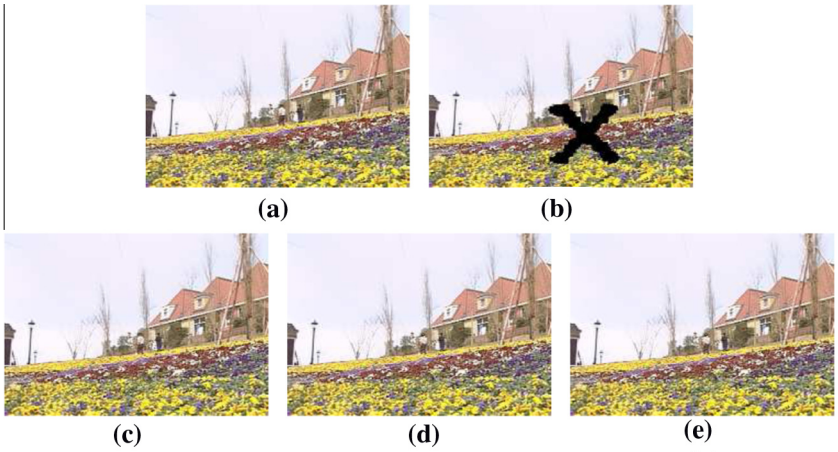

Fig. 8. (a) Original frame. (b) Damaged frame. (c) Regular exemplar-based inpainting result (Frame number 13, MSE = 19.13). (d) Patch fusion exemplar-based inpainting result (Frame number 13, MSE = 18.4). (e) Two-stage (exemplar patch fusion-based method followed by the bandlet-based 3D regularization) inpainting result (Frame number 13, MSE = 11.86).

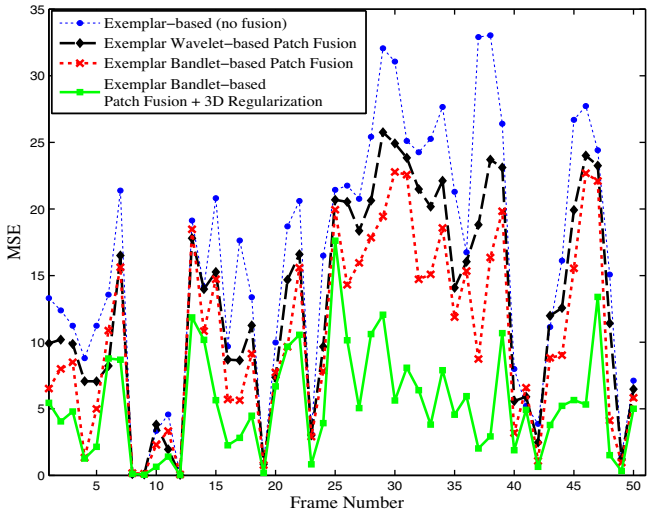

Fig. 9. Objective evaluation of patch fusion and 3D regularization in video inpainting.

the missing region by the best matching patch and also using an effective fusion method [36] based on wavelets.

Similar experiments are carried out in order to evaluate the effectiveness of bandlet-based 3D regularization in the inpainting task. This time, the proposed two-stage video inpainting method is carried-out for the video sequence of Fig. 8. In other words, the damaged video of Fig. 8(b) has been inpainted using spatio-temporal patch-fusion followed by the 3D regularization step in order to refine the results and also to preserve the visual consistency (Fig. 8(e)). The corresponding MSE plots in Fig. 9 show a higher performance for the proposed video completion method compared to using solely the patch fusion scheme or the convectional exemplar-based video inpainting technique presented in Section 3. It is worth mentioning again that the regularization methods are not practical for large regions due to the blur effect they impose on the resulting frames [7]. However, as presented here a precise combination of a regularization-based method and an exemplar-based method can result in a higher accuracy. The majority of the run-time of our algorithm is spent on bandlet transform which lacks an optimized implementation since it is relatively new. Therefore, it is not straightforward to discuss the complexity in a precise way that can be presented in this paper. With the preset implementation the run-time may be around 3 h to finish a completion task for a typical video employed in our experiments. The run-time improvement is a challenge to be addressed in a future work.

