DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-193

Video inpainting of complex scenes based on local statistical model

Voronin V.V.[(a)], Sizyakin R.A.[(a)], Marchuk V.I.[(a)], Yigang Cen[(b)], Galustov G.G.[(c)], Egiazarian K.O.[(d)]; (a) Don State Technical university, Dept. of Radio-Electronics Systems, Gagarina 1, Rostov on Don, Russian Federation; (b) Institute of Information Science, Beijing Jiaotong University, Beijing, China; (c) Southern Federal University, Nekrasovski 44, Taganrog, Russian Federation; (d) Dept. of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 1, Tampere, Finland FI-33720

Abstract

This paper describes a framework for temporally consistent video completion. Proposed method allow to remove dynamic objects or restore missing or tainted regions present in a video sequence by utilizing spatial and temporal information from neighboring scenes. The algorithm iteratively performs following operations: achieve frame; update the scene model; update positions of moving objects; finding a set of descriptors that encapsulate the information necessary to reconstruct a frame; replace parts of the frame occupied by the objects marked for remove with use of a 3D patches. In this paper, we extend an image inpainting algorithm based texture and structure reconstruction by incorporating an improved strategy for video. Our algorithm is able to deal with a variety of challenging situations which naturally arise in video inpainting, such as the correct reconstruction of dynamic textures, multiple moving objects and moving background. Experimental comparisons to state-of-the-art video completion methods demonstrate the effectiveness of the proposed approach.

Introduction

The problem of automatic video reconstruction in general, and automatic object removal and modification in particular, is beginning to attract the attention of many researchers. This is problem refers to a field of computer vision that aims to remove objects or restore missing or tainted regions present in a video sequence by utilizing spatial and temporal information from neighboring scenes. Video signals often contain static images which may hide some useful information. There are a lot of examples of such images like different channel logos, date, time or subtitles that are superimposed on the video with further coding. A possible application of video inpainting techniques may be the concealment of errors and lost blocks in decoded bit streams caused by lossy compression performed by a video coder and media data transmission artifacts. In some cases there may be sophisticated video removal (of undesired static or dynamic objects) by completing the appropriate static or dynamic background information on the video sequence. Here, the term object refers to a connected region of pixels. The example of such object can be a moving car or person, the defect caused by a scratch on the film or the entire background scene.

The task of video repairing is related to the problem of image inpainting. The generic goal of replacing areas of arbitrary shapes and sizes in images by some other content was first presented by Masnou and Morel in [1]. This method used level-lines to disocclude the region to inpaint. The term “inpainting” was first introduced by Bertalmio et al. in [2]. A number of algorithms for automatic still image completion have been proposed in the literature [2-5]. There has also been some preliminary work on frame-by-frame partial differential equations (PDEs) based video inpainting [3]. This method does not take into account the temporal information that a video provides, and its application is thereby limited. This approach is justified in removing the defects of the film. Many types of defects appear only on one frame, and absent from his neighbors. The virtue of this method is its simplicity. This method is applicable only to small objects, its application to large areas leads to unsatisfactory results.

Another group of an image inpainting methods is related field of texture synthesis. For texture synthesis the region can be much larger with the main focus being the filling in of two-dimensional repeating patterns that have some associated stochasticity i.e. textures. Structural properties, such as edges of an objects, are extracted from the spatial domain and used to complete an object with its structural property extended [5]. Some of an image inpainting techniques can complete holes based on both spatial and frequency features [6].

Subsequently, a vast amount of research was done in the area of image inpainting, and to a lesser extent in video inpainting. The only difference is the necessity to maintain temporal continuity in addition to the spatial continuity. The temporal information is either considered by using a segmentation of video objects (e.g. tracking) or a global coherency in space-time patches.

Existing methods of video inpainting can be divided into several classes:

1) There are approaches similar to methods of static images inpainting. The main varieties: the methods based on solution of partial differential equations in partial derivatives (PDE); the methods based on orthogonal transformations; the methods based on texture synthesis.

