Adjusting videoendoscopic 3D reconstruction results using tomographic data

K.A. Halavataya [1], K.V. Kozadaev [1], V.S. Sadau [1]
[1] BSU – Belarusian State University,
220030, Minsk, Belarus, Nezavisimosti Avenue 4

Abstract

Videoendoscopic and tomographic research are the two leading medical imaging solutions for detecting, classifying and characterizing a wide array of pathologies and conditions. However, source information from these types of research is very different, making it hard to cross-correlate them. The paper proposes a novel algorithm for combining results of videoendoscopic and tomographic imaging data based on 3D surface reconstruction methods. This approach allows to align separate parts of two input 3D surfaces: surface obtained by applying bundle adjustment-based 3D surface reconstruction algorithm to the endoscopic video sequence, and surface reconstructed directly from separate tomographic cross-section slice projections with regular density. Proposed alignment method is based on using local feature extractor and descriptor algorithms by applying them to the source surface normal maps. This alignment allows both surfaces to be equalized and normalized relative to each other. Results show that such an adjustment allows to reduce noise, correct reconstruction artifacts and errors, increase representative quality of the resulting model and establish correctness of the reconstruction for hyperparameter tuning.

Keywords: image reconstruction techniques, medical and biological imaging, image processing.

Citation: Halavataya KA, Kozadaev KV, Sadau VS. Adjusting videoendoscopic 3D reconstruction results using tomographic data. Computer Optics 2022; 46(2): 246-251. DOI: 10.18287/2412-6179-CO-910.

Acknowledgments: The work was partially funded and financially supported by the World Federation of Scientists National Scholarship Programme.

Introduction

Videoendoscopic research and tomographic imaging techniques are two of the most important and widespread methods for diagnosing and treating a wide array of conditions. While these two imaging techniques are different in both base physical principle of image acquisition and in the representation of their results, combining the two forms a unique challenge and opportunity of providing new types of visual representations that can be used to enhance the diagnostics process. One of the specific diagnostic areas that could benefit from such an approach is gastrointestinal tract imaging, which is used to detect, diagnose, classify and plan treatment of numerous pathologies, most notable adenomas and other types of growths which may pose a risk of developing cancer.

Cancer is currently the second leading cause of death globally – about 1 in 6 deaths occur due to cancer, according to the 2020 report of World Health Organization. Gastrointestinal cancer (including colon, rectum intestine, stomach and other gastrointestinal tract organs) makes up one of the largest cancer danger groups, mostly because it may develop without symptoms and can only be observed on screening procedures [1]. Both videoendoscopic research and tomographic imaging are the most prominent tools in early detection, diagnosis and treatment of these types of cancer.

These two research methods are not mutually exclusive; in fact, both methods have their own advantages and disadvantages, and for situations where maximum amount of information is required about specific case (i.e. about any kind of object of interest found during screening), it is very likely that both these methods need to be employed complementary. Therefore, algorithms for combining videoendoscopic research data with tomographic imaging results can be used to provide a composite 3D model representation.

This paper presents a combined algorithm for building 3D surface representations of specific areas of interest based on both videoendoscopic data and tomography data, where tomography-based surface is used as a reference for geometric structure of the area, while videoendoscopic reconstruction results are adjusted towards this reference structure in order to correct reconstruction artifacts and errors. The resulting model is therefore more representative of the area as a whole, combining both visual representation obtained by videoendoscopic research and spatial representation obtained by tomography.

1. 3D reconstruction from videoendoscopic data

Videoendoscopic research is one of the most important in-vivo diagnosis methods that allows for diagnosing and treatment of wide array of pathologies, most notable potentially oncological objects.

Videoendoscopes are specialized medical devices for visual (i.e. visible-light spectrum) observation of cavities in internal organs, often equipped with additional hardware that also allows taking biological samples for re-

search, injecting various types of drugs, and performing different therapeutic and surgical tasks. Modern videoendoscopes employ a powerful directed light source (usually LED-based) that illuminates the area of interest within the patient's body, after which reflected light is captured by a digital imaging sensor (CMOS or CCD) on the distal end of the endoscope. Captured image is then transmitted in digital format using regular wiring in the flexible endoscope tube. This signal is later displayed in real-time on an external monitor, and is additionally recorded as a video sequence [2].