4.2. Comparison with state-of-the-art methods

The performance of video inapinting/completion methods is generally evaluated subjectively. However, we use MSE to evaluate

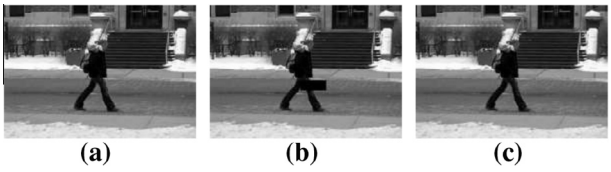

Fig. 10. (a) Original frame. (b) Damaged frame. (c) Proposed method completion result (Frame number 22, MSE = 8.18).

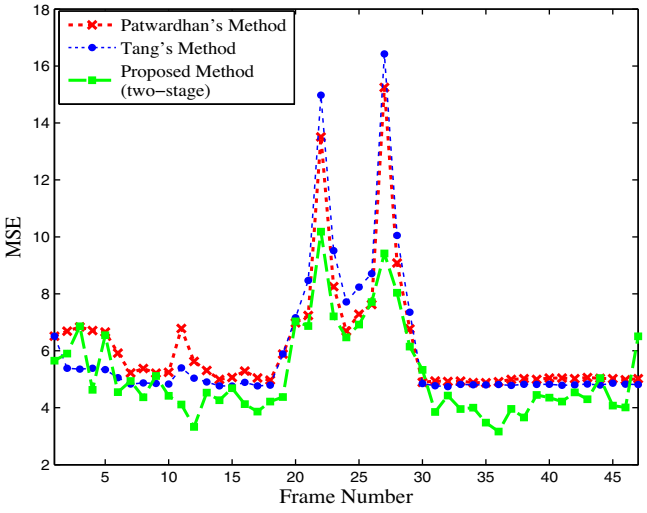

Fig. 11. Objective evaluation of the proposed video completion method. Average frame MSE is 6.11, 6.02, 5.1863 for Patwardhan et al. [9], Tang et al. [18], and the proposed two-stage method, respectively.

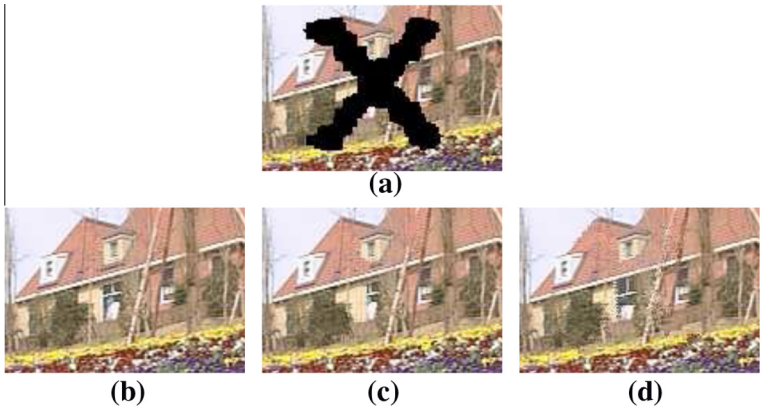

Fig. 12. A sample frame inpainted by three different methods. (a) Damaged frame. (b) The proposed algorithm result. (c) Completion result of [9]. (d) Completion result of [18]. (For a better illustration the images are cropped from left, right and bottom).

the effectiveness of our method as done in our previous experiments. A manual damage is produced on an original video sequence. Then, the result of the completion method on the damaged video is compared with the original video sequence by computing the MSE value for the corresponding frames of the original and the completion result video sequences. Fig. 10(a) shows a frame of the video chosen for evaluation which is damaged as in Fig. 10(b) and then completed as in Fig. 10(c). The green plot in Fig. 11 shows the MSE graph of all the 47 frames of the original video and the completion result sequence using the proposed method. For almost all the frames, the MSE value is low, indicating visually pleasing completion results.