2) Methods that use the space-time recovery. Provide good quality restoration, but usually quite costly in terms of computation.

3) Methods, separating the original video sequence to a set of layers (in simple case background and foreground), each layer of restoring individually and perform compound-treated layers.

The fact that video inpainting dealing with moving objects in time and must consider not only the spatial continuity of such objects, but also their temporal continuity. In this regard, a simple application of inpainting approaches designed for images sequentially to each frame leads to unsatisfactory results. One of problems is the appearance of so-called "ghosts" [7]. A small change of lighting or the movement of surrounding pixels can lead to a significant change in the result of recovery.

The problem of video inpainting can be divided into the following categories [7]: stationary video with moving objects; non-stationary video with still objects; non-stationary video with moving objects (could be occluded), including all camera motions.

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-193

A method for space-time completion of damaged areas in a video as a global optimization problem was proposed in [8]. This work extends the technique of nonparametric sampling [9]. The filling of the hole can be performed based on globally optimized method also. The approach proposed by Wexler et al. [10] solves the video inpainting problem of static camera videos based on the optimization of a well-defined energy function. The inpainting is performed by optimizing the cost function expressing the local consistencies by using a weight of the global completion quality provided by each possible pixel value. The main disadvantage of this approach is the assumption that objects move in a periodic manner and also they do not significantly change scale. Also, the camera is static for this method and the processing time is high even for low resolutions videos.

A probabilistic video inpainting method has been proposed in [11]. In this method define "epitomes" as patch based probability models that used to synthesize data in the areas of video damage or object removal. The approach is very computational complexity and more suitable for low-resolution videos.

A method for repairing damaged video has been proposed in [12] but it is semi-automatic approach. The user has to manually draw the boundaries of the different depth layers of the sequence. For reconstruction moving objects use synthesis process and calculation interpolated trajectory. A related algorithm, also combining motion layer estimation and segmentation, has been reported in [13]. The complexity of the search for the best matching patches has been reduced in [14] by using an extension of the patch matching algorithm to the video.

The method proposed in [15] is a global optimizing inpainting approach with low computational complexity by tracking every pixel, but this approach can handle only translational or periodic objects motions. The approach proposed in [16] searches the optimal displacements so-called shift map which is a vector fields of the correspondences between missing pixels and their corresponding unoccluded values.

The method introduced in [17] reconstructs the motion of people in videos based on the patches similarity in terms of texture and motion. This approach allows reducing the time complexity of the patches matching search based on 2D skeleton model of each tracked person in the video. It is shows correct results only for cyclic object motions.

Other approaches based on transfer of motion fields into the missing area by propagating motion vectors [18] or by using motion similarities between patches [19]. These methods are likely to suffer from smoothing artifacts after few frames making the approach not well suited for completion over a high number of frames.

The approach based on video segmentation into moving foreground objects and background has been introduced by Patwardhan et al. in [20] by extending examplar-based image inpainting approach. To inpaint the stationary background a relatively simple spatio-temporal priority scheme is employed where undamaged pixels are copied from frames temporally close to the damaged frame, followed by a spatial filling in step which replaces the damaged region with a best matching patch so as to maintain a consistent background throughout the sequence. This algorithm provides high-quality visual recovery, but demanding of computing resources in the search for similar patches.

This approach was extended to process video sequences in [21] where the authors have attempted to provide both spatial and temporal continuity. Searching of similar patch is performed not only on the current frame, but throughout the video sequence, or in some bounded area of it. In [22-24] have been made some tries to use various optimizations: object tracking, mosaic images, separation of video sequence to set of moving.