The input data that is used for further processing by proposed methods and algorithms are represented by video sequences captured during the procedure. While some of the proposed methods may be used in real time during the procedure itself, usually visual analysis is best performed separately, because of the reasonable time constraints on endoscopic procedures. Most of the modern video endoscopes record full-colored digital images with two primary resolutions – High Definition (HD) of 1280×1024 pixels, or High Definition Plus (HD+) of 1600×1200 pixels, with framerate of 25, 30 or 60 Hz.

Video sequences obtained during videoendoscopic research have a limited representative quality, especially when researching a specific area of interest. It is usually hard or impossible to perform linear measurements of the objects of interest, or observe the larger area than what's directly visible from endoscopic imaging system. One of the solutions for this problem is building 3D representations based on videoendoscopic research data based on structure from motion family of algorithms.

In our research, we perform 3D reconstruction using methods previously proposed in [3]. The approach is based on wide-angle spherical projection model [4], as well as adaptive feature descriptors and detectors [5]. It includes the following steps:

1. Frame extraction,
2. Frame filtering,
3. Point matching,
4. Point match filtering,
5. Sparse point cloud reconstruction,
6. Iterative dense point cloud reconstruction,
7. Polygonal model reconstruction,
8. Texturing.

Video frame extraction and filtering is used to establish a minimum required baseline of projections that will be used for further structure from motion reconstruction. Using proposed detectors and descriptors, keypoints are located and matched across the filtered frames. Since some of the keypoint matches may be invalid, they are additionally filtered. After that, proposed modified bundle adjustment based on spherical wide-angle projection is used in order to build initial sparse point cloud. The cloud density is then iteratively increased, repeated until the error starts to stably increase across consecutive iterations. After dense point cloud is obtained, polygonal model is interpolated using screened Poisson reconstruction algorithm. Since dense point cloud coordinates on the polygonal model can be traced back to specific two-dimensional image points, it's possible to use them as reference for model texturing by back-projecting image pixel colors onto the polygonal model itself, thus establishing its texture.

3D reconstruction from videoendoscopic data allows to retain the original visual look of the objects of interest. This is especially important in diagnosis, since visible patterns of potentially cancerous objects are generally used to classify and distinguish between different types of polyps, adenomas and other growth kinds (e.g. Kudo's Pit Pattern classification [6]). However, due to specific nature of source data for videoendoscopic 3D reconstruction, the resulting representation has numerous disadvantages:

1. 3D reconstruction based on structure from motion algorithms is generally very sensitive to noise;
2. Texturing the surfaces may produce visual artifacts where certain parts of source two-dimensional projections are stretched or shrunk along a certain direction;
3. 3D reconstruction in the areas with insufficient spatial information available (i.e. on edges or in areas only visible on a small number of frames) will produce anomalies, "floating" keypoint regions, tearing, gaps, unexpected peaks and other types of spatial artifacts.

As such, enhancing spatial characteristics of resulting models is a very important problem. It should also be noted that using heuristic post-processing methods can potentially lead to loss of important data, so usefulness of artifact correction techniques for such models is limited.

2. 3D reconstruction from tomographic data

Medical tomographic data typically refers to visual representations obtained using magnet resonance imaging and X-ray-based computer tomography techniques.