We compared our approach to two well-known video completion methods introduced in [9,18]. Fig. 12 shows a sample frame of a video sequence processed by these two methods as well as by our technique. We performed the same MSE graph generation i.e., computing MSE for the completion results and the original sequence. The produced graphs are depicted in Fig. 11. The graphs and the computed average MSE values of all the frames indicate a high performance for our proposed method compared to these two methods. Despite the crucial importance of temporal consistency in video completion, to the best of our knowledge, none of the existing techniques have been evaluated objectively in the literature in this sense. This is due to the fact that there is no standard temporal quality measurement framework designated for video inpainting. Here, we employ the spatio-temporal most apparat distortion (STMAD) model to analyse our approach with regards to temporal consistency [37]. In fact, the extension of the still image-based most apparat distortion (MAD) model [38] by taking the motion information between frames into account is the main idea of STMAD. Table 1 presents STMAD values obtained for the completed videos by the three different techniques. The STMAD is calculated between the inpainted video and the original one of Fig. 10. The obtained values are normalized to the range of 0 to 1 and then they are subtracted from 1. Hence, a higher value in the table indicates a better consistency. As the table indicates, our approach has the highest value for STMAD and consequently the best temporal consistency among the other methods. This high performance is largely credited to the effective role of bandlets in the patch-fusion scheme in the spatio-temporal completion and the 3D regularization and a good combination of these two different stages.

Table 1
Temporal consistency evaluation. STMAD obtained for each resulting video using different video completion techniques.

Method	Patwardhan et al. [9]	Tang et al. [18]	Ours
STMAD	0.501	0.484	0.601

5. Conclusions

We have presented a video inpainting approach that effectively benefits from the geometric features represented by bandlets. The conventional exemplar-based video completion is modified and followed by a 3D regularization in order to perform the inpainting task. The patch search is carried out using the pixel values and instantaneous motion information. Then, the best matching patches are blended by a bandlet-based fusion framework to fill in the border patch. The fusion procedure employs the geometric flows and texture structures revealed by the bandlet transform. Afterwards, since some patches remain unchanged in the generated video, a 3D regularization based on bandlets refines the inpainting results. This is performed by enforcing the sparseness of the bandlet image representation through a minimization over the bandlet coefficients. The minimization is done iteratively by a soft-thresholding scheme in the video volume.

Unlike many existing video completion methods, our approach does not require background/foreground segmentation, decomposition of motion layers, tracking and/or optical-flow mosaics computation. Moreover, the experimental results indicate a high performance of our video inpainting approach in preserving the spatio-temporal consistency, and consequently in reconstructing the videos visually pleasingly.

References

[1] T.F. Chan, J. Shen, Mathematical models for local nontexture inpaintings, SIAM J. Appl. Math. 62 (2001) 1019–1043.
[2] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in: Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2000, pp. 417–424.
[3] M. Bertalmio, Strong-continuation, contrast-invariant inpainting with a third-order optimal pde, IEEE Trans. Image Process. 16 (2006) 1934–1938.
[4] M. Bertalmio, A.L. Bertozzi, G. Sapiro, Navier–Stokes, fluid dynamics, and image and video inpainting, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. I355–362.
[5] M. Ghoniem, Y. Chahir, A. Elmoataz, Geometric and texture inpainting based on discrete regularization on graphs, in: Proc. 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 1349–1352.
[6] H. Grossauer, O. Scherzer, Using the complex Ginzburg-Landau equation for digital inpainting in 2D and 3D, in: Proc. of the 4th International Conference on Scale Space Methods in Computer Vision, in: LNCS, vol. 2695, Springer, 2003, pp. 225–236.
[7] A. Criminisi, P. Perez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Trans. Image Process. 13 (2004) 1200–1212.
[8] K. Patwardhan, G. Sapiro, M. Bertalmio, Video inpainting of occluding and occluded objects, in: Proc. IEEE International Conference on Image Processing (ICIP), 2005, pp. II69–72.