The main drawbacks of the known methods come from the fact that the most of them are unable to recover the curved edges of object. It should be also noted that these methods often blur image in the recovery of large areas with missing pixels. Additionally, the methods suffer from the lack of global visual coherence especially for large holes. For most of the methods, both periodic motion for each occluded object and accurate segmentation of moving objects and static background are often necessary to provide pleasing results. Otherwise, segmentation errors lead to severe artifacts. Furthermore, if the inpainting of moving objects is performed independently of the background, the blending between the completed foreground and the background can look unnatural. Most of these methods are computationally very demanding and inappropriate for implementation on modern mobile platforms.

In this paper we propose a framework for video reconstruction, aimed at achieving high-quality results in the context of film post-production. Our proposed method builds on existing exemplar-based techniques and extends them to process videos. We propose to use a set of descriptors that encapsulate the information of periodical motion of objects necessary to reconstruct missing/corrupted frames. For background restorations used set of 3D patches.

The proposed video inpainting method

A discrete frame defined on a $I \times J$ rectangular grid is denote $\left\{Y_{i,j}^t\right\}\left(i=\overline{1,I}, j=\overline{1,J},\ t=\overline{t,T}\right)$ and can be represented as follows: $Y_{i,j}^t = (1-M_{i,j}^t)\cdot S_{i,j}^t + M_{i,j}^t \cdot R_{i,j}^t$, where $S_{i,j}^t$ are the true image pixels; $M=\left[M_{i,j}^t\right]$ is a binary mask of the distorted values of pixels (1 - corresponds to the missing pixels, 0 - corresponds to the true pixels); $R_{i,j}^t$ are missing pixels; I is the number of rows, J is the number of columns and T is the number of the frames.

In this article we will discuss the video inpainting proposal put forward by Patwardhan et al. [21]. The special feature of this method is the ability to restore the video, shot by a moving camera. In fact, this method is a generalization of the exemplar based method in case of video sequences that adds to the spatial continuity the time. Recovery area may be different: moving object, static object and other. It can be background or foreground object. It can be blocked by other objects or can block them. The algorithm includes preprocessing stage and the two work phases. At the preprocessing stage is performed a rough segmentation of each frame in the foreground and the background. After this step some regions can still be empty - for its restoration is used to search for similar blocks of the current frame. This algorithm has some disadvantages. Searching patches in the texture restoration requires significant computational complexity to restore large texture areas. The exemplar-based methods use a non-parametric sampling model and texture synthesis. We will tackle this problem by first stage restoration using a set of descriptors that encapsulate the information of periodical motion of objects. The diagram of the proposed method is shown at Figure 1.