Computer tomography (CT) employs X-ray scanning to obtain multiple cross-sectional slice projections of the human body. Most common scanner construction is spiral CT, which uses one or several X-ray tubes and detector plates placed on the opposite ends of a rotating mechanism, so that they can be rotated around the central axis of the scanned area where patient is placed. X-rays produced by the tubes travel towards the detector plate and are partially absorbed along the way as they encounter different kinds of obstacles; moreover, absorption amount depends on the density of the substance along the ray path. Intensities of the passing X-rays are captured on a discrete sensor grid of the detector plate to produce a two-dimensional radiographic image. This way, multiple radiographic images are obtained during the scan. After that, an algorithmic reconstruction method (usually based on a combination of Fourier projection-slice theorem and one of the algebraic reconstruction techniques) can be used to match tube and detector positions for each radiographic image to produce a series of projected rays [7] that were used to generate the image and refine them to

find out the density of a given point in 3D modelling space, usually represented as a voxel grid.

Magnetic resonance imaging (MRI) tomography is based on the principle of nuclear magnetic resonance, where subatomic particles with non-zero spin exposed to an external magnetic field start to resonate, i.e. precess perpendicular the magnetic field direction with frequency depending on the field strength. Most commonly, magnetic fields in MRI cause resonance on frequencies corresponding to radio wave frequency range. After that, to produce a tomographic image, a series of radio frequency bursts and magnetic field adjustments are employed. The relationship between frequencies of produced bursts, their repetition times and particle resonance frequencies allows to perform the scan in various modes. Weighting inputs from different types of produced signals allows to differentiate between various types of tissue, depending on its substance composition. Sensors placed at fixed or rotating positions allow forming multiple cross-sectional images, which can be later combined using algebraic reconstruction methods, two-dimensional and three-dimensional Fourier transforms and various other algorithms. Similar to CT imaging, the results are usually represented as 3D voxel grid.

Voxel grid representation of scanned objects allow to produce cross-sectional images along any axis, typically along one of the anatomical planes that allows to examine different internal areas of human body.

Numerous techniques exist that allow to perform surface reconstruction based on voxel grid data. These algorithms typically determine the boundaries of the object of interest based on the density or composition similarity in order to distinguish surface tissue from other areas [7, 8].

Representation based on voxel grid also has numerous disadvantages:

1. Some types of tissue could be indistinguishable from the surroundings because of the similar density or substance composition. The biggest problem among these types of tissue are specific kinds of adenomas. Detecting these kinds of anomalies using tomographic scanning is usually impossible without using various complex contrasting techniques, whereas videoendoscopic scan would allow to distinguish them based on their color, which isn't available in tomographic scan;
2. The large amount of information produced by a single tomographic scan in a voxel grid representation can be very hard to analyze. For instance, in typical digestive tract screening it is necessary to check for adenomas on the inner surface of the tract, which would require to form several complex cross-section projections with irregular angles to observe at any given point;
3. Whole-body tomographic scans typically tend to have lower resolution. This might cause the specialists to miss smaller objects that would otherwise be clearly visible during videoendoscopic screening. As a consequence, even for adenomas detected using tomographic scans, the information on them is not enough to fully diagnose and plan the treatment – usually, in such cases, additional videoendoscopic examinations need to be performed in order to classify the growth based on its visible pattern and establish its spatial parameters.

3. Combining tomographic and videoendoscopic 3D surface reconstruction results

Typically, the problem of medical 3D reconstruction in context of analyzing various internal organs and parts of human body is associated with 3D reconstruction based on tomographic scans. Indeed, the representation of a scanned object as a voxel grid allows to produce cross-sectional images and not only observe the external boundaries of a certain area of interest, but also evaluate its internal composition. On the other hand, this kind of representation also has various disadvantages.

Videoendoscopic imaging is generally completely different from tomographic, because the representations obtained by video endoscope correspond to visible spectrum examination of the surfaces of internal organs, presented as a sequence (video) of wide-angle full-color images. This kind of representation generally tends to be much easier to interpret and understand by medical specialists than tomographic scan projections and cross-sections. Moreover, in gastrointestinal tract diagnosis a single videoendoscopic screening allows to observe the entire inner surface of the tract in one go, with high enough spatial resolution to detect and analyze various objects of interest, while controlled nature of the screening allows the specialist to focus on specific areas in order to obtain a more complete view of their surroundings. However, it is an invasive procedure with more complex time limits and constraints, and the representations obtained by videoendoscopic scans are limited to cavities that can be reached by the endoscope.