[9] K.A. Patwardhan, G. Sapiro, M. Bertalmio, Video inpainting under constrained camera motion, IEEE Trans. Image Process. 16 (2007) 4545–4553.
[10] T.K. Shih, N.C. Tang, W.-S. Yeh, T.-J. Chen, W. Lee, Video inpainting and implant via diversified temporal continuations, in: Proc. 14th annual ACM International Conference on Multimedia, 2006, pp. 133–136.
[11] T.K. Shih, N.C. Tang, J.-N. Hwang, Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity, IEEE Trans. Circuits Syst. Video Technol. 19 (2009) 347–360.
[12] Y.-T. Jia, S.-M. Hu, R.R. Martin, Video completion using tracking and fragment merging, Visual Comput. 21 (2005) 601–610.
[13] Y. Wexler, E. Shechtman, M. Irani, Space-time completion of video, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 463–476.
[14] Y. Zhang, J. Xiao, M. Shah, Motion layer based object removal in videos, in: Proc. Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTIONS'05), 2005, pp. 516–521.
[15] J. Jia, Y.-W. Tai, T.-P. Wu, C. Tang, Video repairing under variable illumination using cyclic motions, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 832–839.
[16] J. Jia, C.-K. Tang, Image repairing: robust image synthesis by adaptive nd tensor voting, in: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2003, pp. 643–650.
[17] T. Shiratori, Y. Matsushita, X. Tang, S.B. Kang, Video completion by motion field transfer, in: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 411–418.
[18] N. Tang, C.-T. Hsu, C.-W. Su, T. Shih, H.-Y. Liao, Video inpainting on digitized vintage films via maintaining spatiotemporal continuity, IEEE Trans. Multimedia 13 (2011) 602–614.
[19] X. Li, Y. Zheng, Patch-based video processing: a variational bayesian approach, IEEE Trans. Circuits Syst. Video Technol. 19 (2009) 27–40.
[20] V. Cheung, B.J. Frey, N. Jojic, Video epitomes, in: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 42–49.
[21] O. Guleryuz, Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising-part i: theory, IEEE Trans. Image Process. 15 (3) (2006) 539–554.
[22] O.G. Guleryuz, Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising-part ii: adaptive algorithms, IEEE Trans. Image Process. 15 (3) (2006) 555–571.
[23] Z. Xu, J. Sun, Image inpainting by patch propagation using patch sparsity, IEEE Trans. Image Process. 19 (5) (2010) 1153–1165.
[24] A. Mosleh, N. Bouguila, A.B. Hamza, A video completion method based on bandlet transform, in: Proc. IEEE International Conference on Multimedia and Expo (ICME), Barcelona, Spain, 2011, pp. 1–6.
[25] A. Mosleh, N. Bouguila, A. Ben Hamza, Video completion using bandlet transform, IEEE Trans. Multimedia 14 (6) (2012) 1591–1601.
[26] S. Mallat, G. Peyre, A review of bandlet methods for geometrical image representation, Numer. Algorithms 44 (2007) 205–234.
[27] E.L. Pennec, S. Mallat, Sparse geometric image representations with bandelets, IEEE Trans. Image Process. 14 (2005) 423–438.
[28] E.L. Pennec, S. Mallat, Bandelet image approximation and compression, SIAM Multiscale Model. Simul. 4 (2005) 992–1039.
[29] S. Mallat, G. Peyre, Surface compression with geometric bandelets, ACM Trans. Graphics 24 (2005) 601–608.
[30] S. Mallat, G. Peyre, Orthogonal bandlets bases for geometric image approximation, Commun. Pure Appl. Math. 61 (2008) 1173–1212.
[31] D.L. Donoho, J.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (3) (1994) 425–455.
[32] A. Wong, J. Orchard, A nonlocal-means approach to exemplar-based inpainting, in: Proc. IEEE International Conference on Image Processing (ICIP), 2008, pp. 2600–2603.
[33] Z. Zhang, R. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proc. IEEE 87 (8) (1999) 1315–1326.
[34] X.Q.J. Yan, G. Xie, Z. Zhu, B. Chen, A novel image fusion algorithm based on bandelet transform, Chin. Opt. Lett. 5 (2007) 569–572.
[35] J. Starck, M. Elad, D. Donoho, Redundant multiscale transforms and their application for morphological component separation, Adv. Imaging Electron Phys. 132 (2004) 287–348.
[36] G. Pajares, J.M. de la Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (2004) 1855–1872.
[37] P. Vu, C. Vu, D. Chandler, A spatiotemporal most-apparent-distortion model for video quality assessment, in: Proc. IEEE International Conference on Image Processing (ICIP), 2011, pp. 2505–2508.
[38] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010) 011006-1–011006–21.