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-193

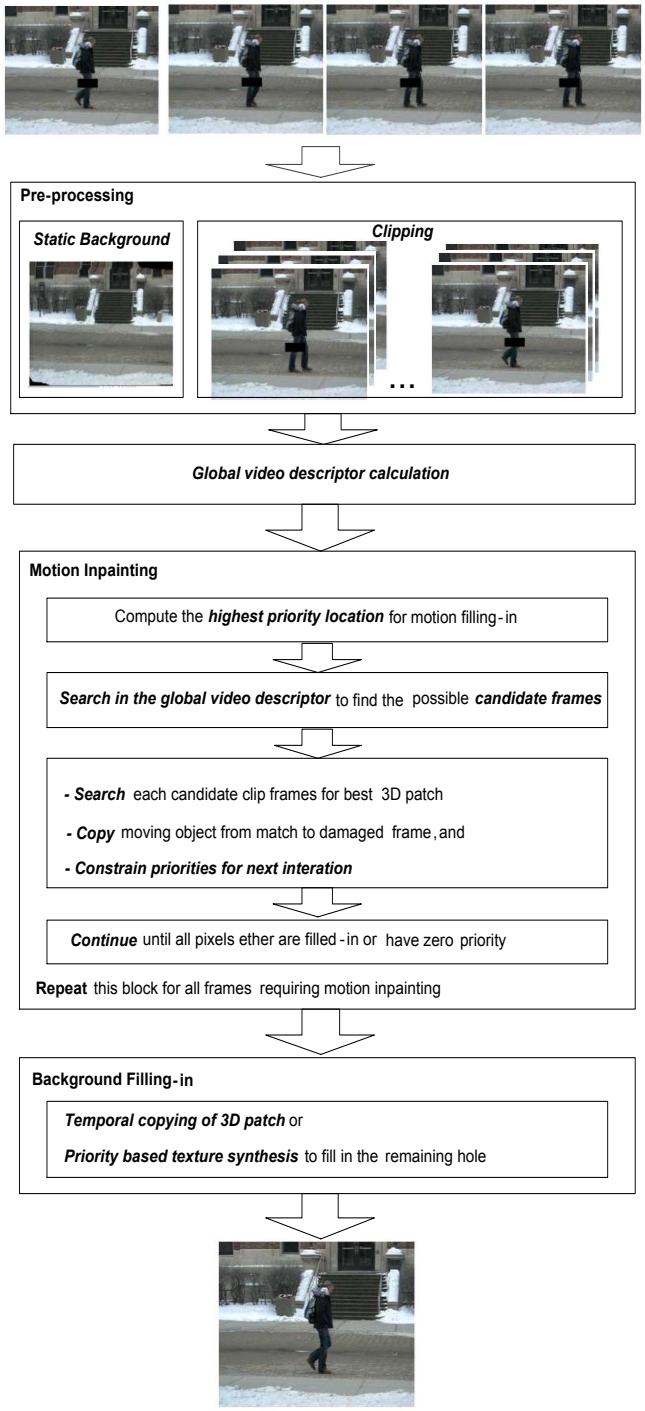


Figure 1. Algorithm of video inpainting method

Proposed approach allow to remove objects or restore missing or tainted regions present in a video sequence by utilizing spatial and temporal information from neighboring scenes. The algorithm iteratively performs following operations: achieve frame; update the scene model; update positions of moving objects; finding a set of descriptors that encapsulate the information necessary to reconstruct a frame; replace parts of the frame occupied by the objects marked for remove with use of a 3D patches.

At the preprocessing stage is performed a rough segmentation of each frame in the foreground and the background. The main idea of the proposed approach is to, rather than directly attempt to interpolate missing pixels, estimate, based on all available spatio–temporal information, the value of a set of descriptors that encapsulate the information necessary to reconstruct missing/corrupted frames.

In order to estimate the values of descriptors, we will collect the values of all the descriptors corresponding to the F_{th} clip in a vector $\mathbf{f}_k$. For this purpose we use the global descriptor is computed by applying a bank of 3D spatiotemporal filters on the frequency spectrum of a video sequence for integrates the information about the motion and scene structure. In this article we use a descriptor put forward by Solmaz et al. [25] which describes an approach to classify realistic videos of different actions.

The descriptor is generated by applying a bank of 3D spatio-temporal filters on the frequency spectrum of a sequence (Fig. 2). The bandpass nature of these filters alleviates the need for motion compensation. Furthermore, as opposed to the approaches which apply bag-of-features model, the approach preserves the spatial and temporal information, as we perform quantization in fixed spatio-temporal sub-volumes after application of each filter on the frequency spectrum and taking the inverse Fourier transform. As the filter responses for all filters on all sub-volumes are concatenated, the ordering and the length of each feature vector are identical for all video clips. The framework of this approach is shown in Fig. 3.