As such, there is no general approach for combining these two types of input data. Most of the existing research papers on the topic of combining and adjusting various 3D representation types do not consider using 3D models obtained by visual spatial reconstruction from endoscopic data [9, 10], instead focusing mostly on working with more streamlined CT and MRI 3D imaging data. Moreover, 3D reconstruction from monocular endoscopic visual representation is a relatively novel approach [3, 4, 11], so combining it with other types of research is a promising way of increasing visual quality and spatial accuracy of the resulting model.

A proposed solution is performing 3D reconstruction from videoendoscopic data in order to obtain a 3D model that can be later traced and cross-correlated with specific points in tomographic modeling space. However, this is not something that can be performed directly, since surface reconstruction based on videoendoscopic data produces a surface which needs to be matched against specif-

ic parts of cross-sectional projection from the tomographic voxel grid. In order to perform the match, 3D surface generation must be used in order to transform voxel-based representation of the object into a parametric surface that can be later matched against the results of videoendoscopic image 3D reconstruction [12].

The general schema of base image acquisition model is presented in fig. 1.

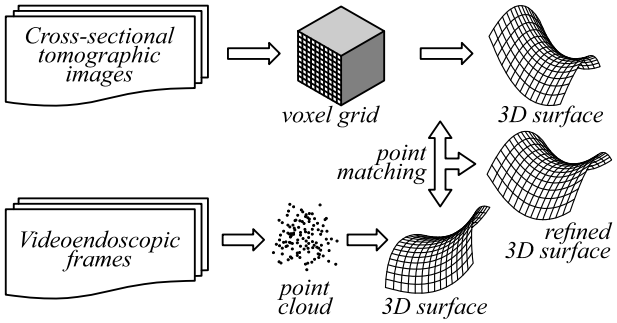

Fig. 1. Combined tomographic and videoendoscopic image acquisition model

Cross-sectional tomographic images are reconstructed into a voxel grid. After that, surface reconstruction algorithms are used to determine the boundaries of object of interest based on the density or composition similarity in order to distinguish surface tissue from other areas. This produces a reconstructed 3D surface of object of interest.

Likewise, videoendoscopic image frames are reconstructed based on the algorithms discussed earlier to create sparse and dense 3D point clouds, and then use point clouds as references to backproject image data onto the modelling space and reconstruct a continuous 3D surface.

After both 3D surfaces are obtained, point matching can be used in order to establish similar regions across the reconstructed surfaces and combine them. The combination can be used to compare one type of reconstruction against the other, using it as baseline reference, or in order to normalize artifacts and anomalies across both surfaces to produce a combined refined result.

The key stage of the proposed algorithm is matching two surfaces based only on their spatial configuration, since color information is unavailable for tomographic images. In order to perform this matching, we propose an algorithm based on local feature matching across smoothed surface normal maps.

Normal map is a two-dimensional image representation of any surface, where intensities of a single pixel across different color channels correspond to the respective normal vector to the surface at this point. In other words, if normal map is projected towards the surface as a texture, each point on the surface of this texture will have a three-channel intensity, corresponding to color, with one fixed channel (usually blue) where intensity defines a two-coordinate direction vector that would be normal (perpendicular) to the represented surface at this point. That means that each point of the normal map corresponds to a specific surface point in the modelling space, while its color corresponds to the curvature of the surface at this point. An example of 3D surface reconstructed from videoendoscopic data and its corresponding normal map is presented in fig. 2. The input of the matching algorithms are surfaces S_v with normal map $N[S_v]$ obtained from videoendoscopic 3D reconstruction, and S_t with normal map $N[S_t]$ obtained from tomographic imaging.

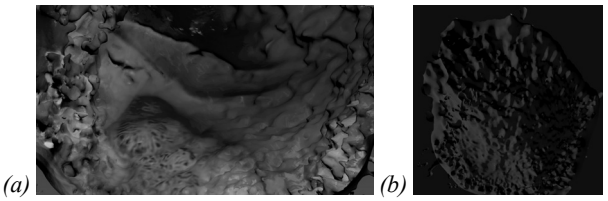

Fig. 2. (a) 3D surface model obtained during reconstruction from videoendoscopic images; (b) its corresponding normal map

Finer detail on the 3D surface (smaller bumps and pits) correspond to high-frequency component on the normal map. Because these high-frequency components can be introduced by noise or other types of artifacts, it is generally preferable to avoid matching across these points. For this reason, the first stage of surface alignment and matching is smoothing. To perform smoothing, we propose using simple two-dimensional Gaussian kernel convolution G across both normal maps $(G * N[S_v])_{ij}$ and $(G * N[S_t])_{ij}$ (fig. 3).

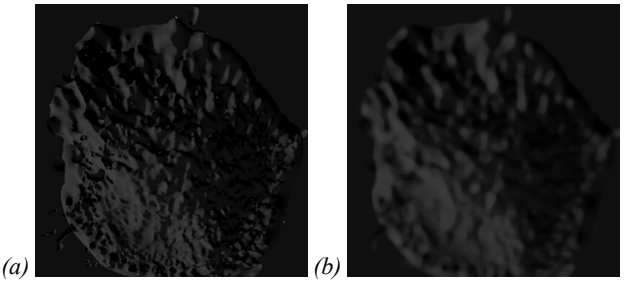

Fig. 3. Gaussian blur applied to normal map to eliminate high-frequency components: (a) before smoothing; (b) after smoothing

After normal maps are smoothed, it's possible to align them using local feature matching. We propose to use the same adaptive circumference-based keypoint algorithms that were developed for 3D reconstruction from videoendoscopic images [3, 5], with parameters adapted to working with colored normal maps. Feature extraction and matching allows to establish a number of common points across both smoothed normal maps, which means that these points correspond to the same area of space. This can be used to align smoothed normal map obtained by 3D surface from videoendoscopic reconstruction and from tomographic reconstruction. To eliminate false matches, we use distance ratio test based on Lowe's simulated probability distribution function for potentially incorrect matches [13] and filter out points that have matching descriptors but greatly differ in geometry. While these points might be indicative of actual discrepancies between the two surfaces that need to be corrected later, it is

important not to use them for surface alignment since it may lead to errors in refinement later.

After establishing a number of aligned keypoints $M \subset \mathbb{N}_0 \times \mathbb{N}_0$ (where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, i.e. a set of non-negative integers), a homographic projection matrix H is calculated to align one normal map respective to the other, minimizing projection error $\delta\vec{e}_{ij}$:

$$(i,j) \in M : (G * N[S_v])_{ij} = (H \cdot (G * N[S_t]))_{ij} + \delta\vec{e}_{ij}. \quad (1)$$

Both of the resulting surfaces are then back-projected into 3D modelling space.

In order to perform artifact correction, a weight coefficient $w \in [0;1]$ is introduced. Each point of the non-smoothed projected normal map obtained from videoendoscopic 3D reconstruction $N[S_v]$ is then adjusted based on its difference between non-smoothed aligned tomographic 3D surface normal map:

$$\left(N[S_v]_{res}\right)_{ij} = N[S_v]_{ij} + w\left(\left(H \cdot N[S_t]\right)_{ij} - N[S_v]_{ij}\right). \quad (2)$$

Weight coefficient w can be used to adjust the relative influence of videoendoscopic and tomographic surface. For, videoendoscopic normal maps remain almost unchanged. For $w \to 1$, videoendoscopic normal maps are almost completely replaced by normal maps from surface obtained using 3D reconstruction from tomography.

Example of the resulting adjusted surface for different values of w are presented in fig. 4.