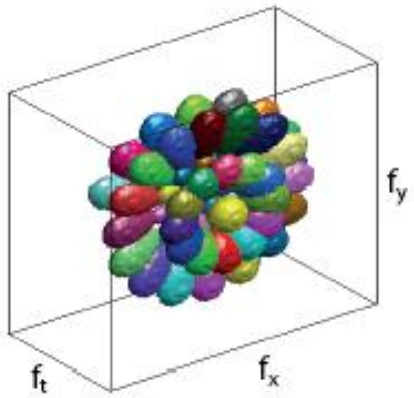


Figure 2. A bank of 3D spatio-temporal filters

The frequency spectrum computed for a video clip could capture both scene and motion information effectively, as it represents the signal as a sum of many individual frequency components. In a video clip, the frequency spectrum can be estimated by computing the 3-D discrete Fourier transform (DFT). The motion is an important element which can be representative of the type of performed action in a scene. The frequency spectrum of a two-dimensional pattern translating on an image plane lies on a plane, the orientation of which depends on the velocity of the pattern.

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-193

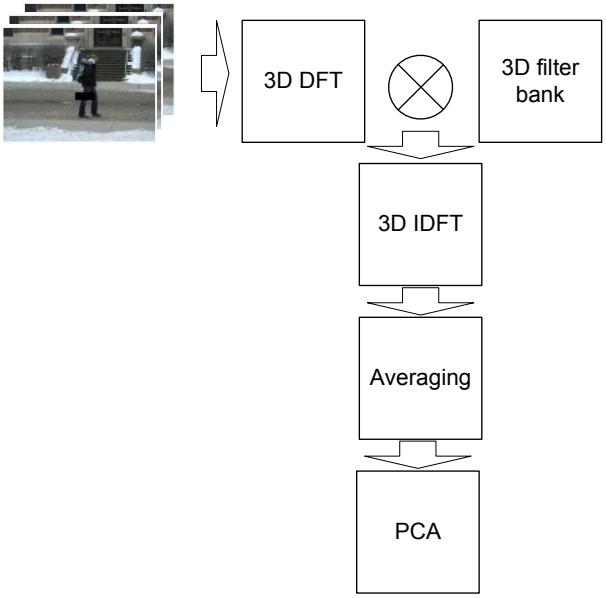


Figure 3. Algorithm of global video descriptor calculation

The global video descriptor allows finding frames with similar movement. The foreground objects to be inpainted exhibits repetitive motion 3D patches (Fig. 4). The partial objects are first completed with the appropriate object templates selected by minimizing a window-based dissimilarity measure. Between a window of partially-occluded objects and a window of object templates from the database, we define the dissimilarity measure as the Sum of the Squared Differences (SSD) in their overlapping region. The first step in treatment is the restoration of moving foreground objects, which "overlap" the restored area. After that there is recovery of the remaining area by copying blocks from adjacent frames. After this step some regions can still be empty - for its restoration is used to search for similar blocks of the current frame. This 3D patch searching is implemented using the following steps (refer to Fig. 5).

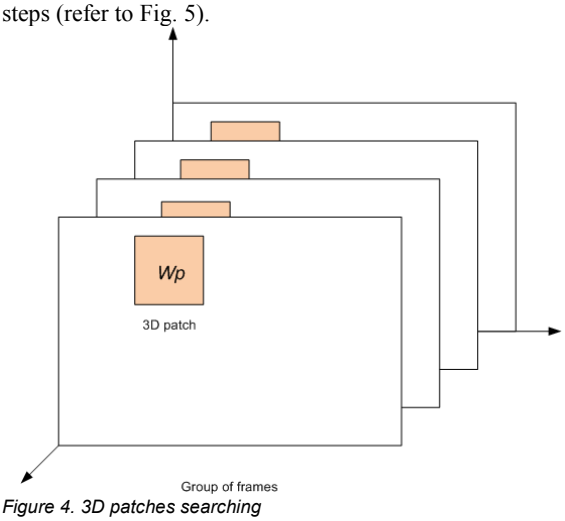


Figure 4. 3D patches searching

Pseudo-code for completing moving object in one clip (F):

- $\psi_p = getHighestPriorityPatch(F)$;
- $While\ (\psi_P\ exist)$:

  $\psi_{Pcand} = Search(\psi_P)$;

  $(f_1....f_n) = getCandidates(\psi_{Pcand})$;

  $\psi_q = getMatch(F, \psi_P, (f_1....f_n))$;

  $copy\ (\psi_q, \psi_P)$;

  $constraintPriority\ (\psi_P)$;

  $updateConfidence\ (\psi_P)$.