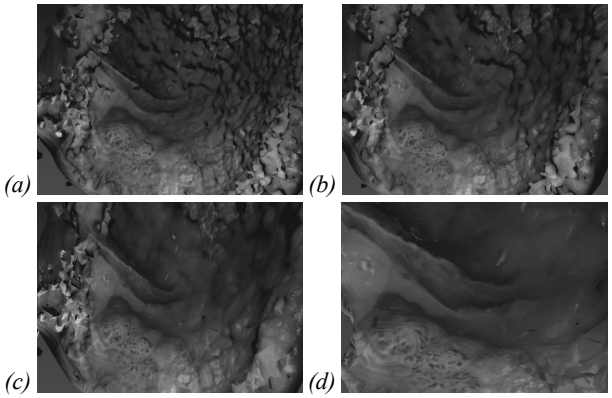

Fig. 4. Adjusted videoendoscopic surface for different weight values: (a) w = 0, equivalent to videoendoscopic surface model; (b) w = 0.3; (c) w = 0.6; (d) w = 0.8 – tearing and artifacts are mostly removed, but surface configuration is simplified

As seen from the fig. 4, weight coefficient needs to be carefully considered when performing the process. Lower values tend to retain most of the tearing, artifacts and high-frequency errors produced by structure from motion reconstruction, usually caused by misalignment of one or more cameras during the bundle adjustment. On the other hand, higher values, while helpful for alleviating artifacts, tend to simplify the geometry too much, essentially "flattening" the surfaces and thus making it harder to distinguish finer detail.

The results of 3D reconstruction and its further adjustments based on proposed algorithms are intended for use by medical endoscopy specialists, and different levels of detail and artifact tolerance might be required based on specific diagnostic or therapeutic goals. For this reason, it is useful to present the results not as a single resulting model, but as a range of models for different weight coefficient values, so that medical specialists may choose what form of visual representation works best for specific case.

One open problem of using 3D reconstruction and adjustment algorithms is establishing the correctness of the results [4, 9]. In traditional photogrammetry-based 3D reconstruction methods, correctness is evaluated by comparing it to an existing base; however, establishing a base in videoendoscopic imaging is way harder, since direct measurements are not possible, and secondary research methods usually have lower spatial definition. For this reason, correctness and visual quality of the resulting model are usually evaluated by medical experts. Such an evaluation has an inherent bias since it is highly subjective and non-quantifiable. Therefore, researching new ways of confirming spatial correctness of the resulting models is an important direction of further research on the topic.

Conclusion

3D reconstruction techniques provide a unique and powerful tool for medical data visualization. While tomographic reconstruction methods are generally accepted as the standard for 3D data representation, 3D reconstruction from video sequences obtained during videoendoscopic research provides a more understandable representation. Both of these techniques have their advantages, disadvantages and usage scenarios, but both endoscopic and tomographic research are generally used complementary in tandem in order to obtain more information about specific regions of interest.

As such, combining the results of these two types of 3D reconstruction helps alleviate some of the shortcomings of each method by creating a geometrically and spatially correct representation of objects of interest, as guaranteed by tomography, while preserving finer detail and visible-spectrum overview of the areas provided by video endoscopy, as well as correcting spatial artifacts introduced by reconstruction based on structure from motion family of algorithms.

A method proposed in this paper is based on aligning the surfaces based on local feature descriptor matching across their two-dimensional normal map representations. After the surfaces are aligned, it is possible to adjust both surfaces to incorporate spatial data of tomographic reconstruction with finer detail and visible texturing of videoendoscopic reconstruction. This provides a unique type of visualization that can be helpful in analyzing and diagnosing various types of objects observed during videoendoscopic screening and tomographic scanning.

References

[1] WHO report on cancer: setting priorities, investing wisely and providing care for all. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.