Repeat the above procedure for each frame requiring moving object completion.

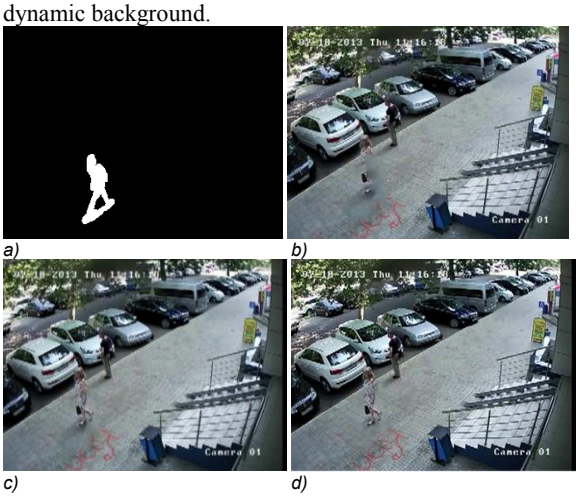
Figure 5. Pseudocode for the motion inpainting step

The proposed method has the following advantages over currently existing techniques: it leads to a non–iterative, computationally attractive algorithm that optimizes the use of (global) spatio/temporal and dynamics information and has a moderate computational burden; it is not restricted to the case of periodic motion, static background or stationary cameras; it can be used to extrapolate frames, that is extend a given video sequence, and, in the case of dynamic textures.

Experimental results

The effectiveness of the presented scheme is verified on the test frames of a video sequence with missing pixels presented. After applying the missing mask, all frames have been inpainted by the proposed method and state-of-the-art methods [8, 14].

In this example we will consider the problem of inpainting dynamic textures, e.g. sequences whose frames are relatively unstructured, but possessing some overall stationary properties. In Figures 6-9 examples of video completion (a - the image with a missing pixels, b - the restoration by the Wexler method, c - the restoration by the Newson method, d - the restoration by the proposed method) are shown. It should be observed that our technique compares favorably even in the presence of the moderate dynamic background.



a) b)

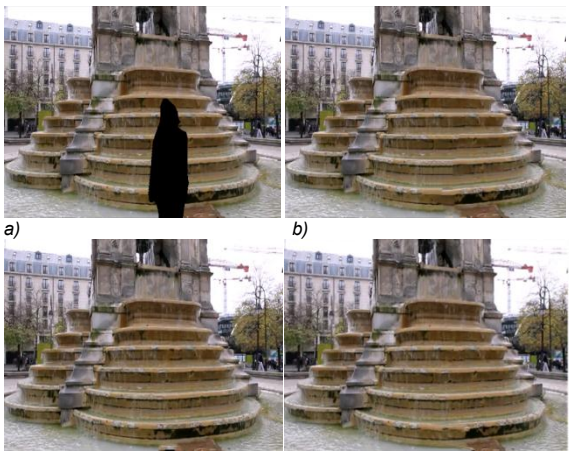
c) d)

Figure 6. Examples of video inpainting

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-193

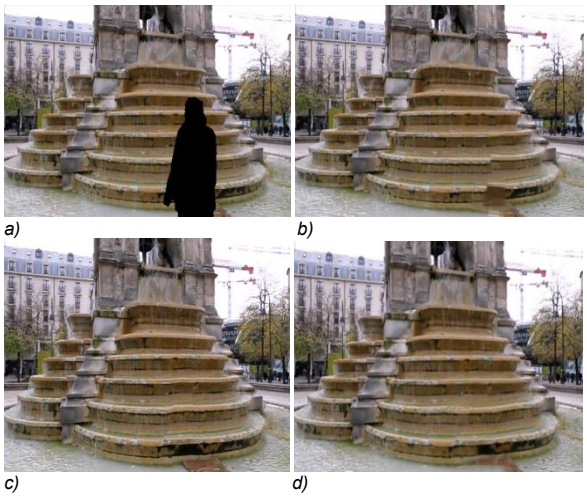
a) b) c) d)

Figure 7. Examples of video inpainting



a) b) c) d)

Figure 8. Examples of video inpainting



a) b) c) d)

Figure 9. Examples of video inpainting

Conclusion

The paper presents an video inpainting algorithm of complex scenes based on the texture and structure reconstruction of video sequence. The background is filled-in by extending spatial texture synthesis techniques based on local statistical model. Examples presented in this paper demonstrate the effectiveness of the algorithm in restoration of static background and moving foreground of the video sequences having different geometrical characteristics.

Acknowledgment

The reported study was supported by the Russian Foundation for Basic research (RFBR), research projects №15-07-99685, №15-01-09092, №15-37-21124.

References

[1] S. Masnou and J.-M. Morel, “Level lines based disocclusion,” in Int. Conf. Image Processing (ICIP), 1998.

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in Proc. ACM SIGGRAPH, pp. 417–424, 2000.

[3] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, “Navier-stokes, fluid dynamics, and image and video inpainting,” in Proc. IEEE Computer Vision Pattern Recognition, vol. 1, pp. 355–362, 2001.

[4] C. Ballester, V. Caselles, and J. Verdera, “Dissoclusion by joint interpolation of vector fields and gray levels,” SIAM Multiscale Model. Simul., vol. 2, pp. 80–123, 2003.

[5] A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based inpainting,” IEEE Trans. Image Process., vol. 9, no. 9, pp. 1200–1212, 2004.

[6] T.F. Chan, J. Shen, "Mathematical models of local non-texture inpaintings," SIAM J. Appl. Math, vol. 62(3), pp. 1019-1043, 2002.

[7] Timothy K. Shih, Nick C. Tang, and Jenq-Neng Hwang, “Exemplar-Based Video Inpainting Without Ghost Shadow Artifacts by Maintaining Temporal Continuity,” IEEE transactions on circuits and systems for video technology, vol. 19, no. 3, pp. 347-360, 2009.

[8] Y. Wexler, E. Shechtman, and M. Irani, “Space-time video completion,” in Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 120–127, 2004.

[9] A.A. Efros and T.K. Leung, “Texture synthesis by non-parametric sampling,” presented at the IEEE Int. Conf. Computer Vision, Corfu, Greece, 1999.

[10] Y. Wexler, E. Shechtman, and M. Irani, “Space-time completion of video,” IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), pp. 463–476, 2007.

[11] V. Cheung, B. J. Frey, and N. Jojic, “Video epitomes,” in IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 42–49, 2005.

[12] J. Jia, T. Wu, Y. Tai, and C. Tang, “Video repairing under variable illumination using cyclic motions,” in Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 364–371, 2004.

DOI: 10.2352/ISSN.2470-1173.2016.15.IPAS-193

[13] Y. Zhang, J. Xiao, and M. Shah, "Motion layer based object removal in videos," in Proc. Workshop on Applications of Computer Vision, pp. 516–521, 2005.

[14] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Perez, "Towards fast, generic video inpainting," in Proceedings of the 10th European Conference on Visual Media Production, pp. 7:1–7:8, 2013.

[15] Y. Shen, F. Lu, X. Cao, and H. Foroosh, "Video completion for perspective camera under constrained motion," in IEEE Int. Conf. on Image Proc. (ICIP), vol. 3, pp. 63–66, 2006.

[16] Y. Hu and D. Rajan, "Hybrid shift map for video retargeting," in Int. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 577–584, 2010.

[17] T. Shih, N. Tan, J. Tsai, and H.-Y. Zhong, "Video falsifying by motion interpolation and inpainting," in Int. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8, 2008.