[2] Dremel HW. General principles of endoscopic imaging. In Book: Ernst A, Herth FJF, eds. Principles and practice of interventional pulmonology. New York, NY: Springer; 2013. DOI: 10.1007/978-1-4614-4292-9_2.
[3] Halavataya K, Sadov V. Algorithm for 3D scene reconstruction from videoendoscopic examination data [In Russian]. News of Science and Technologies 2019; 3: 10-18.
[4] Halavataya K, Sadau V. Model of image acquisition for 3D scene reconstruction from videoendoscopic imaging data [In Russian]. Polotsk State University Proceedings 2019; 12: 43-49.
[5] Halavataya K. Local feature descriptor indexing for image matching and object detection in real-time applications Pattern Recognit Image Anal 2020; 30(1): 18-23. DOI: 10.1134/S105466182001006X.
[6] Kudo S, Tamura S, Nakajima T, Yamano H, Kusaka H, Watanabe H. Diagnosis of colorectal tumorous lesions by magnifying endoscopy. Gastrointest Endosc 1996; 44(1): 8-14. DOI: 10.1016/S0016-5107(96)70222-5.
[7] Herman GT. Fundamentals of computerized tomography: Image reconstruction from projections. 2nd ed. New York: Springer; 2010. ISBN: 978-1-85233-617-2.
[8] Nagai Y, Ohtake Y, Suzuki H. Tomographic surface reconstruction from point cloud. Comput Graph 2015; 46: 55-63. DOI: 10.1016/j.cag.2014.09.034.
[9] Baek J, Pelc NJ. A new method to combine 3D reconstruction volumes for multiple parallel circular cone beam orbits. Med Phys. 2010; 37(10): 5351-5360. DOI: 10.1118/1.3484058.
[10] Basha MAA, AlAzzazy MZ, Ahmed AF, et al. Does a combined CT and MRI protocol enhance the diagnostic efficacy of LI-RADS in the categorization of hepatic observations? A prospective comparative study. Eur Radiol 2018; 28(6): 2592-2603. DOI: 10.1007/s00330-017-5232-y.
[11] Resindra A, Yusuke M, Okutomi M, Suzuki S. Whole stomach 3D reconstruction and frame localization from monocular endoscope video. IEEE J Transl Eng Health Med 2019; 7: 1-10. DOI: 10.1109/JTEHM.2019.2946802.
[12] Khan U, Yasin A, Abid M, Shafi I, Khan SA. A methodological review of 3D reconstruction techniques in tomographic imaging. J Med Syst 2018; 42(10): 190. DOI: 10.1007/s10916-018-1042-2.
[13] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004; 60: 91-110. DOI: 10.1023/B:VISI.0000029664.99615.94.

Authors' information

Katsiaryna Aliaksandrauna Halavataya, (b. 1993), graduated from Belarusian State University in 2015, defended her Ph.D. qualificative work in 2020. Works as Associate Professor of the Intelligent Systems department of Belarusian State University. Research interests: computer vision, 3D reconstruction from images, machine learning, computer graphics, virtual and augmented reality. E-mail: *katerina-golovataya@yandex.ru* .

Konstantin Vladimirovich Kozadaev (b. 1983) graduated from Belarusian State University in 2005, majoring in Physics, Laser Spectroscopy. In 2009 defended his PhD thesis for the degree of Candidate of Physical and Mathematical Sciences. In 2020 attained the degree of Doctor of Physical and Mathematical Sciences. Currently working as a Vice-Rector for Academic Affairs and Internationalization of Education at Belarusian State University, and as Professor of Intelligent Systems department of Faculty of Radiophysics and Computer Technologies, Belarusian State University. Research interests include automation of a physical experiment, information technologies and laser plasma physics.
E-mail: *kozadaevkv@bsu.by* .

Vasiliy Sergeevich Sadau, (b. 1950), graduated from Minsk Radio Engineering Institute in 1974, defended his Ph.D. qualificative work in 1985. Works as Professor of the Intelligent Systems department of Belarusian State University. Research interests: computer steganography, intelligent systems, emotion recognition, integrated electronics.
E-mail: *sadov@bsu.by* .

Code of State Categories Scientific and Technical Information: 28.23.15
Received April 17, 2021. The final version – September 8, 2021.