[18] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), pp. 1150–1163, 2006.

[19] T. Shiratori, Y. Matsushita, X. Tang, and S. Kang, "Video completion by motion field transfer," in Int. Conf. on Computer Vision and Patternn Recognition (CVPR), pp. 411–418, 2006.

[20] K.A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting of occluding and occluded objects," in Proc. IEEE Int. Conf. Image Process., vol. 2, pp. 69–72, 2005.

[21] K.A. Patwardhan, G. Sapiro, and M. Bertalmío, "Video inpainting under constrained camera motion," IEEE Trans. Image Process., vol. 16 (2), pp. 545–553, 2007.

[22] J.-F. Aujol, S. Ladjal and S. Masnou, "Exemplar-based inpainting from a variational point of view," SIAM J. Math. Anal., vol. 44, pp. 1246–1285, 2010.

[23] F. Cao, Y. Gousseau, S. Masnou, P. Pérez, "Geometrically Guided Exemplar-Based Inpainting," SIAM J. Imaging Sci., vol. 4(4), pp. 1143–1179, 2011.

[24] V.V. Voronin, V.I. Marchuk, N.V. Gapon, A.V. Zhuravlev, S. Maslennikov, S. Stradanchenko, "Inpainting for videos with dynamic objects using texture and structure reconstruction," Proc. SPIE 9497, Mobile Multimedia/Image Processing, Security, and Applications, 94970Y, 2015.

[25] B. Solmaz, S.M. Assari, M. Shah, "Classifying web videos using a global video descriptor," Machine Vision and Applications, vol. 24 (7), pp. 1473-1485, 2013.

Author Biography

Viacheslav Voronin was born in Rostov (Russian Federation) in 1985. He received his BS in radio engineering from the South-Russian State University of Economics and Service (2006), his MS in radio engineering from the South-Russian State University of Economics and Service (2008) and his PhD in technics from Southern Federal University (2009). Voronin V. is member of Program Committee of conference SPIE. His research interests include image processing, inpainting and computer vision.

Sizyakin Roman was born in Rostov in 1989. He recieved his BS in electrical engineering from South-Russian university of economics and service in 2011. He recieved the MS degree in technics from Don State Technical University (Russian Federation) in 2013. At the time he is a PhD student at Don State Technical University. His research interests lay in the areas of digital image processing and computer vision.

Vladimir Marchuk was born in 1951. He received the D.Tech. degree in technics from Southern Federal University (Russian Federation) in 2006. Since 2006, he has been a Professor. His research interests are in the areas of applied statistical mathematics, signal and image processing.

Yigang Cen was born in 1978. He received the Ph.D. degree at School of Computer & Information Technology, Beijing Jiaotong University. His research interests are in the areas of applied mathematics, signal processing, and digital logic.

Gennady Galustov was born in 1941. He defended PhD thesis in cybernetics in 1971, and a doctorate thesis in computer-aided simulation of biological signals in 1991 (both in TRTI). He was a researcher and lecturer in the theory of random processes TRTI. His research interests are in the field of development of the theory of electrical signals, the synthesis of random processes, research and treatment of biomedical signals.

Karen Egiazarian (SM'96) was born in Yerevan, Armenia, in 1959. He received the M.Sc. degree in mathematics from Yerevan State University in 1981, the Ph.D. degree in physics and mathematics from Moscow State University, Moscow, Russia, in 1986, and the D.Tech. degree from the Tampere University of Technology (TUT), Tampere, Finland, in 1994. He has been Senior Researcher with the Department of Digital Signal Processing, Institute of Information Problems and Automation, National Academy of Sciences of Armenia. Since 1996, he has been an Assistant Professor with the Institute of Signal Processing, TUT, where he is currently a Professor, leading the Spectral and Algebraic Methods in DSP group. His research interests are in the areas of applied mathematics, signal processing, and digital logic.